

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Самарский государственный технический университет»

На правах рукописи

МОСИН ВЛАДИМИР ГЕННАДЬЕВИЧ

**МЕТОДОЛОГИЯ И ИНСТРУМЕНТАРИЙ
ДЕТЕКЦИИ АНОМАЛИЙ В УПРАВЛЕНИИ КАЧЕСТВОМ
ПРОЦЕССОВ АВТОМОБИЛЬНОЙ ОТРАСЛИ**

2.5.22. Управление качеством продукции. Стандартизация.
Организация производства

Диссертация на соискание учёной степени
доктора технических наук

Научный консультант:
доктор технических наук, профессор
Козловский Владимир Николаевич

Самара — 2025

ОГЛАВЛЕНИЕ

Введение	10
Глава 1. Концептуальный анализ основных категорий управления качеством в контексте технологий больших данных	24
1.1. Обзор концепций, методов и стандартов управления качеством	25
1.1.1. Исторически сложившиеся концепции управления качеством	25
1.1.2. Основные инструменты управления качеством	30
1.1.3. Стандарты управления качеством.....	34
1.2. Типология процессов в управления качеством.....	35
1.2.1. Общая типология	35
1.2.2. Измеримость и стабильность процессов в сфере управления качеством.....	38
1.3. Детекция аномалий в измеримых нестабильных процессах	42
1.3.1. Продукт как совокупность процессов.....	42
1.3.2. Алгоритмический подход к детекции аномалий в нестабильных процессах	46
1.3.3. Детекция аномалий по времени и по локации	48
1.3.4. Модель как альтернативная версия стандарта.....	50
1.3.5. Обобщающие свойства модельного подхода.....	54
1.4. Процесс в формате MLC (Moment, Location, Characteristic)	56
1.4.1. Форматы данных ML (Moment, Location) и LM (Location, Moment)	58

1.4.2. Отступления от форматов ML/ LM	61
1.4.3. Обобщение форматов ML/ LM на множество характеристик	63
1.5. Выводы по главе 1	64
Глава 2. Методика МССР (Modeling, Challenge, Calibration, Production) как средство алгоритмической детекции аномалий в нестабильных процессах	67
2.1. Признаковые описания объектов	69
2.2. Моделирование процесса	80
2.2.1. Разбиение данных на этапы истории и обучения	80
2.2.2. Пул моделей.....	81
2.2.3. Моделирование по моментам времени	82
2.2.4. Моделирование по локациям	88
2.3. Критические границы отклонений	94
2.3.1. Критические границы отклонений от прогноза по времени ..	94
2.3.2. Критические границы отклонений от прогноза по локациям.....	98
2.3.3. Калибровочные коэффициенты по совокупности условий ..	100
2.4. Сравнение алгоритмической детекции с экспертной оценкой	102
2.4.1. Один эксперт	102
2.4.2. Группа экспертов	105
2.4.3. Коррекция алгоритма.....	107
2.5. Пилотная эксплуатация алгоритма.....	108
2.5.1. Первый момент эксплуатации	109
2.5.2. Второй и последующие моменты эксплуатации	109
2.5.3. Обобщение на несколько характеристик.....	113
2.6. Аудит локаций.....	114
2.7. Выводы по главе 2.....	115
Глава 3. Экспериментальное исследование возможностей и ограничений методики МССР на синтетических данных .	118

3.1. Синтез данных в формате ML/ LM	120
3.1.1. Репер кубической кривой	121
3.1.2. Кривая тренда, построенная на базовом репере	123
3.1.3. Семейство кривых тренда, построенное по семейству реперов	125
3.1.4. Синтез нормального шума	128
3.1.5. Портреты локаций и портреты моментов.....	131
3.2. Моделирование по времени	133
3.2.1 Применение функции <code>modeling_by_moments</code>	134
3.2.2. Выбор наилучшей схемы моделирования	136
3.3. Моделирование по локациям	138
3.3.1. Применение функции <code>modeling_by_locations</code>	139
3.3.2. Выбор наилучшей схемы моделирования	140
3.4. Синтез экспертного мнения	143
3.4.1. Отклонения лучшей модели от реальных значений.....	144
3.4.2. Синтез экспертного сообщества.....	146
3.4.3. Мнение экспертного сообщества	150
3.5. Бинаризация отклонений по времени	151
3.5.1. Применение функции <code>calibration</code>	153
3.5.2. Вариация экспертного мнения.....	154
3.5.3. Сопоставление с экспертной оценкой	157
3.6. Бинаризация отклонений по локациям	158
3.7. Пилотная эксплуатация детектора аномалий.....	162
3.7.1. Детекция аномалий по времени.....	163
3.7.2. Детекция аномалий по локациям	164
3.7.3. Календарь аномалий	165
3.8. Выводы по главе 3.....	166
Глава 4. Детекция аномалий в процессе автомобильного производства	168

4.1. Производство как социальный процесс.....	170
4.1.1. Роль организационных аспектов процесса производства.....	172
4.1.2. Обучение персонала как один из факторов качества производства	173
4.1.3. Роль организации рабочего места в управлении качеством.	173
4.1.4. Мотивация сотрудников в контексте управления качеством производственного процесса	174
4.1.5. Социальные процессы как важная часть процесса производства.....	175
4.2. Данные об исполнении операции установки уплотнителей на стекла автомобиля	176
4.2.1. Когортный анализ производительности	176
4.2.2. Описание данных	178
4.3. Детекция аномалий по времени в процессе производства	180
4.3.1. Моделирование по времени процесса производства	180
4.3.2. Калибровка отклонений по времени в процессе производства.....	183
4.3.3. Пилотная эксплуатация детектора аномалий по времени в процессе производства	184
4.4. Детекция аномалий по локациям в процессе производства	189
4.4.1. Моделирование по локациям процесса производства	189
4.4.2. Калибровка отклонений по локациям в процессе производства.....	190
4.4.3. Пилотная эксплуатация детектора аномалий по локациям в процессе производства	191
4.5. Об интерпретации аномалий, выявленных в процессе производства	194
4.6. Выводы по главе 4.....	197

Глава 5. Детекция аномалий в процессе сервисного обслуживания автомобилей	199

5.1. Сервисное обслуживание как процесс потребления услуги	202
5.1.1. Специфика потребления сервисных услуг	202
5.1.2. Временные паттерны в процессах потребления	203
5.1.3. Роль сервисного обслуживания в общей оценке качества продукта	205
5.2. Данные о работе сети сервисных центров.....	205
5.2.1. Предварительное преобразование данных к формату ML/ LM.....	206
5.2.2. Парето-анализ сервисных центров.....	210
5.2.3. Описание данных	213
5.3. Детекция аномалий по времени в процессе сервисного обслуживания для числа оформленных заявок.....	216
5.3.1. Моделирование по времени процесса сервисного обслуживания для числа оформленных заявок.....	216
5.3.2. Калибровка отклонений по времени в процессе сервисного обслуживания для числа оформленных заявок	217
5.3.3. Пилотная эксплуатация детектора аномалий по времени в процессе сервисного обслуживания для числа оформленных заявок.....	218
5.4. Детекция аномалий по локациям в процессе сервисного обслуживания для числа оформленных заявок.....	222
5.4.1. Моделирование по локациям процесса сервисного обслуживания для числа оформленных заявок	222
5.4.2. Калибровка отклонений по локациям в процессе сервисного обслуживания для числа оформленных заявок	223
5.4.3. Пилотная эксплуатация детектора аномалий по локациям в процессе сервисного обслуживания для числа оформленных заявок.....	224
5.5. Детекция аномалий по времени в процессе сервисного обслуживания для суммарной стоимости гарантийных работ.....	228

5.5.1. Моделирование по времени процесса сервисного обслуживания для суммарной стоимости работ.....	228
5.5.2. Калибровка отклонений по времени в процессе сервисного обслуживания для суммарной стоимости работ.....	230
5.5.3. Пилотная эксплуатация детектора аномалий по времени в процессе сервисного обслуживания для суммарной стоимости работ	231
5.6. Детекция аномалий по локациям в процессе сервисного обслуживания для суммарной стоимости гарантийных работ.....	235
5.6.1. Моделирование по локациям процесса сервисного обслуживания для суммарной стоимости работ.....	235
5.6.2. Калибровка отклонений по локациям в процессе сервисного обслуживания для суммарной стоимости работ.....	237
5.6.3. Пилотная эксплуатация детектора аномалий по локациям в процессе сервисного обслуживания для суммарной стоимости работ	238
5.7. Об интерпретации аномалий, выявленных в процессе сервисного обслуживания	241
5.8. Выводы по главе 5.....	245
Глава 6. Детекция аномалий в процессе маркетингового сопровождения автомобильного бренда.....	246
6.1. Маркетинговое сопровождение автомобильного бренда как процесс потребления контента	248
6.1.1. Коммуникационные площадки для продвижения нарратива автомобильных брендов.....	250
6.1.2. Этапы маркетинговых кампаний на цифровых площадках .	251
6.1.3. Таргетинг рекламной кампании в интернет-маркетинге	253
6.2. Данные о переходах на канал автомобильного бренда.....	255
6.3. Детекция аномалий по времени в процессе маркетингового сопровождения	258

6.3.1. Моделирование по времени процесса производства	259
6.3.2. Калибровка отклонений по времени в маркетинговом процессе	260
6.3.3. Пилотная эксплуатация детектора аномалий по времени в процессе маркетинга	261
6.4. Детекция аномалий по локациям в процессе маркетингового сопровождения	265
6.4.1. Моделирование по локациям процесса маркетингового сопровождения	265
6.4.2. Калибровка отклонений по локациям в процессе маркетингового сопровождения	267
6.4.3. Пилотная эксплуатация детектора аномалий по локациям в процессе маркетингового сопровождения.....	268
6.5. Об интерпретации аномалий, выявленных в процессе маркетингового сопровождения	271
6.6. Выводы по главе 6.....	275

Глава 7. Интеграция методологии и инструментария детекции аномалий в теорию и практику управления качеством.....277

7.1. Управление нестабильными процессами на основе выявленных аномалий	279
7.1.1. Неформальный пример.....	280
7.1.2. Расчёт прироста эффективности аудита	284
7.2. Детекция аномалий в нестабильных процессах и принцип непрерывного улучшения.....	287
7.2.1 Философия Kaizen Масааки Имаи.....	288
7.2.2. Динамическая стандартизация измеримого нестабильного процесса	289
7.2.3. Синтез данных	290
7.2.4. Операциональное вмешательство в синтетический процесс	

и перерасчёт величины критического отклонения	295
7.3. Аспекты практической значимости концепции детекции аномалий в нестабильных процессах	299
7.3.1. Напоминание: практическая значимость контрольных карт Шухарта	300
7.3.2. Детекция аномалий как эволюция контрольных карт	302
7.4. Выводы по главе 7	305
Заключение.....	307
Список литературы.....	314
Список иллюстративного материала.....	343
Приложение А. Теоретические основы стохастического прогнозирования.....	356
Приложение В. Библиотеки анализа больших данных.....	414
Приложение С. Ключевые алгоритмы методики МССР ...	432
Приложение Д. Календари выявленных несоответствий в нестабильных процессах одного из ведущих отечественный автопроизводителей.....	452
Приложение Е. Акты внедрения.....	488

ВВЕДЕНИЕ

Актуальность темы исследования

В деятельности любой компании качество корпоративной системы управления является приоритетным, поскольку напрямую влияет на успешность стратегического развития. Ключевым инструментом для улучшения деятельности является система менеджмента качества (СМК), основанная на международных стандартах ISO 9000, при этом в автомобильной промышленности определяющим стандартом СМК выступает IATF 16949, а национальный стандарт ГОСТ Р ИСО 9001-2015 адаптирует международные требования в области качества к отечественным условиям. Соблюдение положений стандартов в области качества является обязательным требованием, учитывая их важность для признания организации и её возможности работать как на внутреннем, так и на международных рынках.

Управление качеством особенно важно в отраслях промышленности с высокой конкуренцией, где производятся сложные технические изделия, например, современном машиностроении, в частности — автомобилестроении. Современный автопроизводитель рассматривает свой продукт не только как совокупность технических характеристик, но и как сложную систему, включающую в себя весь жизненный цикл автомобиля, от маркетинговых исследований до организации сервисного обслуживания и формирования потребительской лояльности к бренду. Автомобилестроение характеризуется разветвлённой цепочкой поставок компонентов и предприятий, выполняющих кооперационные технологические процессы, при этом стремительная цифровизация процессов управления качеством предполагает освоение новых компетенций

в области применения цифровых методов планирования, организации, мониторинга и управления качеством продукции. В контексте маркетинговых инициатив, сервисного обслуживания, а также в условиях аддитивного производства, процессы, могут быть лишены стабильности, и при этом именно сегодня, в условиях тотальной цифровизации автопроизводители располагают практически неограниченными массивами данных, отражающих динамику этих процессов: количество заявок на гарантийное обслуживание в различных сервисных центрах; статистика просмотров контента на онлайн-платформах бренда; данные с активных датчиков контроля технологических производственных линий. Однако применение традиционных статистических методов, таких как контрольные карты Шухарта, разработанных для анализа стабильных процессов, оказывается либо принципиально невозможным, либо приводит к получению результатов, характеризующихся неприемлемой степенью погрешности, что исключает их практическую ценность.

В машиностроении накапливается колоссальный объем данных, генерируемых на различных этапах жизненного цикла продукции. Датчики, станки с числовым программным управлением, автоматизированные комплексы с системой предиктивного управления, а также производственные системы диагностирования непрерывно поставляют информацию о параметрах технологических процессов, состоянии узлов и агрегатов, качестве выпускаемой продукции. Однако, как это ни парадоксально, основная проблема заключается не в дефиците данных, а в неспособности эффективно их анализировать. Существующие методы мониторинга и аналитики, зачастую, не справляются с обработкой таких массивов разнородной и сложной информации. Данные могут быть зашумлены, неполны, противоречивы и представлены в различных форматах. Традиционные статистические подходы к управлению качеством оказываются недостаточно мощными для выявления скрытых закономерностей и зависимостей. Кроме того, данные отражающие ход машиностроительного производства характеризуются высокой многофакторностью и нелинейно-

стью. Сложно выявить взаимосвязи между множеством параметров, влияющих на качество и надёжность продукции.

Таким образом, можно сделать существенный вывод о наличии в современной науке и практике управления качеством системного пробела и противоречия, которые образуют отраслевую проблему.

Пробел заключается в том, что широкоизвестные традиционные инструменты контроля и мониторинга показателей качества в условиях взрывного роста объемов электронных данных отражающих характеристики протекания процессов системы менеджмента машиностроительного (автомобильного) производства не всегда в состоянии полноценно и адекватно дать оценку функционированию, результативности и эффективности деятельности.

Противоречие заключается в том, что в актуальных стандартах менеджмента качества, таких как IATF 16949:2016, на уровне требований декларируется обязательность применения инженерных методик управления качеством, таких, например, как SPC, MSA и т.д., в то время как фундамент указанных методов выстроен, исходя из статистических инструментов, подразумевающих высокий уровень стабильности процессов.

Для устранения пробела и преодоления противоречия необходимы новые подходы к мониторингу данных, сочетающие в себе статистические методы, машинное обучение и экспертные знания специалистов. Только тогда можно будет превратить поток данных в процессах системы менеджмента качества автосборочного предприятия в ценный ресурс для повышения конкурентоспособности. Решение выделенной отраслевой научно-технической проблемы является ключевым фактором необходимым для дальнейшего развития отрасли и внедрения концепции «Индустрии 4.0». Сегодня недостаёт 1) целостной концепции, 2) методологии и 3) инструментария, позволяющих эффективно использовать имеющиеся информационные ресурсы управления качеством. Настоящая диссертационная работа восполняет этот пробел, предлагая концептуальные, методологические и алгоритмические основы для управ-

ления качеством процессов в машиностроении (автомобилестроении), с учётом гарантированности управленческих решений вне зависимости от уровня стабильности процессов. Таким образом, в диссертации предлагается переход от базового понятия стабильности процесса к понятию его предсказуемости в управлении качеством. Успешное применение аналитики данных позволяет перейти от реактивного управления к проактивному в системе менеджмента, предсказывая и предотвращая проблемы до их возникновения. Это, в свою очередь, приведёт к сокращению издержек, повышению эффективности производства и улучшению качества выпускаемой продукции.

Актуальность темы диссертационного исследования обусловлена необходимостью повышения результативности и эффективности управления качеством в условиях современной машиностроительной отрасли в общем и автомобилестроении в частности за счёт системного развития комплексных инструментов детекции аномалий в управлении качеством, то есть, специализированной системы статистического управления, предполагающей использование алгоритмов машинного обучения (таких как алгоритмы классификации и анализа временных рядов), которые обучаются на исторических данных, представляющих предсказуемое поведение как стабильных, так и нестабильных процессов, разрабатываются с учётом специфики автомобильного производства и способны эффективно распознавать отклонения от прогнозируемого поведения процессов системы менеджмента даже в условиях высокой вариативности данных. Анализ больших данных, поступающих из систем мониторинга качества, позволяет выявлять скрытые зависимости и нелинейные корреляции, зачастую недоступные для традиционных методов статистического анализа, при этом алгоритмы машинного обучения способны эффективно обнаруживать аномальные повторяющиеся участки (паттерны) в этих данных, что способствует оперативному вмешательству и предотвращению развития аномальных состояний. С практической точки зрения, актуальность темы диссертационного исследования подтверждается тем, что применение детекции аномалий

в управлении качеством в автомобилестроении повышает оперативность выявления отклонений (так как автоматизированный анализ данных обеспечивает быстрое обнаружение аномалий (несоответствий), минимизируя время реакции на проблемы) и снижает затраты на контроль качества (так как автоматизация процессов детекции аномальных состояний снижает потребность в ручном мониторинге и инспекциях, сокращая затраты на рабочую силу и оборудование).

Степень разработанности темы

Различные подходы к вопросам управления качеством, в частности — концепции, в той или иной степени апеллирующие к статистическим методам широко представлены в исследованиях как зарубежных, так и отечественных авторов.

Основополагающие научные работы таких авторов, как Э. Деминг, Дж. Джуран, П. Друкер, К. Исикава, Н. Кано, Р. Каплан, Ф. Котлер, Ф. Кросби, Г. Тагути, Н. Талеб, Ф. Тейлор, А. Фейгенбаум и В. Шухарт, оказали значительное влияние на формирование теоретической базы данного диссертационного исследования.

Научно-практическая направленность работы определена на основе исследований ведущих отечественных ученых, среди которых Ю. П. Адлер, В. Н. Азаров, Г. Г. Азгальдов, И. З. Аронов, В. А. Барвинок, В. Я. Белобрагин, Б. В. Бойцов, В. В. Бойцов, В. А. Васильев, В. Г. Версан, Г. П. Воронин, А. В. Гличев, В. А. Лapidус, В. В. Окрепилов, И. И. Чайка и другие.

Важнейшие научно-прикладные аспекты исследования рассматриваются в работах Д. В. Антипова, В. Ф. Безъязычного, С. А. Васина, В. Е. Годлевского, О. А. Горленко, С. Я. Гродзенского, В. П. Димитрова, А. Я. Дмитриева, В. В. Ефимова, А. В. Зажигалкина, А. Г. Ивахненко, В. А. Качалова, В. Я. Кершенбаума, Ю. С. Ключкова, В. Н. Клячкина, В. Н. Козловского, П. А. Лонциха, С. В. Мищенко, С. Н. Николаева, И. Н. Омельченко, К. Г. Пивоварова, Е. В. Плахотниковой, М. А. Поляковой,

С. В. Пономарева, В. Б. Протасьева, С. В. Пугачева, М. И. Розно, Т. А. Салимовой, Е. Г. Семеновой, Л. Е. Скрипко, А. Г. Сулова, Х. А. Фасхиева, А. И. Хаймовича, И. Н. Хаймович, Ю. К. Чернова, А. Д. Шадрина, А. П. Шалаева, В. Л. Шпера, В. В. Щипанова, Г. Л. Юнака и других российских исследователей.

Цель диссертационной работы

Цель работы заключается в повышении результативности и эффективности системы менеджмента качества автопроизводителя в условиях отраслевого развития информатизации и цифровизации посредством разработки нового методологического инструментария статистического управления качеством при помощи детекции аномалий в процессах автомобильной отрасли.

Научно-прикладные задачи

Для достижения поставленной цели в настоящей диссертационной работе решаются следующие научно-прикладные задачи:

1. Научно-технический анализ проблемы, связанной с выявлением фактов оценки стабильности процессов в системе менеджмента качества автопроизводителя и их контролем, а также выделение группы процессов, обладающих свойствами нестабильности.
2. Формирование и реализация научной концепции методологии детекции аномалий в управлении качеством процессов, дающей универсальное представление о процессах и подходах к определению нормы и аномалии на основании измеримых характеристик в автомобильной отрасли.
3. Разработка методологии детекции аномалий процессов системы менеджмента качества автопроизводителя, позволяющей давать оценку «норма-аномалия» любому его состоянию на основании наблюдаемой характеристики без апелляции к экспертной оценке.

4. Разработка и реализация научно-программного комплекса моделей и алгоритмов детекции аномалий в процессах автомобильной отрасли, входящих в систему менеджмента качества.
5. Апробация и практическая реализация полученного инструментария детекции аномалий в процессах системы менеджмента качества автопроизводителя.

Область исследования

Область исследования соответствует направлениям исследований паспорта научной специальности 2.5.22. Управление качеством продукции. Стандартизация. Организация производства в части пунктов:

4. Инновации при разработке, развитии, цифровизации систем менеджмента качества (СМК) предприятий и организаций;
8. Разработка научно-практического статистического инструментария управления качеством;
9. Разработка и совершенствование научных инструментов оценки, мониторинга и прогнозирования качества продукции и процессов;
14. Развитие основных положений и содержания Всеобщего Управления Качеством (TQM), и других концепций управления качеством;
25. Разработка моделей описания, методов и алгоритмов решения задач проектирования производственных систем, организации производства и принятия управленческих решений в цифровой экономике.

Объект исследования

Объектом исследования являются комплексные системы статистического мониторинга качества процессов менеджмента предприятий автомобильной отрасли, в условиях информатизации и цифровизации.

Предмет исследования

Предметом исследования являются статистические методы организации мониторинга качества процессов менеджмента предприятий автомобильной отрасли.

Научная новизна работы

Научная новизна заключается в разработке системных научно-практических инструментов и комплексов моделей, составляющих методологию детекции аномалий в управлении качеством автопроизводителя, направленных на повышение результативности и эффективности системы менеджмента качества, в условиях отраслевого развития информатизации и цифровизации и включает следующее:

1. Разработана концепция методологии детекции аномалий в управлении качеством процессов менеджмента автопроизводителя, которая в отличие от заложенных норм и требований в стандарте IATF 16949:2016 по применению инструментов SPC определяет новый подход к анализу данных, основанный на предиктивном описании процессов системы менеджмента качества. Пункт 14 паспорта научной специальности 2.5.22.
2. Разработан комплекс научно-программных инструментов по обработке данных о процессах СМК автомобильной отрасли, отличающийся возможностью генерировать синтетические данные о процессах и их экспертных оценках, направленный на решение задачи тестирования инструментов детекции аномалий в условиях различных искусственно смоделированных экспертиз. Пункт 4 паспорта научной специальности 2.5.22.
3. Разработан комплекс методов и инструментов выявления аномалий в процессах, который позволяет детектировать аномальные состояния в характеристиках измеримых процессов системы менеджмента качества автомобильной отрасли. Пункт 8 паспорта научной специальности 2.5.22. Комплекс включает в себя:

3.1. Метод моделирования временных рядов, отличающийся возможностью

одновременного исследования семейства процессов, протекающих в совокупности локаций (метод направлен на детекцию аномалий процесса по отношению к его состояниям в предыдущие моменты времени).

3.2. Метод моделирования состояния семейства процессов, протекающих в совокупности локаций, в фиксированные моменты времени (метод направлен на детекцию аномалий процесса по отношению к его состояниям в других локациях).

3.3. Статистический критерий оценки выраженности аномалий в характеристиках измеримых процессов, основанный на определении локтевой точки, который, во-первых, применим к любым законам статистических распределений, в отличие от критерия 6 сигм, применимого только к нормальным распределениям, а во-вторых, не апеллирует к экспертной оценке.

4. Разработан комплекс научно-программных инструментов, позволяющий реализовать детекцию аномалий в управлении качеством процессов системы менеджмента предприятий автомобильной отрасли, отличающийся универсальностью применения к анализу процессов, имеющих как стабильные, так и нестабильные характеристики протекания. Пункт 9 паспорта научной специальности 2.5.22.
5. Предложен метод принятия решений построенный на основе рационального подхода, отличающийся от инструментария управления, реализованного на основе выборочного контроля, полным анализом всей совокупности статистических данных о процессах и направленный на исключение человеческого фактора. Пункт 25 паспорта научной специальности 2.5.22.

Теоретическая значимость

Теоретическая значимость диссертационного исследования заключается в создании методологии детекции аномалий в управлении качеством процессов предприятий автомобильной отрасли, направленной на повышение результативности и эффективности системы менеджмента качества и обеспечивающей

системное развитие положений теории управления качеством, а также установлением и углублением связей между базовыми положениями, образующими принципы системы менеджмента качества — процессным подходом и принятием решений основанных на свидетельствах. Предложенный методологический инструментарий расширяет арсенал статистических инструментов мониторинга качества процессов системы менеджмента качества автомобильной отрасли в условиях развития информатизации и цифровизации.

Разработанная методология обладает высокой степенью универсальности. В рамках автомобильной отрасли она позволяет изучать с точки зрения соответствия или аномалий норме самые разнообразные процессы: маркетинг, проектирование, производство, сервисное техническое обслуживание, что обеспечивает широкий охват характеристик, обеспечивающих качество автомобильной продукции. Вместе с тем, область применения комплекса разработанных методов не ограничивается автомобилестроением и может быть расширена на другие отрасли машиностроения. Существенную значимость работы составляет аспект, связанный с развитием отраслевого стандарта IATF 16949:2016 и его российского аналога ГОСТ Р 58139-2018 в части обязательного применения инженерных инструментов статистического управления процессами системы менеджмента (SPC), за счёт включения предложенного в работе инструментария, образующего методологию детекции аномалий в процессы СМК предприятий автомобильной отрасли.

Практическая значимость

Практическая значимость результатов диссертации состоит в разработке комплекса научно-прикладных решений, повышающих результативность и эффективность мониторинга процессов системы менеджмента качества в автомобильной отрасли. Полученные в диссертационной работе результаты вошли в устойчивую отраслевую практику крупных автомобилестроительных предприятий. В частности, внедрены инструменты автоматизированного поиска и

устранения нарушений, связанных с необоснованным ремонтом автомобильной техники на предприятиях фирменной сервисной сети автопроизводителя, а также инструментарий поиска позиций дефектов автомобилей в эксплуатации, имеющих наибольший прирост за отчётный анализируемый период времени. Также в работе предложены инструменты анализа стабильности ключевых параметров качества автомобилей, для серийной и перспективной продукции автопроизводителя. В практику процесса испытаний при проектировании автомобильной техники внедрён инструмент детекции аномалий, позволяющий выявлять несоответствия продукции по ключевым параметрам качества. Предложенный в работе метод принятия решения о качестве автомобилей по ключевым параметрам обеспечивает рационализацию выводов по результатам статистического контроля при проектировании, производстве и эксплуатации продукции. Комплексные научно-технические решения внедрены в организациях: ПАО «КАМАЗ», г. Набережные Челны; АО «АВТОВАЗ», г. Тольятти. Предложенные научно-прикладные инструменты доказали свою результативность и эффективность при проведении мониторинга качества автомобилей по ключевым параметрам. При применении инструментов мониторинга достигнут уровень снижения трудоёмкости анализа данных о качестве продукции на 40% по сравнению с ранее применяемыми методами и подходами. Применение разработанных технических решений существенно расширяет базу инструментов управления качеством в рамках действующей системы менеджмента автопроизводителя.

Методология и методы исследования

Для решения поставленных в работе задач использовались принципы всеобщего управления качеством (TQM), принципы Деминга и Шухарта, методы теории систем и системного анализа, статистические методы управления качеством, методы стохастического прогнозирования. Базы данных о процессах проектирования, производства, сервиса и маркетингового сопровождения автомобильного бренда реализованы в среде SQL. Программный инструментарий

для решения задачи алгоритмической детекции аномалий реализован в среде Jupyter Notebook на языке Python при помощи библиотек numpy, pandas, sklearn. Визуализации выполнены с использованием библиотек matplotlib и seaborn.

На защиту выносятся:

1. Концепция методологии детекции аномалий в управлении качеством процессов менеджмента предприятий автомобильной отрасли.
2. Комплекс научно-программных инструментов и алгоритмов генерации синтетических данных для тестирования инструментов, составляющих содержательную часть методологии детекции аномалий в управлении качеством процессов менеджмента.
3. Методы моделирования временных рядов, реализующие задачи детекции аномалий процессов менеджмента качества по отношению к их состояниям в предыдущие моменты времени и по отношению к их состояниям в других локациях.
4. Критерий определения аномальных состояний в процессах, входящих в СМК предприятий автомобильной отрасли, позволяющий детектировать аномалии в условиях отсутствия экспертной оценки, и исключающий, тем самым, человеческий фактор при принятии управленческих решений.
5. Комплекс научно-программных инструментов для детекции аномальных состояний в процессах, входящих в систему менеджмента предприятий автомобильной отрасли.
6. Метод принятия управленческих решений на основе полного анализа всей совокупности статистических данных о процессах, входящих в СМК предприятий автомобильной отрасли, и алгоритмически выявленных в них аномалиях.
7. Результаты комплексной апробации методологии детекции аномалий в процессах, входящих в СМК автомобильной отрасли.

Личный вклад автора

Все результаты диссертационной работы получены автором самостоятельно. Под научным руководством автора (или при его преобладающем участии) разработана методология и инструментарий детекции аномалий в процессах автомобильного производства, сервисного обслуживания и маркетингового сопровождения автомобильного бренда. Все направления исследований диссертационной работы, формулировки проблем и постановки задач обсуждались с научным консультантом — д. т. н., профессором В. Н. Козловским, что отражено в совместных публикациях, в которых основные результаты принадлежат диссертанту.

Связь работы с научными программами, темами, грантами

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (тема №FSSE-2023-0003) в рамках государственного задания Самарского государственного технического университета.

Степень достоверности и апробация результатов

Основные положения настоящей диссертационной работы в виде научных работ и тезисов докладывались на международных и всероссийских конференциях и семинарах. Результаты диссертационного исследования нашли своё отражение в сборниках научных трудов и были доложены автором на международных и всероссийских научно-практических конференциях, в том числе: Московский авиационно-космический салон МАКС-2021. II Международная конференция. Москва, 2021 г.; VIII Международная научно-практическая конференция «Инновации в информационных технологиях, машиностроении и автотранспорте». Кемерово. 2024 г.; IV всероссийская научно-практическая конференция «Энергетика, управление и автоматизация: инновационные ре-

шения проблем», Санкт-Петербург, 2024 г.; X Международная научно-практическая конференция «Актуальные вопросы экономики и управления: Теоретические и прикладные аспекты», Донецк, 2025 г.

Публикации

Основные результаты диссертации представлены в 50 научных трудах, из них — 1 монография; 33 статьи, опубликованные в рецензируемых периодических изданиях, рекомендованных ВАК, 5 статей в иностранных научных изданиях, 11 статей, опубликованных в других научных изданиях.

Структура и объем работы

Диссертация состоит из введения, семи глав, заключения, списка литературы и пяти приложений. Общий объем диссертации составляет 491 страницу, включая 122 рисунка, 76 таблиц, список литературы содержит 249 наименований.

ГЛАВА 1. КОНЦЕПТУАЛЬНЫЙ АНАЛИЗ ОСНОВНЫХ КАТЕГОРИЙ УПРАВЛЕНИЯ КАЧЕСТВОМ В КОНТЕКСТЕ ТЕХНОЛОГИЙ БОЛЬШИХ ДАННЫХ

В связи с глобальными изменениями в индустрии производства промышленных изделий и услуг, в основном связанных с цифровой трансформацией, ростом сложности продуктов и проектов, а также с повсеместным внедрением технологий гибкого управления и давлением на жёсткие стандарты со стороны изменчивой среды, в настоящее время в компаниях, ориентированных на постоянный улучшение качества, происходит сдвиг в сторону использования аналитического потенциала современных информационных систем с применением аппарата машинного обучения и интеллектуального анализа данных [16, 28, 45, 69]. Основным приоритетом и целью деятельности таких компаний является поставка ценности клиенту, что обязывает поставщика услуг анализировать все [59, 170, 223], в том числе — недокументированные потребности потребителя, связанные с эксплуатацией изделий производителя [33, 34, 96, 97, 99, 102, 104, 105, 107, 113, 117, 122, 242, 243]. Сложные и неопределённые потребности, которые часто вообще не поддаются документированию, либо их документирование затруднено, создают значительную неопределённость в процессе управления качеством продукта, в силу того, что это затрудняет формирование чётких стандартов и критериев качества. В таких условиях традиционные методы контроля качества становятся неэффективными, что требует от компаний поиска нестандартных и инновационных решений, одним из которых является интеллектуальный анализ данных [49, 82, 100, 111, 114, 129, 134, 140, 145, 149, 179, 183], позволяющий обрабатывать огромные массивы информации, что, в свою очередь, позволяет адаптировать продукцию и услуги к изменяющимся требованиям.

Вместе с тем, в условиях цифровой трансформации [142] по-прежнему

остаётся актуальным обращение к исторически известным концепциям управления качеством, которые формировались на протяжении десятилетий и служат надёжной основой для понимания того, как цифровые технологии могут интегрироваться в современные практики управления. Классические методологии, использующие статистические методы управления качеством, подчёркивают важность непрерывного улучшения процессов и ориентации на результат, а цифровая трансформация требует не только новых навыков, но и осознания традиционных принципов управления качеством [18, 19, 46, 101, 141, 147, 226], которые помогают создать структурированный подход к внедрению инноваций [223], определяя ключевые показатели, с которыми цифровые решения должны согласовываться.

1.1. Обзор концепций, методов и стандартов управления качеством

В этом разделе мы дадим обзор некоторых известных концепций качества, принятых в системе менеджмента качества, и полученных в результате их применения методов управления качеством, а также принятые в ряде стран (в том числе — в Российской Федерации) стандарты, регламентирующие это понятие.

1.1.1. Исторически сложившиеся концепции управления качеством

Ведущими идеологами управления качеством принято считать группу исследователей (в большей степени являвшихся практиками нежели академическими учёными), живших в разное время и относившихся к разным культурам, которые внесли значительный вклад в понимание и практическое руководство качеством. К ним обычно относят П. Кросби, У. Деминга, Дж. Джурана, К. Ишикаву, Г. Тагучи и А. Фейгенбаума [18, 19, 26, 27, 31, 36, 63, 64, 72, 73, 85, 88, 90, 151, 152, 168, 202, 207, 215].

Концепция Деминга. Эдвард Деминг (1900–1993 гг.), американский учёный,

статистик и консультант по менеджменту качества, видимо, был первым в историческом контексте, кто утверждал, что для управления процессами в промышленности, а также для последующей идентификации ошибок и отказов следует использовать имеющуюся статистическую информацию. Фундаментальной основой для Деминга являлось понятие рассеяния или вариабельности, которое, с его точки зрения, связано с производственными и человеческими факторами и обусловлено специальными и/или общими причинами. Согласно его представлениям, знание о теории вариабельности включает в себя понимание того, что всё, что измеряется состоит как из «системных» отклонений, получающихся по причине наличия системы (это то, что он называл общими причинами), и «несистемных отклонений» (они в его терминологии назывались особыми причинами), которые приводят к непредсказуемости процесса; а для повышения качества нужно уметь распознавать разницу между ними для устранения «особых причин», то есть возвратить процесс в управляемое или стабильное состояние. Кроме того, Деминг известен, благодаря доработанному им циклу У. Шухарта (1891–1967 гг.), который теперь называется циклом Шухарта — Деминга PDSA (Plan, Do, Check, Act) [203].

Концепция Джурана. Концепция Джозефа Джурана (1904–2008 гг.) американского инженера и консультанта по менеджменту, известна его знаменитой трилогией качества — подходе к управлению, который состоит из трёх процессов: планирование качества, контроль качества и улучшение качества. Основным для Джурана было представление о том, что качество не может иметь случайный характер, оно должно планироваться. Он описал ключевые элементы стратегического планирования качества [73]:

1. идентификация потребителей и их потребностей;
2. определение оптимальных целей;
3. создание системы измерений качества;
4. планирование процессов в соответствии с целями, определёнными выше;
5. как следствие — снижение количества ошибок.

Несмотря на то, что акцент в концепции Джурана делался на организационных мерах (прежде всего — образовательных, связанных с обучением персонала), сам Джуран с большим пиететом относился к статистическим методам Э. Деминга и активно использовал их в своей практике; он был одним из первых, кто в задачах управления качеством применил статистический принцип Парето, который со временем стал одним из ключевых методов определения основных факторов качества.

Концепция 6σ. Эта концепция была разработана в корпорации Motorola в 1986 году после того, как на неё стали массово поступать жалобы в связи с большим количеством производственного брака, выявленного в продукции. Свою популярность методология шести сигм обрела после 90-х годов, когда она стала основной стратегией General Electric. Её суть в использовании методов управления качеством, в первую очередь, статистических методов, требование использования измеримых целей и результатов, а также создание специальных рабочих групп на предприятии, осуществляющих проекты по устранению проблем и совершенствованию процессов. Концепция 6σ выстраивается на трёх принципах [72, 207]:

1. для успешного ведения бизнеса необходимо постоянно стремиться к установлению устойчивого и предсказуемого протекания процессов;
2. показатели KPI (Key Performance Indicators), характеризующие протекание процессов производства и бизнес-процессов, должны быть измеряемыми, контролируруемыми и улучшаемыми, а также отражать изменения в протекании процессов;
3. для достижения постоянного улучшения качества необходимо вовлечение персонала организации на всех уровнях, особенно высшего руководства.

Согласно концепции 6σ при реализации проектов совершенствования процессов следует использовать последовательность этапов DMAIC (Define, Measure, Analyze, Improve, Control):

1. определение целей проекта и запросов потребителей;
2. измерение процесса, чтобы определить текущее выполнение;
3. анализ и определение коренных причин дефектов;
4. улучшение процесса, сокращающего дефекты;
5. контроль дальнейшего протекания процесса.

Концепция Фейгенбаума. Идею TQM (Total Quality Managment) предложил американский исследователь и практик в области обеспечения качества А. Фейгенбаум (1920–2014 гг.) [230]. Хорошо известная и широко распространённая в практических аспектах производства товаров и услуг пирамида Фейгенбаума представляет собой кортеж, состоящий из определённых видов контроля, последовательно осуществляемых на различных этапах жизненного цикла продукции. Он включает пять уровней:

1. Первый уровень — подготовка к проектированию. Выбор методов контроля качества, оценка качества продукции различных поставщиков, разработка планов приёма материалов и оборудования, контроль измерительных приборов, предварительная оптимизация затрат на качество.
2. Второй уровень — проектирование системы обеспечения качества. Планирование системы обеспечения качества, испытание образцов изделий, определение уровня их надёжности, оценка эффективности различных методов контроля, анализ стоимости затрат на обеспечение качества.
3. Третий уровень — этап активизации системы. Разработка технологии контроля качества, обратная связь и контроль качества, разработка системы информации о качестве продукции.
4. Четвёртый уровень — этап реализации системы по стадиям. Контроль новых проектов, входной контроль материалов, контроль качества производственных процессов, анализ и улучшение производственных процессов.
5. Пятый уровень — завершающий. Комплексный контроль качества.

В соответствии с концепцией Фейгенбаума контроль качества рассматривается как вмешательство во все фазы производственного процесса: от требований потребителя, через проектирование и производство до доставки продукта потребителю.

Концепция Тагучи. В основе концепции Тагучи лежат статистические методы, разработанные японским инженером и статистиком Геничи Тагучи (1924–2012 гг.) для улучшения качества производимых товаров. Концепция Тагучи утверждает в частности, что программа постоянного повышения качества включает непрерывное уменьшение отклонений рабочих характеристик продукта относительно заданных величин, а для идентификации параметров продукта или процесса, влияющих на снижение отклонений в эксплуатации, должны использоваться статистически планируемые эксперименты [19].

Приведённый выше обзор исторически сложившихся и получивших признание концепций управления качеством, безусловно, неполон: мы ничего не сказали о концепции гибкого управления Agile, философии бережливого производства Lean и других направлениях, которые с той или иной степенью успешности применяются в различных компаниях и сферах экономической деятельности [153, 224]. Мы не претендовали на полноту, и цель нашего обзора состояла не в том, чтобы привязать свою работу (в комплементарной или, наоборот, в полемической форме) к той или иной концепции, а в том, чтобы определить место настоящей диссертации в широком спектре различных философий, теорий и концепций управления качеством.

В связи с этим, нельзя не отметить, что несмотря на существенные, подчас принципиальные различия упомянутых нами концепций, ни одна из них не отрицает важности статистических методов, и, в той или иной форме, в них присутствуют обязательные к исполнению положения:

1. показатели качества должны быть измеримыми;
2. должен производиться непрерывный мониторинг этих показателей;

3. результаты мониторинга должны подвергаться статистическому анализу.

Именно эти идеи мы реализуем в дальнейшем: в методологической (см. главу 2) и технологической (см. главы 3–6) частях диссертационной работы.

1.1.2. Основные инструменты управления качеством

Известно, что существует семь признанных методов, используемых специалистами в области управления качеством [6, 19, 38, 56, 154, 162, 212, 225, 226, 236, 237], каждый из которых играет свою роль и может называться инструментом качества.

1. Контрольный листок. Инструмент для сбора данных и их автоматического упорядочения для облегчения дальнейшего использования собранной информации.
2. Стратификация. Инструмент, позволяющий произвести селекцию данных в соответствии с различными факторами.
3. Гистограмма. Инструмент, позволяющий зрительно оценить распределение статистических данных, сгруппированных по частоте попадания данных в определённый интервал.
4. Анализ Парето. Инструмент, позволяющий объективно представить и выявить основные факторы, влияющие на исследуемую проблему и распределить усилия для её решения.
5. Причинно-следственная диаграмма Исикавы. Инструмент, который позволяет выявить наиболее существенные факторы, влияющие на конечный результат.
6. Диаграмма разброса. Инструмент, позволяющий определить вид и тесноту связи двух рассматриваемых параметров процесса.
7. Контрольная карта. Инструмент, позволяющий отслеживать ход протекания процесса и воздействовать на него, предупреждая его отклонения от предъявленных к процессу требований.

Они были отобраны из множества статистических методов союзом японских учёных и инженеров, а профессор Токийского университета и крупнейший специалист в области управления качеством К. Исикава (1915–1989 гг.) опубликовал книгу «Руководство по контролю качества», где утверждал, что с помощью семи инструментов качества можно решить до 90% проблем в производственных организациях.

Отметим, что общепризнанные формулировки этих семи инструментов были получены более полувека назад, в районе 1960–1970 гг., однако сложно не заметить, насколько они близки (иногда — до полного совпадения понятий и терминов) к инструментам анализа больших данных и машинного обучения, то есть, к областям человеческой деятельности, возникшим совсем недавно¹. Действительно: несколько старомодное выражение «контрольный листок», идущее первым номером в списке семи инструментов качества, после его корректной детализации как инструмента сбора, обработки и хранения информации означает не что иное как базу данных (или другой источник для получения информации, который, в конечном счёте, всё равно реализуется в виде базы), из которой необходимая информация извлекается при помощи SQL-запросов. Идущая вторым номером стратификация — это, в терминологии Data Science, способ ограничения дата-фрейма при помощи процедур локализации (по признакам, объектам и/или условиям), который позволяет получать из одного и того же набора данных подмножества, обладающие заданными характеристиками, необходимыми для дальнейшего анализа. Третий номер в списке инструментов, гистограмма распределения буквально в том же самом смысле воспринимается и используется в анализе больших данных как инструмент визуального наблюдения распределений, определения наиболее частотных зна-

¹ Исключение составляет лишь идущая под пятым номером диаграмма Исикавы, которая не имеет отношения к анализу больших данных, а относится, скорее, к описанию процесса и анализу его последовательных этапов.

чений, наличия или отсутствия статистических выбросов и т. д. Анализ Парето, отмеченный в списке семи инструментов четвёртым пунктом, в том же виде применяется современными аналитиками данных и решает ту же самую задачу: отбор наиболее значимых признаков и/или объектов и удаление второстепенных данных, образующих белый шум и затрудняющих анализ. Пятый пункт мы пропустим в силу сделанного выше замечания, а шестая позиция, диаграмма разброса точно так же применяется для визуализации меры парной корреляционной связи в библиотеках программирования, которые в анализе больших данных и машинном обучении применяются повсеместно на этапе, предшествующем построению моделей. Наконец, седьмая позиция, где расположены контрольные карты Шухарта, представляет для нас особый интерес, поэтому мы остановимся на ней подробнее.

Понятие контрольных карт в рамках ряда публикаций впервые ввёл американский учёный и консультант У. Шухарт [226]. В основе методологии контрольных карт лежит понимание технологического процесса как случайного стационарного процесса, подверженного воздействию дестабилизирующих факторов, которые должны быть устранены в результате деятельности технолога. Действуя в рамках этой парадигмы, операторам следует стремиться к тому, чтобы привести процесс в стабильное состояние (или, в терминах управления качеством — добиться управляемости и изолировать от внешнего воздействия), чтобы затем, путём дополнительной настройки параметров процесса, обеспечить количественно допустимый уровень качества его работы. Цель построения контрольной карты — выявление точек выхода процесса из стабильного состояния для последующего установления причин появившегося отклонения и их устранения. Характеристика процесса всегда имеет изменчивость вследствие воздействия различных факторов, но в стабильном состоянии значения характеристики лежат в определённом коридоре, называемой зоной системной вариабельности процесса.

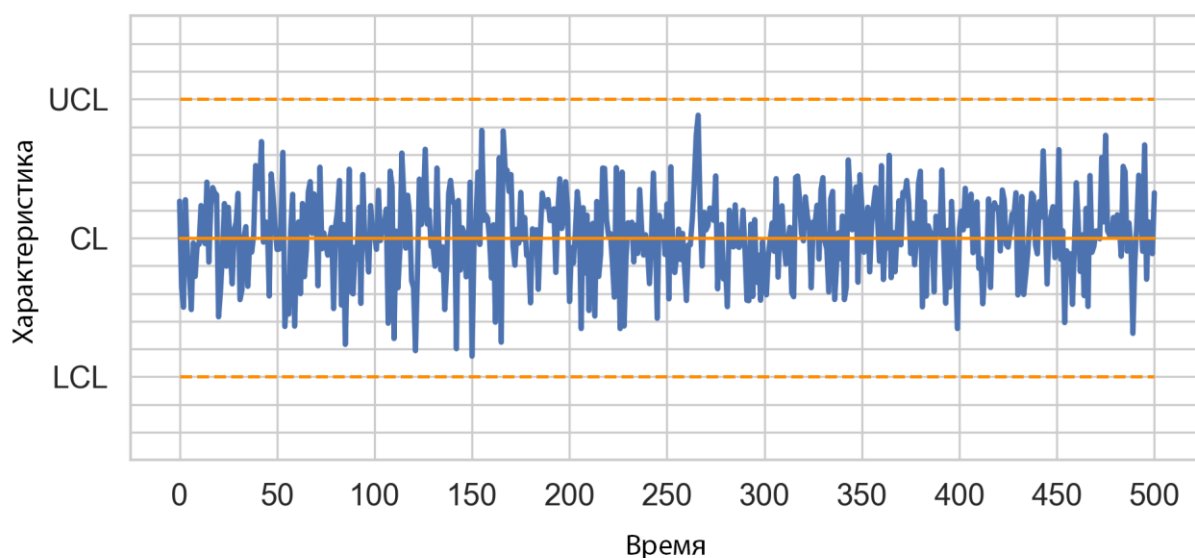


Рисунок 1.1 — Карта Шухарта. Нормальное протекание стабильного процесса

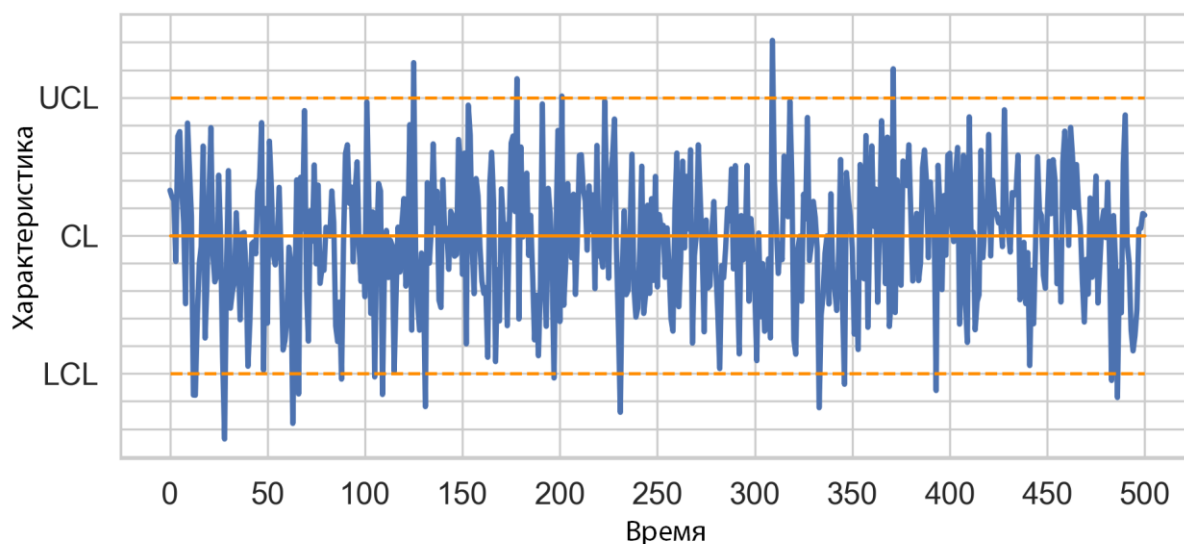


Рисунок 1.2 — Карта Шухарта. Выход стабильного процесса за контрольные границы

На картах присутствуют: CL — центральная линия; LCL — нижняя контрольная граница; UCL — верхняя контрольная граница; где контрольные границы фиксируют коридор стабильного состояния процесса. Если все наблюдаемые значения измеряемой характеристики лежат в пределах этого коридора, то протекание процесса считается нормальным (см. рисунок 1.1), а ситуации, когда какие-то значения выходят за их пределы (см. рисунок 1.2), сигнализируют о нарушении стабильности и управляемости процесса.

Возвращаясь к начатой нами выше параллели между традиционными

инструментами управления качеством и методами анализа больших данных, отметим, что в настоящей диссертационной работе мы, в числе прочего, получим некий инструмент (см. главу 2), обобщающий подход Шухарта на произвольные нестабильные процессы.

1.1.3. Стандарты управления качеством

На сегодняшний день большинство развитых стран мира имеют свои национальные институты и системы стандартизации:

Великобритания — BSI (British Standards Institution), система была основана в 1901 г.;

США — NIST (The National Institute of Standards and Technology), система была основана в 1901 г.;

Германия — DIN (Deutsches Institut für Normung), система была основана в 1917 г.;

Россия — ГОСТ (Государственные стандарты), система была основана в 1925 г.;

Франция — AFNOR (Association Francaise de Normalisation), система была основана в 1926 г.

Кроме того, в результате многостороннего соглашения в 1946 г. была создана Международная организация по стандартизации ISO (International Organization for Standardization), которая за время своего существования опубликовала более 24 тыс. международных стандартов, охватывающих практически все разделы технологий и производства, в том числе и управление качеством: соответствующие требования описываются стандартами ISO 9000. В настоящее время действует версия стандартов ISO 9000, включающая следующие основные документы:

ISO 9000:2015 «Системы менеджмента качества. Основные положения и словарь»;

ISO 9001:2015 «Системы менеджмента качества. Требования» (содержит минимальный набор требований для получения сертификата соответствия стандарту ISO 9000);

ISO 9002:2016 «Системы менеджмента качества. Руководство по применению»;

ISO 9004:2018 «Менеджмент с целью достижения устойчивого успеха организации. Подход с позиций менеджмента качества» (содержит расширенные рекомендации по постоянному улучшению деятельности организации).

С 2001 г. в России действует серия отечественных стандартов ГОСТ Р ИСО 9000, повторяющая стандарты ISO 900X. Карты Шухарта реализованы в ГОСТ Р ИСО 7870-2-2015 «Статистические методы. Контрольные карты. Часть 2. Контрольные карты Шухарта», который идентичен международному стандарту ISO 7870-2:2013.

1.2. Типология процессов в управления качеством

Управление качеством как научная дисциплина охватывает широкий спектр процессов, обеспечивающих соответствие продукции или услуг установленным требованиям и ожиданиям потребителей [59, 90, 161, 168, 203]. Типология этих процессов, используемая в данной научной области, позволяет систематизировать и оптимизировать деятельность организации для достижения устойчивого качества.

1.2.1. Общая типология

В целом весь спектр процессов, входящих в сферу управления качеством, может быть описан как следующий набор различных по своей природе и целям групп процессов [31, 64, 90, 175, 221]: основные процессы (производственные или операционные); вспомогательные процессы; управляющие процессы; совершенствующие процессы; CRM-процессы.

А. Основные процессы. Эти процессы напрямую связаны с созданием продукта или оказанием услуги, они включают в себя все этапы преобразования сырья, материалов или информации в результат, представляющий ценность для потребителя [26, 175, 221, 223]. Важнейшей характеристикой основных процессов является их стабильность, обеспечивающая предсказуемость характеристик выпускаемой продукции. К таким процессам относятся:

1. Проектирование и разработка: определение характеристик продукта, разработка чертежей, спецификаций и прототипов.
2. Закупки и поставки: выбор поставщиков, заключение договоров, контроль качества поставляемых материалов и компонентов.
3. Производство (оказание услуг): непосредственное изготовление продукта или предоставление услуги, включающее технологические операции.
4. Логистика и дистрибуция: транспортировка, хранение и доставка готовой продукции потребителю.

В. Вспомогательные процессы. Эти процессы обеспечивают корректное функционирование основных процессов, предоставляя необходимые для этого ресурсы и инфраструктуру [58, 151]. Они не создают продукт или услугу напрямую, но оказывают существенное влияние на их качество. Например:

1. Управление персоналом: набор, обучение, мотивация и оценка персонала, создание благоприятных условий труда.
2. Управление операционной инфраструктурой: обеспечение необходимым оборудованием, зданиями, коммуникациями и другими ресурсами.
3. Управление финансами: планирование бюджета, а также контроль затрат и финансовый анализ.
4. Управление цифровыми технологиями: обеспечение функционирования информационных систем, защита данных.
5. Техническое обслуживание и ремонт: поддержание оборудования в рабочем состоянии, предотвращение поломок и простоев.

С. Процессы управления. Процессы этой группы направлены на планирование, организацию, контроль и улучшение деятельности организации в целом, включая основные и вспомогательные процессы [63, 170, 229, 230]. Они обеспечивают текущую согласованность действий различных подразделений и достижение стратегических целей в области качества продукта.

1. Стратегическое планирование: определение целей в области качества, разработка стратегий и планов по их достижению.
2. Управление рисками: идентификация, оценка и управление рисками, которые могут повлиять на качество продукции или услуг.
3. Управление изменениями: планирование и реализация изменений в процессах, направленных на повышение качества.
4. Управление документацией: разработка, утверждение, распространение и хранение документации, регламентирующей процессы управления качеством.

Д. Процессы анализа и улучшения. Эти процессы направлены на улучшение деятельности организации путём сбора данных, их анализа, а также разработки корректирующих или предупреждающих действий [36, 85]. Они являются ключевым элементом системы управления качеством и обеспечивают её непрерывное и постоянное совершенствование.

1. Мониторинг и измерение: сбор данных о качестве продукции, процессов и удовлетворённости потребителей.
2. Анализ данных: выявление тенденций, закономерностей и проблем, связанных с качеством.
3. Улучшение процессами: разработка и реализация мероприятий по устранению выявленных проблем и повышению общей эффективности процессов.
4. Удовлетворённость потребителей: анализ информации о потребителях, выявление их потребностей и ожиданий.

5. Внутренние аудиты: проверка соответствия всей системы управления качеством установленным требованиям.

Е. Процессы, ориентированные на потребителя (CRM-процессы).

Эти процессы направлены на установление и поддержание долгосрочных отношений с потребителями, повышение их лояльности и удовлетворённости, они включают в себя все этапы взаимодействия с потребителем, от привлечения до послепродажного обслуживания [33, 34, 64, 96, 97, 99, 113, 102, 103, 105, 107, 113, 117, 122, 242, 243].

1. Маркетинг и продажи: привлечение новых клиентов, продвижение продукции или услуг, заключение договоров.
2. Обслуживание клиентов: предоставление консультаций, ответы на вопросы, обработка жалоб и предложений.
3. Управление взаимоотношениями с клиентами (CRM): сбор и анализ информации о клиентах, разработка индивидуальных предложений, персонализация обслуживания.
4. Обратная связь с потребителями: сбор и анализ отзывов о продукции или услугах, выявление проблем и потребностей.

Данная типология всех процессов является основой для построения эффективной системы управления качеством, позволяет достигать высокого качества продукции или услуг, удовлетворять потребности потребителей и обеспечивать устойчивое развитие.

1.2.2. Измеримость и стабильность процессов в сфере управления качеством

Приведенная выше типология была нужна нам для того, чтобы продемонстрировать огромное разнообразие процессов, входящих в систему управления качеством, и все это разнообразие можно разделить по двум принципам: по принципу измеримости процесса и по принципу его стабильности.

Например, набор, обучение, а также мотивация персонала являются неизмеримыми процессами, а оценка персонала — напротив, пример измеримого процесса, так как подразумевает наличие объективной метрики эффективности сотрудника. Создание благоприятных условий труда — неизмеримый процесс и т. д. Процессы группы А (основные процессы) являются стабильными (точнее говоря, они должны быть стабильными, и нарушение стабильности служит основным маркером, свидетельствующим о проблемах с качеством продукта или услуг), а защита данных — пример нестабильного процесса, так как DDoS-атака (от англ. Distributed Denial of Service — «распределённый отказ в обслуживании») всегда внезапна и непредсказуема. Привлечение новых клиентов или продвижение продукта на рынке — тоже нестабильные процессы, так как зависят от внешних факторов, неподвластных лицу (или подразделению), отвечающему за качество этих процессов.

В связи с этим, в рамках настоящего диссертационного исследования мы предлагаем следующую классификацию процессов.

Прежде всего, все процессы, множество которых было описано выше, мы предлагаем разделить на измеримые и неизмеримые процессы, и далее нас будут интересовать исключительно и только измеримые процессы, так как только перевод процессов в шкалы числовых характеристик позволяет работать с ними при помощи методов статистического анализа. А кроме того, все измеримые процессы мы будем разделять на стабильные, то есть такие, у которых наблюдаемые значения измеряемой характеристики демонстрируют незначительные колебания около одного и того же среднего значения [38, 56, 212, 236], и нестабильные, то есть, процессы, не обладающие этим свойством, и, как правило, отличающиеся неким неконстантным и даже, скорее всего, нелинейным трендом.

Таким образом, типология, которой будем придерживаться мы, такова:

А. Деление с позиций измеримости

А1. Измеримые процессы. Это процессы, которые имеют числовые

метрики или характеристики, и которые наблюдаются с течением времени в одной или нескольких локациях. Например:

1. Технологические процессы: металлообработка, сборка, монтаж и т. д. — все это подвергается жёстким и объективным измерениям.
2. Процессы продаж: количество заключённых сделок, их распределение по классам продукции или услуг и т. д. — измеряются и протоколируются.
3. Сервисные процессы: общее число обработанных заявок на гарантийное обслуживание, суммарные затраты на устранение дефектов по гарантии, распределение дефектов по категориям и т. д. — эти характеристики тоже допускают объективное измерение.

A2. Неизмеримые процессы. Процессы, которые либо не допускают выражения в числовых шкалах в принципе, либо такое выражение, будучи формально возможным, не отражает при этом сути процесса и является бессмысленным набором цифр. Например:

1. Стратегическое планирование: нет никакого смысла подсчитывать число проведённых заседаний или количество принятых планов.
2. Процессы проектирования: проектировщики занимаются инженерным творчеством, то есть, процессом, который с большим трудом поддаётся регламентированию, и вряд ли попытки измерить творческий процесс можно назвать хорошей идеей.
3. Создание комфортных условий на рабочих местах: какую-либо метрику комфортности нельзя предложить даже теоретически.

В. Деление с позиций стабильности

Понятие стабильности мы далее будем применять только к измеримым процессам, поэтому наша дальнейшая типология предполагает обязательный префикс A1.

A1.B1. Измеримые стабильные процессы. Это измеримые процессы,

характеристики которых с течением времени изменяются незначительно и колеблются около постоянного среднего значения.

1. Технологические процессы: металлообработка, сборка, монтаж и т. д. — во-первых, измеримы, а во-вторых, они, по определению, должны всегда приводить к одним и тем же результатам, то есть, они стабильны.

При этом подавляющее большинство процессов, не относящихся к категории технологических, не являются стабильными.

A1.B2. Измеримые нестабильные процессы. Это процессы, которые являются объективно измеримыми, но нестабильны по своей природе, то есть обладают неконстантным и чаще всего нелинейным трендом.

1. Процессы производства, рассматриваемые с социальной точки зрения: эффективность того или иного сотрудника (или подразделения) является измеримой величиной, но в социальном контексте (в отличие, как это отмечалось выше, от технологического) производственные процессы не являются стабильными, так как на них оказывают влияние уже не технологические, а ряд социальных факторов (самочувствие работника, взаимоотношения в трудовом коллективе, условия труда и т. д.).
2. Сервисное обслуживание: этот процесс является измеримым, в качестве метрики можно использовать, например, число заявок на сервис по гарантии, обработанных в течение дня, но это число нестабильно, и не может быть стабильным, потому что оно зависит от внешних факторов, прежде всего, от потребностей самих владельцев автомобилей, которыми невозможно управлять.
3. Маркетинговая кампания: охват аудитории, количество просмотров, объем переходов на сайт или канал — все это измеримые показатели, но эффективность маркетинговой кампании по определению нестабильна, так как она зависит от большого числа внешних факторов, таких как сезонность, экономическая ситуация, действия конкурентов и т. д.

4. Взаимодействие с поставщиками: время ответа поставщика на запрос или процент брака в партии — это измеримые величины, которые можно принять в качестве характеристик процесса, однако, стабильность этих показателей может колебаться из-за различных факторов: загруженности поставщика, проблем с логистикой, изменений в ценах и т. д.

Предметом нашего интереса будут измеримые нестабильные процессы, то есть те процессы, которые в проведённой нами классификации относятся к категории A1.B2. Везде в дальнейшем, говоря о каком-либо процессе, мы, если это специально не будет оговорено, будем иметь в виду именно такие процессы.

1.3. Детекция аномалий в измеримых нестабильных процессах

В этом разделе мы рассмотрим подход к продукту, характерный более для научного исследования некоего неизвестного ранее явления, чем для набора правил, инструкций и предписаний по производству. С одной стороны, это принципиально усложнит анализ процессов и результатов по сравнению с рассмотренными нами выше, в разделе 1.1, традиционными концепциями статистического управления качеством. Но с другой стороны, мы получим возможность контролировать процессы, не предполагающие стабильности, и, тем не менее, исключительно важные с точки зрения качества конечного продукта.

1.3.1. Продукт как совокупность процессов

Прежде всего, мы хотим разделить такие концептуально базовые понятия как изделие и продукт, так как хорошо известно, что в современных условиях под продуктом понимается широкий, едва ли не всеобъемлющий спектр явлений (от продуктов питания до туристических или образовательных продуктов) [72, 207]. Возможно, это продиктовано спецификой связей в обществе потребления, в котором все разнообразие общественных отношений строится на торговле и извлечения прибыли, возможно — какими-то другими причинами, но

так или иначе, разнообразие продуктовых типов огромно. Мы рассмотрим несколько максимально далёких друг от друга примеров с тем, чтобы подчеркнуть свою мысль: продукт — это нечто большее, нежели изделие.

Пример 1. Рассмотрим, например, продукт, который условно назовём «Брендовые кроссовки». Вообще говоря, это просто обувь для спорта или повседневного использования. Она производится на высокотехнологичных предприятиях, проходит приёмку, обладает целым рядом эксплуатационных характеристик: прежде всего, конечно же прочностью, износоустойчивостью, способностью отталкивать влагу и т. д. и т. п.; и, безусловно, процесс её производства исключительно важен, потому что в процессе производства и возникает то самое изделие, вокруг которого формируется продукт.

Но идентичен ли продукт изделию? Ответ отрицательный. Не стоит верить людям, которые утверждают, что «не готовы платить за этикетку» и приписывают значение имени бренда при покупке товара. Маркетинговые исследования чётко показывают, что покупатели в основном не способны переключиться на неизвестные марки даже в том случае, если аналог ничем не уступает брендам и дешевле их более чем на 20%, так как видят в нем высокую эмоциональную ценность.

Пример 2. Второй пример — некий образовательный курс (все равно, по какой программе) на какой-либо платформе. Это образовательный продукт и, на первый взгляд, в этом примере трудно произвести фрагментацию и отделить изделие от продукта, но это, разумеется, только для тех, кто далёк от образования. Нам же очевидно, что изделием в данном случае служит контент: записанные заранее видео, подготовленные заранее печатные материалы, онлайн-тесты и т. д. Однако, то, что мы называли образовательным изделием (хотя так и не принято выражаться) — это ещё не продукт. Для того, чтобы возник образовательный продукт, необходимо взаимодействие между студентом и куратором, потому что образование по самой своей сути состоит именно в специфической форме коммуникации, в ходе которой студент повышает свою

профессиональную квалификацию.

Фактически, так как студент уже оплатил своё образование ещё до начала курса, речь идёт о постпродажном обслуживании изделия, и только в тот момент, когда это обслуживание осуществляется, возникает собственно образовательный продукт. В этом примере студент платил не за контент, а за сервис вокруг контента.

Пример 3. Рассмотрим, наконец, с этих позиций такой объект как смартфон. Нам хорошо знакомо это устройство: компактный универсальный компьютер с полноценными пользовательскими интерфейсами и интерфейсами связи. Его производство — одна из технологических вершин, достигнутых цивилизацией, и для нас смартфоны, которые мы видим в магазинах или на страницах электронных витрин представляют собой весьма широкую совокупность возможностей за определённую цену. Иногда это сдобрено ореолом бренда, иногда рекламой, а иногда той самой, приятной и соблазнительной функцией, ради которой мы тратим гораздо больше запланированного бюджета. Так или иначе, это уже готовый продукт.

Разделить изделие и продукт в этом примере гораздо сложнее, чем в рассмотренных нами выше ситуациях, но разделение все же есть. Можно было бы сказать, что смартфон как изделие — это то, что мы держим в руках, включая предустановленную операционную систему, а смартфон как продукт определяется его интерфейсами, и в этом смысле он близок предыдущему примеру с образовательным курсом, но это было бы не совсем верно, потому что давление бренда в случае со смартфоном настолько велико, что он, скорее, походит на пример с кроссовками.

Перечень таких и других подобных им ситуаций неограничен, но уже приведённых нами примеров вполне достаточно для того, чтобы отметить, что создание продукта подразумевает наличие, как минимум, трёх процессов:

1. процесса производства, в результате которого возникает изделие;

2. процесса постпродажного обслуживания, который определяет уровень сервиса вокруг него;
3. маркетингового процесса, который формирует бренд.

При этом мы делаем сильное упрощение: процесс производства — это не один процесс, а сложнейшее переплетение множества технологических, логистических, управленческих и проч. процессов, и то же в равной степени относится к процессам сервиса и маркетинга. Но в данный момент нас интересует не точное описание этого явления, а его структурная основа, состоящая в том, что продукт формируется в ходе процессов, причём, эти процессы имеют разную природу.

Технологические процессы производства нам подконтрольны. Точнее, они были бы нам подконтрольны, если бы мы жили в идеальном мире. В реальности же в каждом узле процесса производства происходят какие-то сбои, что-то идёт не так, и если пустить всё на самотёк, то, в конечном счёте, это неминуемо приведёт к полной разбалансировке производства; а управление качеством в производстве как раз и состоит в выявлении и устранении причин подобных нарушений: технологических, кадровых или организационных. Но так или иначе, мы можем влиять на технологические процессы производства, чтобы удерживать их в границах допустимых отклонений — в этом смысле у нас есть возможность контролировать процессы, причём, повторимся: в идеале этот контроль является полным.

В процессах сервиса и маркетинга все не так просто. Сервис вокруг изделия — это оказание услуги, а процесс сервиса во многом определяется тем, что это процесс потребления, а не производства. Ещё в большей степени это относится к маркетинговым процессам, в ходе которых также происходит потребление, но на этот раз не материального объекта или услуги, а контента. Контролировать поведение потребителя мы не можем. Точнее говоря, этот контроль желателен, и косвенно он даже всё-таки осуществляется за счёт мар-

кетинговых процессов (вопрос — насколько эффективно, но это другой вопрос), но полного контроля, в отличие от процесса производства нет, и его нельзя добиться даже в идеале. На этот счёт у маркетологов даже есть своеобразная поговорка, продиктованная, видимо, ковбойской мифологией США: «Можно привести лошадь на водопой, но нельзя заставить её пить». Поведение потребителя, и как следствие, все мыслимые и немыслимые показатели в процессах оказания различных сервисных услуг и проведения маркетинговых кампаний зависят от множества факторов, на которые мы не можем повлиять, начиная с неподвластной нам погоды: «Сегодня идёт дождь, и я понимаю, что мне давно нужно съездить в сервисный центр, но сегодня я всё-таки не поеду, а поеду в следующее воскресенье, ничего страшного не произойдёт», — и заканчивая конъюнктурой рынка, которая нам тоже неподвластна.

Таким образом, если в процессе идеального производства все идёт так, как мы решили с самого начала на этапе планирования проекта, то в процессе идеального сервисного обслуживания и, тем более, маркетинга все идёт так, как решил потребитель, и у нас нет прямой возможности потребовать от него выполнения тех или иных плановых показателей.

1.3.2. Алгоритмический подход к детекции аномалий в нестабильных процессах

Подчёркивая разницу между технологическими процессами производства с одной стороны и процессами сервисного обслуживания и маркетинга с другой стороны, отметим следующее:

1. технологические процессы по своей природе детерминированы и непосредственно контролируемы;
2. процессы сервисного обслуживания и маркетинга по своей природе являются процессами потребления, они имеют стохастический характер и контролируются лишь опосредованно;

и именно поэтому к нормальным и аномальным состояниям различных процессов следует подходить с разных позиций.

Ответить на вопрос о том, какое состояние является нормальным в технологическом процессе, а какое — аномальным отклонением от нормы, не составляет большого труда, так как ответы на все такие вопросы уже есть: они даны в техническом задании. Стандартная величина любой измеряемой характеристики процесса производства указана в ТЗ как и допустимые отклонения от неё, и если в результате контрольного измерения мы получаем эту величину плюс-минус небольшое отклонение, то это означает, что технологический процесс находится в нормальном состоянии. Но если задать тот же вопрос относительно состояния сервисного процесса — нормально оно или аномально — мы не сможем дать ответ, заглянув в справочник, так как таких стандартов просто не существует. Нельзя сказать, что такая-то станция технического обслуживания должна в день обрабатывать столько-то дефектов такого-то типа, потому что это не мы, а потребитель нашего продукта решает, что должна станция технического обслуживания, и чего она не должна. Ещё большая неопределённость кроется в процессах маркетинга.

Таким образом, говоря о процессах сервисного обслуживания и маркетингового сопровождения бренда, мы имеем в виду процессы, которые протекают по непонятным, непредсказуемым и неизвестным нам законам, зависящим от множества неподвластных нам факторов, часть которых тоже неизвестна. Звучит пессимистично. Но, тем не менее, все не так уж плохо. Дело в том, что у нас есть возможность наблюдать эти процессы, причём, число наблюдений практически ничем не ограничено — и в этом наше преимущество, потому что в нашем распоряжении есть технологии интеллектуального анализа больших данных [2–5, 7–9, 17, 35, 48, 204, 231], которыми мы, несомненно, воспользуемся, но чуть позже, когда перейдём сначала к методологической, а затем к алгоритмической части настоящей диссертационной работы.

Пока же представим себе некий абстрактный процесс изменения во времени некой абстрактной величины, который мы наблюдаем. Наблюдение в

данном случае является точным синонимом измерения, так как, с научной точки зрения, наблюдать процесс — значит, иметь возможность проводить измерения, и для простоты будем считать, что интервалы между соседними измерениями постоянны. Это значит, что в нашем распоряжении есть временной ряд, а наша задача — выяснить в какие моменты времени этот ряд находился в аномальных состояниях. Вопрос: можем ли мы это сделать, если принять во внимание, что нам вообще ничего не известно о семантике наблюдаемого нами процесса, и все, что у нас есть — это очень длинная строка чисел (или очень высокий столбец). Ответ положительный. Во-первых, детекция аномалий — это одна из типовых задач анализа временных рядов [17, 22, 35, 65, 172, 218], в ней нет ничего необычного, хорошо известно, как она решается, а во-вторых, так поступали очень многие исследователи задолго до нас, причём, в гораздо более запутанных ситуациях. Несовпадение наблюдаемых значений с расчётными прогнозируемыми значениями какого-либо процесса даёт повод выдвинуть гипотезу об аномальности состояния, в котором оказался наблюдаемый процесс, в этом суть предиктивной детекции аномалий — приёма, который широко применяется в, например, астрофизике и других областях науки при изучении недоступных объектов.

При этом отметим, что наша задача гораздо проще, так как от нас не требуется теория, объясняющая аномалию, требуется лишь её констатация без объяснения причин. Нам достаточно иметь алгоритм, позволяющий прогнозировать состояние процесса, после чего у нас появится возможность сравнивать прогнозируемые значения с реальными и оценивать величину отклонения. Этот алгоритм и станет нашим предиктивным детектором.

1.3.3. Детекция аномалий по времени и по локации

Если мы изучаем нестандартизированные процессы, а, как мы выяснили выше, есть довольно-таки широкий класс процессов, относящихся именно к такой категории, то мы, к сожалению, не можем воспользоваться готовыми рекомендациями, и нам нужно самостоятельно принимать решение о том, что есть

норма, и что есть аномалия. Для этого нам нужно иметь достаточное количество наблюдений за процессом, другими словами, достаточно глубокую историю, так как норма и аномалия образуют взаимодополняющую пару и различаются только в сравнении. Норма и аномалия — две философские категории, определяющие порядок и хаос, их невозможно определить в отрыве друг от друга. Под нормой понимается нечто ожидаемое, а аномалия — это отклонение от нормы. Чтобы пояснить свою мысль, рассмотрим ещё несколько примеров из различных областей человеческой деятельности, когда нам приходится, ориентируясь на те или иные критерии констатировать нормальность или аномальность.

Пример 4. Каждому из нас нетрудно представить разговор с кем-либо из своих студентов: «Объясните, что происходит? Вы всегда хорошо учились. Почему сейчас у вас одни двойки?» Мы апеллируем к прошлому студента, и считаем, что происходит что-то ненормальное, потому что наш предыдущий опыт противоречит тому, что мы наблюдаем в данный момент времени. Но можно представить и другой разговор: «У других учащихся нет таких проблем с успеваемостью, как у вас», и это уже совсем другой подход, в котором мы апеллируем не к прошлому, а к другим студентам, то есть не к опыту, а к наблюдаемым значениям характеристики процесса обучения, зафиксированным в других локациях.

Пример 5. Видимо, наиболее показательный пример доставляет процесс погодных изменений. Когда мы говорим о температурных аномалиях, мы имеем в виду, что таких значений не было за всю историю наблюдений, или что они наблюдались крайне редко и не типичны для данной местности. Вместе с тем, мы называем Оймякон полюсом холода не потому что минус сорок — это что-то необычное для Якутии, в этом как раз нет ничего необычного, зимой там всегда минус сорок, а потому что в других регионах так не бывает, или бывает крайне редко. В этом примере, так же, как и в рассмотренном выше, мы снова видим двоякий подход к констатации аномальности: аномалия фиксируется

либо по отношению к прошлому, и это детекция аномалии по времени, либо по отношению к географии, и это детекция аномалии по локации.

Пример 6. Более близкий, и даже непосредственно относящийся к теме настоящей диссертации пример — процесс сервисного обслуживания. Пусть для простоты единственная наблюдаемая характеристика, которая нас интересует — это число заявок, обработанных тем или иным сервисным центром в течение рабочего дня. Мы можем сравнивать наблюдаемое значение с предыдущими днями и делать из этого выводы о нормальности или аномальности состояния процесса в данный момент времени. А в силу того, что сервисный центр у нас не один, их у нас целая сеть, мы можем сравнивать характеристики разных локаций, и это тоже даёт нам сведения о нормальности или аномальности наблюдаемой характеристики, но уже с другой точки зрения.

Итак, на интуитивном уровне, говоря об аномальности какого-либо состояния, мы ориентируемся либо 1) на свой предыдущий опыт: состояние аномально, потому что раньше такого не было; либо 2) на наблюдения в параллельных локациях: состояние аномально, потому что в других местах не так. С позиций методологии сформулированный выше интуитивный критерий детекции аномалий означает следующее: 1) зная значения измеряемой характеристики в предыдущие моменты времени, мы ожидаем увидеть некий результат в данный момент, но видим что-то совершенно другое; либо, 2) зная значения измеряемой характеристики в параллельных локациях, мы ожидаем увидеть некий результат в данной локации, но тоже видим что-то сильно отличающееся от наших ожиданий. Ожидание в данном контексте означает не что иное как прогноз, и нам остаётся только исключить из наших рассуждений интуицию и перейти к рациональному моделированию.

1.3.4. Модель как альтернативная версия стандарта

Рациональное, а не интуитивное моделирование предполагает, что мы оперируем данными, а не ощущениями. В простейшем случае данные представляют

собой одномерный набор наблюдений какой-либо одной характеристики в какой-либо одной локации, то есть, на входе мы имеем временной ряд её значений, а наша задача — выяснить, в какие моменты наблюдаемая характеристика принимала аномальные значения.

Мы считаем, что процесс, о котором мы говорим относится к категории измеримых нестабильных процессов, так как именно они составляют основное содержание настоящей диссертации². При этом мы не располагаем никакими сведениями о его семантике, и таким образом, все, что у нас есть — это ряд чисел, который, впрочем, можно визуализировать (см. рисунок 1.3).

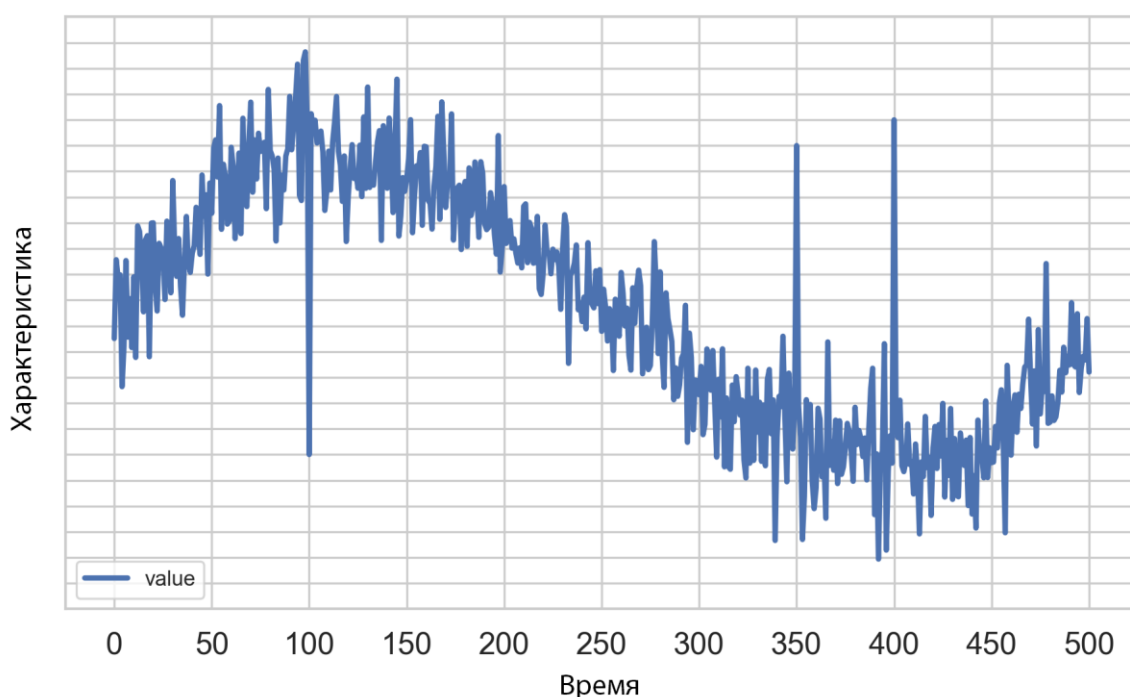


Рисунок 1.3 — Нестабильный процесс

С точки зрения детекции аномалий в ходе протекания этого процесса, первое, что приходит в голову — посчитать среднее наблюдаемых значений и

² Строго говоря, мы не требуем обязательной нестабильности. В принципе, процесс может быть стабильным, или частично стабильным на отдельных участках с потерей стабильности на других участках. Мы акцентируем внимание на нестабильности, так как спрогнозировать нестабильный процесс гораздо сложнее, чем стабильный, и именно эту сложность нам предстоит преодолеть.

использовать эту величину в качестве критерия нормальности. Тогда если очередное значение наблюдаемого процесса окажется слишком далеко от вычисленного нами среднего значения, мы скажем, что оно аномально³. Такой подход практически воспроизводит методологию карт Шухарта [6, 19, 56, 162, 226] с той лишь разницей, что мы не извлекаем стандартное значение центральной линии из технического задания, а вычисляем его сами, но мы точно так же применяем эталонную среднюю линию и две умозрительные границы: верхнюю и нижнюю (см. рисунок 1.4, а также рисунки 1.1 и 1.2).

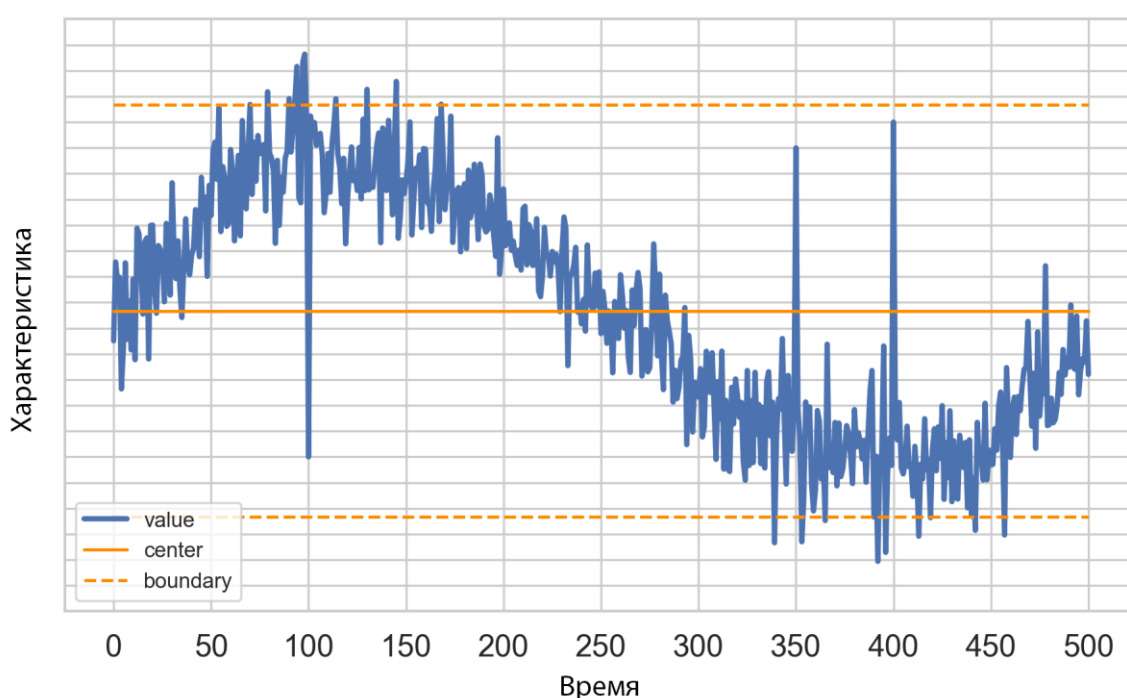


Рисунок 1.4 — Контроль нестабильного процесса по среднему значению

На приведенной выше иллюстрации мы видим, что есть большой набор

³ Можно, конечно, поступить и по-другому: например, засечь исторический максимум или исторический минимум и считать, что все значения наблюдаемой характеристики, находящиеся внутри этого коридора, являются нормой, а те значения, которые выходят за эту границу, аномальны. Но и в случае со средним значением, и в случае с историческим максимумом, и в случае с историческим минимумом — во всех этих трех ситуациях речь идет о фиксации некой константы: норму или аномалию мы констатируем в сравнении с чем-то постоянным и неизменным.

значений, выходящих за контрольные границы: это, во-первых, несколько значений в интервале примерно от 50 до примерно 150, которые нарушают верхнюю допустимую границу, а во-вторых — значения в интервале примерно от 375 до 475, которые нарушают нижнюю границу. Однако, нетрудно заметить, что значения, нарушающие границы, скорее, являются типичными для своих интервалов. Гораздо более подозрительными являются значения с номерами 100, 350 и 400. Несмотря на то, что они находятся внутри контрольного коридора, они слишком не похожи на те значения, которые их окружают, и это приводит к мысли, что, если процесс не является стационарным, а в нашем случае это именно так, то контролировать его по среднему значению — это не самая хорошая идея.

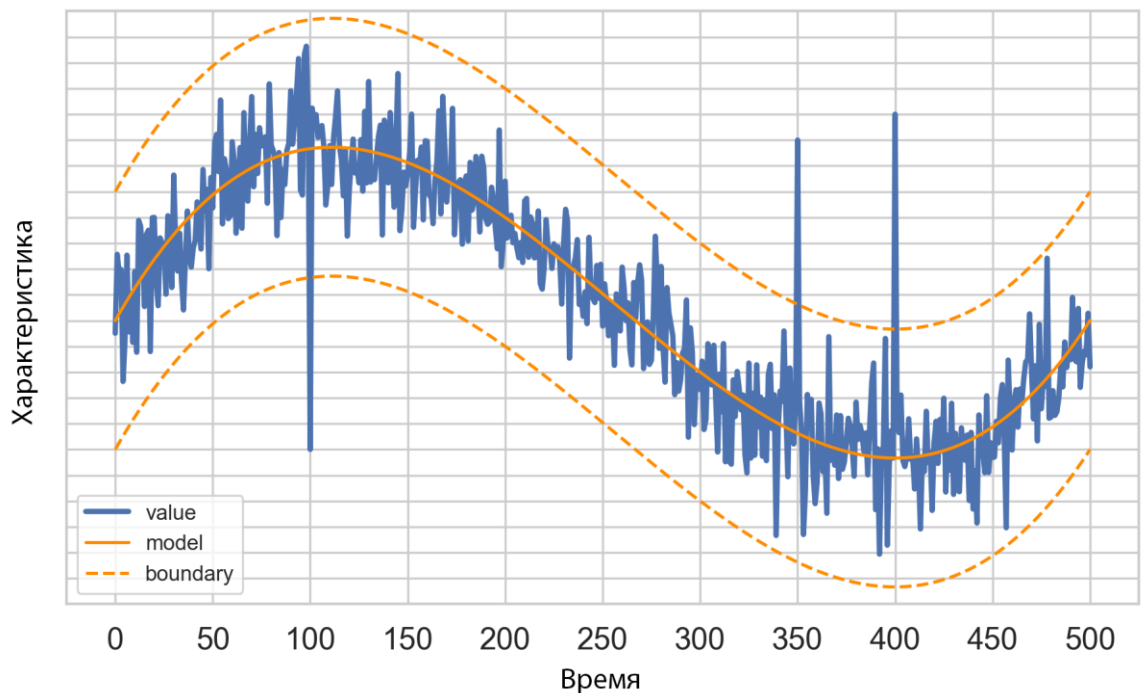


Рисунок 1.5 — Контроль нестабильного процесса по модельным значениям

В качестве альтернативы среднему значению мы можем использовать какую-либо аппроксимирующую модель⁴, и построить контрольный коридор,

⁴ Заметим, что аппроксимирующая модель — это совсем не то же самое, что прогнозирующая модель. В рамках настоящей диссертационной работы нам предстоит строить именно прогнозы, но пока мы берём аппроксимацию, чтобы проиллюстрировать основную идею детекции аномалий.

привязанный к ней, как к центральной линии (см. рисунок 1.5). Тогда значения, детектированные ранее как аномальные, перестают быть таковыми, и наоборот: аномальными оказываются значения, ранее считавшиеся нормальными. Кроме того, существенно уменьшается ширина контрольного коридора, а это демонстрирует более высокую стабильность процесса, о которой так часто говорил Э. Деминг [203], только стабильность понимается не относительно константы, а относительно модели. А самое главное — моделирование снимает вопрос, ответ на который нам не известен, но который, тем не менее, является ключевым в области управления качеством: что есть стандарт для нестабильного по своей природе процесса, и как его определить, особенно если он не был задан изначально на уровне проектирования процесса при помощи соответствующих записей в техническом задании?

1.3.5. Обобщающие свойства модельного подхода

На самом деле, вопрос, заданный в конце предыдущего раздела, имеет ответ в терминах обычной константной стандартизации.

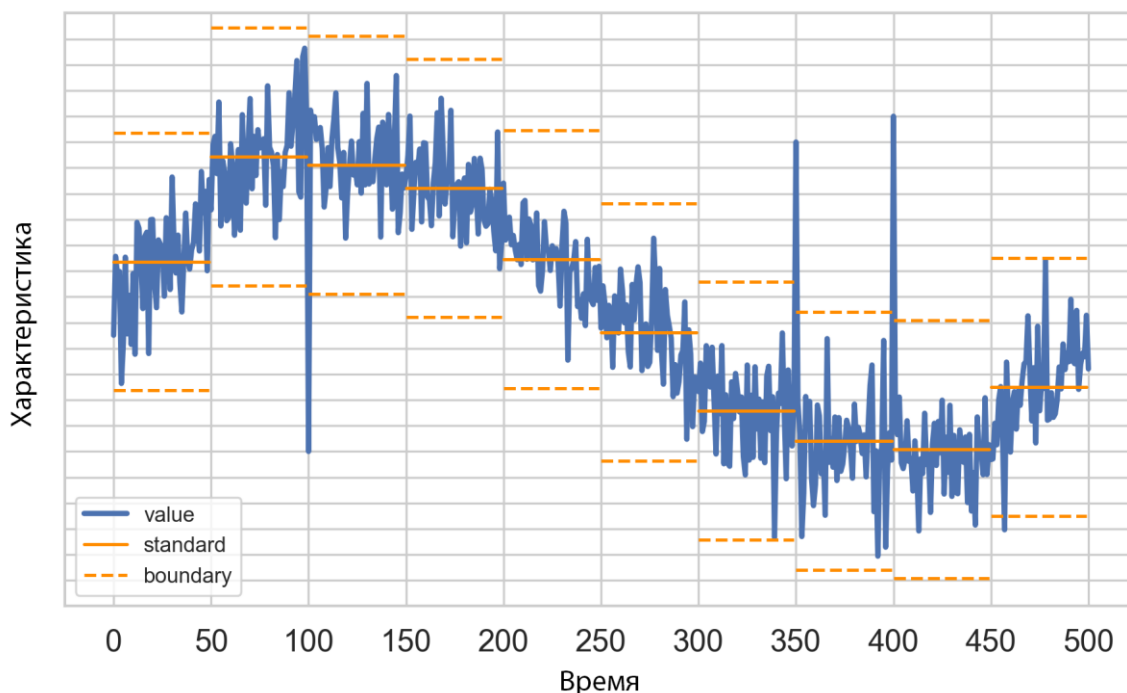


Рисунок 1.6 — Контроль нестабильного процесса по серии константных стандартов

Он состоит в том, что для таких процессов вместо одного константного стандарта нужно использовать серию константных стандартов, разбив множество всех возможных ситуаций на мелкие блоки. Например, в той максимально упрощённой схеме, которую мы сейчас рассматриваем, единственный входящим параметром является время измерения, а это значит, что можно разбить весь временной интервал от 0 до 500 на несколько, допустим, на 10 частичных интервалов и заранее оговорить стандартное значение наблюдаемой характеристики и величину допустимого отклонения на каждом из них (см. рисунок 1.6). Это и будет серия константных стандартов, вполне удовлетворяющих задачам контроля процесса: для такого-то периода времени стандартным является такое-то значение характеристики, для такого-то — другое и т. д.

Если вспомнить пример «Погода» (см. выше, раздел 1.3.3), то нельзя не отметить, что подобным образом поступают, например, в метеорологии, вычисляя среднемесячную температуру для какой-либо местности. Что это, как не серия из двенадцати температурных стандартов, характерных для данной географической локации? Впрочем, тот же пример заставляет нас вспомнить и о локациях, ведь такую серию стандартов нужно указывать для каждой из них, а все это вместе образует уже солидный фолиант с наборами цифр. Хорошо ещё, что метеорологические процессы, как правило, обладают поразительной годовой устойчивостью, и их не приходится переписывать каждый год.

Для процессов, не обладающих столь ярко выраженной сезонной воспроизводимостью, такой подход заведомо является тупиковым, в силу ряда причин, в частности, в силу того, что стандартизация — это не только аналитическая процедура, но также и бюрократическая. Можно, оперируя мощными вычислительными ресурсами, сформировать тот самый фолиант со множеством константных стандартов для всех возможных ситуаций, причём, можно сделать это достаточно быстро, допустим, за сутки, но нельзя за те же сутки принять и утвердить все эти новые, быстро пересчитанные стандарты.

Учитывая сказанное, модельный подход к стандартизации процессов,

которые являются нестабильными по своей природе, представляется предпочтительным. Кроме того, серия константных стандартов, по сути, тоже является моделью, только она описывается не гладкой функцией с красивым графиком (см. рисунок 1.5, где для построения модели использовалась кубическая полиномиальная регрессия), а кусочно-постоянной функцией, которая терпит разрывы в точках сопряжения сегментов (см. рисунок 1.6), то есть, она является частным случаем модели вообще. А если, используя модельный подход, вычислить стандартное состояние стабильного процесса (или процесса, который мы считаем стабильным, хотя, возможно, он и не является таковым), то получится горизонтальная линия (см. рисунок 1.1, 1.2 и 1.4), которая описывается регрессионной полиномиальной моделью нулевой степени, то есть, константой [2–5, 7–9, 13, 17, 48, 52, 172, 200, 218, 240].

1.4. Процесс в формате MLC (Moment, Location, Characteristic)

Напомним, что под наблюдением процесса мы понимаем не праздную рефлексию, а возможность его измерения, причём, мы предполагаем, что 1) измеряется какая-то величина (или несколько величин) в числовой шкале, 2) измерения проводятся через равные промежутки времени, 3) измерения проводятся на некотором множестве локаций. Наблюдать процесс — значит, иметь все эти данные, а сам процесс можно считать неотличимым от результатов измерения, другими словами, мы считаем, что данные о процессе — это и есть процесс.

Наиболее показательный пример, иллюстрирующий такой взгляд на вещи, доставляет «Погода» (см. выше, раздел 1.3.3), так как этот пример близок и понятен каждому из нас по повседневной жизни. Что измеряется в метеорологии? — температура, влажность воздуха, атмосферное давление, скорость и направление ветра и т. д., то есть целый спектр различных характеристик. С какой периодичностью производятся измерения? — каждые 6 часов, это международно признанный метеорологический интервал. В каких лока-

циях они производятся? — на станциях метеорологических наблюдений, распределенных по поверхности планеты более или менее равномерно. Не столь однозначный, но, тем не менее, интересный пример — «Обучение» (см. выше, раздел 1.3.3). Что измеряется при обучении? — успеваемость студента, причём, по целому спектру дисциплин: «Математический анализ», «Линейная алгебра», «Теория вероятностей» и т. д. С какой периодичностью производятся измерения? — чёткой периодичности нет, всё зависит от конкретной рабочей программы и мнения педагога-предметника, хотя, итоговая успеваемость фиксируется в конце каждого семестра, то есть, два раза в год. В каких локациях производятся измерения? — измеряется успеваемость каждого студента, то есть, в рамках процесса обучения локацией является один отдельно взятый студент, несмотря на то, что у него нет никаких фиксированных координат, и он не привязан к какому-либо конкретному месту в учебной аудитории⁵.

Вместе с тем, понимание процесса как совокупности данных имеет свою слабую сторону, и это нужно отметить в самом начале, чтобы впоследствии уже не возвращаться к этому вопросу. Процесс как данные имеет смысл только тогда, когда измеренные характеристики действительно описывают наблюдаемый процесс. На первый взгляд, это совершенно очевидное условие, не требующее отдельных пояснений, однако, практика анализа больших данных показывает, что весьма часто владелец процесса, то есть, лицо или организация, обладающие наиболее глубокими и точными представлениями о нем, предоставляет огромное количество данных, не имеющих к нему никакого отношения (из-за чего приходится проводить дополнительный факторный анализ для определения наиболее значимых компонент), и при этом необходимые и важные сведения оказываются забытыми (что гораздо хуже, потому что тогда приходится запрашивать недостающие сведения, а это дополнительные траты времени и сил, причём, далеко не всегда они бывают результативны).

⁵ Это важное замечание. Оно ещё проявит себя, когда мы будем говорить о близости локаций, так как при этом географическое расположение не будет играть никакой роли.

В дальнейшем, если не оговорено противное, мы будем считать, что набор имеющихся у нас данных корректно описывает наблюдаемый нами измеримый процесс.

1.4.1. Форматы данных ML (Moment, Location) и LM (Location, Moment)

Все это говорит о важности формата данных, которые в ходе работы мы будем подвергать алгоритмической обработке, потому что, отметим ещё раз: с нашей точки зрения, данные, которые описывают процесс, и сам процесс — это одно и то же. Мы будем придерживаться формата ML, который связывает время, место и величину наблюдаемой характеристики.

Когда мы наблюдаем какую-либо характеристику процесса, мы получаем временной ряд, подобный тому, что был рассмотрен выше. Но так происходит в какой-то одной отдельно взятой локации, которых, на самом деле, несколько (см. рисунок 1.7).

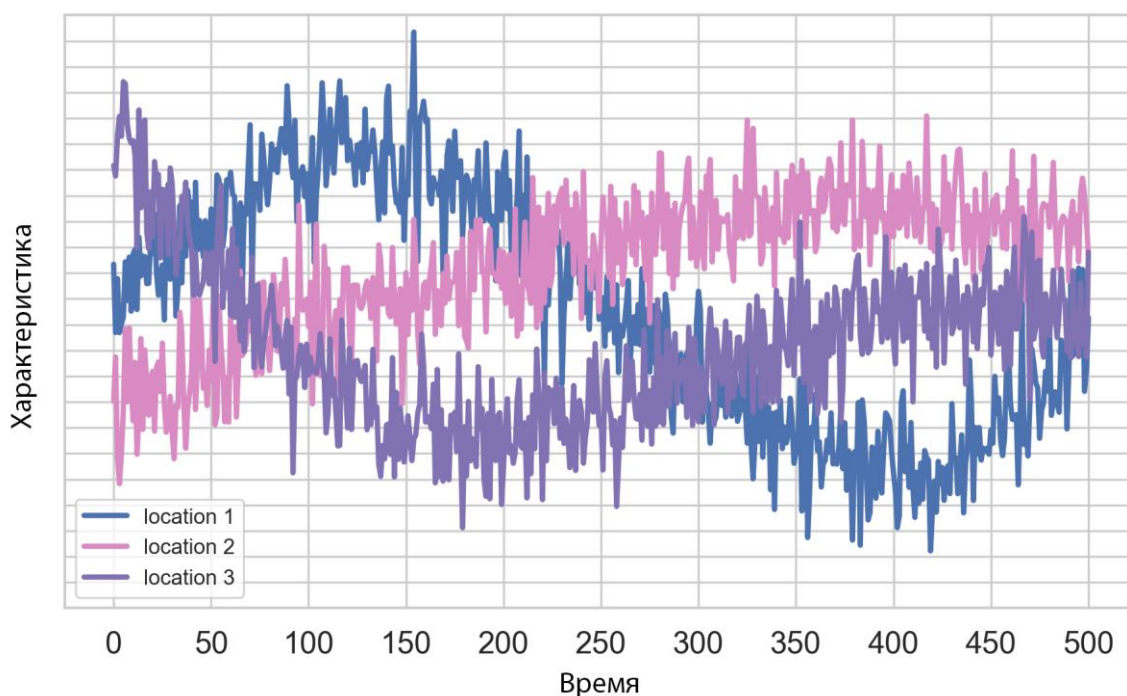


Рисунок 1.7 — Нестабильный процесс в нескольких локациях

Поэтому с визуальной точки зрения, процесс, даже если рассматривать

его в смысле изменения одной единственной характеристики, представляет собой переплетение множества кривых, число которых соответствует числу локаций, в которых он происходит. Отметим, что число локаций, как правило весьма велико, и, более того, владелец процесса может по своему произволу наращивать их до бесконечности, точнее, до того предела, превышение которого уже не приводит к существенным изменениям производственных, сервисных или маркетинговых процедур, то есть, становится избыточным.

Например, при изучении процесса обучения (см. раздел 1.2.3, пример «Обучение»), можно ориентироваться на показатели одной отдельно взятой студенческой группы, и, при желании, расширить исследование до факультета или университета в целом, но дальнейшее расширение уже вряд ли окажется полезным с практической точки зрения, хотя теоретически возможно и оно. То же можно сказать и о метеорологических процессах (см. раздел 1.3.3, пример «Погода») — метеостанций много, но нет смысла ставить их на каждом перекрёстке, потому что тогда их будет слишком много. И так далее.

Итак, допустим, что мы производим множество измерений какой-либо характеристики наблюдаемого нами процесса во множестве доступных нам локаций. Тогда в результате объединения всех измерений у нас возникает таблица чисел, которую в анализе больших данных принято называть датафреймом [2–5, 7–9, 23, 49, 173]. Для обозначения её формата мы будем использовать аббревиатуру ML (Moment, Location), а поскольку мы не делаем разницы между процессами, и данными, которые их описывают, эту таблицу мы также будем называть процессом.

Таблица 1.1 — Процесс в формате ML (Moment, Location)

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент 1	C ₁₁	C ₁₂	C ₁₃	...	C _{1m}
Момент 2	C ₂₁	C ₂₂	C ₂₃	...	C _{2m}
Момент 3	C ₃₁	C ₃₂	C ₃₃	...	C _{3m}
...	...	...	...	...	...
Момент n	C _{n1}	C _{n2}	C _{n3}	...	C _{nm}

Таблица содержит n строк, которые маркируются моментами наблюдений, и m столбцов, промаркированных локациями. И, несмотря на то, что табличное представление не обладает наглядностью отдельно взятого временного ряда, оно имеет целый ряд преимуществ. Во-первых, датафрейм — это базовая единица анализа данных, и большинство стандартных процедур анализа ориентированы именно на работу с датафреймами [4, 5, 7–9, 49, 51, 172, 201, 218], во-вторых, такой формат позволяет компактно в одном месте хранить массив данных о процессе и обращаться к той или иной локации по мере необходимости, а третьих, перечисленное выше не является главным преимуществом такого формата. Главное преимущество в том, что датафрейм, будучи по существу матрицей n на m , элементарно транспонируется, после чего формат ML (Moment, Location) становится форматом LM (Location, Moment).

Таблица 1.2 — Процесс в формате LM (Location, Moment)

	Момент 1	Момент 2	Момент 3	...	Момент n
Локация 1	C_{11}	C_{21}	C_{31}	...	C_{n1}
Локация 2	C_{12}	C_{22}	C_{32}	...	C_{n2}
Локация 3	C_{13}	C_{23}	C_{33}	...	C_{n3}
...	...	...	...	...	...
Локация m	C_{1m}	C_{2m}	C_{3m}	...	C_{nm}

На первый взгляд, это тавтология: мы просто поменяли индексы местами, и то, что раньше обозначалось c_{ij} , теперь обозначается c_{ji} . Формально так оно и есть. Но по существу эта нехитрая операция даёт совершенно разные подходы к моделированию: если в формате ML наблюдениями служат моменты времени, а локации выступают в качестве предикторов модели, и это позволяет осуществлять прогноз по времени (см. раздел 1.3.3, где мы говорили о прогнозах по времени и по локации), то в формате LM, наоборот, наблюдениями служат локации, а предикторами — моменты времени, и это позволяет делать прогноз по локации (см. там же). И, несмотря на то что прогноз по времени обладает своей спецификой, требует особого подхода к кросс-валидации

модели и внимания к другим важным деталям [2–5, 15, 23, 60, 172, 218] нас не может не радовать тот факт, что простейшая операция транспонирования обеспечивает нам реализацию сразу двух концептуальных идей. Мы подробно рассмотрим оба подхода в главе 2, посвящённой общей методике детекции аномальных состояний в нестабильных процессах, входящих в систему менеджмента качества.

1.4.2. Отступления от форматов ML/ LM

Формат ML, безусловно, обладает изящной универсальностью, которую мы только что отметили, и тем не менее, мы не считаем его неприкасаемым. Так, в формате ML объектами служат моменты времени, которые описываются значениями наблюдаемой характеристики в наборе локаций (см. таблицу 1.1). Но момент может обладать и другими свойствами, и их тоже нужно учитывать для построения более точной прогнозирующей модели. Например, мы можем знать, что данный момент времени относится к категории «Понедельник», и это серьёзно отличает его от моментов категории «Воскресенье» или «Суббота», потому что, если изучаемый процесс относится к процессам потребления, то он наверняка обладает выраженной недельной сезонностью [198]. Или момент может попасть на интервенцию⁶, что тоже сильно изменит его свойства. Важную роль могут играть месяц момента или его год, время, прошедшее с какого-то важного для изучаемого процесса события, а также другие признаки. Аналитик, занимающийся прогнозированием процесса, должен учитывать все факторы, доступные ему на момент прогноза, поэтому для усиления прогнозирующей способности модели к набору локаций всегда можно добавить дополнительные столбцы (см. таблицу 1.3).

⁶ В анализе временных рядов интервенцией называют нарушение принятого календарного режима. Например, среда — это рабочий день, но если среда приходится на восьмое марта, то это выходной, и в контексте данного временного ряда восьмое марта оказывается интервенцией.

Таблица 1.3 — Процесс в формате ML с дополнительными признаками

	Признак 1	...	Признак s	Локация 1	...	Локация m
Момент 1	f_{11}	...	f_{1s}	C_{11}	...	C_{1m}
Момент 2	f_{21}	...	f_{2s}	C_{21}	...	C_{2m}
Момент 3	f_{31}	...	f_{3s}	C_{31}	...	C_{3m}
...	...	...	...	...	...	...
Момент n	f_{n1}	...	f_{ns}	C_{n1}	...	C_{nm}

Те же соображения справедливы и для формата LM, в котором объектами выступают локации, а описывающими их признаками — моменты времени, так как локация может обладать и другими, не зависящими от времени признаками. Например, в процессе сервисного обслуживания локациями служат сервисные центры, которые заведомо не являются идентичными: один расположен в мегаполисе, а другой в посёлке городского типа; в работе одного задействовано пятьдесят сотрудников, а в работе другого едва ли наберётся пять — и т. д. Все эти факторы мы тоже можем использовать в качестве дополнительных признаков при описании процесса в формате LM.

Таблица 1.4 — Процесс в формате LM с дополнительными признаками

	Признак 1	...	Признак s	Момент 1	...	Момент n
Локация 1	f_{11}	...	f_{1s}	C_{11}	...	C_{m1}
Локация 2	f_{21}	...	f_{2s}	C_{12}	...	C_{m2}
Локация 3	f_{31}	...	f_{3s}	C_{13}	...	C_{m3}
...	...	...	...	...	...	...
Локация m	f_{m1}	...	f_{ms}	C_{1m}	...	C_{mn}

Заметим, что признаки, дополнительные для моментов времени, не зависят от локаций (понедельник остаётся понедельником во всех локациях сразу), а признаки, дополнительные для локаций не зависят от моментов времени (мегаполис был мегаполисом вчера, является мегаполисом сегодня и останется мегаполисом завтра), поэтому, с точки зрения построения прогнози-

рующей модели, применять к дополненным форматам операцию транспонирования для перехода от прогноза по времени к прогнозу по локации (и наоборот) не имеет никакого смысла — в результате мы получим лишь бессодержательные признаки, которые всё равно не окажут на модель никакого влияния, потому что всегда будут принимать одни и те же значения.

С одной стороны, это, конечно, делает процедуру анализа менее гибкой, но с другой стороны, никогда не следует пренебрегать точностью модели ради простоты метода, следует помнить, что формат ML является базой для анализа процесса, но не является догмой, и при необходимости этот формат можно и даже нужно пополнять дополнительными признаками, которые помогают полнее и точнее описать изучаемые процессы.

1.4.3. Обобщение форматов ML/ LM на множество характеристик

Трудно представить себе процесс, который бы полностью описывался какой-нибудь одной числовой характеристикой, разве что классическое теоретико-вероятностное подбрасывание монеты, но и оно является умозрительным и не рассматривает вполне реальный вариант, когда монета закатилась под диван, и результат подбрасывания остался неизвестен. Любой реальный процесс всегда описывается множеством характеристик (причем, часто их число измеряется десятками или даже сотнями). Каждую из них мы можем описать в формате ML, то есть в виде некой числовой матрицы, но это описание, в силу того, что оно учитывает лишь одну из множества возможных характеристик, заведомо будет односторонним и, следовательно, неполным.

Вместе с тем, точно так же мы можем поступить и со всеми другими характеристиками, которые описывают изучаемый нами процесс, получив, тем самым, набор датафреймов с одинаковой структурой объектов и признаков (см. таблицу 1.5, где под символами ML понимаются таблицы формата таблицы 1.1). После этого процесс в формате MLC можно рассматривать как трёхмерную матрицу с числовыми компонентами C_{ijk} , которые представляют

собой значение k -й характеристики, зафиксированное в j -й локации в i -й момент времени.

Таблица 1.5 — Процесс в формате MLC (Moment, Location, Characteristic)

	Характеристика 1	Характеристика 2	...	Характеристика k
Процесс	ML_1	ML_2	...	ML_k

То же относится и к формату LM, где в качестве объектов выступают локации, а моменты времени играют роль признаков (см. таблицу 1.6, где под символами LM понимаются таблицы формата таблицы 1.2). В случае множества характеристик формат LM преобразуется в формат LMC, который по сути также представляет собой трехмерную матрицу.

Таблица 1.6 — Процесс в формате LMC (Location, Moment, Characteristic)

	Характеристика 1	Характеристика 2	...	Характеристика k
Процесс	LM_1	LM_2	...	LM_k

Различные характеристики процесса, как правило, не связаны между собой или связаны весьма опосредованно, и эта связь неочевидна. Поэтому если одна из них в какой-то момент времени принимает аномальное значение, сильно отличающееся от прогнозируемого, то, скорее всего, это ничего не говорит о поведении других характеристик: они могут быть как аномально отклоняющимися, так и лежащими в допустимых пределах, — но в любом случае, констатируя нормальность или аномальность того или иного состояния процесса, нам нужно ориентироваться на весь доступный нам набор его характеристик для того, чтобы быть уверенным в том, что мы всесторонне рассматриваем анализируемый нами процесс.

1.5. Выводы по главе 1

В завершение этой главы в качестве краткого резюме всего изложенного выше отметим некоторые положения, на которые мы будем опираться в дальнейшей

работе, и которые подразумеваются в последующем тексте как нечто само собой разумеющееся.

1. Продукт — это совокупность процессов, среди которых выделяются три группы: 1) производства, 2) сервисного обслуживания и 3) маркетинговые процессы. Для управления качеством продукта следует управлять процессами всех трёх групп.
2. Технологические процессы производства всегда являются стабильными. Процессы производства, понимаемые в социальном контексте, то есть, рассматриваемые с точки зрения эффективности персонала, а также сервисные и маркетинговые процессы — нестабильны по своей природе.
3. Для нестабильных по своей природе процессов характерно отсутствие заранее заданных стандартов. В качестве эталонного значения процесса мы будем рассматривать не константный стандарт, а модельное значение, которое не является постоянным. Это приводит к идее алгоритмической детекции аномалий в нестабильных процессах, методику которой мы изложим в следующей главе.
4. Управление процессом означает, в частности, его контроль и детекцию аномальных состояний. Аномальность может фиксироваться либо 1) по отношению к прошлым состояниям процесса, либо 2) по отношению к его настоящим состояниям в других локациях.
5. Процесс и данные о процессе — это одно и то же. Любой процесс может быть представлен как совокупность временных рядов, описывающих поведение некоторого набора характеристик в некотором наборе локаций.

Таким образом, материалом, с которым мы будем работать, являются данные, причём, нам предстоит работать с ними не с точки зрения элементарной описательной статистики: гистограмм, средних значений или дисперсий, — а с точки зрения построения сложных регрессионных моделей, обладающих максимально высокой прогнозирующей способностью [2–5, 7–9, 13, 15,

17, 23, 49, 65, 172, 20, 209, 234]. Наши методы вполне согласуются с классическими методами управления качеством на уровне идеологии, так как в основе наших подходов, так же, как и у наших предшественников, лежит математическая статистика [6, 19, 38, 56, 154, 162, 212]. Но на уровне реализации они представляют собой нечто принципиально новое, в силу того, что их можно реализовать лишь с использованием технологий больших данных и машинного обучения [2–5, 7–9, 17, 35, 51, 172–174], что было невозможно в прошлом, и прежде такие подходы не фигурировали даже на уровне постановки задачи.

Идеи, изложенные в этой главе, закладывают основы для дальнейшего исследования и задают цель настоящей диссертационной работы, которая заключается в повышении результативности и эффективности системы менеджмента качества автопроизводителя в условиях отраслевого развития информатизации и цифровизации посредством разработки нового методологического инструментария статистического управления качеством при помощи детекции аномалий в процессах автомобильной отрасли (см. выше: введение, раздел «Цель диссертационной работы»).

ГЛАВА 2. МЕТОДИКА МССР (MODELING, CHALLENGE, CALIBRATION, PRODUCTION) КАК СРЕДСТВО АЛГОРИТМИЧЕСКОЙ ДЕТЕКЦИИ АНОМАЛИЙ В НЕСТАБИЛЬНЫХ ПРОЦЕССАХ

В условиях высокодинамичной производственной и эксплуатационной среды современного автомобилестроения традиционные подходы к управлению качеством, ориентированные на стабилизацию процессов, становятся все менее эффективными [195, 198, 199, 214]. Возникает необходимость в разработке и внедрении новых методологий, способных адаптироваться к нестабильности и изменчивости бизнес-процессов. Применение прогнозирующих моделей и иных алгоритмов машинного обучения представляет собой перспективное направление, позволяющее осуществлять управление качеством в условиях неопределенности.

Сервисные и маркетинговые процессы в автомобильной отрасли, в силу своей подверженности влиянию внешних факторов, таких как потребительские предпочтения, изменения законодательства и действия конкурентов, по определению являются нестабильными [96, 97, 99, 102, 104, 105]. Контрольные карты Шухарта и другие инструменты статистического управления процессами (SPC, англ. — statistical process control) в данном случае демонстрируют ограниченную применимость [180, 186, 188, 198]. Однако, использование алгоритмов машинного обучения позволяет выявлять скрытые закономерности и тренды в данных, что дает возможность прогнозировать изменения спроса, предотвращать отток клиентов и оптимизировать маркетинговые кампании. Не меньшего внимания заслуживает и нестабильность производственных процессов, обусловленная человеческим фактором. Несмотря на строгие технологические регламенты, вариативность в действиях персонала, отклонения в качестве сырья и оборудования могут приводить к браку, задержкам и

другим негативным последствиям. В этой связи, применение машинного обучения для мониторинга и анализа данных с производственных линий, включая информацию о работе оборудования, действиях операторов и параметрах технологического процесса, позволяет выявлять причины отклонений и оперативно принимать корректирующие меры.

Интеграция систем машинного обучения с системами управления качеством позволяет создать замкнутый контур улучшения, обеспечивающий непрерывную адаптацию к изменяющимся условиям [45, 101, 119, 143, 144–148], при этом прогнозирующие модели позволяют не только обнаруживать потенциальные проблемы на ранних стадиях, но и оптимизировать параметры процесса в режиме реального времени, что приводит к повышению эффективности и снижению затрат.

Таким образом, управление качеством нестабильных процессов в автомобилестроении представляет собой новый этап развития данной научной области, требующий переосмысления традиционных подходов и активного внедрения инновационных технологий. Этот подход обеспечивает повышение клиентоориентированности, снижение производственных издержек и повышение устойчивости бизнеса в условиях динамичной рыночной среды. Разработка и внедрение эффективных моделей управления качеством нестабильных процессов открывает новые возможности для повышения конкурентоспособности автомобилестроительных предприятий и улучшения качества выпускаемой продукции, причем, именно использование машинного обучения и предиктивной аналитикой является ключевым элементом новой парадигмы управления качеством в автомобилестроении.

Переходя к методике, при помощи которой мы предполагаем детектировать аномальные состояния нестабильных процессов, отметим некоторые специфические обстоятельства, которые будут важны для понимания всего дальнейшего изложения диссертационной работы в аспектах, касающихся реализации алгоритма.

Прежде всего, мы должны подчеркнуть, что излагаемая в этой главе методика не является инструкцией или обязательным предписанием, которое требует прямого и буквального исполнения. Напротив, она лишь указывает последовательность шагов и обозначает направления, в которых следует проводить работу, допуская при этом весьма значительные отклонения при исполнении каждого шага. Так, на первом же шаге, при формировании пула используемых моделей, у читателя может возникнуть вопрос — почему мы используем именно этот пул моделей, а не какой-то другой; и ответ состоит в том, что мы могли бы пополнить его, например, моделью полиномиальной регрессии, или, наоборот, исключить из него регрессию на ближайших соседях, а в выборе именно такого пула заложено наше представление об эффективности регрессионных моделей, и, с нашей точки зрения, оно хорошо обосновано применительной практикой. Однако, любой другой аналитик может сформировать свой собственный пул — в этом смысле у него полностью развязаны руки. Это относится и ко всем остальным шагам методики, вплоть до финального, когда заведомо возникает вопрос, какой метод проверочной тактики следует применить — с учетом или без учета выраженности аномалий; и в данном случае рекомендация может состоять лишь в том, чтобы в пилотном варианте попробовать оба варианта, после чего выбрать тот метод, который наиболее соответствует задачам бизнеса.

Второе немаловажное замечание относится к тому, что наша методика не привязана к какой-то конкретной реализации, и, хотя в дальнейшем мы будем использовать исключительно язык Python и его библиотеки, следует иметь в виду, что то же самое может быть написано на R или даже на C — важны не коды алгоритма и не язык, на котором они написаны, а идеи, которые будут изложены ниже: моделирование и оценка моделей.

2.1. Признаковые описания объектов

В первой главе настоящей работы, рассуждая о концептуальных аспектах кон-

троля качества в нестабильных процессах, мы пришли к выводу, что нам понадобятся данные в формате ML (когда в качестве объектов выступают моменты времени) и LM (когда в качестве объектов выступают локации). Сейчас мы остановимся на этих форматах чуть подробнее.

Допустим, мы наблюдаем значение одной (и пока только одной) характеристики процесса, производя её регулярные измерения в одной (и только одной) локации в течение какого-либо промежутка времени. В таком формате всё, что мы знаем о последовательности значений измеряемой характеристики — это порядковый номер очередного наблюдения. Данные такого формата образуют временной ряд, а прогнозирование временных рядов представляет собой хорошо развитый раздел машинного обучения [2–5, 7–9, 17, 35, 174]. В анализе временных рядов используется несколько основных методов, каждый из которых имеет свои особенности применения, но в самом широком смысле они делятся на две группы⁷:

1. авторегрессионные модели AR (англ. — autoregressive model), которые основываются на зависимости текущего значения временного ряда от его предыдущих значений;
2. скользящие средние MA (англ. — moving average), использующие средние значения предыдущих измерений.

В данных, описывающих временные ряды, имеется единственный предиктор — порядковый номер наблюдения (см. таблицу 2.1), и можно сразу сказать, что в большинстве случаев это слишком бедное признаковое описание (этот факт часто упоминается в литературе, например, в [17, 23, 174, 218]).

Таблица 2.1 — Формат ML: нуль-мерные данные

	y
1	y ₁

⁷ Эти две группы могут комбинироваться в авторегрессионные интегрированные модели скользящих средних ARIMA (англ. — autoregressive integrated moving average) [5, 7, 48].

	y
2	y_2
3	y_3
...	...
n	y_n

Однако, можно учесть, что второй момент времени характеризуется не только тем, что у него второй порядковый номер, и не только тем, что в этот момент в локации наблюдалось значение y_2 , но ещё и тем, что в предыдущий момент в ней наблюдалось значение y_1 . Третий момент описывается не только своим номером и значением y_3 , но и тем, что в предыдущий момент в локации наблюдалось значение y_2 , и так далее. В силу таких замечаний каждый момент времени (кроме первого) приобретает один дополнительный признак x_1 , а именно: значение характеристики, зарегистрированное в локации в предыдущий момент времени (см. таблицу 2.2). Этот приём называется авторегрессией, он распространён в практике машинного обучения и повсеместно применяется при обучении регрессионных моделей [7–9, 12, 190, 201, 240].

Таблица 2.2 — Формат ML: авторегрессия со сдвигом на одну позицию

	x_1	y
1	—	y_1
2	$x_{21} = y_1$	y_2
3	$x_{31} = y_2$	y_3
...	...	...
n	$x_{n1} = y_{n-1}$	y_n

Рассуждая аналогично, можно произвести ещё один сдвиг и получить авторегрессию с двумя дополнительными признаками (см. таблицу 2.3).

Таблица 2.3 — Формат ML: авторегрессия со сдвигом на две позиции

	x_1	x_2	y
1	—	—	y_1
2	$x_{21} = y_1$	—	y_2
3	$x_{31} = y_2$	$x_{32} = y_1$	y_3
...	...	...	...

	x_1	x_2	y
n	$x_{n1} = y_{n-1}$	$x_{n2} = y_{n-2}$	y_n

Здесь мы используем тот факт, что третий момент времени характеризуется тем, что в предыдущее наблюдение было зафиксировано значение y_2 , а в предыдущее — y_1 , и, таким образом, третий момент времени описывается двумя предикторами. То же можно сказать о четвёртом и всех последующих моментах времени, но нельзя сказать о первом и втором: формат авторегрессии со сдвигом на две позиции вынужденно доставляет первому и второму моменту времени пропуски данных. Это значит, что, увеличивая при помощи авторегрессии число признаков, мы, тем самым, уменьшаем число объектов, которые ими описываются, однако, при достаточно большом числе наблюдений такое уменьшение не является существенным.

Наконец, ничто не мешает провести сдвиг временного ряда на k позиций и получить k -мерное признаковое описание (см. таблицу 2.4).

Таблица 2.4 — Формат ML: авторегрессия со сдвигом на k позиций

	x_1	x_2	x_3	...	x_k	y
1	—	—	—		—	y_1
2	$x_{21} = y_1$	—	—	...	—	y_2
3	$x_{31} = y_2$	$x_{32} = y_1$	—	...	—	y_3
4	$x_{41} = y_3$	$x_{42} = y_2$	$x_{43} = y_1$	...	—	y_4
...	...	...	...	...	...	...
$n-1$	$x_{n-1,1} = y_{n-2}$	$x_{n-1,2} = y_{n-3}$	$x_{n-1,3} = y_{n-4}$	...	$x_{n-1,k} = y_{n-k-1}$	y_n
n	$x_{n1} = y_{n-1}$	$x_{n2} = y_{n-2}$	$x_{n3} = y_{n-3}$	...	$x_{nk} = y_{n-k}$	y_n

k -мерное признаковое описание уже весьма богато точки зрения включённых в анализ предикторов, такие данные позволяют строить и обучать регрессионные модели с хорошей прогнозирующей способностью.

Но ситуация принципиально изменится, если от формата ML мы перейдём к формату LM, когда в качестве объектов выступают локации, а в качестве признаков — моменты времени. В этом случае мы никак не сможем пополнить набор предикторов за счёт самих себя, как мы это сделали выше в случае с временными рядами, и регрессионная задача останется нуль-мерной. Пусть

мы наблюдаем значение одной (и только одной) характеристики во всех локациях сразу, но в один (и только один) момент времени (см. таблицу 2.5). Формально данные о таких наблюдениях выглядят точно так же, как и временной ряд (см. таблицу 2.1), но это только формальный взгляд на вещи.

Таблица 2.5 — Формат LM: нуль-мерные данные

	y
1	y_1
2	y_2
3	y_3
...	...
m	y_m

Принципиальное отличие этих данных от временного ряда состоит в том, что множество локаций, в отличие от множества моментов времени, не обладает естественным отношением порядка: нельзя сказать, что вторая локация следует после первой, а третья — после второй, это лишено всякого смысла. Поэтому, если в случае с форматом ML мы, прибегая к приёму авто-регрессии, можем рассматривать целевую функцию как функцию нескольких переменных в качестве которых выступают все добавленные нами предикторы, то для данных формата LM мы не располагаем такой возможностью, и целевая функция остаётся с нулём переменных, то есть — является тождественной константой. Другими словами, единственный прогноз, который нам доступен в этой ситуации состоит в вычислении среднего значения наблюдаемых значений характеристики процесса, то есть в применении самого примитивного из всех возможных регрессоров: коэффициент детерминации такого регрессора равен нулю (подробнее о метриках эффективности регрессионных моделей, в частности о коэффициенте детерминации см. прилож. А).

Такое положение дел нас не устраивает, мы хотим иметь возможность строить гораздо более точные прогнозы, а, значит, мы хотим располагать гораздо более детальными сведениями, желательно, чтобы они максимально подробно описывали изучаемый нами процесс. Когда мы говорили о данных в

формате ML, мы смогли получить такое описание за счёт авторегрессии. Но что означает авторегрессия с точки зрения структуры данных? Ответ на этот вопрос состоит в том, что в формате ML одну отдельно взятую локацию оказалось возможным описать при помощи истории наблюдений в этой отдельно взятой локации: сначала мы добавили один предыдущий момент истории, затем ещё один, и, действуя таким образом, мы получили достаточно подробное описание. Теперь поставим аналогичный вопрос для формата LM: если отдельно взятая локация описывается при помощи всех доступных нам моментов времени, то как тогда описывается отдельно взятый момент времени? Ответ даётся по аналогии — момент времени описывается при помощи локаций. Таким образом, сравнивая локации между собой для детекции аномальных состояний измеримого нестабильного процесса, в качестве признаков регрессионной задачи можно использовать сведения о состоянии процесса в этих локациях в какие-то другие, предшествовавшие текущему наблюдению моменты времени. Про такие наблюдения мы будем говорить, что мы использовали глубину истории наблюдений, равную одному, двум, трём, четырём и т. д. наблюдениям в зависимости от того сколько предыдущих моментов мы отвели под описание локаций⁸ (см. таблицу 2.6, 2.7, 2.8).

Таблица 2.6 — Формат ML: регрессия с добавлением одного исторического момента

	x_1	y
1	$x_{21} = c_{n-1,1}$	y_1
2	$x_{21} = c_{n-1,2}$	y_2
3	$x_{21} = c_{n-1,3}$	y_3
...	...	...
m	$x_{m1} = c_{n-1,m}$	y_m

Для простоты в таблице 2.6 мы добавили предыдущий момент времени, но,

⁸ Не стоит путать глубину истории наблюдений с глубиной авторегрессии, которую мы рассматривали выше. Это разные способы построения признаков описаний.

немного забегаая вперёд, скажем, что этот единственный добавленный предиктор можно было бы взять исходя из соображений близости моментов, о которой мы поговорим чуть ниже. Аналогично, в таблице 2.7 мы указываем в качестве предикторов предыдущий и пред-предыдущий моменты.

Таблица 2.7 — Формат ML: регрессия с добавлением двух исторических моментов

	x_2	x_1	y
1	$x_{12} = c_{n-2,1}$	$x_{11} = c_{n-1,1}$	y_1
2	$x_{22} = c_{n-2,2}$	$x_{21} = c_{n-1,2}$	y_2
3	$x_{32} = c_{n-2,3}$	$x_{31} = c_{n-1,3}$	y_3
...	...	...	...
m	$x_{m2} = c_{n-2,m}$	$x_{m1} = c_{n-1,m}$	y_m

Наконец, в таблице 2.8 мы добавили k предыдущих моментов времени и получили, тем самым, k -мерную регрессионную задачу.

Таблица 2.8 — Формат ML: регрессия с добавлением k исторических моментов

	x_k	...	x_2	x_1	y
1	$x_{1k} = c_{n-k,1}$	...	$x_{12} = c_{n-2,1}$	$x_{11} = c_{n-1,1}$	y_1
2	$x_{2k} = c_{n-k,2}$	...	$x_{22} = c_{n-2,2}$	$x_{21} = c_{n-1,2}$	y_2
3	$x_{3k} = c_{n-k,3}$	...	$x_{32} = c_{n-2,3}$	$x_{31} = c_{n-1,3}$	y_3
...	...	...	...	...	...
m	$x_{mk} = c_{n-k,m}$	...	$x_{m2} = c_{n-2,m}$	$x_{m1} = c_{n-1,m}$	y_m

В заключение этого раздела затронем тему наглядного представления данных в форматах ML/ LM, которое, при грамотном использовании может привести к улучшению прогнозирующих моделей⁹. Временной ряд, который описывает состояние процесса в одной отдельно взятой локации во все моменты времени (точнее говоря, его визуализацию в виде кривой), мы будем

⁹ Понятия портрета локации и портрета моментов мы импользуем ниже, когда будем говорить о моделирующих схемах алгоритма детекции аномалий. Они понадобятся для схем, работающих попринципу нескольких ближайших моментов или нескольких ближайших локаций.

называть портретом локации (см. рисунок 2.1).

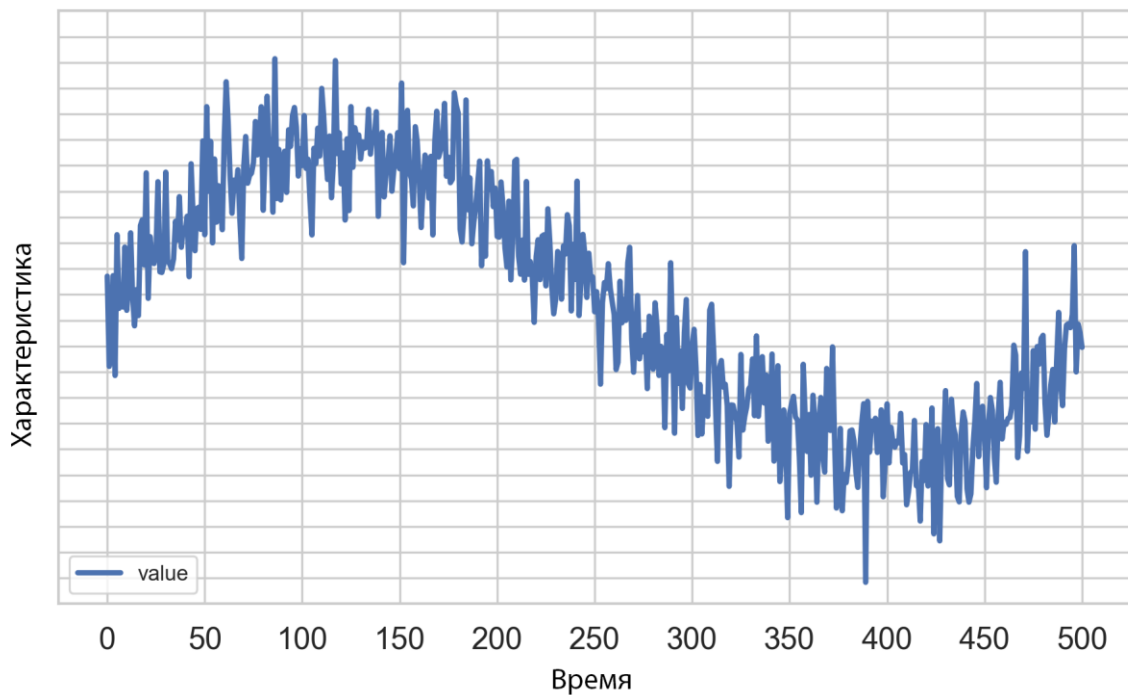


Рисунок 2.1 — Портрет локации

Другими словами, портрет локации — это некая кривая, характеризующая изменение характеристики с течением времени в этой локации. Две локации будут тем более «похожи» друг на друга (с точки зрения процесса изменения измеряемой характеристики), чем более «похожими» будут их портреты. С учётом того, что временной ряд представляет собой просто вектор, близость двух локаций (разумеется, не в географическом, а в аналитическом смысле) может быть измерена при помощи обычной евклидовой метрики, и это может помочь при построении прогнозов в одной отдельно взятой локации, если отталкиваться от знаний о поведении процесса в других локациях, наиболее «похожих» на изучаемую.

Аналогично введём понятие портрета момента. Портретом момента мы будем называть визуализацию состояния процесса во всех локациях сразу в некоторый фиксированный момент времени, другими словами, портрет момента — это тепловая карта (см. рисунок 2.2).

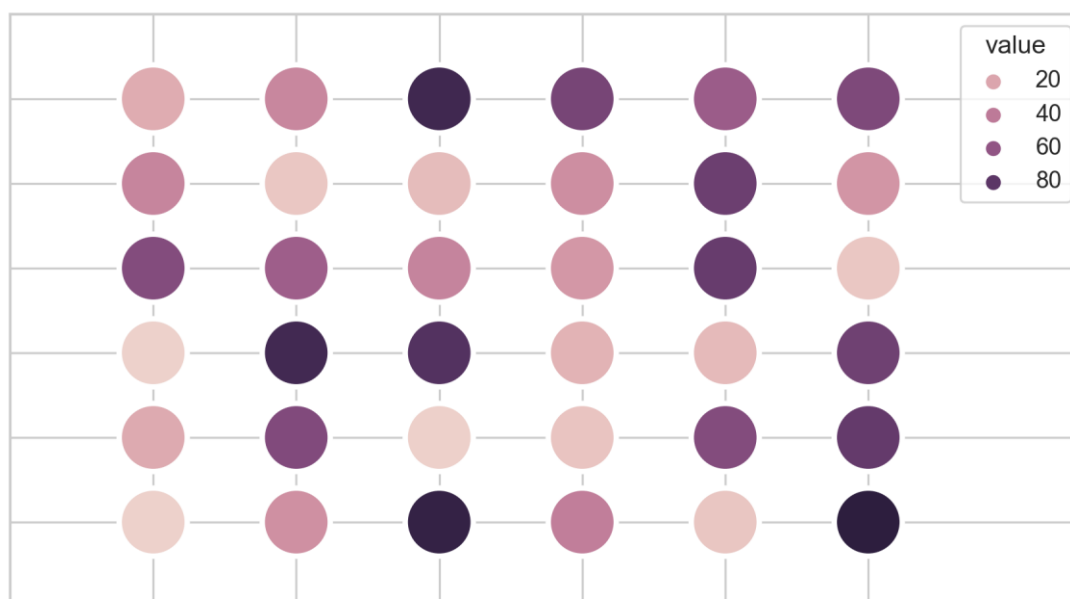


Рисунок 2.2 — Портрет момента

Точно так же, при помощи обычной евклидовой метрики, мы сможем определять близкие друг к другу по значениям измеряемой характеристики моменты времени: моменты будут «похожи» друг на друга, если будут «похожи» их портреты.

Здесь уместна будет аналогия с метеорологическими наблюдениями. Чем характеризуется та или иная географическая локация с точки зрения наблюдаемой среднемесячной температуры? — историей наблюдений, то есть некой кривой температур, характерной для данной местности, и располагая такими кривыми мы сможем отличить Рязань от Самарканда. И чем характеризуется тот или иной месяц с точки зрения среднемесячной температуры? — тепловой картой с обобщёнными сведениями обо всех регионах, и, имея такие карты, мы сможем отличить июнь от декабря¹⁰.

На иллюстрации ниже (см. рисунок 2.3) приведены портреты трёх локаций, среди которых первая и вторая являются близкими друг другу, в то время

¹⁰ Впрочем, строго говоря, таким способом мы не сможем отличить апрель от октября, и в этом, разумеется, есть большой недостаток выбранной нами аналогии. Но, во-первых, этот недостаток носит, скорее, технический характер (его можно преодолеть), а во-вторых, приведённая аналогия была нужна нам только для того, чтобы проиллюстрировать идею.

как третья далека от них обеих.

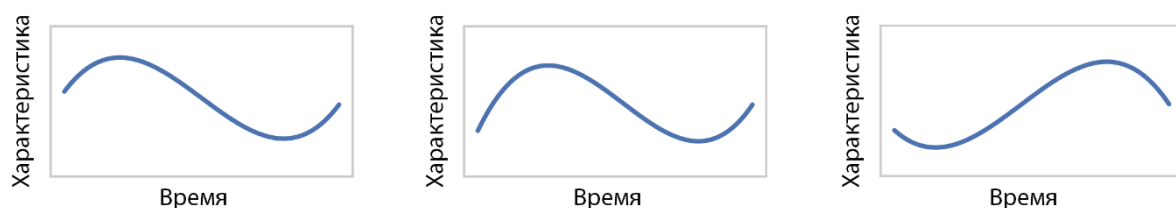


Рисунок 2.3 — Близость локаций

Здесь нужно сделать ещё одно замечание и предостеречь читателя от соблазна измерять близость локаций с формальных позиций. Следует помнить, что мы говорим не о буквальной близости, а о близких описаниях локаций, причём, описание берётся в терминах наблюдавшихся значений измеряемой характеристики. Например, возвращаясь к метеорологическому примеру со среднемесячной температурой, можно сказать, что Москва гораздо ближе к Монреалю, хотя, между ними 7 000 км, чем к Санкт-Петербургу. Другой пример, убедительно показывающий важность такого предостережения, можно обнаружить в ситуации с процессом обучения, которую мы вскользь затрагивали в первой главе, когда говорили о концептуальных основах настоящей диссертации. Представим обычную студенческую группу, которая изучает какую-нибудь в высшей степени интересную дисциплину, например, линейную алгебру. Локациями в этом процессе выступают отдельные студенты. Если смотреть на близость локаций сугубо формально, то, чтобы определить студента, ближайшего к студенту А, нужно взять список группы и указать либо следующего, либо предшествующего ему студента. С формальных позиций по отношению к студенту А они будут самыми близкими, хотя совершенно очевидно, что формализм упорядочивания фамилий по алфавиту не имеет никакого отношения к процессу обучения, и студентами по-настоящему близкими к студенту А (близость, разумеется, здесь понимается не в общечеловеческом смысле, а в смысле близких успехов в линейной алгебре) будут те, кто в течение семестра получал похожие оценки.

Аналогично близости локаций можно говорить о близости моментов

времени: близкими моментами будут те моменты, которые обладают похожими портретами. Близость моментов продемонстрирована на следующей иллюстрации (см. рисунок 2.4), где приведены три тепловые карты: первая и вторая близки друг к другу, а третья сильно отчается от них обеих.



Рисунок 2.4 — Близость моментов

Заметим, что, в отличие от локаций, на которых, как мы уже отмечали, нет естественного отношения порядка, на множестве моментов времени отношение порядка есть: из двух моментов один всегда предшествует другому (разумеется, кроме той вырожденной ситуации, когда они равны). Тем не менее, говоря о близких моментах времени, тоже следует проявлять определённую осторожность. На первый взгляд, ближайшим моментом времени по отношению к настоящему моменту является предшествующий момент — другого мнения просто не может быть. Но если мы, например, изучаем процесс потребления пива населением, и если сегодня пятница, то к нашей сегодняшней пятнице гораздо ближе будет предыдущая пятница, а вовсе не вчерашний четверг.

Эти замечания мы будем активно использовать в дальнейшем, на этапе моделирования процесса, а пока в качестве небольшого резюме к этому разделу отметим, что при изучении и прогнозировании процессов аналитик должен учитывать абсолютно все сведения, доступные ему на момент прогноза [2–5, 7, 174], а это значит, что такие понятия как портрет локации (учитывающий все моменты времени) и портрет момента (учитывающий, в свою очередь все локации) будут нам, безусловно, полезны. Завершая тему визуальной интерпретации данных ML/ LM, отметим, что рисунки 2.1 и 2.2 иллюстрируют

ситуацию, когда процесс наблюдался в 36 локациях в течение 500 моментов времени, и в этом случае глобальные данные о процессе визуализируются при помощи 36 временных рядов и 500 тепловых карт.

2.2. Моделирование процесса

Этап моделирования — это важнейший этап методики детекции аномалий. Именно на этом этапе полностью проявляются технологические особенности анализа больших данных, и применяется весь потенциал машинного обучения для построения прогнозирующих моделей. Если на этом этапе мы получим корректно обученную модель с высокой прогнозирующей способностью [2–5, 7–9, 51, 173], дальнейшая работа детектора не составит большого труда, поэтому сейчас мы опишем этап моделирования максимально подробно.

2.2.1. Разбиение данных на этапы истории и обучения

Пусть у нас имеется n наблюдений какой-либо характеристики процесса в m локациях, причём, наши данные представлены в формате ML, где c_{ij} — наблюдаемые значения характеристики процесса. Разделим все наблюдения в пропорции α (см. таблицу 2.9): первые αn будем считать историческими данными, а оставшиеся $(1 - \alpha)n$ мы будем прогнозировать.

Таблица 2.9 — Разбиение данных в формате ML на этапы истории и обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент 1	c_{11}	c_{12}	c_{13}	...	c_{1m}
Момент 2	c_{21}	c_{22}	c_{23}	...	c_{2m}
...	...	...	...	...	...
Момент αn	$c_{\alpha n,1}$	$c_{\alpha n,2}$	$c_{\alpha n,3}$	...	$c_{\alpha n,m}$
Момент $\alpha n + 1$	$c_{\alpha n+1,1}$	$c_{\alpha n+1,2}$	$c_{\alpha n+1,3}$	...	$c_{\alpha n+1,m}$
...	...	...	...	...	...
Момент n	c_{n1}	c_{n2}	c_{n3}	...	c_{nm}

Наша задача состоит в том, чтобы заполнить позиции из зелёной зоны спрогнозированными значениями (эту процедуру мы опишем в разделах 2.2.3 и

2.2.4), после чего бинаризовать разность Δ_{ij} между истинными и спрогнозированными значениями характеристики, получив, тем самым, на каждой позиции зелёной зоны False или True, где True означает, что в соответствующей локации в соответствующий момент времени наблюдалась аномалия с точки зрения нашего алгоритма детекции, а False — что наблюдалась норма (эту процедуру мы опишем в разделе 2.3). Полученные таким способом алгоритмические метки аномалий мы сможем сравнить с метками, которые будут расставлены экспертной группой, проведя, тем самым, процедуру сопоставления (которую мы подробно обсудим в разделе 2.4), и, если этот челлендж будет успешно пройден, мы будем считать, что детектор аномалий готов к внедрению.

Но прежде всего, нам, безусловно нужен качественный прогноз, к методам построения которого мы переходим.

2.2.2. Пул моделей

Специфика машинного обучения состоит в том, что никто не может заранее сказать, какая именно модель окажется удачной, а какая нет [2–5, 7–9, 48, 60]. Поэтому обычной практикой является использование некоторого заранее определённого набора моделей [3, 17, 234], так называемого пула моделей, прогнозы которых сравниваются между собой посредством какой-либо метрики эффективности [7, 8, 49, 65, 174] (см. приложение А, разделы А.7 и А.8).

Прогноз нестационарного процесса относится к типу регрессионных задач [2–5, 7–9, 173, 218], в этой области имеется довольно-таки широкий спектр моделей (см. приложение А, разделы А.3–А.6), и из них мы выберем следующие семь:

1. `LinearRegression` (линейная регрессия);
2. `DecisionTreeRegressor` (регрессия на решающих деревьях);
3. `RandomForestRegressor` (регрессия на случайных лесах);
4. `GradientBoostingRegressor` (градиентный бустинг над случайными

лесами);

5. SVR (регрессия на методе опорных векторов);
6. KNeighborsRegressor (регрессия на ближайших соседях);
7. MLPRegressor (нейронная сеть многослойного перцептрона).

По качеству предоставляемого решения, как, впрочем, и по частоте использования в задачах машинного обучения на табличных данных, безусловным лидером среди этих пяти моделей является градиентный бустинг [3, 4, 7, 174, 209, 231], однако подчеркнём ещё раз: мы не можем утверждать заранее, что именно эта модель окажется лучшей. Мы будем знать это наверняка только после того как применим все модели нашего пула и вычислим метрики их эффективности. Поэтому во всех последующих рассуждениях этой главы мы будем использовать модель, под которой будет пониматься одна из заявленных выше, а в алгоритмической реализации детектора нам придётся использовать цикл по приведенному выше пулу моделей.

2.2.3. Моделирование по моментам времени

Допустим, мы находимся в моменте времени αt в локации location_1 , и нам нужно спрогнозировать значение наблюдаемой характеристики в этой локации в следующий момент времени $\alpha t + 1$. Есть несколько способов выполнить такой прогноз. Наиболее очевидное решение стоит в том, чтобы использовать локальный формат данных (то есть, относящийся только к данной локации) и применить авторегрессию (см. таблицу 2.10).

Далее в этом разделе мы будем использовать авторегрессию со сдвигом на один момент времени, хотя, разумеется, можно использовать и другие глубины для сдвига, так как глубина авторегрессии является одним из настраиваемых параметров алгоритма (в главах 4–6 глубина принимает значения в пределах от 1 до 10). Здесь мы использовали первый столбец из таблицы 2.6 в качестве целевой функции, а в качестве единственного предиктора взяли его же со сдвигом на одну позицию вниз.

Таблица 2.10 — Авторегрессия на локальных данных

	x_1	y
Момент 1	—	C_{11}
Момент 2	C_{11}	C_{21}
...	...	...
Момент an	$C_{an-1,1}$	$C_{an,1}$
Момент $an+1$	$C_{an,1}$	

Получилась классическая одномерная регрессионная задача, которую можно решить при помощи какой-либо модели регрессии из пула заявленных выше регрессионных моделей и получить предсказанное значение характеристики $p_{an+1,1}$, то есть, то значение, которое мы прогнозируем в локации $location_1$ в момент времени $an+1$. После этого мы можем точно так же поступить с локацией $location_2$, затем — с локацией $location_3$ и так далее, пока не исчерпаем все локации вплоть до $location_m$. В результате у нас возникает набор предсказаний для первого момента, основанный на некоторой модели из пула моделей (см. таблицу 2.11).

Таблица 2.11 — Прогноз по времени для первого момента периода обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $an+1$	$p_{an+1,1}$	$p_{an+1,2}$	$p_{an+1,3}$	...	$p_{an+1,m}$

Теперь допустим, что наступил момент времени $an+1$, нам стали известны истинные значения наблюдаемой характеристики во всех локациях, и мы хотим спрогнозировать их значения в следующий момент времени $an+2$. Для этого нам достаточно скорректировать статусы моментов времени в исходных данных: теперь мы будем считать, что момент времени $an+1$ тоже относится к периоду исторических наблюдений, а период обучения алгоритма и подбора порога бинаризации начинается уже со следующего момента $an+2$ (см. таблицу 2.12). Другими словами, мы сместим разбиение данных на периоды исторических наблюдений и обучения на одну строку вниз.

Таблица 2.12 — Вторая итерация разбиения данных в формате ML на периоды исторических наблюдений и обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент 1	C_{11}	C_{12}	C_{13}	...	C_{1m}
Момент 2	C_{21}	C_{22}	C_{23}	...	C_{2m}
...	...	...	...	...	...
Момент $\alpha n+1$	$C_{\alpha n+1,1}$	$C_{\alpha n+1,2}$	$C_{\alpha n+1,3}$	...	$C_{\alpha n+1,m}$
Момент $\alpha n+2$	$C_{\alpha n+2,1}$	$C_{\alpha n+2,2}$	$C_{\alpha n+2,3}$	...	$C_{\alpha n+2,m}$
...	...	...	...	...	...
Момент n	C_{n1}	C_{n2}	C_{n3}	...	C_{nm}

После этого мы повторяем те же самые действия (строим регрессионные модели, используя ещё одно дополнительное наблюдение в момент времени $\alpha n+1$) и получаем прогноз наблюдаемой характеристики во всех локациях для следующего момента времени $\alpha n+2$ (см. таблицу 2.13).

Таблица 2.13 — Прогноз по времени для первых двух моментов периода обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $\alpha n+1$	$p_{\alpha n+1,1}$	$p_{\alpha n+1,2}$	$p_{\alpha n+1,3}$	...	$p_{\alpha n+1,m}$
Момент $\alpha n+2$	$p_{\alpha n+2,1}$	$p_{\alpha n+2,2}$	$p_{\alpha n+2,3}$	...	$p_{\alpha n+2,m}$

Итерационно наращивая наблюдаемые моменты времени вплоть до $n-1$, мы получим набор прогнозов для периода обучения (см. таблицу 2.14).

Таблица 2.14 — Полный прогноз по времени для периода обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $\alpha n+1$	$p_{\alpha n+1,1}$	$p_{\alpha n+1,2}$	$p_{\alpha n+1,3}$	...	$p_{\alpha n+1,m}$
...	...	...	...	...	...
Момент n	p_{n1}	p_{n2}	p_{n3}	...	p_{nm}

Получив полный набор прогнозов, мы сравниваем истинные значения наблюдаемой характеристики C_{ij} для периода обучения, записанные в таблице 2.6, с прогнозируемыми значениями p_{ij} , записанными в таблице 2.14, и при помощи какой-либо метрики эффективности (средняя квадратичная ошибка

MSE, средняя абсолютная ошибка MAE, коэффициент детерминации R^2) оцениваем прогнозирующую способность модели.

Затем мы переходим к следующей модели из пула регрессионных моделей, сформированных нами в самом начале, и проделав с ней те же самые действия, получаем оценку её прогнозирующей способности, затем — к следующей модели и так далее, пока не переберём весь пул. После этого нам остаётся только сравнить полученные оценки: та из моделей, которая продемонстрирует наилучшую метрику эффективности, признается лучшей моделью, и именно её мы будем использовать в дальнейшей работе для реализации основных задач алгоритма детекции аномалий.

Рассмотренную схему авторегрессионного прогноза по моментам времени можно усилить, так как выше мы опирались исключительно на локальные данные, относящиеся к одной отдельно взятой локации, в тот момент как нам доступны сразу все данные во всех локациях, в которых наблюдается характеристика процесса. Наша идея состоит в том, чтобы расширить набор предикторов регрессионной модели настолько, насколько это вообще возможно, и применять для прогноза характеристики не одномерные, а многомерные регрессионные модели [4, 5, 8, 23, 52, 172]. Другими словами, если выше мы прогнозировали значение характеристики в одной локации по её предыдущему значению в этой локации, то более сложный (но и более точный) вариант состоит в том, чтобы прогнозировать характеристику в одной отдельно взятой локации по полному портрету предыдущего момента времени: одно число мы будем прогнозировать по всей тепловой карте (см. раздел 2.1).

Пусть снова мы находимся в моменте времени αn в локации location_1 , и нам нужно спрогнозировать значение наблюдаемой характеристики в этой локации в следующий момент времени $\alpha n + 1$. Теперь мы будем использовать глобальный формат данных, то есть, относящийся ко всем локациям сразу, а для этого применим авторегрессию со сдвигом на один момент времени (см. таблицу 2.15).

Таблица 2.15 — Авторегрессия на глобальных данных

	x_1	x_2	x_3	...	x_m	y
Момент 1	—	—	—	...	—	C_{11}
Момент 2	C_{11}	C_{12}	C_{13}	...	C_{1m}	C_{21}
...	...	...	...	...	...	...
Момент a_n	$C_{a_n-1,1}$	$C_{a_n-1,2}$	$C_{a_n-1,3}$	...	$C_{a_n-1,m}$	$C_{a_n,1}$
Момент a_n+1	$C_{a_n,1}$	$C_{a_n,2}$	$C_{a_n,3}$	...	$C_{a_n,m}$	

Теперь неформально нашу регрессионную задачу можно сформулировать следующим образом. У первого момента времени не было предыдущего. Второй момент времени характеризовался тем, что предыдущий момент обладал портретом $C_{11}, \dots, C_{1m}$, и при этом во второй момент времени в первой локации наблюдалось значение C_{21} . Третий момент времени характеризовался тем, что предыдущий момент обладал портретом $C_{21}, \dots, C_{2m}$, и при этом во второй момент времени в первой локации наблюдалось значение C_{31} . И так далее. Такие соответствия у нас имеются вплоть до момента времени a_n . Вопрос, на который мы хотим ответить, состоит в том, что если момент времени с номером a_n обладал портретом $C_{a_n,1}, \dots, C_{a_n,m}$, то каково будет значение наблюдаемой характеристики в первой локации в следующий момент времени a_n+1 ? Это классическая регрессионная задача с несколькими предикторами [2–5, 7–9, 60, 201, 218]. Для её построения мы использовали первый столбец из таблицы 3.6 в качестве целевой функции, а в качестве набора предикторов взяли все её столбцы со сдвигом на одну позицию вниз. Все последующие действия: от выбора модели из пула моделей до получения предсказаний $\hat{r}_{ij}$ при помощи лучшей модели — осуществляются точно так же, как и выше.

И, наконец, можно предложить ещё одну, видимо, наиболее разумную схему восстановления регрессии в первой локации¹¹. Идея состоит в том, чтобы использовать для этого не одну локацию и не все, а несколько локаций,

¹¹ Хотя, отметим ещё раз, что в машинном обучении никто не знает заранее, какая именно модель окажется лучшей: любая гипотеза требует подтверждения.

наиболее близких к первой в смысле близости их портретов. Напомним, что портретом локации служит, по определению, временной ряд (см. раздел 2.1, рисунок 2.1), а расстояние между портретами вычисляется при помощи обычной евклидовой метрики, поэтому задача определения ближайших портретов — это несложная задача: нужно просто вычислить все попарные расстояния между портретами и упорядочить их по убыванию. Тогда на первом месте окажется сама локация $location_1$, потому что расстояние от неё до себя самой будет равно нулю. Затем последует первый нетривиальный ближайший сосед, после него — сосед, уступающий первому, но ближайший по сравнению со всеми остальными и так далее. Например, задавшись пятью ближайшими соседями, получим авторегрессию, представленную ниже (см. таблицу 2.16), где номер в скобках указывает на то, что это не номер локации, а её порядковый номер в упорядоченном списке локаций, ближайших к локации $location_1$.

Таблица 2.16 — Авторегрессия на ближайших локациях

	x_1	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	y
Момент 1	—	—	—	—	—	C_{11}
Момент 2	C_{11}	$C_{1(1)}$	$C_{1(2)}$	$C_{1(3)}$	$C_{1(4)}$	C_{21}
...	...	...	...	...	...	...
Момент an	$C_{an-1,1}$	$C_{an-1,(1)}$	$C_{an-1,(2)}$	$C_{an-1,(3)}$	$C_{an-1,(4)}$	$C_{an,1}$
Момент $an+1$	$C_{an,1}$	$C_{an,(1)}$	$C_{an,(2)}$	$C_{an,(3)}$	$C_{an,(4)}$	

Все остальные действия осуществляются точно так же и в конечном счете приводят нас к выбору лучшей модели и к полученным при помощи неё прогнозам p_{ij} для всех локаций во все моменты времени из периода обучения алгоритма (см. таблицу 2.14).

Вопрос о том, какая из предложенных нами трёх схем — авторегрессия на локальных данных, на глобальных данных или на ближайших локациях — является предпочтительной, решается так же, как и вопрос о выборе лучшей модели: при помощи метрики эффективности. Мы вычисляем среднеквадратичную ошибку MSE, среднюю абсолютную ошибку MAE или коэффициент

детерминации R^2 и отдаём предпочтение той схеме, которая получает наилучшую оценку¹². Выбрав наилучший из прогнозов, мы сможем судить о наличии или отсутствии аномалий на основании отклонений реальных значений наблюдаемой характеристики от прогнозируемых: если отклонение слишком велико, то скорее всего в данной локации в данный момент времени наблюдается аномалия (подробнее об этом см. ниже, раздел 2.3), если же отклонение не столь велико, то, скорее всего, аномалии нет, и можно утверждать, что в данный момент времени в данной локации наблюдаемая характеристика не-стабильного процесса находится в пределах нормы.

2.2.4. Моделирование по локациям

Однако, мы помним (см. главу 1, раздел 1.3.3), что аномалия как явление может определяться с двух точек зрения: с одной стороны, это аномалия в сравнении с предыдущими наблюдениями, то есть — аномалия по времени, и именно таким типом аномалий мы занимались в предыдущем разделе, но с другой стороны, аномалия также может возникнуть в сравнении с другими локациями. В этом разделе мы разберём идеи алгоритмов, прогнозирующих значение наблюдаемой характеристики в некоторый момент времени в одной отдельно взятой локации на основании данных о её значениях в других локациях в тот же момент времени, другими словами — изложим алгоритм, позволяющий детектировать аномалии по локациям.

Для детекции аномалий по локациям понадобится переход от формата данных ML к формату данных LM, но он осуществляется элементарно путём транспонирования матрицы данных¹³ (см. таблицу 2.17).

¹² Везде в дальнейшем в рамках настоящей диссертационной работы мы используем метрику R^2 как наиболее универсальную, но следует иметь в виду, что вполне допустимо использовать и другие метрики (подробнее о метриках эффективности прогнозирующих моделей см. приложение А).

¹³ Обратите внимание на индексацию компонент: индексы поменялись местами по сравнению с матрицей данных в формате ML (см. таблицу 2.6).

Таблица 2.17 — Разбиение данных в формате LM на периоды истории и обучения

	Момент 1	...	Момент αn	Момент $\alpha n+1$	...	Момент n
Локация 1	C_{11}	...	$C_{\alpha n,1}$	$C_{\alpha n+1,1}$	...	C_{n1}
Локация 2	C_{12}	...	$C_{\alpha n,2}$	$C_{\alpha n+1,2}$	...	C_{n2}
Локация 3	C_{13}	...	$C_{\alpha n,3}$	$C_{\alpha n+1,3}$	...	C_{n3}
...	...	...	...	...	...	...
Локация m	C_{1m}	...	$C_{\alpha n,m}$	$C_{\alpha n+1,m}$	...	C_{nm}

Далее мы реализуем ту же последовательность рассуждений, что и в предыдущем разделе: нас будет интересовать, что произойдёт, если мы будем использовать локальные данные, что — если глобальные, и что — если подберём несколько предикторов модели из числа наиболее близких к целевой функции. Но поскольку моделирование по локациям существенно отличается от моделирования по времени, мы снова дадим максимально подробное описание всех процедур, не ограничившись ссылкой на аналогию, так как эта аналогия всё-таки не является буквальной. Напомним, что наша цель — заполнить все ячейки периода обучения предсказанными значениями наблюдаемой характеристики процесса, ориентируясь при этом на значения в других локациях.

Начнём с моделирования на локальных данных. Пусть наступил первый момент периода обучения алгоритма, то есть, момент с номером $\alpha n+1$. В формате LM локальные данные означают, что мы оперируем сведениями о значениях характеристики во всех локациях, но только в один момент времени. Пусть мы хотим спрогнозировать значение характеристики в локации $location_k$, отталкиваясь от её значений во всех остальных локациях. У нас нет возможности выполнить какой-либо разумный сдвиг целевой функции, как мы это делали, применяя метод авторегрессии к временным рядам, и, следовательно, на локальных данных мы получаем регрессионную задачу с нулевым числом предикторов (см. таблицу 2.18). Единственный прогноз, который можно сделать в такой ситуации состоит в том, чтобы вычислить среднее значение характеристики по всем остальным локациям, кроме k -й, и указать его в

качестве прогнозируемого значения.

Таблица 2.18 — Регрессия на локальных данных

	у
Локация 1	$C_{a\bar{n}+1,1}$
...	...
Локация k-1	$C_{a\bar{n}+1,k-1}$
Локация k	
Локация k+1	$C_{a\bar{n}+1,k+1}$
...	...
Локация m	$C_{a\bar{n}+1,m}$

Трудно назвать такое решение хорошей моделью. Но зато мы можем поступить так сразу со всеми локациями: прогноз для первой локации — среднее по всем локациям, кроме первой; прогноз по второй — среднее по всем локациям, кроме второй; и так далее. Более того, мы можем поступить так со всеми локациями во все моменты периода обучения, для этого достаточно в таблицу 2.18 поместить другой столбец из таблицы 2.17 (имеются в виду столбцы из зеленого сектора, так как именно их нам нужно спрогнозировать). Таким образом мы сразу получаем полный прогноз по всем локациям для всех моментов времени всего периода обучения¹⁴. Да, он примитивный и, разумеется, неточный. Зато простой.

Таблица 2.19 — Полный прогноз по локациям для периода обучения

	Момент $a\bar{n}+1$	...	Момент n
Локация 1	$q_{a\bar{n}+1,1}$	...	q_{n1}
Локация 2	$q_{a\bar{n}+1,2}$	...	q_{n2}
Локация 3	$q_{a\bar{n}+1,3}$	...	q_{n3}
...	...	...	...
Локация m	$q_{a\bar{n}+1,m}$	...	q_{nm}

¹⁴ Прогноз по локациям мы обозначаем буквой q, чтобы отличать его от прогноза по времени, который мы обозначали выше буквой p.

Итак, анализ на локальных данных позволяет получить какой-то, пусть и не вполне удовлетворительный результат, но нам хотелось бы иметь более точные модели. Отметим, что пока мы ориентировались на локальные данные, у нас даже не было возможности использовать хоть какую-нибудь модель из пула регрессионных моделей, который мы заранее подготовили и оговорили (см. выше, раздел 2.2.2) — просто в силу того, что модели не работают без предикторов.

Посмотрим теперь, что мы сможем извлечь из глобальных данных, если у нас есть возможность использовать сведения обо всех локациях во все моменты времени, доступные на момент прогноза.

Тогда в нашем распоряжении оказываются сразу a_n предикторов¹⁵ (см. таблицу 2.20), что, как мы знаем, обеспечивает модели гораздо более высокую надежность прогноза: каждая локация описывается при помощи сразу всего своего портрета, а это целый временной ряд, начиная с первого наблюдения и заканчивая наблюдением с номером a_n .

Таблица 2.20 — Регрессия на глобальных данных

	x_1	x_2	x_3	...	x_{a_n}	y
Локация 1	c_{11}	c_{21}	c_{31}	...	$c_{a_n,1}$	$c_{a_n+1,1}$
...	...	...	...	...	...	...
Локация k-1	$c_{1,k-1}$	$c_{2,k-1}$	$c_{3,k-1}$	...	$c_{a_n,k-1}$	$c_{a_n+1,k-1}$
Локация k	$c_{1,k}$	$c_{2,k}$	$c_{3,k}$	...	$c_{a_n,k}$	
Локация k+1	$c_{1,k+1}$	$c_{2,k+1}$	$c_{3,k+1}$		$c_{a_n,k+1}$	$c_{a_n+1,k+1}$
...	...	...	...	...	...	...
Локация m	c_{1m}	c_{2m}	c_{3m}		$c_{a_n,m}$	$c_{a_n+1,m}$

Такие данные позволяют строить многомерную регрессионную модель. Отличие от прогнозирования по времени состоит в том, что мы не обязаны прогнозировать именно последнее наблюдение (ни первых, ни последних объектов в

¹⁵ Здесь предикторы $x_1, \dots, x_{a_n}$ означают моменты времени $moment_1, \dots, moment_{a_n}$, а целевая функция y — момент времени $moment_{a_n+1}$.

этом датафрейме нет, так как на локациях отсутствует отношение порядка). Мы можем прогнозировать любую локацию, и, проведя эту процедуру m раз, мы получаем заполненный первый столбец периода обучения алгоритма в таблице 2.17.

После этого наступает момент времени $an+2$, нам становятся известны истинные значения наблюдаемой характеристики в этот момент, и мы можем точно так же построить другую регрессионную модель, используя уже $an+1$ предикторов, в результате чего окажется заполненным второй столбец периода обучения в таблице 2.17. И так далее, вплоть до самого последнего момента времени с номером n . В результате мы получим таблицу 2.19, полностью заполненную нашими прогнозами для периода обучения (только величины этих прогнозов будут, разумеется, отличаться от тех величин, что мы получили вычислением средних значений, потому что на этот раз мы используем более сложные модели, прогнозы которых будут, скорее всего, более точными).

И, наконец, мы можем реализовать ещё одну идею, которую обсуждали выше: применить в качестве предикторов регрессионной модели не все доступные нам моменты времени, то есть, не полные портреты наших локаций, а только несколько ближайших моментов¹⁶.

Итак, пусть мы находимся в моменте времени $an+1$, и пусть нас интересуют не все предыдущие моменты, а пять ближайших (с точки зрения близости их портретов). Тогда датафрейм для решения регрессионной задачи прогнозирования, которая нас интересует в рамках этого раздела, будет иметь следующий вид (см. таблицу 2.21), где индекс в скобках указывает не на порядковый номер момента времени, а на его порядковый номер в упорядоченном списке моментов, ближайших к моменту $an+1$.

¹⁶ Подчеркнем ещё раз, что мы говорим не о хронологической близости, а о близости портретов в смысле евклидовой метрики.

Таблица 2.21 — Регрессия на ближайших моментах

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	Y
Локация 1	$C_{(1)1}$	$C_{(2)1}$	$C_{(3)1}$	$C_{(4)1}$	$C_{(5)1}$	$C_{an+1,1}$
...	...	...	...	...	...	...
Локация k-1	$C_{(1),k-1}$	$C_{(2),k-1}$	$C_{(3),k-1}$	$C_{(4),k-1}$	$C_{(5),k-1}$	$C_{an+1,k-1}$
Локация k	$C_{(1),k}$	$C_{(2),k}$	$C_{(3),k}$	$C_{(4),k}$	$C_{(5),k}$	
Локация k+1	$C_{(1),k+1}$	$C_{(2),k+1}$	$C_{(3),k+1}$	$C_{(4),k+1}$	$C_{(5),k+1}$	$C_{an+1,k+1}$
...	...	...	...	...	...	...
Локация m	$C_{(1)m}$	$C_{(2)m}$	$C_{(3)m}$	$C_{(4)m}$	$C_{(5)m}$	$C_{an+1,m}$

Отметим, что, когда мы говорили о ближайших локациях в прогнозировании временного ряда, мы допустили, что самой близкой будет сама локация, в которой происходит прогноз (её расстояние до самой себя равно нулю). Теперь же, когда мы говорим о прогнозах по локациям, мы не можем взять в качестве ближайшего момента времени момент времени $an+1$, потому что тогда регрессионная задача теряет смысл: прогнозируемое значение окажется просто равным уже известному значению¹⁷. Поэтому при построении регрессии на ближайших моментах первым ближайшим будет первый действительно ближайший в нетривиальном смысле (его расстояние до момента $an+1$ должно быть не нулевым). Если же мы учли все эти нюансы, дальнейшая работа не представляет больших трудностей: обучаем модели, делаем прогнозы и постепенно заполняем таблицу прогнозов по локациям для периода обучения, то есть, всё ту же таблицу 2.19.

Далее, вне зависимости от того, какую из схем исследования мы использовали (нуль-мерную регрессию, регрессию на полных данных или регрессию на ближайших моментах), мы сможем сравнить наши прогнозы, полученные для периода обучения, с реальными значениями наблюдаемой характеристики, и если реальное значение отклоняется от прогнозируемого слишком

¹⁷ Здесь уместно будет отослать читателя к внимательному рассмотрению таблицы 2.18: если $X_{(1)} = Y$, то $C_{an+1,k-1} = C_{(1),k}$, то есть, прогнозируемое значение известно заранее.

сильно, то мы признаем, что в этот момент в этой локации наблюдалась аномалия. Этим мы займёмся позже (см. следующий раздел). Тогда наряду с таблицей 2.19 у нас возникнет таблица, сигнализирующая об аномалиях, ячейки которой будут заполнены значениями True для зафиксированной аномалии и False для ситуаций, когда состояние процесса нормально и её уже можно будет сравнивать с мнением экспертной группы.

2.3. Критические границы отклонений

После этапа моделирования мы получаем две серии прогнозов:

1. прогнозы по времени p_{ij} , которые мы сосредоточили в таблице 2.14;
2. прогнозы по локациям q_{ij} , записанные в таблице 2.19.

В этом разделе мы подробно рассмотрим идею определения критической границы отклонения от прогноза по времени (см. раздел 2.3.1), что же касается ситуации с прогнозами по локациям, то она ничем не будет отличаться, поэтому мы опишем её весьма кратко (см. раздел 2.3.2), лишь в самых общих чертах.

2.3.1. Критические границы отклонений от прогноза по времени

Итак, у нас есть две таблицы. В первой таблице содержатся истинные значения наблюдаемой характеристики в каждой локации в каждый момент времени, относящиеся к периоду обучения нашего алгоритма (см. таблицу 2.22).

Таблица 2.22 — Истинные значения наблюдаемой характеристики для периода обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $an+1$	$C_{an+1,1}$	$C_{an+1,2}$	$C_{an+1,3}$	...	$C_{an+1,m}$
...	...	...	...	...	...
Момент n	C_{n1}	C_{n2}	C_{n3}	...	C_{nm}

Во второй (см. таблицу 2.23) — то, что мы спрогнозировали, пользуясь знаниями о предыдущих моментах времени¹⁸ (см. раздел 2.2.3).

Таблица 2.23 — Прогноз по времени наблюдаемой характеристики для периода обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $an+1$	$p_{an+1,1}$	$p_{an+1,2}$	$p_{an+1,3}$	...	$p_{an+1,m}$
...	...	...	...	...	...
Момент n	p_{n1}	p_{n2}	p_{n3}	...	p_{nm}

Теперь мы хотим узнать, насколько отличаются реально наблюдаемые значения характеристики от спрогнозированных. Это сделать очень легко: достаточно взять разность таблиц 2.22 и 2.23, эта разность покажет величину ошибки, и мы будем знать, в каких локациях и в какие моменты времени наблюдались наибольшие несоответствия. Обозначим величину ошибки

$$\Delta_{ij} = p_{ij} - c_{ij}$$

и сформируем таблицу с отклонениями (см. таблицу 2.24). В ней сосредоточены все необходимые нам сведения о несоответствии между прогнозируемыми и реально наблюдаемыми значениями изучаемой нами характеристики.

Таблица 2.24 — Абсолютные отклонения реальных значений от прогноза по времени

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $an+1$	$\Delta_{an+1,1}$	$\Delta_{an+1,2}$	$\Delta_{an+1,3}$	...	$\Delta_{an+1,m}$
...	...	...	...	...	...
Момент n	Δ_{n1}	Δ_{n2}	Δ_{n3}	...	Δ_{nm}

Отклонение может быть как положительным, так и отрицательным. Причём, его природа такова, что если рассматривать его как случайную величину, то оно подчиняется нормальному закону распределения с нулевым математическим ожиданием [6, 81, 83] (см. рисунок 2.5).

¹⁸ Таблица 2.23 буквально воспроизводит таблицу 2.14. Мы продублировали её здесь, исходя из соображений удобства и наглядности.



Рисунок 2.5 — Распределение отклонений

Мы можем зафиксировать какую-нибудь величину отклонения как некую критическую границу и считать, что если отклонение превышает эту границу, то наблюдаемое значение характеристики является аномальным. Но мы так же вправе изменить своё мнение, увеличив или уменьшив границу (см. рисунок 2.6). Чем больше значение границы, тем меньшее число наблюдений будет характеризоваться как аномальные, и наоборот: чем меньше граница, тем большее количество наблюдений будут признаны аномальными (аномалии соответствуют закрашенным областям на рисунок 2.6).

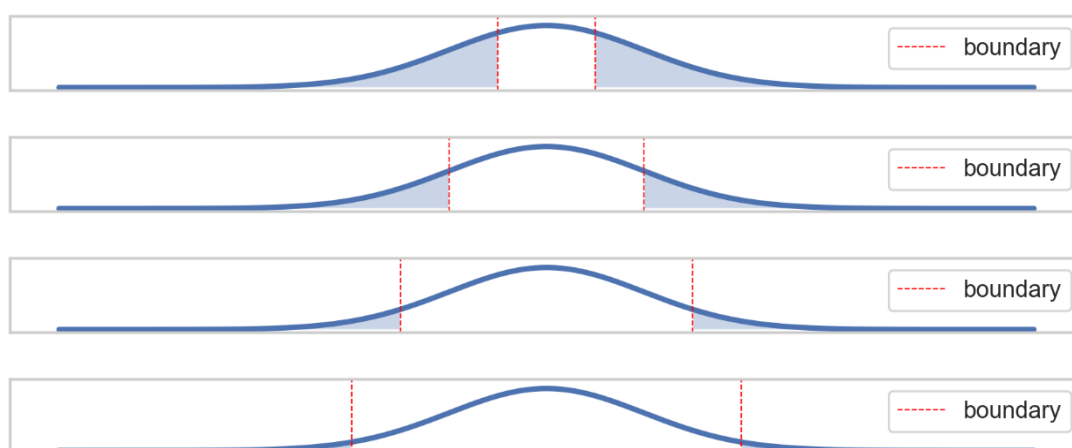


Рисунок 2.6 — Вариация порога аномалий

Тем самым, возникает параметрическое семейство гипотез об аномальности того или иного состояния наблюдаемой характеристики, зависящее от одного параметра: допустимой границы отклонений. В качестве параметра можно использовать буквальное отклонение, измеренное в тех же единицах, что и изучаемая характеристика, но удобнее (по крайней мере, с точки зрения универсальности калибровки и возможности сравнивать её с калибровками других характеристик) привязать отклонение к величине среднеквадратичного отклонения, то есть, к величине сигмы этого распределения.

Здесь уместно вспомнить знаменитое правило трёх сигм, породившее

даже целое направление в теории управления качеством (см. выше, главу 1, где это направление упоминалось под именем 6σ). Правило трёх сигм состоит в том, что если случайная величина распределена по нормальному закону, то событие, при котором она примет значение, отклоняющееся от математического ожидания больше чем на 3σ , является практически невозможным (а поскольку отклонения возможны как налево, так и направо, в сумме получаются как раз те самые 6σ , что и оправдывает приведённое выше название одной из парадигм управления качеством). Уместно так же вспомнить, что правило трёх сигм является эмпирическим, и поэтому ничто не мешает нам говорить об отклонении на 2σ или на 4σ , или на 2.5σ , то есть использовать вместо тройки любой коэффициент. Именно этот коэффициент будет играть роль варьируемого параметра в параметрическом семействе наших гипотез об аномальности. Поскольку в этом разделе мы говорим об аномалиях по времени, обозначим коэффициент при сигме k и сформулируем утверждение об аномальности наблюдаемых состояний процесса в более формализованном виде: если $\Delta_{ij} > k\sigma(\Delta)$, то в i -й момент времени в j -й локации наблюдалось аномальное значение характеристики процесса.

Отметим ещё раз, что пока это утверждение даже не является утверждением в математическом смысле этого слова. Пока это гипотеза, зависящая от величины параметра k , и нам ещё предстоит подобрать подходящее значение этого коэффициента на этапе сопоставления найденных алгоритмом отклонений с мнением экспертной группы (см. далее, раздел 2.4). В зависимости от выполнения или невыполнения условия аномальности, описанного приведённым выше неравенством, возникает ещё одна таблица гипотетических булевых значений (см. таблицу 2.25).

Таблица 2.25 — Индикатор аномалий по времени для периода обучения

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $an+1$	$(T' F')_{an+1,1}$	$(T' F')_{an+1,2}$	$(T' F')_{an+1,3}$	...	$(T' F')_{an+1,m}$
...	...	...	...	...	...

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент n	$(T' F')_{n1}$	$(T' F')_{n2}$	$(T' F')_{n3}$	...	$(T' F')_{nm}$

Здесь символ $(T'|F')_{ij}$ означает истину (True), если условие аномальности выполнено, и ложь (False), если оно не выполнено. Но выполнение условия аномальности зависит от калибровочного коэффициента k , поэтому у нас возникает не одна, а серия таких булевых таблиц. Например, если мы варьируем коэффициент от 2 до 5, придавая ему только целые значения, таких таблиц будет 4 (и, вообще говоря, они все будут разными), а если вариация происходит с шагом 0.1, то их будет уже 40 (и все они тоже будут разными).

2.3.2. Критические границы отклонений от прогноза по локациям

Аналогично происходит поиск отклонений от прогнозов по локациям. Теперь на входе у нас тоже имеются две таблицы, причём, одна из них получается простым транспонированием таблицы 2.22, в которой представлены истинные значения изучаемой характеристики, наблюдаемые на периоде обучения, но вторая, вообще говоря, никак не связана с таблицей спрогнозированных значений 2.23, потому что она получается за счёт совсем других моделей: не тех, что давали прогноз по времени, а тех, что давали прогноз по локациям.

Вот так устроена таблица истинных значений (см. таблицу 2.26, и заметим, что это просто зеленый блок таблицы 2.17):

Таблица 2.26 — Транспонированные истинные значения для периода обучения

	Момент $\alpha n+1$	Момент $\alpha n+2$	Момент $\alpha n+3$	...	Момент n
Локация 1	$C_{\alpha n+1,1}$	$C_{\alpha n+2,1}$	$C_{\alpha n+3,1}$	...	C_{n1}
...	...	...	...	...	...
Локация m	$C_{\alpha n+1,m}$	$C_{\alpha n+2,m}$	$C_{\alpha n+3,m}$	...	C_{nm}

С другой стороны, мы можем представить в виде таблицы прогнозы по локациям наблюдаемой нами характеристики для периода обучения и получить таблицу точно такого же формата, содержащую спрогнозированные при помощи моделей значения:

Таблица 2.27 — Прогноз по локациям для периода обучения

	Момент $\alpha n+1$	Момент $\alpha n+2$	Момент $\alpha n+3$	...	Момент n
Локация 1	$q_{\alpha n+1,1}$	$q_{\alpha n+2,1}$	$q_{\alpha n+3,1}$	...	q_{n1}
...	...	...	...	...	...
Локация m	$q_{\alpha n+1,m}$	$q_{\alpha n+2,m}$	$q_{\alpha n+3,m}$	...	q_{nm}

Точно так же, как и выше, когда мы работали с прогнозом по времени, мы берём разность наблюдаемых и истинных значений:

$$\Theta_{ij} = q_{ij} - c_{ij}$$

и получаем набор отклонений прогнозируемых при помощи моделей значений наблюдаемой характеристики процесса от реальных значений для периода обучения алгоритма (см.таблицу 2.28).

Таблица 2.28 — Отклонения реальных значений от прогноза по локациям

	Момент $\alpha n+1$	Момент $\alpha n+2$	Момент $\alpha n+3$	...	Момент n
Локация 1	$\Theta_{\alpha n+1,1}$	$\Theta_{\alpha n+2,1}$	$\Theta_{\alpha n+3,1}$	...	Θ_{n1}
...	...	...	...	...	...
Локация m	$\Theta_{\alpha n+1,m}$	$\Theta_{\alpha n+2,m}$	$\Theta_{\alpha n+3,m}$	...	Θ_{nm}

Мы снова формулируем условие аномальности при помощи калибровочного коэффициента: мы будем гипотетически считать, что наблюдаемое значение в i -й момент времени в j -й локации является аномальным, если отклонение наблюдаемого значения от прогнозируемого превышает некоторый порог: $\Theta_{ij} > k_2 \sigma(\Theta)$. При этом второй калибровочный коэффициент k_2 не обязан совпадать k (то есть, в частности, при некоторых условиях они могут и совпасть, но скорее всего не совпадут). Коэффициент k_2 тоже является варьируемой величиной, так же, как и коэффициент k , но они изменяются независимо друг от друга.

После этого мы, так же как и выше, получаем таблицу индикаторов аномалий, то есть, таблицу булевых значений в зависимости от выполнения или невыполнения условия аномальности:

Таблица 2.29 — Индикатор аномалий по локациям для периода обучения

	Момент $\alpha n+1$	Момент $\alpha n+2$	Момент $\alpha n+3$	...	Момент n
Локация 1	$(T'' F'')_{\alpha n+1,1}$	$(T'' F'')_{\alpha n+2,1}$	$(T'' F'')_{\alpha n+3,1}$	...	$(T'' F'')_{n1}$
...	...	...	...	...	...
Локация m	$(T'' F'')_{\alpha n+1,m}$	$(T'' F'')_{\alpha n+2,m}$	$(T'' F'')_{\alpha n+3,m}$	...	$(T'' F'')_{nm}$

Заметим, что булевы таблицы 2.29 и 2.25 не получаются друг из друга транспонированием, пара момент-локация может оказаться аномальной по времени и не оказаться таковой по локациям, и наоборот¹⁹. Это разные таблицы, полученные при помощи разных моделей. Значения истинности в таблице 2.29 зависят от величины калибровочного коэффициента k_2 , тем самым, мы получаем серию таблиц, которые образуют параметрическое семейство гипотез об аномальности, и в качестве параметра этого семейства выступает коэффициент k_2 .

2.3.3. Калибровочные коэффициенты по совокупности условий

Итак, у нас есть две серии гипотез: гипотетические аномалии по времени, зависящие от калибровочного коэффициента k , и гипотетические аномалии по локациям, зависящие от калибровочного коэффициента k_2 . Для того, чтобы выдвинуть общую гипотезу нам нужно каким-то образом соединить две гипотезы в одну, но как это сделать: при помощи объединения или при помощи пересечения условий аномальности? Мы считаем, что ситуацию следует признавать аномальной, если выполняется хотя бы одно из двух условий аномальности, а это означает объединение условий. Другими словами, если ситуация является аномальной хотя бы в одном из двух рассматриваемых нами смыслов, то мы будем считать, что это аномальная ситуация. Формально, это означает, что для

¹⁹ Мы даже подчеркнули это при помощи разных обозначений: в одном случае мы использовали один штрих, а в другом — два штриха.

получения совокупной булевой таблицы нам нужно применить операцию сложения двух таблиц по правилу сложения логических выражений:

$$\begin{aligned} \text{True}' + \text{True}'' &= \text{True} \\ \text{True}' + \text{False}'' &= \text{True} \\ \text{False}' + \text{True}'' &= \text{True} \\ \text{False}' + \text{False}'' &= \text{False} \end{aligned}$$

Таблицы 2.25 и 2.29 обладают разной конфигурацией, однако их конфигурации нетрудно согласовать при помощи транспонирования (ниже мы транспонировали их к виду 2.29, но можно было бы точно так же транспонировать их к виду 2.25, это ничего бы не поменяло в дальнейших рассуждениях²⁰). После сложения получается булева таблица по совокупности двух типов аномалий:

Таблица 2.30 — Совокупный индикатор аномалий на этапе обучения

	Момент $\alpha n+1$	Момент $\alpha n+2$	Момент $\alpha n+3$	...	Момент n
Локация 1	$(T F)_{\alpha n+1,1}$	$(T F)_{\alpha n+2,1}$	$(T F)_{\alpha n+3,1}$	...	$(T F)_{n1}$
...	...	...	...	...	...
Локация m	$(T F)_{\alpha n+1,m}$	$(T F)_{\alpha n+2,m}$	$(T F)_{\alpha n+3,m}$	...	$(T F)_{nm}$

Эта таблица описывает двухпараметрическое семейство гипотез аномальности, она зависит от двух разных калибровочных коэффициентов: от k и от k_2 . Например, если мы будем варьировать границу k , придавая ей последовательно какие-нибудь 10 значений, и границу k_2 , придавая ей последовательно 20 значений, то мы получим 200 булевых таблиц 3.30, и, по сути — сформулируем 200 гипотез о норме и аномалиях, но какая из них будет наиболее близка к реальности, мы сможем узнать лишь на следующем этапе методики МССР, на этапе сопоставления алгоритмического решения с экспертным

²⁰ Можно было бы даже выстроить их в векторы, сохраняя общий порядок следования элементов, согласно индексам ij .

мнением (см. раздел 2.4).

2.4. Сравнение алгоритмической детекции с экспертной оценкой

Варьируя калибровочные границы, мы не утверждаем аномальность наверняка, мы всего лишь формулируем гипотезы об аномальности, которые требуют подтверждения (или опровержения) для того, чтобы из статуса гипотез перейти в статус истинных (или ложных) утверждений. Критерием истины, как известно, служит практика, а практические сведения об аномальности того или иного состояния процесса мы можем получить только в профессиональной среде той предметной области, которую мы изучаем, от специалистов, обладающих высокой квалификацией, другими словами — сверившись с мнением экспертного сообщества.

Этап сопоставления предусмотрен в методике МССР для сравнения алгоритмической детекции, которая получается на предыдущем шаге, с экспертной оценкой, которая не имеет никакого отношения к алгоритму, и можно считать, что она у нас есть априори. В этом разделе мы рассмотрим несколько сценариев проведения сопоставления, и покажем, при каких обстоятельствах этот этап считается успешно пройденным (или, наоборот, проваленным).

2.4.1. Один эксперт

Заметим, что алгоритмическую детекцию аномалий можно рассматривать как своеобразный бинарный классификатор, ведь по итогам своей работы он производит разбиение множества всех наблюдений на два класса: класс аномальных наблюдений и класс нормальных наблюдений. Своеобразие этого классификатора состоит в том, что классификация производится без предварительной разметки: на этапе обучения наш алгоритм «не знает» какое из наблюдений является нормальным, а какое — аномальным. Тем не менее, путём сложных рассуждений (см. выше, разделы 2.2 и 2.3) и ещё более сложных вычислений, алгоритм формирует такую бинарную разметку. Открытым остаётся лишь вопрос о том, насколько эта разметка удачна, но, чтобы ответить на

него у нас имеется широкий спектр метрик классификации: от незамысловатой метрики асугасу до учитывающих одновременно Recall и Precision метрик F1 и ROC AUC [2–5, 7–9, 23, 60, 172] (см. приложение А, где мы подробно разбираем метрики эффективности классификаторов). Для этого нужно иметь всего лишь два вектора: вектор предсказанных значений y_pred (а это набор компонент булевой таблицы 2.30) и вектор истинных значений y (осталось только понять, что такое истинные значения).

Если мы рассматриваем схему экспертной оценки с участием всего одного эксперта, то истинные значения — это и есть его мнение в виде нулей и единиц. Сопоставление алгоритмической и экспертной оценки удобно реализовать в виде следующей таблицы (см. таблицу 2.31).

Таблица 2.31 — Сопоставление алгоритмической и экспертной оценки

	y_pred	y
1	y_pred_1	y_1
2	y_pred_2	y_2
3	y_pred_3	y_3
...	...	...
N	y_pred_N	y_N

Здесь в столбцы развёрнуты компоненты булевых таблиц, полученных от алгоритма и от эксперта соответственно. Порядок, в котором следуют элементы несущественен, важно лишь то, что он должен быть одинаковым для обоих столбцов. Сами компоненты представляют собой логические выражения истинности True или False, однако чаще используют принятые в бинарной классификации 0 и 1, где 0 соответствует булеву значению False, а 1 — значению True. Отталкиваясь от таблицы 2.31, мы можем вычислить любую метрику бинарной классификации.

Метрика Accuracy. Это просто процент совпадений [7, 23, 60, 172, 174, 200, 234]. Метрика считается очень легко и ещё легче интерпретируется: высокий процент совпадений — хорошая классификация, низкий — плохая. Однако,

эта метрика лукавая, и доверять ей нельзя. Дело в том, что она хорошо работает только для сбалансированных классов, когда их представители присутствуют примерно в равной пропорции [7, 172]. В нашей ситуации, когда единицей мы помечаем аномалию, а нулём — норму, подавляющее большинство представителей таблицы 2.31 будут нулями, и если наш алгоритм вернёт все нули, то его метрика ассигасу будет очень высокой. Но будет ли это означать, что алгоритм хорошо распознал аномалии? — конечно, нет: он вообще ничего не распознал. Именно в силу природы изучаемой ситуации, в силу того, что аномалий не может быть много (их всегда мало, на то они и аномалии), удобная и легко интерпретируемая метрика ассигасу не подходит для наших задач.

Метрики Precision и Recall. Это более продвинутые метрики, причём, одну из них (Recall) вполне можно использовать для самостоятельной оценки. Напомним (см. приложение А), что Precision (точность) — это доля действительных аномалий, среди наблюдений, которые наш детектор обозначил как аномалии, а Recall (полнота) — это доля выявленных аномалий среди всех действительных аномалий [7, 48, 172]. Обычно считают, что эти метрики обладают равной значимостью. Причем, если мы каким-то образом повышаем метрику Precision, то при этом понижается метрика Recall, и наоборот: если мы повышаем Recall, то понижается Precision. В идеале классификатор должен был бы обладать высокими значениями обеих этих метрик, но повысить их одновременно невозможно.

Метрики F1 и ROC AUC. Наконец, метрики F1 и ROC AUC, учитывающие одновременно и Precision, и Recall, являются универсальными [7, 23, 60, 172, 174, 209] и вполне подходят для нашей задачи.

Мы помним (см. раздел 2.3.3), что, на самом деле, наш алгоритм возвращает не какой-то один конкретный вектор предсказаний `y_pred`, а двухпараметрическое семейство таких векторов, где в качестве параметров выступают калибровочные границы отклонений от прогнозов времени и по локациям.

Можно обозначить это семейство так:

$$y\_pred = y\_pred(k, k_2).$$

Теперь, когда у нас есть вектор истинных значений и метрики для сравнения предсказаний с истинными метками, мы производим цикл по двумерной сетке параметров для оптимизации метрики классификации и выделяем из семейства $y\_pred(k, k_2)$ оптимальное значение y_pred_{opt} .

Отметим, что именно в этот момент мы произвели настройку калибровочных границ (это те значения k и k_2 , которые доставляют оптимум выбранной нами метрике классификации), именно в этот момент мы из семейства гипотетических аномалий выбрали тот набор, который наиболее соответствует мнению эксперта (напомним, что пока у нас один эксперт), и именно эти границы мы будем использовать дальше, на этапе пилотной эксплуатации (см. раздел 2.5).

2.4.2. Группа экспертов

Однако, как правило, в предметной области есть не один эксперт, их, как минимум, несколько. И, как правило, их мнения не бывают полностью совпадающими: то есть, в большинстве типичных ситуаций, их мнения, конечно, совпадут, но, если ситуация нетипична, то вполне возможно, что кем-то она будет распознана как аномальная, а кем-то другим — как проявление нормы, хотя и с некоторыми оговорками.

Мы формируем экспертную группу, которая даёт свою оценку состояниям процесса, наблюдаемым на этапе обучения (см. таблицу 2.32).

Таблица 2.32 — Сопоставление алгоритмической оценки и оценки экспертной группы

	y_pred	y_1	y_2	...	y_M
1	y_pred_1	y_1_1	y_2_1	...	y_M_1
2	y_pred_2	y_1_2	y_2_2	...	y_M_2
3	y_pred_3	y_1_3	y_2_3	...	y_M_3
...	...	...	...	...	...

	y_pred	y_1	y_2	...	y_M
N	y_pred _N	y_1 _N	y_2 _N	...	y_M _N

Здесь каждый вектор y_i — это мнение i -го эксперта экспертной группы, а всего в группу входят M экспертов.

Самый простой способ провести сравнение алгоритма с экспертной группой состоит в том, чтобы агрегировать мнения отдельных экспертов путем прямого голосования и получить один столбец экспертных меток «норма/аномалия», после чего ситуация в точности воспроизведёт ту, что была рассмотрена нами в предыдущем разделе, когда у нас был один эксперт.

Но это неинтересно. Гораздо интереснее поставить экспертов и алгоритм в равные условия (выше это было не так, так как мнение эксперта априорно считалось истинным) и провести действительно челлендж (напомним, что слово *challenge* присутствует в аббревиатуре МССР, и одно из значений этого слова — оспаривание, отстаивание мнения). Это совсем другой сценарий, и мы остановимся на нем подробнее.

Идея состоит в том, что мнение каждого отдельного эксперта мы не будем считать истинным. В качестве истинной разметки мы примем агрегирование всех мнений прямым голосованием, а мнение отдельного эксперта будем считать гипотетическим, как и мнение нашего алгоритма. Для алгоритма мы подбираем калибровочные границы отклонения прогоном по двумерному циклу, оптимизируя тем самым, алгоритмический результат, а затем сравниваем качество классификации алгоритма и каждого отдельного эксперта по какой-либо метрике. Тем самым, возникает конкурирующая пара: с одной стороны, это алгоритм со своей гипотезой о норме и аномалии, а с другой стороны — один из экспертов со своим мнением. В качестве критерия истины выступает общая разметка экспертной группы, и в паре алгоритм-человек побеждает тот, чья метрика оказывается выше. Если после сравнения с каждым из

экспертов окажется, что алгоритм побеждал чаще²¹, то мы считаем, что алгоритм успешно прошёл челлендж и готов к этапу пилотной эксплуатации (см. ниже, раздел 2.5).

2.4.3. Коррекция алгоритма

Отдельно следует оговорить ситуацию, когда на этапе сопоставления алгоритм детекции аномалий показал неудовлетворительные результаты. Во-первых, это не катастрофа. Катастрофа произошла бы если бы мы допустили такой алгоритм к внедрению и, тем самым, не улучшили бы, а, наоборот: ухудшили бы показатели бизнес-процесса. Неудача на этапе сопоставления — это самая обычная рабочая ситуация, аналогичная неудовлетворительным результатам испытания какого-либо материального изделия. Это вовсе не повод бросить разработку и зафиксировать свои убытки (моральные, материальные и финансовые), а объективная причина внести необходимые изменения в технологические процедуры изготовления материального объекта. Алгоритм детекции аномальных состояний не является материальным объектом, но также обладает целым рядом характеристик, которые можно и при необходимости — нужно изменять.

Если алгоритм не прошёл сопоставления, мы возвращаемся на этап моделирования. Даже если просто комбинаторно перебирать различные методы формирования данных (локальные данные, глобальные данные, данные на ближайших моментах или локациях, которые мы рассматривали выше, в разделах 2.2.3 и 2.2.4), мы сможем получить уже достаточно большое разнообразие алгоритмических решений. Но можно пойти дальше и добавить к вариантам формата данных тюнинги²² [2–5, 7–9, 17, 240] участвующих в моделировании моделей из пула семи моделей, которые мы используем. А можно пойти

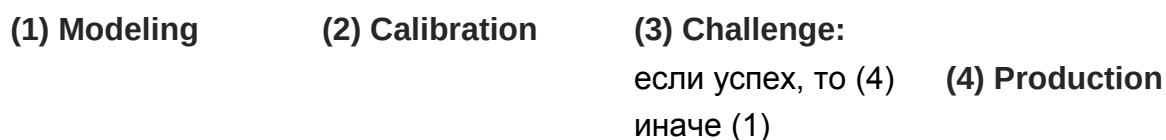
²¹ Это, конечно, необычайно жёсткое требование по отношению к алгоритму. Его можно смягчить, заменив «чаще» на «не реже».

²² Тюнингом модели на сленге Data Science называют подбор её гиперпараметров с целью повысить ее прогнозирующую способность.

ещё дальше и изменить сам пул моделей. В конце концов, кто сказал, что нам следует ограничить себя только перечисленными выше моделями. Можно применить приёмы полиномиальной регрессии и варьировать при этом ещё один дополнительный параметр: степень полинома.

Наконец, можно вспомнить, что помимо изучаемой характеристики процесса любая локация и любой момент времени может обладать и другими признаками (об этом мы говорили ещё при обсуждении концептуальных идей, в разделе 1.3.2), и начать изменения алгоритма с добавления новых дополнительных предикторов к нашим моделям, то есть, провести так называемый *feature engineering* (англ. — преобразование признаков) моделей нашего алгоритма [7, 15, 48, 173, 209].

Различных вариантов действий, которые можно использовать для улучшения детектора, необозримо много, и заранее дать однозначную рекомендацию, которая заведомо приведёт к успеху, просто невозможно. В любом случае разработка детектора аномалий по методике МССР представляет собой итерационную процедуру:



Как и материальное изделие, которое не сразу запускается в серию, а постепенно совершенствуется путём весьма длительных и разнообразных испытаний, алгоритм детекции аномалий тоже может в процессе разработки неоднократно повторять цикл моделирование — калибровка — сопоставление до тех пор, пока не будет допущен к внедрению и эксплуатации.

2.5. Пилотная эксплуатация алгоритма

Допустим, алгоритм прошёл этап сопоставления успешно. Возможно, на это ушло множество итераций (см. выше, раздел 2.4.3), но в результате он всё-таки показал приемлемое качество детекции аномалий и допущен к внедрению. В

этом разделе мы опишем процесс его пилотной эксплуатации.

2.5.1. Первый момент эксплуатации

В момент времени moment_n нам доступна полная история наблюдений характеристики процесса во всех локациях, включая n -е наблюдение, и то, что мы знаем о процессе в этот момент, представлено в следующей таблице (см. таблицу 2.33).

Таблица 2.33 — Данные в формате ML в первый момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент 1	C_{11}	C_{12}	C_{13}	...	C_{1m}
Момент 2	C_{21}	C_{22}	C_{23}	...	C_{2m}
...	...	...	...	...	...
Момент $n-1$	$C_{n-1,1}$	$C_{n-1,2}$	$C_{n-1,3}$	...	$C_{n-1,m}$
Момент n	C_{n1}	C_{n2}	C_{n3}	...	C_{nm}

Используя методы моделирования по времени, мы прогнозируем значения наблюдаемой характеристики в следующий момент, согласно алгоритму, изложенному нами выше (см. раздел 2.2.3, где излагалась процедура моделирования по времени). Получаем прогноз по времени на второй момент эксплуатации, полученный в первый момент эксплуатации (см. таблицу 2.34), который представляет собой одну строку (ведь в отличие от этапа обучения, теперь мы мы прогнозируем только один момент).

Таблица 2.34 — Прогноз по времени на второй момент эксплуатации, полученный в первый момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $n+1$	$P_{n+1,1}$	$P_{n+1,2}$	$P_{n+1,3}$	...	$P_{n+1,m}$

2.5.2. Второй и последующие моменты эксплуатации

Во второй момент наши сведения пополняются ещё одной строкой (см. таб-

лицу 2.35), и нам становятся известны истинные значения наблюдаемой характеристики в $(n+1)$ -й момент времени.

Таблица 2.35 — Данные в формате ML во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент 1	C_{11}	C_{12}	C_{13}	...	C_{1m}
Момент 2	C_{21}	C_{22}	C_{23}	...	C_{2m}
...	...	...	...	...	...
Момент n	C_n	C_n	C_n	...	C_n
Момент $n+1$	$C_{n+1,1}$	$C_{n+1,2}$	$C_{n+1,3}$	...	$C_{n+1,m}$

Первое, что мы делаем — сравниваем реальные значения из таблицы 2.35 с тем, что мы прогнозировали в таблице 2.34: берём разность, получаем таблицу отклонений (см. таблицу 2.36).

Таблица 2.36 — Абсолютные отклонения прогнозов по времени во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $n+1$	$\Delta_{n+1,1}$	$\Delta_{n+1,2}$	$\Delta_{n+1,3}$	...	$\Delta_{n+1,m}$

Затем используем тот факт, что калибровочная граница по времени нам уже известна и производим булеву бинаризацию отклонений: если отклонение выше границы, то это аномалия (мы заменяем его на True), если ниже — то норма (мы заменяем его на False). Получаем индикаторы аномалий по времени (см. таблицу 2.37).

Таблица 2.37 — Индикаторы аномалий по времени во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент $n+1$	$(F' T')_{n+1,1}$	$(F' T')_{n+1,2}$	$(F' T')_{n+1,3}$	...	$(F' T')_{n+1,m}$

Кроме того, используя моделирование по локациям, (см. раздел 2.2.4, где излагалась эта процедура), мы получаем прогнозируемые значения во всех локациях во второй момент эксплуатации (см. таблицу 2.38). Отметим, что, в отличие от прогнозов по времени, прогнозы по локациям строятся в момент, когда

очередные сведения уже получены, то есть, не на следующий момент наблюдений, а в тот же момент.

Таблица 2.38 — Прогноз по локациям на второй момент эксплуатации, полученный во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент n+1	$q_{n+1,1}$	$q_{n+1,2}$	$q_{n+1,3}$	...	$q_{n+1,m}$

Снова берём разность наблюдаемых и спрогнозированных значений, получаем отклонения по локациям (см. таблицу 3.39), которые позволят нам на следующем шаге интерпретировать их в виде бинарной метки (см. таблицу 2.40).

Таблица 2.39 — Абсолютные отклонения по локациям для наблюдаемой характеристики во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент n+1	$\Theta_{n+1,1}$	$\Theta_{n+1,2}$	$\Theta_{n+1,3}$	...	$\Theta_{n+1,m}$

Калибровочная граница по локациям нам тоже уже известна после этапа сопоставления, поэтому мы проводим бинаризацию получившихся отклонений по локациям: все значения, которые превышают найденную калибровочную границу, мы объявляем аномалиями (и заменяем на True), а все оставшиеся значения считаем нормой (и заменяем на False). В результате этих действий получаем индикаторы аномалий, которые были зафиксированы по отношению к другим локациям (см. таблицу 2.40).

Таблица 2.40 — Индикаторы аномалий по локациям во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент n+1	$(F'' T'')_{n+1,1}$	$(F'' T'')_{n+1,2}$	$(F'' T'')_{n+1,3}$	...	$(F'' T'')_{n+1,m}$

Складываем таблицы индикаторов 2.37 и 2.40 по правилу сложения логических выражений:

$$\text{True}' + \text{True}'' = \text{True}$$

$$\text{True}' + \text{False}'' = \text{True}$$

False' + True'' = True

False' + False'' = False

Получаем совокупные индикаторы аномалии во всех локациях во второй момент эксплуатации (см. таблицу 2.41).

Таблица 2.41 — Совокупные индикаторы аномалий во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент n+1	$(F T)_{n+1,1}$	$(F T)_{n+1,2}$	$(F T)_{n+1,3}$	...	$(F T)_{n+1,m}$

Ориентируясь на значения таблиц 2.36, 2.37, 2.39, 2.40 и 2.41 формируем отчёт об аномалиях, проявившихся во второй момент эксплуатации: если требуется просто отметить факт аномалии, то берём его из таблицы совокупных индикаторов 2.41, если требуется указать тип аномалии, то детализируем его за счёт таблиц 2.37 и 2.40, если же требуется указать ещё и выраженность аномалии, мы добавляем данные из таблиц 2.36 и 2.39.

Отчёт — это завершающий шаг анализа второго момента эксплуатации, в отчёте содержатся сведения о том, в каких локациях было зафиксировано аномальное значение наблюдаемой характеристики, каков был тип диагностированной аномалии и какова была её выраженность.

Мы понимаем, что вслед за вторым моментом эксплуатации последует третий момент, и нам нужно подготовиться к тому, чтобы сформировать отчёт по третьему моменту. Для этого нам, находящимся во втором моменте эксплуатации, нужен прогноз на следующий момент. Фактически, в нашем распоряжении есть таблица, аналогичная таблице 2.33, только в ней теперь на одну строку больше (см. таблицу 2.42).

Таблица 2.42 — Данные в формате ML во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент 1	C ₁₁	C ₁₂	C ₁₃	...	C _{1m}
Момент 2	C ₂₁	C ₂₂	C ₂₃	...	C _{2m}
...	...	...	...	...	...

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент n	$C_{n,1}$	$C_{n,2}$	$C_{n,3}$	...	$C_{n,m}$
Момент n+1	$C_{n+1,1}$	$C_{n+1,2}$	$C_{n+1,3}$	...	$C_{n+1,m}$

Используя моделирование по времени, получаем прогнозируемые значения (см. таблицу 2.43).

Таблица 2.43 — Прогноз по времени на третий момент эксплуатации, полученный во второй момент эксплуатации

	Локация 1	Локация 2	Локация 3	...	Локация m
Момент n+2	$P_{n+2,1}$	$P_{n+2,2}$	$P_{n+2,3}$	...	$P_{n+2,m}$

Теперь мы готовы к третьему моменту эксплуатации: мы обработаем его точно так же, как только что обработали второй, и по результатам получим отчёт с нужной степенью детализации обо всех аномалиях, которые наступят в третий момент эксплуатации.

Действия раздела 2.5.2 повторяются в каждый последующий момент до тех пор, пока не закончится этап пилотной эксплуатации.

2.5.3. Обобщение на несколько характеристик

Все сказанное выше относилось к ситуации, когда наблюдалась одна характеристика процесса. Обобщение на несколько характеристик не представляет никаких логической трудностей: достаточно сделать то же самое, подав на вход алгоритму данные о поведении других характеристик.

Изменения коснутся лишь последней стадии: формирования отчёта. Каждую характеристику диагностируем с двух точек зрения: аномальность по времени и аномальность по локациям, в результате чего получаем таблицу булевых значений, которая и является индикатором аномалий. Если же характеристик несколько, то у нас возникает несколько таких таблиц, и мы должны признать аномальным любое состояние, когда зафиксирована аномалия хотя бы по одной из наблюдаемых характеристик, то есть, снова выполнить сложение булевых таблиц по правилу сложения логических выражений (см. выше,

разделы 2.3.3 и 2.5.2).

Подробный отчёт детектора будет содержать записи следующей структуры (см. таблицу 2.44).

Таблица 2.44 — Структура отдельной записи в отчёте детектора аномалий

Момент	Номер момента или дата/ время
Локация	Имя локации
Характеристика	Название характеристики
Тип аномалии	По времени или по локациям
Выраженность аномалии	Величина отклонения от модели

2.6. Аудит локаций

Мы много раз выше по тексту диссертации утверждали, что методика МССР универсальна, и что процесс, который анализируется на предмет нормы/ аномалии может иметь любую природу. Это действительно так. Но с практической точки зрения возникает естественный вопрос: зачем нам вообще нужно знание о том, что в такой-то момент времени в такой-то локации наблюдаемая характеристика приняла аномальное значение. Ответ — затем, что тогда у нас появляется возможность проанализировать причины аномалии и, если она нежелательна, то принять меры к их устранению, а если аномальное значение носит положительный характер (например, аномально высокая прибыль нравится всем), то попытаться распространить её причины на другие локации и другие моменты времени.

Например, в пятой главе настоящей диссертационной работы мы рассмотрим приложение методики МССР к данным о работе сервисных центров одного из ведущих отечественных автопроизводителей, используя в качестве наблюдаемых характеристик 1) суточный объем заявок на гарантийное обслуживание, 2) суточную стоимость гарантийного обслуживания, а в качестве локаций — сервисные центры гарантийного обслуживания. Управление сервисной работой предприятия, в частности, включает в себя контроль работы сервисных центров и их непрерывный аудит, который, как правило, проводится

выборочным методом [6, 19, 38, 56, 154, 225, 236]. И с этой точки зрения знание о том, в каких именно локациях наблюдаются аномальные значения рабочих характеристик, имеет большое значение, так как именно это знание позволяет проводить аудит именно в тех локациях, которые демонстрирую существенные отклонения от нормы, и не отвлекаться на аудит локаций, работающих нормально (а заодно не отвлекать от работы людей, которые демонстрируют нормальные результаты).

Работающий на реальном процессе детектор в каждый момент времени помечает некоторые из локаций как аномальные, причём, возвращает не только бинарный факт норма/аномалия, но и выраженность аномальных состояний. Это позволяет специалисту, отвечающему за качество процесса, в каждый момент времени формировать рейтинг аномалий, упорядоченный по убыванию выраженности, что, в свою очередь, даёт возможность проводить аудит в соответствии с этим рейтингом, начиная с локаций, где аномальность выражена наиболее сильно. Такой подход к аудиту локаций обеспечивает высокую результативность проверок, которая, по некоторым оценкам, можеткратно превышать обычный аудит, основанный на методе случайной выборки. Мы подробно разберём этот вопрос в заключительной седьмой главе настоящего диссертационного исследования, после того как рассмотрим несколько практических кейсов применения нашей методики выявления аномальных состояний в нестабильных процессах.

2.7. Выводы по главе 2

Идеи и методы, изложенные в этой главе, являются ключевыми аспектами настоящей диссертации. Именно алгоритмика нашего детектора аномальных состояний составляет её основу, и здесь мы дадим беглый обзор её основных положений.

1. Все данные о протекании процесса разбиваются на три периода: период

исторических наблюдений, который необходим для обучения моделей, период обучения, на котором определяются критические величины отклонений прогнозируемых значений от истинных, и период пилотной эксплуатации, который нужен для верификации алгоритма.

2. Важно подобрать оптимальную схему прогнозирующего моделирования, варьируя при этом целый ряд параметров: модель, глубину признаков, локальность данных и т. д., включая гиперпараметры регрессионных моделей прогнозирования. В результате этого подбора исследователь имеет:
 - а) лучшую из прогнозирующих схем для прогнозирования по времени и
 - б) лучшую из прогнозирующих схем для прогнозирования по локациям.
3. Когда лучшие схемы становятся известными (это происходит после этапа моделирования), следует подобрать критические величины отклонений реальных значений наблюдаемой характеристики от спрогнозированных. Результатом этого подбора являются границы а) Δ_{cr} для прогнозирования по времени и б) Θ_{cr} для прогнозирования по локациям.
4. После этого, на этапе пилотной эксплуатации при помощи сравнения отклонений с критическими величинами каждая пара момент-локация получает две метки норма/ аномалия: а) по отношению к предыдущим наблюдениям, б) по отношению к другим локациям.
5. Этап пилотной эксплуатации нужен для верификации алгоритма детекции аномалий. Если доля выявленных аномалий на этом этапе существенно не совпадёт с долей аномалий, выявленных на этапе обучения, придётся все начинать с самого начала и проводить другое разбиение данных на три периода (см. п. 1).
6. Если этап обучения и сопоставления результатов с экспертной оценкой успешно пройден, а этап пилотной эксплуатации продемонстрировал согласованность результатов с этапом обучения, детектор допускается к промышленной эксплуатации.
7. В промышленной эксплуатации алгоритм возвращает не только факт наличия/ отсутствия аномалии в каждой локации в каждый момент времени, но

и её выраженность, что позволяет формировать рейтинг выявленных аномалий и проводить проверочные мероприятия в строго определённой последовательности: начиная с наиболее аномальных и далее по убыванию выраженности.

Таким образом, изложенная в этой главе алгоритмика составляет методологическую основу новых статистических методов для детекции аномалий в процессах, являющихся существенно нестабильными. В отличие от традиционных методов статистического анализа стабильных процессов [6, 19, 38, 56, 154, 162, 212], новая алгоритмика весьма сложна и требует привлечения методов машинного обучения и анализа больших данных [2–5, 7–9, 13, 15, 17, 23, 49, 65, 172, 20, 209, 234], но в результате она позволяет исследовать со статистических позиций весьма широкий класс процессов, который ранее был недоступен (см. выше, раздел 1.2 главы 1).

ГЛАВА 3. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ И ОГРАНИЧЕНИЙ МЕТОДИКИ МССР НА СИНТЕТИЧЕСКИХ ДАННЫХ

Использование синтетических данных представляет собой перспективный подход в рамках концепции всеобщего управления качеством (TQM), в частности, в контексте улучшения процессов и предиктивной аналитики. Роль синтетических данных в управлении качеством определяется целым рядом аспектов.

Первый из них состоит в преодолении ограничений реальных данных. В ряде отраслей и бизнес-процессов доступность релевантных данных для анализа качества ограничена в силу различных факторов [41, 77, 151, 208, 220], включая проблематику конфиденциальности (например, персональные данные клиентов, коммерческая тайна), высокие затраты на сбор и обработку, либо редкость наблюдаемых дефектов (например, в высокотехнологичном производстве). В таких ситуациях синтетические данные выступают в роли комплементарного ресурса, позволяющего преодолеть дефицит информации и обеспечить репрезентативность выборки для статистического анализа и моделирования.

В качестве еще одного из аспектов следует отметить проверку гипотез и моделирование сценариев. Синтетические данные позволяют реализовать подход «что-если» (англ. — what-if analysis) для оценки потенциального влияния различных факторов на показатели качества продукции и процессов. Это особенно актуально при внедрении новых технологических решений [103, 111, 127, 132], изменении параметров производственного процесса, либо при оценке рисков, связанных с нестабильностью поставок сырья и компонентов [80, 85, 97, 135]. Использование синтетических данных для моделирования различных сценариев позволяет заблаговременно выявлять потенциальные

проблемы и разрабатывать контрмеры, снижающие вероятность возникновения дефектов и отклонений.

Далее отметим, что синтетические данные часто бывают необходимы для обучения моделей машинного обучения и искусственного интеллекта, что напрямую связано с темой настоящей диссертационной работы [17, 51, 66, 67, 179, 195, 201]. Алгоритмы машинного обучения, используемые для решения задач классификации, регрессии и кластеризации в сфере управления качеством (например, автоматизированный контроль качества на основе компьютерного зрения, предиктивное обслуживание оборудования, прогнозирование отказов), требуют значительных объемов размеченных данных для достижения высокой точности и надежности. Синтетические данные могут быть использованы для дополнения реальных данных и создания сбалансированных обучающих выборок, что особенно важно при работе с редкими событиями (например, обнаружение аномалий в данных измерений) [2, 7, 172, 181, 192].

Итак, мы можем сказать, что эксперимент на синтетических данных — это весьма мощный инструмент, позволяющий исследовать возможности и ограничения методов машинного обучения [2–5, 7–9, 23, 49, 65]. Мы можем искусственно создать ситуации, которых никогда не дождались бы, ориентируясь только на реальные данные, полученные из практических источников (или ожидание наше затянулось бы: на десятилетия или больше). Мы можем усилить тот или иной эффект в синтетических данных, добиваясь его яркого проявления, или наоборот исключить ситуации, которые в данный момент нас не интересуют. Мы можем неограниченно наращивать объем данных, тогда как получение реальных данных часто бывает сопряжено со значительными трудностями или требует длительной паузы для их накопления. И так далее, и тому подобное: в каком-то смысле мы можем создавать и как угодно корректировать целый мир (конечно, не буквально, а оставаясь в рамках изучаемой предметной области).

Наиболее уместную, на наш взгляд, аналогию синтетическим данным

доставляет испытательный полигон, на котором конструкторы исследуют возможности и ограничения некоего опытного образца автомобиля (танка, бронетранспортёра и т. д.), который реализован в единичных экземплярах, но ещё не запущен в серию. Никому не приходит в голову выпускать опытный образец на дороги общего пользования и наблюдать его характеристики в реальных условиях: во-первых, это опасно для окружающих, а во-вторых, конструкторы не смогут испытать своё изделие в экстремальных режимах и определить его пределы прочности просто потому, что в реальности такие ситуации возникают крайне редко. Полигон строится специально для того, чтобы искусственно их симитировать.

Точно так же и синтетические данные: прежде чем подавать на вход алгоритму реальные сведения, мы хотим испытать его в разных режимах, оценить его возможности и понять, в каких условиях он будет возвращать приемлемые результаты, а в каких его вообще нельзя использовать.

3.1. Синтез данных в формате ML/ LM

Как мы уже неоднократно отмечали выше, ещё на этапе концептуального обсуждения наших основных идей, процесс — это характеристики процесса, измеренные во времени (моменты) и в пространстве (локации), другими словами, мы не делаем разницы между процессом и его описанием в виде набора данных в формате ML. Мы будем синтезировать процесс, который наблюдался в течение $n\_mom = 100$ моментов времени в $n\_loc = 9$ локациях²³. При этом мы хотим добиться того, чтобы процесс демонстрировал ярко выраженный нелинейный тренд. Другими словами, в этом разделе мы собираемся синтезировать такие данные, чтобы 1) в какие-то одни периоды наших наблюде-

²³ Разумеется, и число наблюдений, и число локаций являются изменяемыми, мы можем поменять их значения. Но в этом разделе мы будем использовать именно такие константы: 100 и 9 выбраны нами, потому что это точные квадраты, в связи с чем нам будет удобно визуализировать временные ряды и тепловые карты (см. рисунки 3.7 и 3.8).

ний значения наших характеристик возрастали, а в какие-то другие — убывали, и чтобы при этом 2) временные ряды в разных локациях не противоречили друг другу, а вели себя согласованно (что является вполне естественным требованием, так как логично предположить, что несмотря на какую-то разницу между локациями, процесс который мы в них наблюдаем, подчиняется какой-то общей для всех локаций логике). Для этого мы будем использовать фрагменты кубических парабол²⁴.

3.1.1. Репер кубической кривой

Хорошо известно, что через четыре точки проходит единственная кубическая парабола. Эти четыре точки образуют так называемый репер кубической кривой: если горизонтальные переменные реперных точек фиксированы, то разные реперы задают разные кривые, и наоборот (см. рис 3.1).

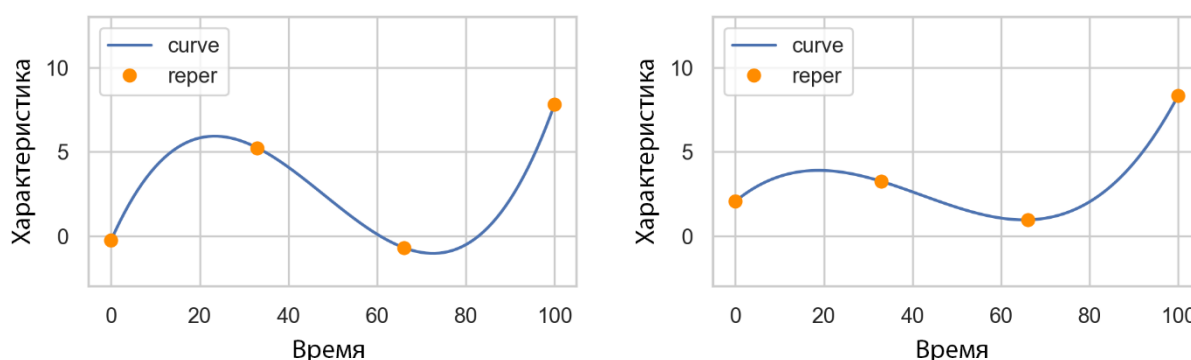


Рисунок 3.1 — Кубические параболы из одного семейства с разными реперами

Нам нужно 9 таких кривых (разных, но не сильно отличающихся друг от друга), тогда их вертикальные координаты дадут нам столбцы данных в формате ML (или, что то же самое, строки данных в формате LM), точнее —

²⁴ Можно было бы использовать и другие базовые функции, для описания поведения характеристик. Например, если бы мы решили синтезировать процессы с постоянным экспоненциальным ростом, мы, разумеется взяли бы для этих целей экспоненты, а не полиномы. Но тогда мы потеряли бы этапы убывания, а нам интересно посмотреть, как будет вести себя наш алгоритм на смене тренда, поэтому мы остановились на выборе кубической параболы, которая доставляет такую возможность, оставаясь при этом весьма простой для моделирования.

тренды столбцов (или строк) до наложения на них случайного шума.

Так как кривая тренда полностью определяется своим репером, для получения синтетических временных рядов нам достаточно зафиксировать горизонтальные координаты реперных точек, а вертикальные варьировать, оставаясь в разумных пределах, чтобы создавать всё-таки подобные друг другу, а не противоречивые временные ряды. Горизонтальные переменные мы выбираем так, чтобы они были равномерно распределены²⁵ в промежутке от 0 до числа наблюдений (см. рисунок 3.1, где горизонтальные позиции двух внутренних реперных точек равны 33 и 66 соответственно). Таким образом, для получения нужного нам семейства реперов у нас есть фиксированный список для значений горизонтальных переменных x , а для получения семейства вертикальных переменных мы сначала формируем фиксированный список y_base , который задаёт «прародителя» семейства (все представители семейства будут «похожи» на него, но сам он в это семейство не войдёт). Теперь при помощи генератора случайных чисел мы создаём ординаты реперных точек. Заметим, что и базовые ординаты прародителя, и величину разброса мы подбирали вручную, ориентируясь на визуальные качества получающихся иллюстраций и опыт работы с данными о реальных процессах. В результате мы синтезировали ситуации, близкие именно к той реальности, ради которой нам понадобились синтетические данные. Если бы анализируемые нами в дальнейшем реальные данные имели другой характер, мы подобрали бы входные параметры синтеза иначе, и подогнали бы его под другой вариант «реальности». В этом и состоит прелесть синтетических данных: синтезировать можно все что угодно. Теперь, если прогнать код в цикле по числу локаций $n\_loc = 9$, то на выходе мы получим девять реперов (см. рисунок 3.2, здесь точки, относящиеся к одному реперу, изображены одним цветом, хотя это и не очень заметно).

²⁵ Требование равномерности не является обязательным. Главное, чтобы выбранные абсциссы реперных точек не изменялись при переходе от одного репера к другому. Мы используем равномерный шаг, исходя из соображений наглядности.

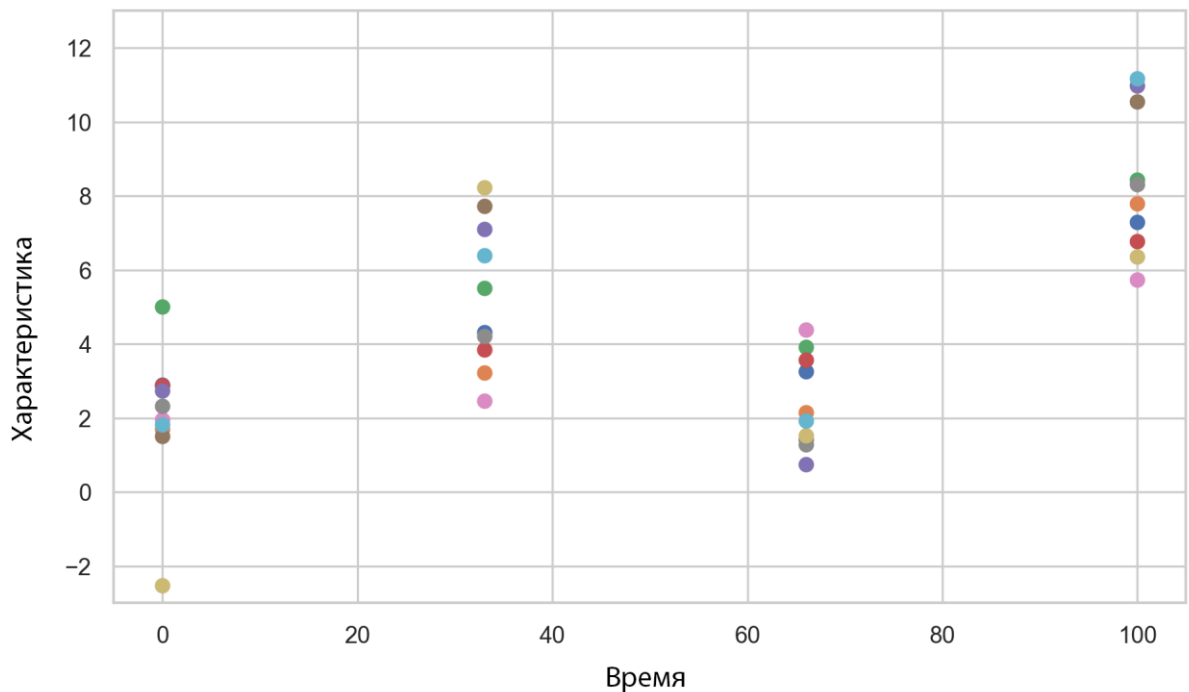


Рисунок 3.2 — Семейство реперов, порождающее семейство трендов

Остаётся построить по каждому реперу кривую и вывести её в датафрейм.

3.1.2. Кривая тренда, построенная на базовом репере

Для построения кубической параболы по четырём точкам мы воспользуемся идеями полиномиальной регрессии [2–5, 7–9, 49, 65, 172], хотя следует сразу отметить, что сейчас наши модели ни в коем случае нельзя рассматривать как прогнозирующие. Они ничего не прогнозируют, они лишь восстанавливают значения функции, относительно которой мы предполагаем, что она является полиномом третьей степени, в промежутках между реперными точками, это будут аппроксимирующие модели, а не прогнозирующие [15, 48, 52, 234].

Сначала мы покажем, как формируется кривая тренда на основе репера нашего базового прародителя (напомним, что `y_base = [2, 5, 2, 8]`). При помощи метода `DataFrame` из библиотеки `pandas` мы формируем пустой датафрейм `points` и последовательно добавляем к нему столбцы `'x'` (в котором записаны абсциссы репера), `'x^2'` (в котором записаны квадраты абсцисс), `'x^3'` (в котором записаны кубы абсцисс) и `'y'` (в котором записаны ординаты точек базового репера), используя при этом методы библиотеки `pandas` (см.

приложение В, где мы разбираем основные конструкции языка Python для анализа данных и машинного обучения, раздел В.1). Получаем датафрейм, выводим его на экран (см. таблицу 3.1).

Таблица 3.1 — Данные для формирования кривой тренда по базовому реперу

	x	x ²	x ³	y
1	0	0	0	2
2	33	1089	35937	5
3	66	4356	287496	2
4	100	10000	1000000	8

После этого при помощи модели `LearRegression` из библиотеки `sklearn` мы решаем регрессионную задачу с левой частью X и правой частью y, в результате чего получаем обученную модель (см. приложение В, где даны краткие описания моделей из библиотеки `sklearn`). Далее нам понадобятся все промежуточные значения горизонтальной переменной, заполняющие все целочисленные позиции от 1 до 100 (это наши синтетические моменты наблюдения за синтетическим процессом), именно к ним мы будем применять обученную только что модель. А помимо них нам понадобятся ещё их квадраты и кубы, потому что модель должна принимать на вход три переменных, а не одну. Поэтому мы формируем ещё один датафрейм, который в отличие от предыдущего содержит не четыре абсциссы, а 100.

После этого остаётся применить метод `predict` — это и будут искомые вертикальные координаты полиномиальной кривой третьей степени, проходящей, через заданный нами репер (см. приложение В, раздел В.2). Исходя из соображений наглядности, мы добавляем его последним столбцом в датафрейм, после чего он приобретает свой окончательный вид (см. таблицу 3.2).

Таблица 3.2 — Сформированные данные кривой тренда по базовому реперу

	x	x ²	x ³	y
1	0	0	0	2.000000
2	1	1	1	2.319371
3	2	4	8	2.620281

	x	x ²	x ³	y
...	...	...	...	...
98	99	9801	970299	7.544917
100	100	10000	1000000	8.000000

В его первом столбце представлены абсциссы, а в последнем — ординаты кривой тренда, сформированной в той предпосылке, что она проходит через базовый репер. Совсем нетрудно убедиться, что первая строка таблицы 3.2 совпадает с первой строкой таблицы 3.1, сотая строка совпадает с четвертой, тридцать третья — со второй, а шестьдесят шестая строка — с третьей. Визуально это означает прохождение кривой в точности через точки репера (см. рисунок 3.3). В этот момент мы восстановили все промежуточные значения кубической параболы, расположенные между реперными точками. Можно было бы прогнать цикл по более детализированной сетке, например, с шагом 0.1, или наоборот, по более грубой, например, с шагом 5, но для наших целей вполне подходят именно целочисленные элементы итераций, так как их легко интерпретировать как номера синтетических «наблюдений».

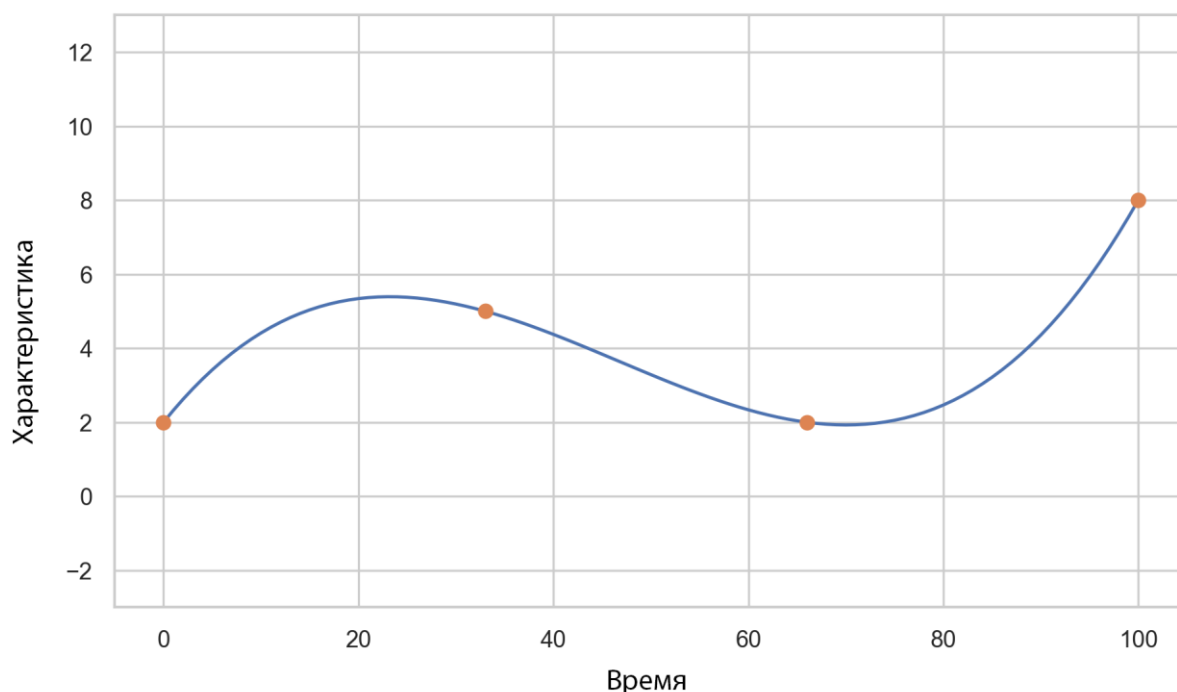


Рисунок 3.3 — Кривая тренда, проходящая через базовый репер

3.1.3. Семейство кривых тренда, построенное по семейству реперов

Изменяя по вертикали положения реперных точек, подобно тому как мы это делали выше (см. раздел 3.1.1) при помощи нормально распределенных случайных смещений, мы получаем разные кривые, принадлежащие, тем не менее, к одному семейству трендов. Ординаты этих кривых образуют набор трендов для нужных нам синтетических данных, мы выводим их во внешний файл. Далее запускаем цикл, воспроизводя действия предыдущего раздела в применении каждый раз к новому случайно сгенерированному реперу, и в конце каждой итерации заносим результат как столбец в датафрейм трендов. На выходе из цикла мы получаем заполненный датафрейм, столбцы которого представляют собой синтетические тренды временных рядов²⁶, которые мы будем изучать в дальнейшем, в следующих разделах этой главы. Ниже приведён пример²⁷ такого датафрейма (см. таблицу 3.3)

Таблица 3.3 — Синтетические тренды в формате ML

	Локация 1	Локация 2	Локация 3	...	Локация 9
Момент 1	2.451758	2.122593	3.010519	...	0.784744
Момент 2	2.649399	2.170495	3.204024	...	1.001408
Момент 3	2.834729	2.217924	3.386436	...	1.203654
Момент 4	3.008039	2.264890	3.557993	...	1.391822
...	...	...	...	...	...
Момент 99	8.249321	5.674356	5.691825	...	5.788233
Момент 100	8.649414	5.709504	5.941837	...	6.232037

Поскольку мы полностью владеем нашей синтетической ситуацией и можем изменять свой синтетический «процесс» произвольное количество раз, мы

²⁶ Особо подчеркнём, что это ещё не сами временные ряды, а их тренды, свободные от какого-либо случайного шума. Сами временные ряды мы получим позже, наложив на такие тренды синтетический шум (см. ниже, раздел 3.1.4).

²⁷ Мы говорим, что это всего лишь пример данных, а не сами данные, потому что при повторном запуске цикла мы получим другие значения в силу случайности реперов, задающих тренды. Именно поэтому в процессе синтеза мы выводим синтетические данные во внешние файлы, чтобы обращаться повторно (и, вообще, столько раз, сколько это понадобится) к тем же самым данным, а не к новым, заново синтезированным.

производим повторные запуски цикла синтеза трендов, и это даёт нам множество всевозможных вариантов поведения наблюдаемой характеристики синтетического процесса в девяти локациях, в которых мы его «наблюдаем» (см. рисунок 3.4).

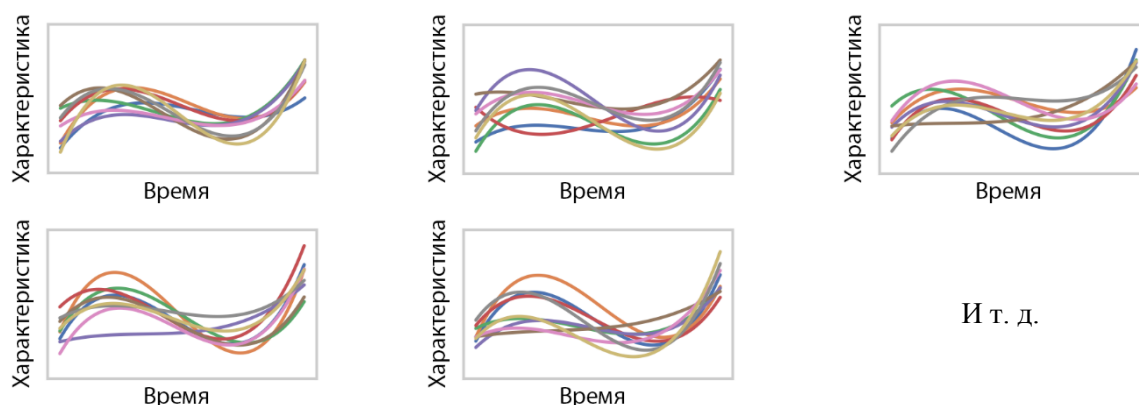


Рисунок 3.4 — Подобные семейства трендов

При каждом запуске цикла формируется новое семейство реперов (см. выше, раздел 3.1.1, в частности, см. рисунок 3.2), оно, в свою очередь, порождает новое семейство трендов, отличное от других семейств, но, тем не менее, обладающее весьма похожими свойствами. Для того, чтобы получить что-то принципиально новое, нужно изменить базовый репер, и это, разумеется, нетрудно сделать, но мы не станем его менять, потому что мы все равно синтезируем данные, лишённые какой бы то ни было семантики. Нам совершенно безразлично, что именно описывают кривые, приведённые выше как кривые трендов какого-то процесса. Нам важно изучить возможности методики МССР (см. выше, главу 2), поэтому гораздо важнее параметры нормального шума, которыми обладают наблюдаемые характеристики.

Таким образом, датафрейм, записанный в таблице 3.3, становится заготовкой для дальнейших исследований: мы можем накладывать на полученное семейство кривых случайные отклонения, имитируя, тем самым, реальные наблюдения реальной характеристики некоего реального процесса, но оставаясь при этом в рамках одних и тех же закономерностей (см. рисунок 3.5).

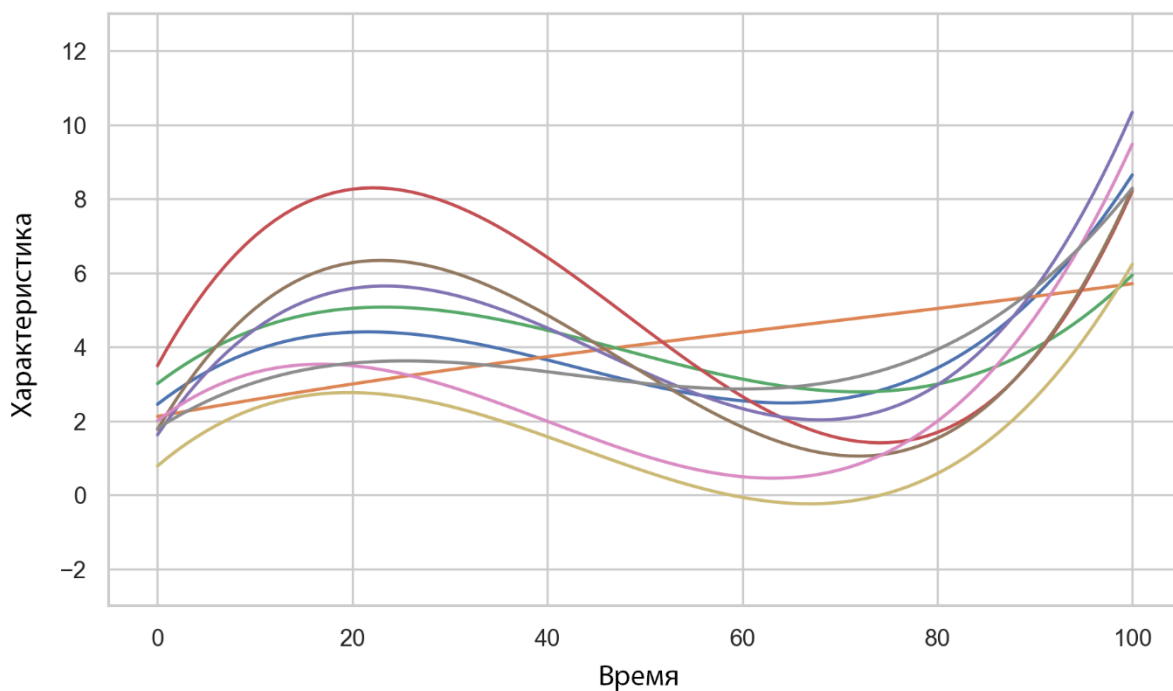
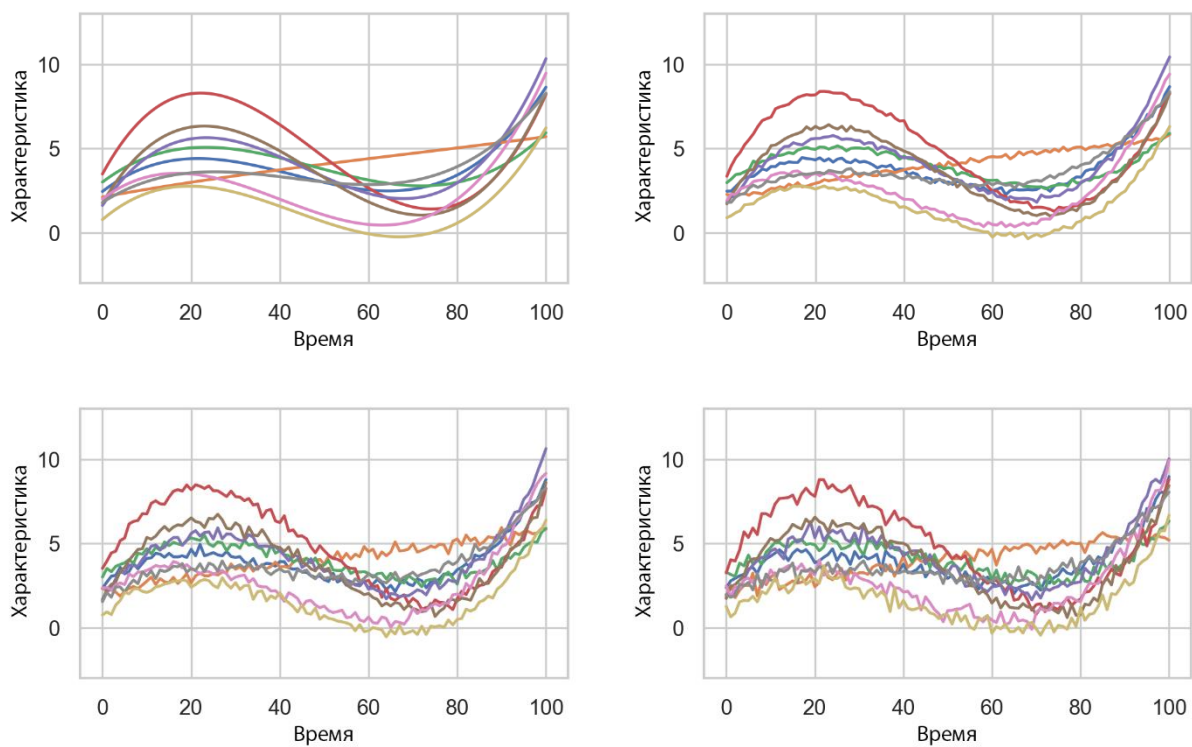


Рисунок 3.5 — Семейство трендов

Везде в дальнейшем в рамках этой главы мы будем работать именно с этим семейством трендов.

3.1.4. Синтез нормального шума



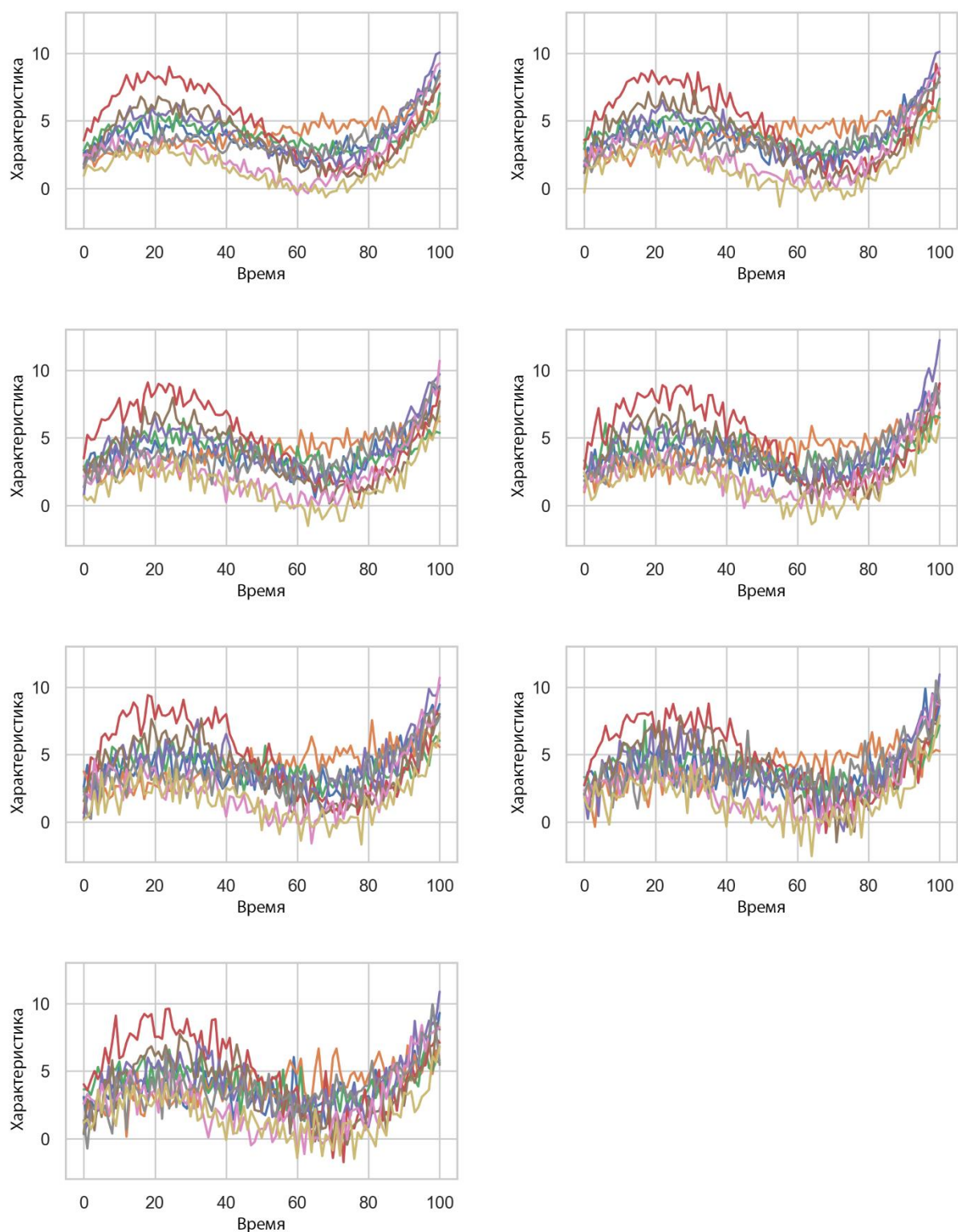


Рисунок 3.6 — Синтетические временные ряды с различной величиной шума

Дисперсия нормального шума, который мы накладываем на тренды, является важной величиной, мы варьируем её в пределах от 0 до 1 с шагом 0.1, в результате чего получаем 11 синтетических «процессов» различной степени загрязнённости нормальным шумом, начиная с абсолютно чистого процесса.

Визуализация синтетических данных представлена на рисунке 3.6, где первая иллюстрация обладает шумом $\sigma = 0$, последняя — шумом $\sigma = 1$, а все промежуточные отвечают увеличению шума с шагом 0.1 . Мы можем прочитать каждый из таких синтезированных наборов данных и получить датафрейм с данными, готовыми для дальнейшего использования. Например, синтетические данные в формате ML с дисперсией нормального шума $\sigma = 0.5$ представлены ниже в таблице 3.4.

Таблица 3.4 — Синтетические данные в формате ML с дисперсией нормального шума $\sigma = 0.5$

	Локация 1	Локация 2	Локация 3	...	Локация 9
Момент 1	2.556096	2.852326	1.925826	...	0.579310
Момент 2	2.963065	1.956280	3.053146	...	1.568918
Момент 3	2.904049	2.634007	2.977713	...	1.292128
Момент 4	3.040992	2.495277	3.145948	...	1.594622
...	...	...	...	...	...
Момент 99	7.990258	5.805157	5.505562	...	5.746973
Момент 100	9.230322	5.896291	5.240052	...	6.980886

Операция транспонирования переводит эти данные из формата ML в формат LM (см. таблицу 3.5).

Таблица 3.5 — Синтетические данные в формате LM с дисперсией нормального шума $\sigma = 0.5$

	Момент 1	Момент 2	Момент 3	...	Момент 100
Локация 1	2.556096	2.963065	2.904049	...	9.230322
Локация 2	2.852326	1.956280	2.634007	...	5.896291
Локация 3	1.925826	3.053146	2.977713	...	5.240052
Локация 4	4.213096	3.274485	4.606357	...	8.533112
Локация 5	1.223536	1.477708	2.487260	...	8.955087
Локация 6	1.667161	2.572267	3.235227	...	9.090019
Локация 7	2.115373	2.735169	1.585930	...	10.390180
Локация 8	2.257823	2.335698	2.392725	...	8.200170
Локация 9	0.579310	1.568918	1.292128	...	6.980886

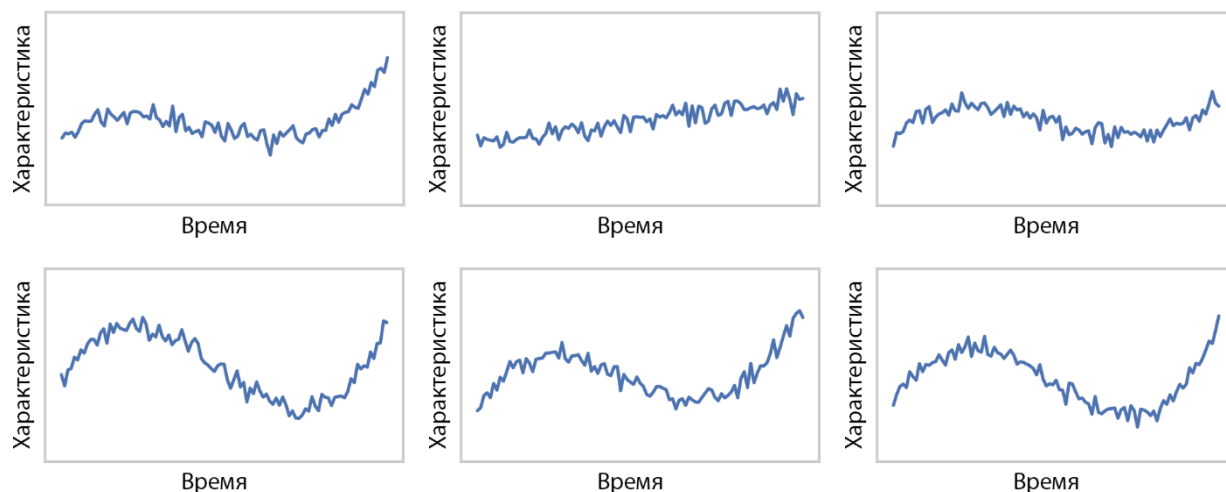
Отметим, что, если таблица 3.4 с данными в формате ML вытянута

сверху вниз (мы предполагаем, что история наблюдений синтетического процесса достаточно глубока), то таблица 3.5 с данными в формате LM наоборот — вытянута по горизонтали (число локаций, в которых проводится наблюдение синтетического процесса, сравнительно невелико).

3.1.5. Портреты локаций и портреты моментов

Напомним (см. главу 2, раздел 2.1), что локация описывается своей историей наблюдений, а портретом локации мы зовываем временной ряд предыдущих значений наблюдаемой характеристики, который может быть изображён в виде кривой.

Для иллюстрации портретов мы взяли процесс с уровнем шума $\sigma = 0.5$ и построили соответствующий кривые для имеющихся у нас девяти локаций (см. рисунок 3.7). Каждая кривая из этой серии представляет собой портрет соответствующей локации с точки зрения наблюдаемой характеристики²⁸, а совокупность кривых образует совокупность портретов, среди которых есть как более «похожие» друг на друга (то есть, обладающие меньшей разностью в смысле евклидовой метрики), так и менее «похожие».



²⁸ Здесь уместно ещё раз привести аналогию с климатическими характеристиками: если измеряемая характеристика — это температура воздуха, то все соответствующие кривые характеризуют историю наблюдаемой температуры в одной отдельно взятой локации за весь период наблюдений.

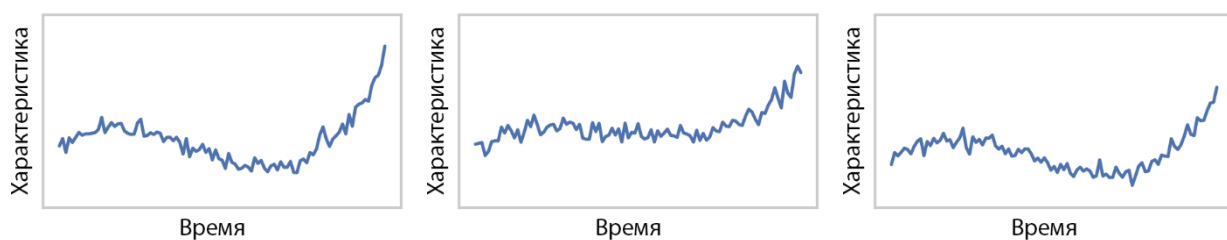


Рисунок 3.7 — Портреты девяти локаций синтетического процесса

Напомним также, что момент времени описывается совокупностью значений наблюдаемой характеристики, которые были зафиксированы сразу во всех локациях в данный момент. Эти значения могут быть изображены в виде тепловых карт, которые мы называем портретами моментов времени: на них отдельные локации изображаются в виде точек на плоскости, а величина наблюдаемой в данный момент характеристики в каждой такой точке индексируется насыщенностью цвета. Таких портретов моментов времени ровно столько, сколько было наблюдений в истории. Для иллюстрации мы взяли тот же синтетический процесс и проиллюстрировали его при помощи ста портретов²⁹, характеризующих сто моментов наблюдений (см. рисунок 3.8).



²⁹ Продолжая аналогию с климатическими процессами, отметим, что каждую иллюстрацию с этой серии можно воспринимать как сообщение о температуре в той или иной местности: чем насыщеннее цвет точки, тем жарче в данной локации в тот момент времени, который отвечает соответствующему моменту наблюдений.

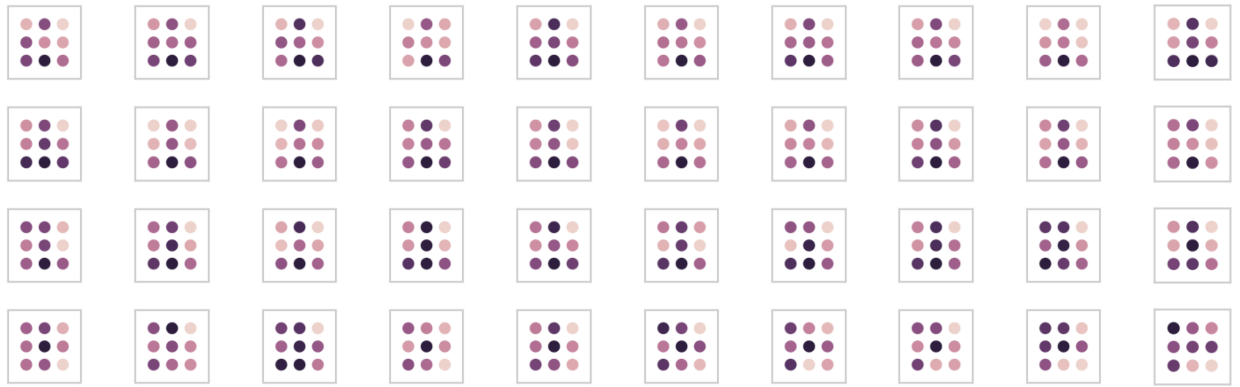


Рисунок 3.8 — Портреты ста моментов синтетического процесса

Важно понимать, что точки тепловых карт никак не привязаны к географии местности или геометрии помещения — они вообще не привязаны к реальным пространственным координатам локаций (которых, согласно семантике наблюдаемого процесса, может и не быть). Рисуя тепловые карты процесса, мы лишь иллюстрируем ту мысль, что, если моменты времени обладают естественным порядком, вследствие чего портреты локаций визуализируются в виде набора кривых, горизонтальная переменная которых естественным образом изменяется слева направо (см. рисунок 3.7), то локации не обладают никаким естественным порядком, они образуют неупорядоченное множество, которое проще всего изобразить в виде набора точек плоскости, так как точки плоскости представляют собой один из простейших примеров неупорядоченного множества. Глядя на эти карты, не следует думать, что движение от первой локации, расположенной в левом верхнем углу, к девятой локации, расположенной в правом нижнем углу, происходит в юго-восточном направлении. Портреты моментов, изображённые в виде тепловых карт — это абстракции. Как, впрочем, и портреты локаций, изображённые в виде кривых (просто к последним все мы привыкли с детства, со времён изучения школьной математики, и поэтому они не вызывают отторжения).

3.2. Моделирование по времени

Мы, разумеется, не станем рассматривать все процессы, синтезированные нами в предыдущем разделе, так как сейчас перед нами стоят совсем другие

задачи. Возможно, это имеет какую-то исследовательскую ценность (уровень шума $\sigma = 0$ даёт гладкие временные ряды, а на противоположном полюсе, при уровне шума $\sigma = 1$ наблюдается практически хаотичное чередование значений характеристики), но, тем не менее, мы остановимся на средней по уровню шума ситуации и везде далее в этой главе мы будем считать, что $\sigma = 0.5$.

3.2.1 Применение функции `modeling_by_moments`

Итак, в нашем распоряжении есть 100 наблюдений в 9 локациях (портреты моментов и портреты локаций мы изобразили выше, в разделе 3.1.5. Алгоритм функции `modeling_by_moments`, который является одним из ключевых алгоритмов диссертации, описан в приложении С (см. раздел С.1), и прежде всего нам нужно указать её параметры:

`path_in` — путь к входному файлу данных;

`history_end` — номер момента времени, в который заканчивается период исторических наблюдений;

`challenge_end` — номер момента времени, в который заканчивается период обучения;

`autoregression_dive_max` — максимальная глубина авторегрессии;

`model_for_fit` — модель, которая используется для обучения;

`path_out` — путь к выходному файлу.

Мы отведём на исторический период первую половину всех наблюдений, на период обучения их третью четверть, и на период пилотной эксплуатации оставшуюся четвертую четверть наблюдений³⁰. Таким образом, полагаем:

³⁰ Исследователь вправе изменить границы этапов. Например, уменьшив историю, отвести на исторические наблюдения первую треть наблюдений, или, наоборот, увеличив историю, дать ей первые две трети. Важно лишь помнить, что не следует доводить границы до абсурдных значений: история должна быть достаточно продолжительной, чтобы модели

```
history_end = 50
challenge_end = 75
```

При исполнении функции моделирования по времени мы будем строить авторегрессионные модели (см. главу 2, раздел 2.2.3) с различной глубиной авторегрессии от 1 до 10. Поэтому указываем:

```
autoregression_dive_max = 10
```

И, наконец, мы будем использовать пул из 7 регрессионных моделей, среди которых будут как несложные линейные модели, так и древовидные, и метрические, и даже одна нейросетевая модель (см. приложение А, где мы подробно обсуждаем различные виды регрессионных моделей). Этот пул мы сначала оформим в виде списка `models_list`, который нам понадобится для организации циклов по пулу моделей:

```
models_list = [LinearRegression(), DecisionTreeRegressor(), RandomForestRegressor(), GradientBoostingRegressor(), SVR(), KNeighborsRegressor(), MLPRegressor()]
```

А после этого применение функции `modeling_by_moments` осуществим в цикле по пулу моделей:

```
for model in models_list:
    modeling_by_moments(path_in = 'путь к входному файлу',
                        history_end = 50,
                        challenge_end = 75,
                        autoregression_dive_max = 10,
                        model_for_fit = model,
                        path_out = 'путь к выходному файлу')
```

На выходе из цикла возникает 7 директорий (по числу использованных моделей), в каждой из которых оказывается по 10 файлов `csv` (по глубине авторе-

могли на ней обучаться, и обучение тоже должно быть не слишком коротким, чтобы можно было делать неголословные выводы об эффективности той или иной модели.

грессии от 1 до 10), и каждый такой файл содержит 25 предсказаний в 9 локациях. Например, в директории LR, которая соответствует модели `LinearRegression`, находится файл `autoregression_dive_5.csv`, в котором записаны предсказания, сделанные линейной регрессионной моделью на локальных данных в формате ML с глубиной авторегрессии, равной пяти³¹. Директория DTR соответствует древесной модели `DecisionTreeRegressor`, в ней тоже находятся 10 файлов с предсказаниями на различных уровнях авторегрессии. И так далее.

Отметим, что процедура моделирования является самой сложной с алгоритмической точки зрения (нам даже пришлось вынести описание алгоритма в отдельное приложение, чтобы не перегружать основной сюжет диссертации) и самой продолжительной по времени: обучение некоторых моделей может занимать до нескольких часов, а с учётом того, что наш алгоритм использует не одну модель, а целый пул из нескольких моделей, продолжительность исполнения алгоритма увеличивается кратно. Но после того, как этот этап закончен, в нашем распоряжении есть все необходимое, чтобы выбрать наилучшую модель и схему наилучшего моделирования.

3.2.2. Выбор наилучшей схемы моделирования

У нас есть истинные значения измеряемой характеристики синтетического процесса, зарегистрированные на этапе обучения во всех локациях, в которых наблюдается наш «процесс» (ещё раз напомним, что это таблица из 25 строк и 9 столбцов, где строки отвечают моментам времени из периода обучения, а столбцы — локациям, в которых ведётся «наблюдение»).

Помимо этого, у нас есть ещё большое количество предсказаний нашего

³¹ Хотя вообще — там находятся 10 файлов, начиная с `autoregression_dive_1.csv` и заканчивая `autoregression_dive_10.csv`, а файл `autoregression_dive_5.csv` мы выбрали просто для примера.

алгоритма: мы применяли 7 моделей, каждую в 10 режимах авторегрессии, следовательно, у нас есть 70 различных (вполне возможно, что и не сильно отличающихся, но точно не одинаковых) предсказаний для этапа обучения. Чтобы выбрать из всех схем моделирования наилучшую, нам нужно сравнить истинные значения с предсказанными значениями, используя при этом какую-либо метрику эффективности регрессионной модели (см. приложение А, раздел А.6, где мы подробно рассматривали метрики прогнозирующей эффективности регрессионных моделей), например, при помощи хорошо известного [2–5, 7–9, 23, 52, 60, 65, 172, 173, 201, 204, 231, 234] и хорошо себя зарекомендовавшего коэффициента детерминации³². Для этого мы применим метод `r2_score` из модуля `metrics` библиотеки `sklearn` (см. приложение В, раздел В.2, где мы привели обзор методов анализа данных на Python).

На иллюстрации ниже (см. рисунок 3.9) по горизонтали отложены глубины авторегрессии, а по вертикали — значение коэффициента детерминации. Мы видим, что наилучшие среди всех значения демонстрируют древесная модель `RandomForestRegressor` и метрическая модель `KNeighborsRegressor`, но это только если подходить к выбору наилучшей схемы формально. Дело в том, что:

1. они достигают своих наиболее высоких значений при наибольших глубинах авторегрессии, а, следовательно, и при наибольших размерностях регрессионных задач;
2. обе эти модели являются чрезвычайно затратными точки зрения времени их обучения.

³² Можно было бы применить и другие метрики эффективности — такие, например, как средняя квадратичная или средняя абсолютная ошибки. Мы применяем коэффициент детерминации, потому что он не никак зависит от размерности и смыслового содержания целевой переменной, а это значит, что при помощи него мы в конечном счете сможем сравнивать результаты, полученные в этой главе на синтетических (то есть, безразмерных) данных, с результатами других глав, где методика МССР будет применяться к реальным данным, обладающими вполне конкретными единицами измерения.

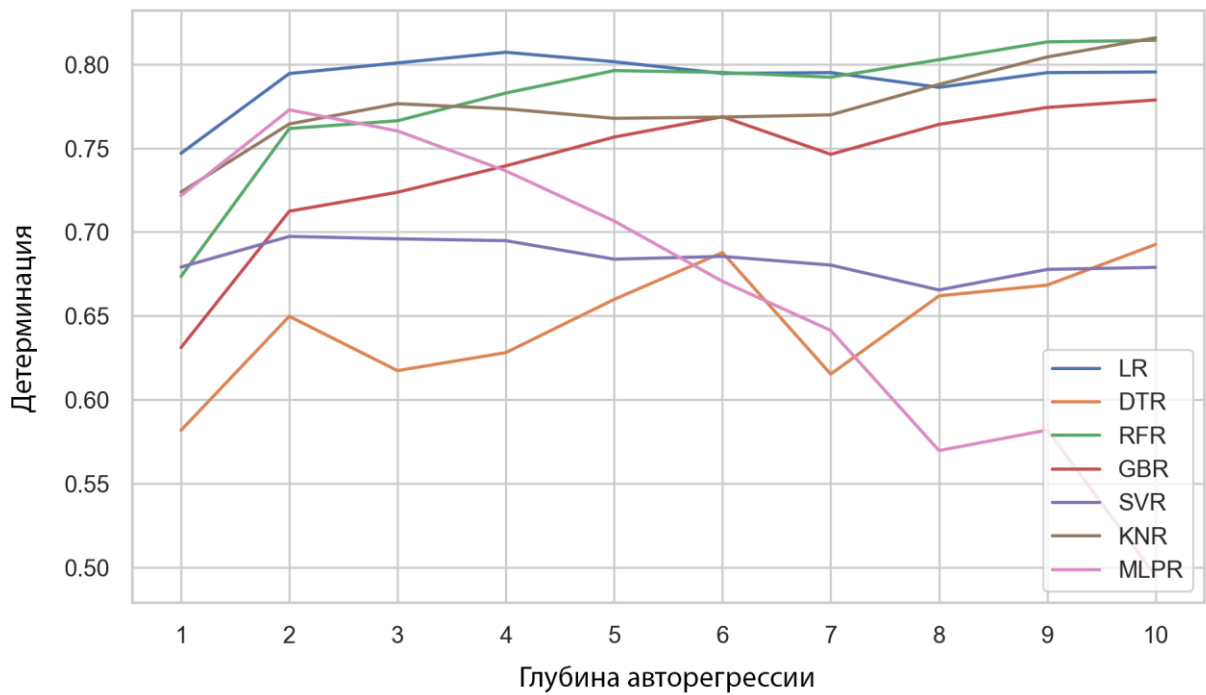


Рисунок 3.9 — Сравнительный анализ схем моделирования по времени

Учитывая эти обстоятельства (а также, принимая во внимание хорошо известный среди разработчиков принцип KISS — англ. keep it simple stupid, что в вольном переводе означает «Не усложняй» или «Не стоит использовать сложные модели тогда, когда можно обойтись простыми» [2–5, 7–9, 23, 52, 48, 172, 200]), мы примем в качестве лучшей схемы моделирования самую простую модель из нашего пула моделей: линейную регрессионную модель `LinearRegression` на глубине авторегрессии, равной четырём (см. рисунок 3.9). Она лишь немного уступает в точности прогноза как древесной модели `RandomForestRegressor`, так и метрической модели `KNeighborsRegressor`, при этом для достижения хорошей прогнозирующей способности требует гораздо меньшей глубины авторегрессии, а кроме того, хорошо известно, что очень быстро учится [2–5, 7–9, 13, 52, 174].

3.3. Моделирование по локациям

В этом разделе мы продолжим изучение синтетических данных с уровнем загрязнённости `sigma = 0.5`, которое мы начали выше, но на этот раз с

другой точки зрения. Теперь мы будем апеллировать не к прошлым состояниям процесса, а к его состояниям в других локациях. Для этого нам понадобятся данные в формате LM, где наблюдениями выступают локации, а признаковые описания даются моментами времени (напомним, что данные в формате LM очень легко получаются из данных в формате ML при помощи простой операции транспонирования, см. главу 1, раздел 1.4.1).

3.3.1. Применение функции `modeling_by_locations`

Описание функции `modeling_by_locations` приведено в приложении С (см. раздел С.2). Её алгоритм несколько проще, алгоритма моделирования по времени, потому что нам не приходится использовать приём авторегрессии, однако мы не станем сейчас останавливаться на особенностях его внутренней структуры, отнеся эти особенности к техническим деталям, и предоставив интересующемуся читателю самостоятельно разобраться в специфических различиях функций `modeling_by_moments` и `modeling_by_locations` (для этого можно обратиться к разделам С1 и С2 настоящей диссертации). С точки зрения непосредственного использования, этот алгоритм, так же как и алгоритм моделирования по времени, возвращает прогнозируемые значения наблюдаемой характеристики процесса для всех моментов этапа обучения (хотя и с принципиально иной точки зрения) и формально отличается от него лишь тем, что вместо глубины авторегрессии мы используем в качестве одного из входящих параметров глубину истории наблюдений. Параметры функции таковы:

`path_in` — путь к входному файлу данных;

`history_end` — номер момента времени, в который заканчивается период исторических наблюдений;

`challenge_end` — номер момента времени, в который заканчивается период обучения;

`history_dive_max` — максимальная глубина истории наблюдений;

`model_for_fit` — модель, которая используется для обучения;
`path_out` — путь к выходному файлу.

Мы используем те же самые границы этапов исторических наблюдений и обучения, что и выше, когда мы занимались прогнозированием по времени:

```
history_end = 50  
challenge_end = 75
```

Вместо глубины авторегрессии указываем максимальную глубину истории наблюдений, которую будем варьировать от 1 до 10:

```
history_dive_max = 10
```

Пул регрессионных моделей берём тот же, что и в предыдущем разделе (и так же оформляем его в виде списка для запуска цикла):

```
models_list = [LinearRegression(), DecisionTreeRegressor(), RandomForestRegressor(), GradientBoostingRegressor(), SVR(), KNeighborsRegressor(), MLPRegressor()]
```

Применяем `modeling_by_locations` в цикле по пулу моделей:

```
for model in models_list:  
    modeling_by_moments(path_in = 'путь к входному файлу',  
                        history_end = 50,  
                        challenge_end = 75,  
                        history_dive_max = 10,  
                        model_for_fit = model,  
                        path_out = 'путь к выходному файлу')
```

В результате — на выходе из цикла мы получаем 7 директорий (по числу задействованных моделей), в каждой из которых находится по 10 файлов `csvs` (по глубине истории наблюдений), причем, каждый такой файл содержит 25 предсказаний (по числу моментов времени на этапе челледжа) для каждой из 9 локаций (столько локаций мы синтезировали в самом начале).

3.3.2. Выбор наилучшей схемы моделирования

Так же как и выше, у нас есть истинные значения характеристики, которые были зарегистрированы на этапе обучения во всех девяти локациях, где наблюдается наш «процесс» (только теперь это транспонированная таблица из 9 строк и 25 столбцов, где строки отвечают локациям, в которых ведётся наблюдение, а столбцы — моментам времени). После этапа моделирования по локациям у нас есть 70 таблиц с предсказаниями (так как у нас работают 7 моделей, каждая из которых отрабатывает на 10 глубинах истории наблюдений), и мы хотим сравнить предсказанные значения с истинными значениями при помощи выбранной нами метрики эффективности регрессионной модели, то есть, при помощи коэффициента детерминации, так как именно его мы использовали для прогнозов по времени (см. выше, раздел 3.2).

Совокупные результаты исследования различных схем моделирования по локациям представлены на иллюстрации ниже (см. рисунок 3.10).

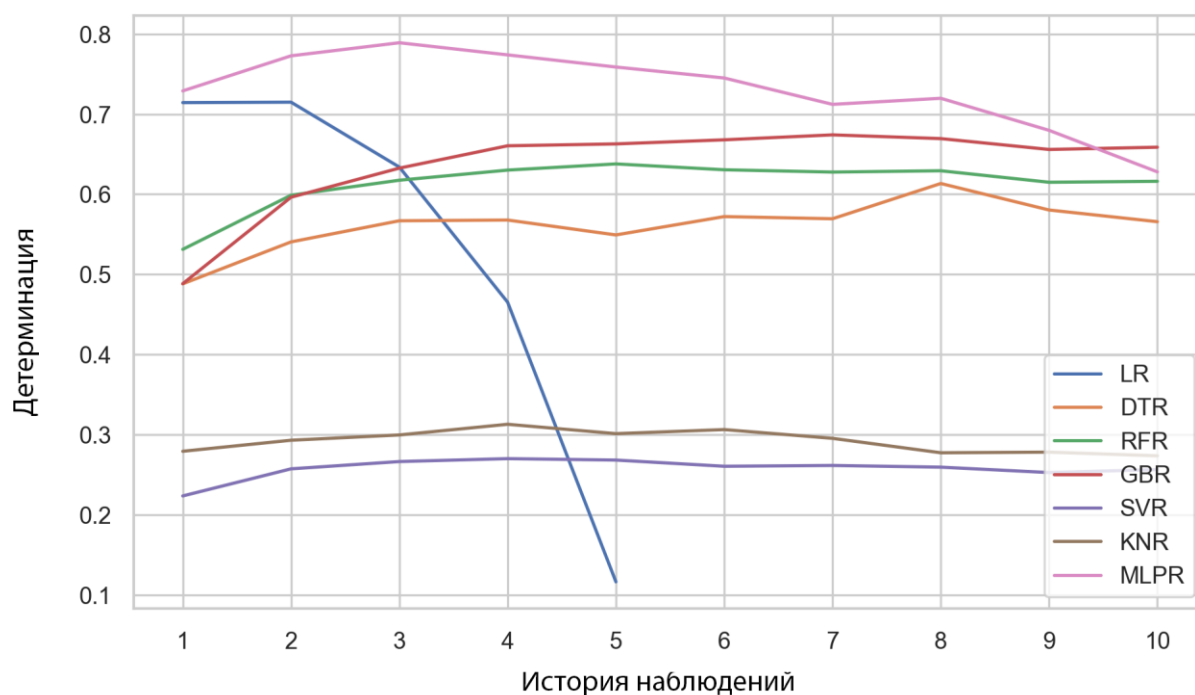


Рисунок 3.10 — Сравнительный анализ схем моделирования по локациям

Здесь по горизонтали отложены глубины истории наблюдений, которые изменяются от 1 до 10, а по вертикали — значение коэффициента детермина-

ции. Так же как и в предыдущем разделе, где мы иллюстрировали прогнозирование по времени, теперь, исполняя прогнозирование по локациям, мы приводим семь кривых по числу использовавшихся нами моделей: от линейной модели `LinearRegression` до нерасеиваемой модели `MLPRegressor`, включая древовидные модели `DecisionTreeRegressor` и `RandomForestRegressor` и построенные на случайных лесах знаменитый градиентный бустинг.

Победитель в этом семействе моделирующих схем очевиден, но при этом весьма интересно поведение линейной модели `LinearRegression`: ее предсказательная способность стремительно падает с увеличением глубины наблюдений³³. Этот феномен объясняется довольно легко [7, 9]. Дело в том, что алгоритм линейной регрессионной модели предполагает псевдорешение переопределенной линейной системы, которая возникает только если число уравнений в системе больше числа переменных (образно говоря, матрица системы должна быть вытянутой по вертикали). Если же число уравнений равно числу переменных или превышает его, то система перестаёт быть переопределенной (в случае равенства она вообще является крамеровской) и у неё возникает точное решение, которое абсолютно точно аппроксимирует обучающие данные, но на любых новых данных, на которых и требуется спрогнозировать целевую переменную, даёт прогноз, который вообще не соответствует действительности. Именно этот эффект и наблюдается на графике. Пока глубина истории наблюдений мала (например, для глубин 1 или 2), прогнозирующая способность линейной модели является одной из лучших, уступая только многослойному перцептрону `MLPRegressor`. Но с ростом глубины увеличивается число предикторов модели (то есть, число переменных в линейной системе), тогда как число наблюдений (то есть, число уравнений в системе) остаётся тем же — их всего лишь 9. Система с ростом числа предикторов становится все

³³ На самом деле, приближаясь к глубине истории 10, коэффициент детерминации линейной модели падает до отрицательных значений порядка -10^4 . Мы обрезаем график, чтобы увидеть различия, потому что иначе графики других моделей сливались бы визуальным образом в одну горизонтальную линию.

менее переопределенной и все более крамеровской, в результате чего линейная модель быстро теряет прогнозирующую способность³⁴. Интересно так же то, что в целом все модели из нашего пула моделей дают меньшее значение метрики эффективности по сравнению с моделированием по времени (см. рисунок 3.9), а отдельные из них, такие как SVR или KNeighborsRegressor, демонстрируют существенно низшие результаты.

Однако, целью наших исследований было вовсе не объяснение всех феноменов и артефактов моделирования, а выбор наилучшей модели с наивысшей прогнозирующей способностью, и наши выводы таковы: по итогам этапа моделирования по локациям мы признаем наилучшей модель MLPRegressor на глубине истории наблюдений 3. Именно её мы будем использовать для калибровки отклонений (см. разделы 3.5 и 3.6) и далее, на этапе пилотной эксплуатации (см. раздел 3.7).

3.4. Синтез экспертного мнения

В этом разделе мы синтезируем экспертное сообщество, которое даст нам управляемую нами экспертную оценку для состояний процесса во всех имеющихся у нас 9 локациях по всем 25 моментам наблюдений из периода обучения. Управляемость будет состоять в том, что у нас появится некий параметр, изменение которого позволит нам синтезировать разные мнения, другими словами — здесь мы получим однопараметрическое семейство экспертных сообществ. Имея таких управляемых «экспертов», в следующих разделах мы сможем поставить наш алгоритм в разные условия и посмотреть, насколько он справляется с разными экспертными мнениями, и вообще — сможет ли он пройти процедуру сопоставления с экспертным мнением, и можно ли будет допустить его к дальнейшей работе на этапе пилотной эксплуатации.

³⁴ Своего минимума кривая метрики линейной регрессии достигает как раз при глубине, равной 9, и этот минимум имеет порядок -10^{22} .

3.4.1. Отклонения лучшей модели от реальных значений

У нас имеются истинные значения наблюдаемой характеристики из периода обучения алгоритма с 50 по 74 наблюдение (см. таблицу 3.5)

Таблица 3.5 — Истинные значения периода обучения в формате ML

	Локация 1	Локация 2	Локация 3	...	Локация 9
Момент 51	2.314840	4.327486	3.919289	...	0.415913
Момент 52	3.758483	4.354091	4.391781	...	-0.081912
Момент 53	3.213714	3.960747	2.468318	...	0.592205
Момент 54	2.522497	4.190926	3.529557	...	0.152768
...	...	...	...	...	...
Момент 74	2.310891	5.339450	2.915688	...	0.108403
Момент 75	2.151387	4.483955	3.093085	...	-1.150825

И имеются значения, предсказанные для периода обучения лучшей из наших моделей³⁵ (см. таблицу 3.6).

Таблица 3.6 — Предсказанные значения для периода обучения в формате ML

	Локация 1	Локация 2	Локация 3	...	Локация 9
Момент 51	3.515374	3.809291	4.411191	...	0.534790
Момент 52	3.150480	4.098408	4.530384	...	0.626117
Момент 53	3.315349	4.343403	4.491420	...	0.152590
Момент 54	3.248649	4.187048	4.216375	...	0.306551
...	...	...	...	...	...
Момент 74	3.105256	4.398571	3.301345	...	-0.048120
Момент 75	2.991011	4.415448	3.219534	...	-0.099135

Теперь мы можем взять разность между истинными значениями и предсказанными, получив, тем самым, набор отклонений (см. таблицу 3.7).

Таблица 3.7 — Отклонения предсказанных значений от истинных в формате ML

	Локация 1	Локация 2	Локация 3	...	Локация 9
Момент 51	-1.200534	0.518195	-0.491902	...	-0.118877
Момент 52	0.608003	0.255684	-0.138603	...	-0.708029
Момент 53	-0.101635	-0.382656	-2.023102	...	0.439615
Момент 54	-0.726152	0.003878	-0.686817	...	-0.153783

³⁵ Напомним, что это была модель LinearRegression с глубиной авторегрессии 4.

	Локация 1	Локация 2	Локация 3	...	Локация 9
...	...	...	...	...	...
Момент 74	−0.794364	0.940879	−0.385657	...	0.156523
Момент 75	−0.839624	0.068506	−0.126449	...	−1.051690

Мы записываем отклонения в датафрейм `delta`. Заметим, что эти значения по своей природе являются ошибками измерения, которые неизбежно возникают в любом экспериментальном исследовании, и, как правило, подчиняются нормальному закону распределения, также известному как распределение Гаусса [6, 81, 83, 160, 163, 164, 206]. На иллюстрации ниже нами представлена гистограмма распределения отклонений, описанных таблицей 3.7 (см. рисунок 3.11). Ее, разумеется, нельзя назвать строго нормальной или, чтобы избежать путаницы в терминах (ведь понятие нормальности мы в рамках настоящей диссертации используем как антоним понятия аномальности), строго гауссовой. Тем не менее, распределение весьма близко к колоколу кривой Гаусса, и это легко подтверждается критерием Шапиро.

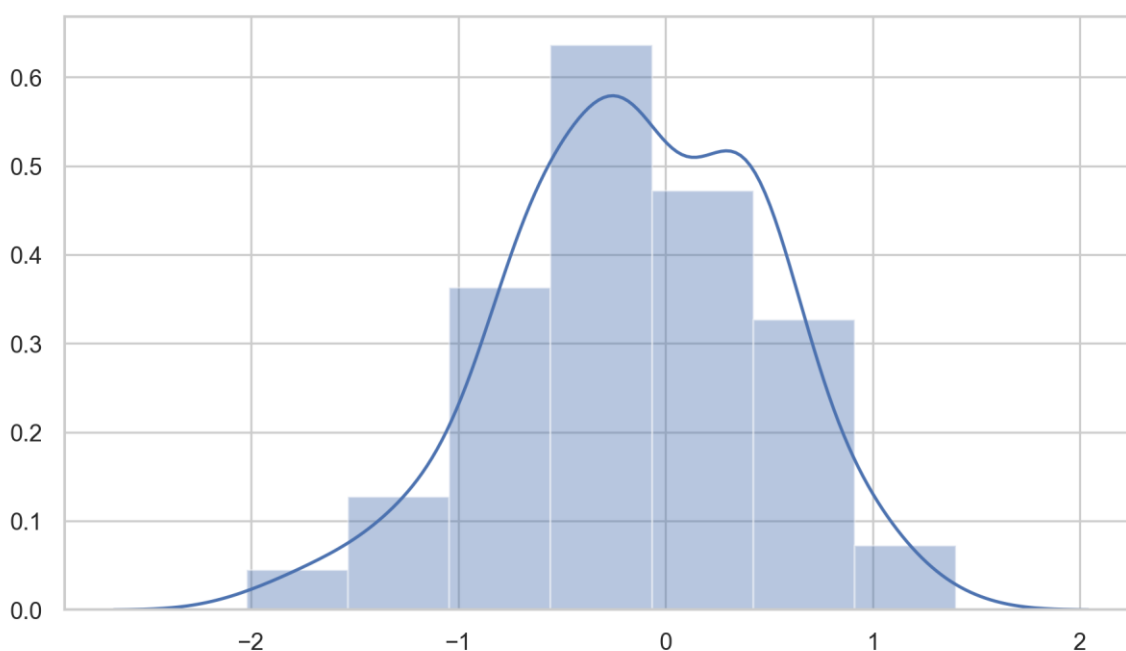


Рисунок 3.11. — Распределение отклонений `delta`

Важной характеристикой этого распределения является его разброс, который определяется сигмой — ведь именно величина отклонения даёт нам повод го-

ворить об аномальности (или нормальности) очередного наблюдения. Вычисление сигмы — это совсем несложная процедура, и выполняя ее, мы получаем средний разброс отклонений прогнозов нашей лучшей модели от реальных наблюдений:

`sigma = 0.628955`

3.4.2. Синтез экспертного сообщества

Как мы только что отмечали, отклонения **delta** подчиняются нормальному закону распределения, а это значит, что отклонение, превышающее 3σ маловероятно [6, 19, 20, 25, 38, 81, 56, 83]. Следовательно, вряд ли найдётся такой эксперт, который отклонение, большее чем 3σ сочтёт нормальным. С другой стороны, если отклонение будет меньше σ , то опять-таки трудно представить себе эксперта, который сказал бы, что такое отклонение аномально. Таким образом, мнения экспертов относительно нормальности или аномальности наблюдаемого отклонения могут и должны различаться, но пороговое значение, превышение которого эксперт сочтёт аномалией, для любого эксперта лежит где-то в пределах от σ до 3σ . Если отклонение меньше σ то ни один здравомыслящий эксперт не сочтёт его аномальным, и если отклонение больше 3σ , ни один эксперт не посчитает его нормальным. Эти соображения мы положим в основу нашего метода синтеза экспертного сообщества.

Экспертное сообщество неоднородно, в нем заведомо присутствуют разные мнения относительно того что есть норма, и что есть аномалия, и с этой точки зрения нам следует выделить две категории экспертов. К первой категории мы отнесём условных ястребов³⁶, то есть, тех, кто считает аномальным самое незначительное отклонение реальности от прогноза. Ко второй категории мы, наоборот, отнесём тех экспертов, которые даже весьма значительные (но

³⁶ Ястреб отличается высокой степенью нетерпимости. Малейшее отклонение он готов объявить аномалией.

всё-таки не беспредельные) отклонения считают одним из проявлений нормы. Они не склонны к резким заявлениям и предпочитают как можно реже говорить об аномальности того или иного состояния процесса. Их мы назовём условными голубями³⁷.

Формализуя эти размышления, мы будем считать ястребами тех экспертов, для которых критический порог отклонения лежит в пределах от σ до 2σ , а голубями — тех, для которых критический порог лежит в пределах от 2σ до 3σ . Это, разумеется, большая условность. Например, эксперт с порогом толерантности 1.9σ согласно нашей типологии, является ястребом, а эксперт с порогом толерантности 2.1σ — голубем, хотя между ними нет большой разницы, и правильнее было бы назвать их обоих центристами. Но для наших целей вполне достаточно разделения экспертов на два лагеря.

Далее мы будем формировать синтетическое экспертное сообщество, изменяя один параметр: долю ястребов. Этот параметр изменяется в пределах от 0 до 1. Нулевая доля ястребов означает, что все наши эксперты — голуби, то есть, ни один эксперт нашего синтетического экспертного сообщества не сочтёт отклонение аномальным, если оно меньше 2σ . Единичная доля ястребов означает, что среди наших экспертов нет ни одного голубя, и, значит, среди них могут оказаться такие, кто признает отклонение, меньшее чем 2σ , аномально высоким. Заметим, что роли голубей и ястребов оказываются несимметричными в силу несимметричности кванторов общности и существования: квантор общности диктует нам фразу «ни один эксперт не сочтёт», а квантор существования — фразу «могут найтись такие, кто признает». Этот факт несимметричности ещё проявит себя в ниже, когда мы будем проводить сравнение экспертной оценки с алгоритмической детекцией аномальных состояний (см. раздел 3.5).

³⁷ Голубь толерантен. Он снисходительно относится к высоким отклонениям и не считает их поводом для объявления аномалии.

Итак, для начала мы зададимся начальными параметрами. Мы синтезируем экспертное сообщество, в котором будет 1000 экспертов, с равными долями ястребов и голубей, и начальные параметры процедуры синтеза — доля ястребов `hawk_rait` и число экспертов `n_experts` — таковы:

```
hawk_rait = 0.5  
n_experts = 1000
```

Нам понадобится список, в который мы будем заносить мнения экспертов в формате 1 или 0 (1, если эксперт признал состояние процесса аномальным, 0 — если посчитал его нормой) Мы заводим список `expert_list`, объявляя его пустым списком:

```
expert_list = []
```

Отметим, что в конечном счёте это будет список списков, то есть, списком, каждый элемент которого снова является списком, содержащим бинарное мнение очередного эксперта. Сначала будем учитывать мнение ястребов. Для этого мы в цикле по числу ястребов в нашем сообществе случайным образом выбираем критический порог отклонения `delta_cr` из интервала ястребов (то есть, от σ до 2σ) и формируем список экспертных оценок `expert`, занося в него 1 и 0 в зависимости от превышения (или его отсутствия) критического порога:

```
for i in range(round(hawk_rait*n_experts)):  
    delta_cr = np.random.uniform(sigma, 2*sigma)  
    expert = []  
    for d in delta_list:  
        if abs(d) > delta_cr:  
            expert.append(1)  
        else:  
            expert.append(0)  
    expert_list.append(expert)
```

На выходе из этого цикла получается список, в котором 500 элементов

(именно столько в нашем экспертном сообществе ястребов), каждый из которых, в свою очередь является списком из 225 чередующихся каким-то образом 0 и 1 (именно столько у нас было наблюдений с учётом того, что наблюдения проводились в течение 25 моментов времени из периода обучения алгоритма в 9 локациях, и каждое из таких наблюдений является либо нормой, либо аномалией с точки зрения очередного эксперта). Тем самым, мы учли мнение всех ястребов.

Для учёта мнения голубей поступаем точно так же. Доля голубей — это дополнение доли ястребов до единицы `1-hawk_rait`, поэтому новый параметр не требуется, а промежуток, из которого мы случайным образом выбираем значение критического порога отклонения — это промежуток от 2σ до 3σ . В остальном изменений никаких нет:

```
for i in range(round((1-hawk_rait)*n_experts)):
    delta_cr = np.random.uniform(2*sigma, 3*sigma)
    expert = []
    for d in delta_list:
        if abs(d) > delta_cr:
            expert.append(1)
        else:
            expert.append(0)
    expert_list.append(expert)
```

Заметим, что после исполнения цикла по ястребам список `expert_list` уже не был пустым, в нем уже было 500 записей. На выходе из цикла по голубям имеем уже полностью заполненный список экспертных мнений: в нем 1000 записей, каждая из которых является списком длиной 225, содержащим 0 или 1. Первый 500 записей описывают мнение ястребов, последние 500 записей — мнение голубей.

Исходя из соображений наглядности, мы реализуем мнения синтетических экспертов в виде таблицы (см. таблицу 3. 8), в которой 1000 строк (каждая строка — это эксперт) и 225 столбцов (каждый столбец — это пара момент-локация на этапе обучения алгоритма). Первые 500 строк описывают мнение

ястребов, и поэтому в них больше единиц, последние 500 строк описывают мнение голубей, и поэтому в них меньше единиц. В этот момент мы синтезировали экспертное сообщество, состоящее из 1000 экспертов, каждый из которых «дал свою оценку» всем состояниям нашего опять-таки синтетического процесса, причём, наше синтетическое экспертное сообщество таково, что в этом сообществе в равных долях представлены ястребы и голуби.

Таблица 3.8 — Синтетическое экспертное сообщество

	Наблюдение 1	Наблюдение 2	...	Наблюдение 225
Эксперт 1	1	1	...	1
Эксперт 2	1	0	...	1
Эксперт 3	1	1	...	1
...	...	...	...	...
Эксперт 998	0	0	...	0
Эксперт 999	0	0	...	0
Эксперт 1000	0	0	...	0

Изменяя долю ястребов, мы получим другое экспертное сообщество, например, если доля ястребов равна 0.1, то в таблице 3.8 первые 100 строк будут содержать больше 1, а последние 900 — меньше 1, а если доля ястребов равна 0.9, то наоборот: первые 900 строк будут содержать больше 1, а последние 1000 — меньше 1. И так далее, для любого промежуточного значения изменяемого параметра. Порядок следования (сначала ястребы, а потом голуби, или наоборот, или в порядке случайного чередования) не имеет никакого значения, потому что мнение каждого отдельного эксперта нас, вообще говоря, не интересует. Нас интересует мнение всего сообщества.

3.4.3. Мнение экспертного сообщества

Мнение экспертного сообщества получается прямым голосованием: если простое большинство экспертов считает наблюдение аномальным, значит, это аномалия, если нет — то это норма. А поскольку мы реализовали мнение экспертного сообщества в виде датафрейма, организовать такое голосование совсем не сложно.

Каждый столбец датафрейма с экспертными оценками — это наблюдение в какой-то момент в какой-то локации, который оценили 1000 раз. Если среди оценок больше половины составляют единицы, то экспертное сообщество сошлось во мнении, что это наблюдение аномально, если единиц меньше половины, совокупное мнение сообщества — наблюдается норма. Сумма по столбцу — это число экспертов, проголосовавших за аномалию. Если она превышает половину численности экспертов, то признаем аномалию, в противном случае объявляем норму. Таким образом, каждое наблюдение (а наблюдение — это, напомним, пара момент-локация из периода обучения алгоритма) получает свою итоговую оценку: это 0 или 1 в зависимости от мнения большинства экспертов (см. таблицу 3.9).

Таблица 3.9 — Совокупное мнение о наличии или отсутствии аномалий синтетического экспертного сообщества, ориентированного на прогноз по времени

	Локация 1	Локация 2	Локация 3	...	Локация 9
Момент 51	0	0	0	...	0
Момент 52	0	0	0	...	0
Момент 53	0	0	0	...	0
Момент 54	0	1	0	...	0
...	...	...	...	...	...
Момент 74	0	0	0	...	1
Момент 75	0	0	0	...	0

Читателя не должно смущать, что в таблицу 3.9 он видит практически одни нули. Так и должно быть. Единицы там есть (в других столбцах и строках, не вошедших в макет страницы), но их очень мало. Наоборот, было бы противоестественно, если бы их было много, ведь речь идёт о детекции аномалий, а аномалии — это какие-то редкие, или даже исключительные состояния, и было бы странно, если бы было наоборот.

3.5. Бинаризация отклонений по времени

Итак, у нас есть мнение экспертного сообщества, реализованное в виде таб-

лицы нулей и единиц³⁸. Но мы не можем сравнить его с мнением нашего детектора аномалий, потому что алгоритма нет своего мнения: он может вычислить отклонения реальных наблюдений от своих прогнозов, но он не знает, какой порог толерантности ему использовать, чтобы признать какое-то наблюдение аномальным. У него и не может быть своего мнения, потому что он алгоритм. Мнение есть у каждого отдельного эксперта, потому что он человек, и мнение есть у экспертного сообщества, потому что оно состоит из людей и учитывает мнение отдельных экспертов, а у алгоритма нет своего мнения.

Теперь перед нами стоят две важные задачи: 1) научить наш алгоритм подстраиваться под мнение нашего экспертного сообщества, и 2) выяснить, насколько удачно алгоритм способен подстроиться под разные мнения (мнение сообщества с преобладанием ястребов, мнение сообщества с преобладанием голубей и т. д. с учётом всех промежуточных вариантов).

Для решения первой задачи мы будем бинаризовать отклонения, прогоняя порог бинаризации по сетке в определённых нами выше разумных пределах от σ до 3σ . Каждый раз мы будем получать бинарный вектор: на низких значениях критического порога в нем будет больше единиц, на высоких — меньше, и каждый такой вектор мы будем сравнивать с истинным бинарным вектором, которым мы располагаем после того, как получили мнение экспертного сообщества (см. таблицу 3.9). Фактически при этом речь идёт об оценке бинарного классификатора (это одна из тем, которую мы затрагиваем в приложении А, посвящённом теоретическим основам стохастического прогнозирования, более подробно см. раздел А7), которая может производиться при помощи целого ряда метрик, начиная с незамысловатой метрики ассурсу и вплоть до сложных метрик, учитывающих одновременно как точность, так и полноту

³⁸ Причём, поскольку сообщество являемся синтетическим, мы можем влиять на его мнение, изменяя долю ястребов: чем больше будет ястребов в сообществе, тем чаще будут встречаться единицы в бинарном векторе совокупных оценок. И наоборот.

классификатора, и это прежде всего — метрики F1 и ROC AUC. Тот критический порог, при котором будут наблюдаться наилучшие значения метрик классификатора, мы передадим нашему алгоритму в качестве рабочего порога, после чего он будет знать, превышение какого именно порога следует считать аномалией, и именно этот порог он будет использовать на этапе пилотной эксплуатации.

3.5.1. Применение функции `calibration`

Функция `calibration` требует на вход три датафрейма (см. приложение С, в котором мы приводим листинг этой функции):

`df_expert_true` — датафрейм в формате ML, который индексирует наличие или отсутствие аномалии для каждой пары момент-локация из периода обучения алгоритма с точки зрения экспертного сообщества (по тексту выше это таблица 3.9);

`df_challenge` — датафрейм в формате ML с истинными значениями наблюдаемой характеристики процесса из периода обучения (это таблица 3.5);

`df_challenge_pred` — датафрейм в формате ML с прогнозируемыми значениями наблюдаемой характеристики процесса из периода обучения (это таблица 3.6).

В теле функции выполняется перебор с достаточно мелким шагом по сетке от σ до 3σ всевозможных значений критического порога отклонений реальности от прогноза и подбирается такая величина `delta_cr`, которая максимизирует метрики бинарной классификации, а выражаясь образно — бинаризация по этому порогу переводит датафрейм абсолютных отклонений `abs(delta)` (см. таблицу 3.7) в бинарный датафрейм, максимально похожий на датафрейм с мнениями экспертного сообщества `df_expert_true` (см. таблицу 3.9). Тем самым, применяя функцию `calibration`, мы решаем первую из заявленных

нами выше задач: мы показываем алгоритму тот критический порог, который наилучшим образом подстраивает его мнение под мнение экспертного сообщества.

3.5.2. Вариация экспертного мнения

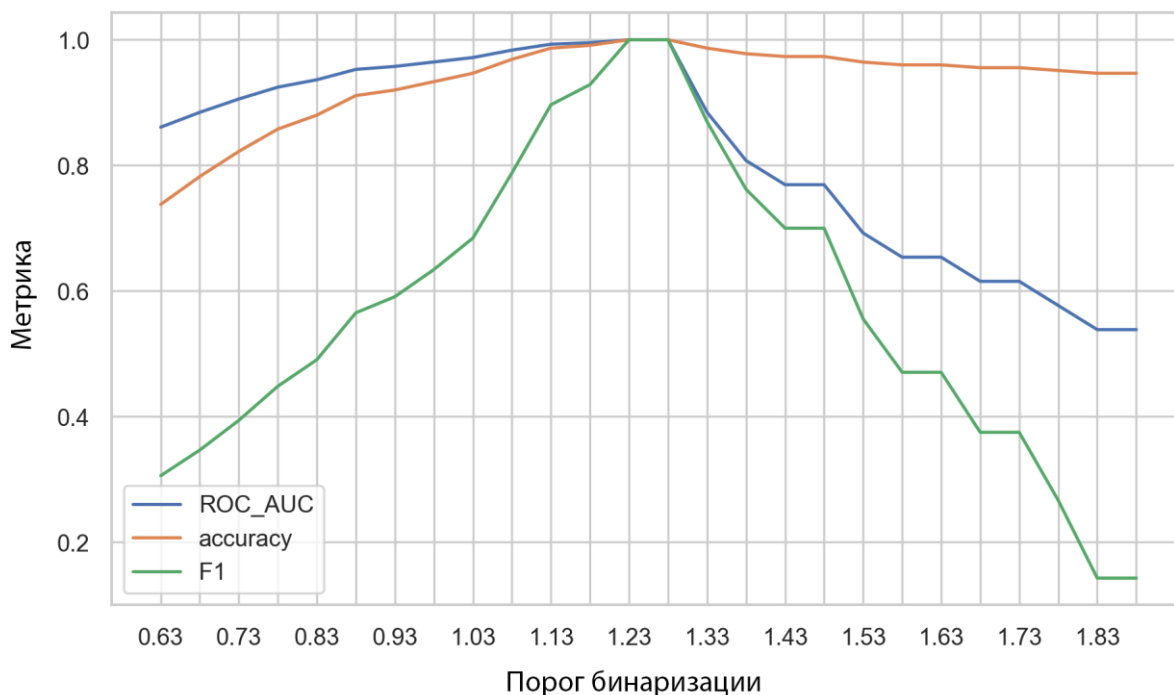


Рисунок 3.12 — Калибровка отклонений в формате ML по отношению к сбалансированному экспертному сообществу

Однако не будем забывать, что в этой главе мы живём в синтетическом мире, в котором можно менять все условия (в следующих главах мы будем работать с реальными данными, и у нас уже не будет такой возможности), в частности — мы можем варьировать мнение экспертного сообщества.

Синтезируя сбалансированное экспертное сообщество, в котором ястребы и голуби представлены в равных пропорциях, мы получаем его мнение в виде бинарной таблицы, (по тексту выше это таблица 3.9). Мы используем её и ещё два датафрейма, полученные нами выше (см. таблицу 3.5 и 3.6), в качестве аргументов функции `calibration`, и в цикле по порогу бинаризации получаем визуализацию (см. рисунок 3.12), которая демонстрирует степень совпадения мнения нашего алгоритма с мнением экспертного сообщества. Здесь

по горизонтали отложены значения величины критического порога отклонения реальности от прогноза, а по вертикали — соответствующие значения метрик бинарного классификатора. При этом функция `calibration` возвращает наилучшие значения метрик и критический порог отклонения, на котором они возникают (см. рисунок 3.12):

```
best_ROC_AUC = 0.995439
best_accuracy = 0.999312
best_F1 = 1.0
best_delta_cr = 1.228955
```

а это значит, что при указанном выше значении критического порога `best_delta_cr` мнение нашего алгоритма практически полностью совпадает с мнением экспертного сообщества (метрики ROC AUC, ассигасу и F1 близки к своему теоретическому максимуму, то есть — к единице).

Изменим мнение экспертного сообщества в пользу ястребов. Для этого зададим в процедуре синтеза экспертов долю ястребов, равную 0.9 (см. рисунок 3.13).

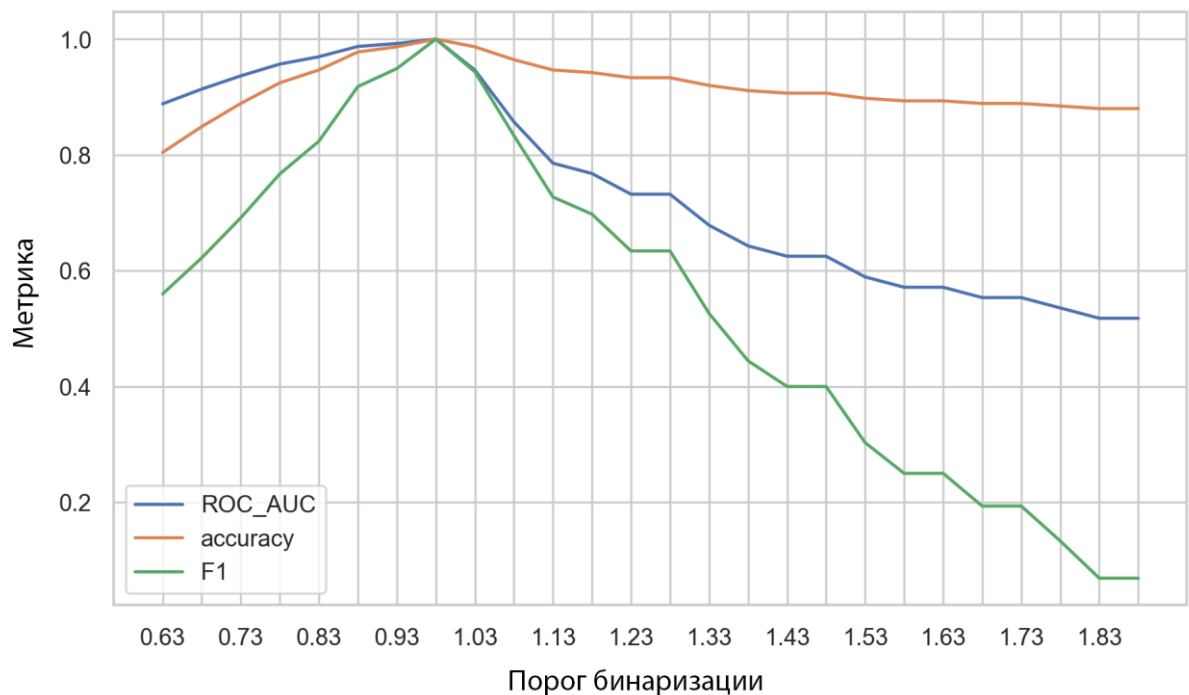


Рисунок 3.13 — Калибровка отклонений в формате ML по отношению к экспертному сообществу с преобладанием ястребов

Мы видим, что алгоритм отреагировал на изменившееся мнение: его критический порог сместился налево и теперь он будет пометать как аномальные наблюдения с меньшими отклонениями от прогноза (см. рисунок 3.13). При этом функция `calibration` возвращает:

```
best_ROC_AUC = 1.0
best_accuracy = 1.0
best_F1 = 0.997541
best_delta_cr = 0.978955
```

Совпадение с мнением экспертного сообщества снова оказывается почти полным, а критический порог становится гораздо ниже, то есть наш алгоритм занимает теперь гораздо более ястребиную позицию.

И, наконец, изменив соотношение в экспертном сообществе в пользу голубей (при доле ястребов, равной 0.1), увидим уже вполне ожидаемую картину (см. рисунок 3.14): оптимальный порог бинаризации, который указывает на критическое значение отклонений, сместился направо.

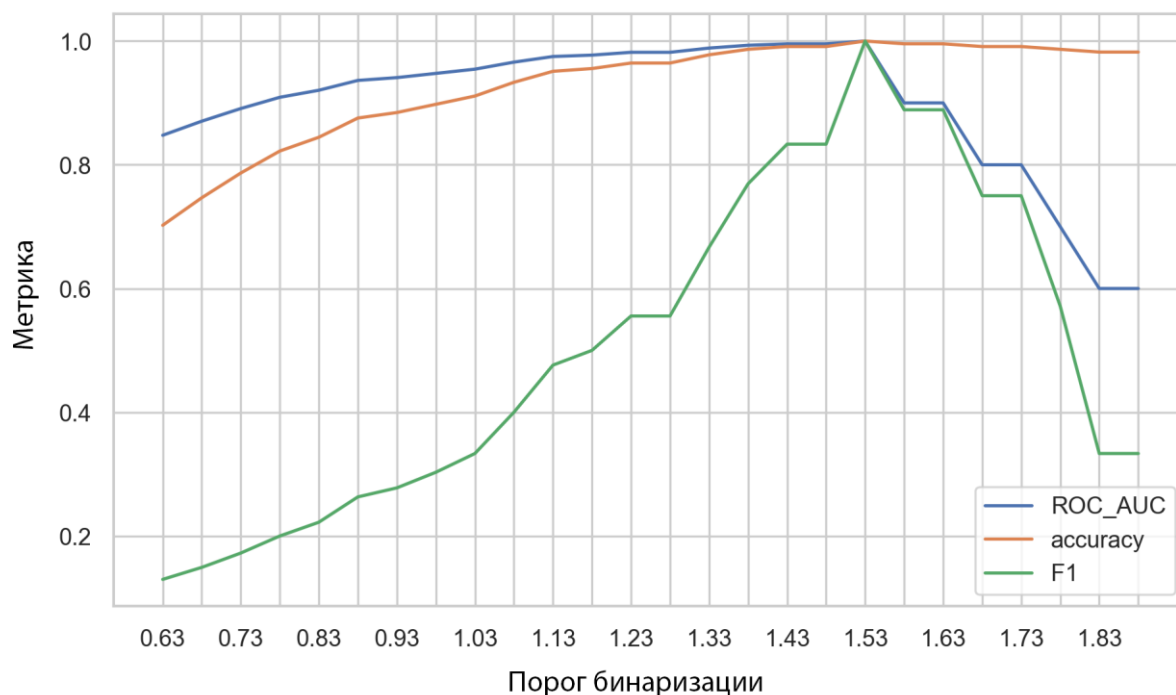


Рисунок 3.14 — Калибровка отклонений в формате ML по отношению к экспертному сообществу с преобладанием голубей

Если при преобладании ястребов максимальные значения метрик эффективности бинарной классификации смещались налево, то теперь ситуация обратная. Алгоритм стал более толерантен, его критический порог сместился направо и теперь он будет признавать аномальными только те состояния процесса, которые не признать таковыми уже просто нельзя, слишком уж сильно они отклоняются от прогнозируемых значений. Формальный вывод функции `calibration` при этом таков:

```
best_ROC_AUC = 0.995283
best_accuracy = 0.995555
best_F1 = 1.0
best_delta_cr = 1.528955
```

Мы снова видим практически полное совпадение мнений алгоритма и экспертного сообщества, а это значит, что мы научили наш алгоритм подстраиваться под экспертов и выполнили, тем самым, вторую задачу из двух, заявленных нами в начале этого раздела.

Теперь напомним одно замечание, которое мы седлали чуть выше, когда говорили о разделении экспертов на ястребов и голубей: мы сказали тогда, что кванторы общности и существования играют несимметричные роли, и что у нас ещё будет возможность это увидеть. Три иллюстрации, представленные выше демонстрируют наш тезис со всей очевидностью, так как графики всех трёх метрик оказываются существенно несимметричны относительно точек своих максимальных значений, и это происходит именно потому, что роли голубей и ястребов здесь несимметричны: у голубей гораздо меньше возможностей детектировать аномалию, и они делают это значительно реже.

3.5.3. Сопоставление с экспертной оценкой

Одно из значений английского слова `challenge` — бросать вызов. Теперь, после того, как мы научили наш алгоритм вычислять критический порог отклонений, у него появилось собственное мнение относительно того, что следует считать нормой, а что следует признать аномалией по отношению ко времени, поэтому

теперь он, в некотором смысле, не отличается от экспертов, каждого со своим мнением. Признавая в качестве неоспоримой истины мнение всего экспертного сообщества в целом (а оно, напомним, получается прямым голосованием), мы можем противопоставить мнение нашего алгоритма мнению каждого из экспертов (тем самым, наш алгоритм бросает вызов экспертному корпусу) и выяснить, чьё мнение ближе к истине. У нас есть тысяча экспертов и один алгоритм. Мнение алгоритма совпадает с истинным мнением на уровне $ROC\_AUC > 0.99528$ (см. выше, раздел 3.5.2), и мы можем узнать, мнение скольких экспертов окажется более истинным по отношению к этой метрике³⁹. Проводя цикл по имеющимся у нас экспертам (которых, напомним, у нас 1000 штук), получаем счетчик числа экспертов, оценка которых превышает оценку нашего алгоритма. Счётчик показывает:

```
count = 57
```

Значит, мнение нашего алгоритма как минимум не уступает мнению 943 экспертов из тысячи — этап сопоставления успешно пройден. Более того, оказалось, что у нашего алгоритма огромный запас прочности: он мог бы проиграть ещё 543 раза и все равно выиграл бы, так как все равно детектировал бы аномальные состояния процесса лучше, чем большинство экспертов.

Теперь мы окончательно подготовили алгоритм детекции аномалий по времени к выводу в эксплуатационный режим (см. ниже, раздел 3.7), но остаётся ещё одна незавершённая часть нашего детектора: алгоритм детекции аномалий по локациям.

3.6. Бинаризация отклонений по локациям

Процедура бинаризации отклонений по локациям в точности воспроизводят

³⁹ Мы используем метрику ROC AUC, а не другие метрики, потому что это обычная практика в оценке классификаторов. С тем же успехом можно было бы применить метрику F1. Но метрике ассигасу нельзя доверять, если речь идёт о несбалансированных классах [3, 7, 51] (а классы нормы и аномалии сильно не сбалансированы в пользу нормы).

рассмотренную выше процедуру бинаризации отклонений по времени, поэтому мы не станем их подробно разбирать, отметим лишь несколько ключевых моментов.

Функция `calibration` принимает на вход данные в формате LM, причём, сведения об истинных значениях наблюдаемой характеристики получаются из соответствующих сведений в формате ML простым транспонированием (см. таблицу 3.10, которая является транспонированной таблицы 3.5).

Таблица 3.10 — Истинные значения на этапе обучения в формате LM

	Момент 51	Момент 52	Момент 53	...	Момент 75
Локация 1	2.314840	3.758483	3.213714	...	2.151387
Локация 2	4.327486	4.354091	3.960747	...	4.483955
Локация 3	3.919289	4.391781	2.468318	...	3.093085
Локация 4	5.075944	3.785200	3.083876	...	0.856539
...	...	...	...	...	...
Локация 8	3.960097	2.900776	2.692202	...	3.359666
Локация 9	0.415913	−0.081912	0.592205	...	−1.150825

Однако для таблицы предсказанных значений, которую мы передаём в датафрейм это не так (см. таблицу 3.11, которая не совпадает с таблицей 3.6 с точностью до транспонирования).

Таблица 3.11 — Предсказанные значения на этапе обучения в формате LM

	Момент 51	Момент 52	Момент 53	...	Момент 75
Локация 1	3.334817	2.286247	2.506115	...	2.658549
Локация 2	4.126926	4.304209	3.398638	...	4.344793
Локация 3	4.551446	4.015344	4.146537	...	3.132890
Локация 4	5.215135	4.821593	4.239362	...	0.956412
...	...	...	...	...	...
Локация 8	3.231870	3.495957	2.842418	...	2.722934
Локация 9	0.738757	0.508342	0.293830	...	0.206733

Прогнозирование по времени и по локациям осуществляется разными моделями (напомним, что в моделировании по времени мы применяли линейную модель `LinearRegression` с глубиной авторегрессии, равной четырём

наблюдениям, а в моделировании по локациям — нейросетевую модель `MLPRegressor` с глубиной истории наблюдений, равной трём). Это две совершенно самостоятельные таблицы, связанные друг с другом лишь семантикой хранящихся в них сведений, но не методами, которыми они были получены.

Как следствие, таблица отклонений для прогнозов по локациям тоже не совпадает с таблицей отклонений для прогнозов по времени (см. таблицу 3.12, которая не получается транспонированием из таблицы 3.7).

Таблица 3.12 — Отклонения предсказанных значений от истинных в формате LM

	Момент 51	Момент 52	Момент 53	...	Момент 75
Локация 1	−1.019977	1.472237	0.707600	...	−0.507161
Локация 2	0.200560	0.049882	0.562108	...	0.139162
Локация 3	−0.632156	0.376437	−1.678220	...	−0.039805
Локация 4	−0.139191	−1.036393	−1.155485	...	−0.099874
...	...	...	...	...	...
Локация 8	0.728228	−0.595181	−0.150217	...	0.636732
Локация 9	−0.322844	−0.590254	0.298375	...	−1.357558

Сигма отклонений реальности от прогнозов также другая, хотя и не сильно отличающаяся от соответствующего параметра прогнозирования по времени. Если тогда у нас получалось `sigma = 0.628955`, то теперь:

`sigma = 0.660583`

Дальнейшие действия ничем не отличаются от уже описанных нами выше в предыдущем разделе: мы точно так же синтезируем экспертное сообщество, варьируем его мнение и анализируем реакцию алгоритма на мнение экспертов.

На сбалансированном экспертном сообществе с равными долями голубей и ястребов — критический порог отклонений, любое превышение которого означает аномалию составляет `best_delta_cr = 1.310583` (см. рисунок 3.15).

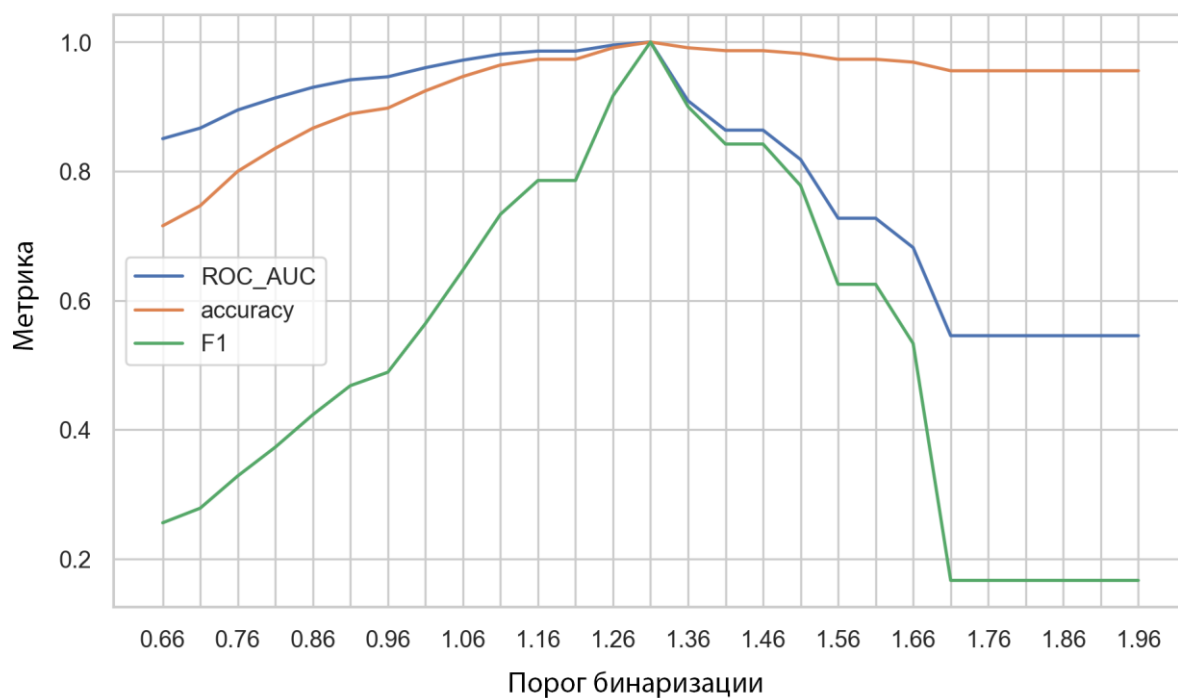


Рисунок 3.15 — Калибровка отклонений в формате LM
по отношению к сбалансированному экспертному сообществу

Для преобладания ястребов $\text{best\_delta\_cr} = 1.010583$ (см. рисунок 3.16).

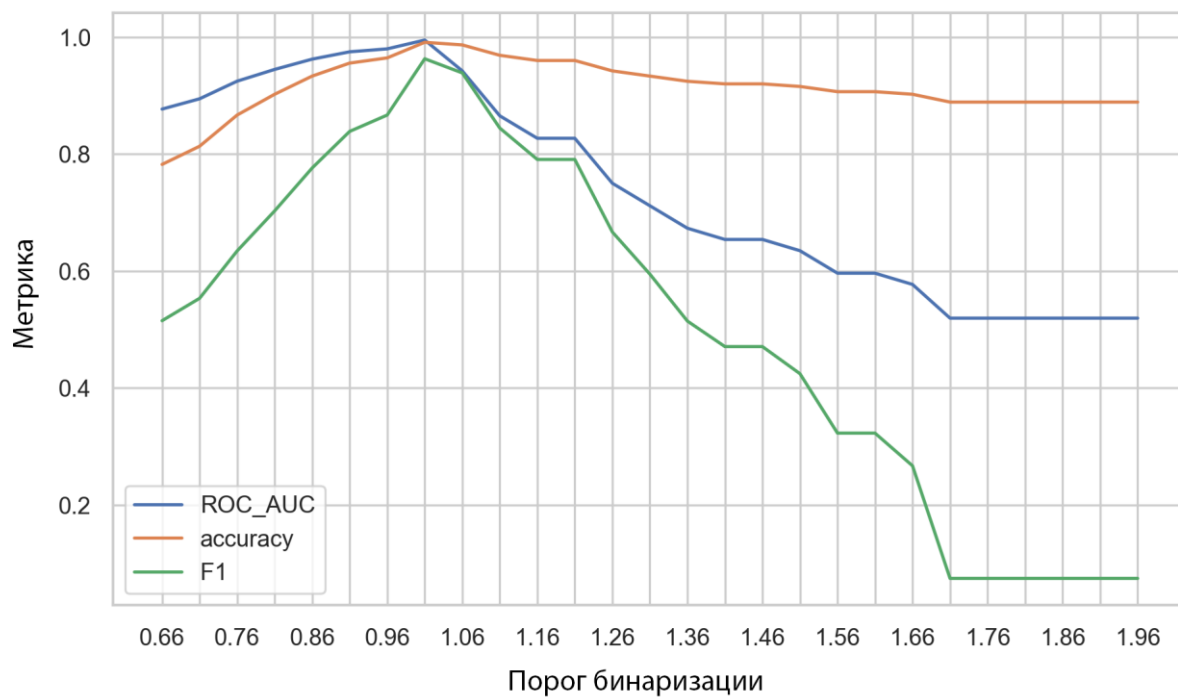


Рисунок 3.16 — Калибровка отклонений в формате LM
по отношению к экспертному сообществу с преобладанием ястребов

Если преобладают голуби, то $\text{best\_delta\_cr} = 1.560583$ (см. рисунок 3.17).

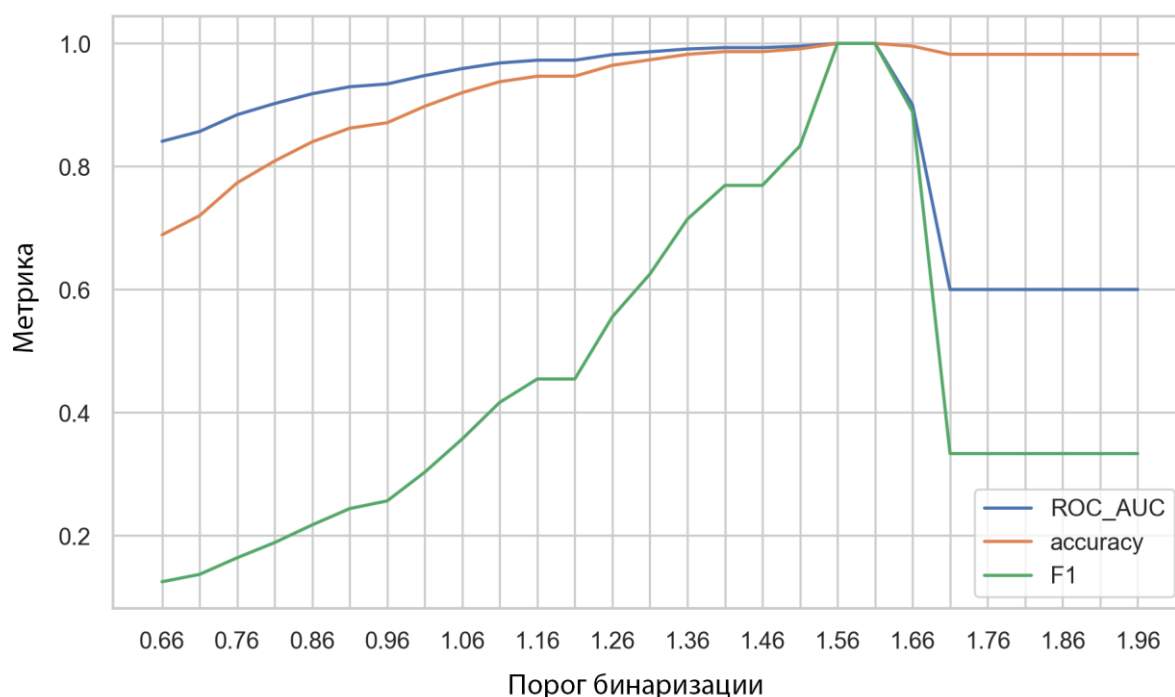


Рисунок 3.17 — Калибровка отклонений в формате LM по отношению к экспертному сообществу с преобладанием голубей

Мы снова наблюдаем ожидаемые смещения критических величин отклонений: для сообщества с преобладанием ястребов смещение происходит налево, а для сообщества с преобладанием голубей — направо.

Кроме того, мы, так же, как и выше, проводим сопоставление с экспертным мнением и убеждаемся, что алгоритм успешно его преодолевает с большим запасом прочности, уступая лишь 62 экспертам из тысячи экспертов синтетического экспертного корпуса.

3.7. Пилотная эксплуатация детектора аномалий

Далее в этом разделе мы будем считать, что наше экспертное сообщество является сбалансированным⁴⁰, голубей и ястребов в нем примерно поровну. Тогда по итогам этапа сопоставления алгоритмической детекции с экспертным

⁴⁰ Хотя, мы могли бы ориентироваться на преобладание голубей или ястребов. Это не имеет никакого значения. Мы зафиксировали именно сбалансированность только лишь для определённости, тем более, что в реальном (а не синтетическом) мире такая определённость присутствует всегда: экспертное сообщество именно таково, каково оно есть на самом деле,

мнением мы располагаем следующими сведениями о том, в каком режиме нам нужно эксплуатировать наш детектор:

1. При прогнозировании состояний нашего процесса по времени мы должны использовать модель `LinerRegression` с глубиной авторегрессии 4, критический уровень допустимых отклонений прогноза от реальности составляет `best_delta_cr = 1.228955`.
2. При прогнозировании состояний процесса по локациям мы должны использовать модель `MLPRegressor` с глубиной истории наблюдений 3, критический уровень допустимых отклонений прогноза от реальности составляет `best_delta_cr = 1.310583`.

Именно эти параметры алгоритма мы будем использовать на этапе пилотной эксплуатации алгоритмического детектора аномальных состояний.

3.7.1. Детекция аномалий по времени

Функция детекции аномалий по времени (см. приложение С, раздел С.4, где приведён листинг её кода и комментарии к нему) принимает на вход историю наблюдений и критический порог отклонений реальности от прогноза, а возвращает пары момент-локация для всех зафиксированных аномалий.

Список таких пар образует список аномалий, который мы предъявляем в виде таблицы (см. таблицу 3.13):

Таблица 3.13 — Список аномалий по времени на синтетических данных на этапе пилотной эксплуатации

Момент	Локация
<code>mom_77</code>	Аномалия в локациях: <code>loc_4</code>
<code>mom_80</code>	Аномалия в локациях: <code>loc_4</code> , <code>loc_7</code>
<code>mom_81</code>	Аномалия в локациях: <code>loc_7</code>
<code>mom_84</code>	Аномалия в локациях: <code>loc_8</code>
<code>mom_86</code>	Аномалия в локациях: <code>loc_9</code>

и другого у нас не будет.

Момент	Локация
mom_89	Аномалия в локациях: loc_4, loc_7
mom_91	Аномалия в локациях: loc_4, loc_5, loc_6, loc_7, loc_8, loc_9
mom_92	Аномалия в локациях: loc_8
mom_93	Аномалия в локациях: loc_1
mom_94	Аномалия в локациях: loc_9
mom_95	Аномалия в локациях: loc_8
mom_97	Аномалия в локациях: loc_1, loc_2
mom_98	Аномалия в локациях: loc_3, loc_8
mom_99	Аномалия в локациях: loc_8

Отметим, что на этапе эксплуатации мы анализировали состояние процесса в 9 локациях на протяжении 25 моментов времени, другими словами, у нас было 225 наблюдений. Из них, если учесть результаты, приведённые выше в календаре аномалий, 23 наблюдения признаны аномальными, что составляет более 10% от общего числа наблюдений. Это, разумеется, огромный процент. Но давайте вспомним, какое экспертное сообщество мы синтезировали — в нем даже голуби были готовы признать аномальным отклонение величиной всего 2σ (а это очень небольшое отклонение для нормально распределённых ошибок, недаром же по классике аномальными признаются отклонения, превышающие 3σ), а ястребы вообще подозревали аномалию, начиная с σ . Наши синтетические эксперты предъявляли к наблюдениям слишком жёсткие требования, поэтому совершенно неудивительно, что детектор, будучи подогнан под мнение экспертного сообщества, возвратил такой высокий процент аномальных наблюдений. Он просто воспроизвёл мнение экспертного корпуса.

3.7.2. Детекция аномалий по локациям

Функция детекции аномалий по локациям (см. приложение С, раздел С.4, где приведён листинг её кода и комментарии к нему) так же, как и функция детекции аномалий по времени, принимает на вход историю наблюдений (ту же самую) и критический порог отклонений реальности от прогноза (другой), а возвращает пары момент-локация, в которых была зафиксирована аномалия. Такие пары образуют список аномалий по локациям (см. таблицу 3.14):

Таблица 3.14 — Список аномалий по локациям на синтетических данных на этапе пилотной эксплуатации

Момент	Локация
mom_75	Аномалия в локациях: loc_5
mom_83	Аномалия в локациях: loc_6
mom_84	Аномалия в локациях: loc_1
mom_87	Аномалия в локациях: loc_7
mom_89	Аномалия в локациях: loc_1
mom_91	Аномалия в локациях: loc_0, loc_2
mom_92	Аномалия в локациях: loc_4
mom_94	Аномалия в локациях: loc_7
mom_97	Аномалия в локациях: loc_1, loc_7
mom_99	Аномалия в локациях: loc_1, loc_1, loc_2

На этот раз из 225 наблюдений на этапе эксплуатации было зафиксировано 13 аномалий, что составляет примерно 6% от общего числа наблюдений. Это тоже много, хотя и меньше, чем в случае с аномалиями по времени.

3.7.3. Календарь аномалий

Имея две различные точки зрения на аномальность того или иного состояния (с одной стороны, возможна аномалия по времени, а с другой стороны — по локациям), мы формируем окончательный вывод при помощи логического оператора OR (см. выше, главу 2, раздел 2.5), то есть, признаем состояние процесса аномальным, если хотя бы один из наших подходов к детекции аномальных состояний сигнализирует об его аномальности.

Таблица 3.15 — Общий список аномалий на синтетических данных на этапе пилотной эксплуатации

Момент	Тип аномалии	Локация
mom_75	Аномалия по локациям:	loc_5
mom_77	Аномалия по времени:	loc_4
mom_80	Аномалия по времени:	loc_4, loc_7
mom_81	Аномалия по времени:	loc_7
mom_83	Аномалия по локациям:	loc_6
	Аномалия по времени:	loc_8

Момент	Тип аномалии	Локация
mom_84	Аномалия по локациям:	loc_1
mom_86	Аномалия по времени:	loc_9
mom_87	Аномалия по локациям:	loc_7
mom_89	Аномалия по времени:	loc_4, loc_7
	Аномалия по локациям:	loc_1
mom_91	Аномалия по времени:	loc_4, loc_5, loc_6, loc_7, loc_8, loc_9
	Аномалия по локациям:	loc_0, loc_2
mom_92	Аномалия по времени:	loc_8
	Аномалия по локациям:	loc_4
mom_93	Аномалия по времени:	loc_1
mom_94	Аномалия по времени:	loc_9
	Аномалия по локациям:	loc_7
mom_95	Аномалия по времени:	loc_8
mom_97	Аномалия по времени:	loc_1, loc_2
	Аномалия по локациям:	loc_1, loc_7
mom_98	Аномалия по времени:	loc_3, loc_8
mom_99	Аномалия по времени:	loc_8
	Аномалия по локациям:	loc_1, loc_2

По совокупности обоих подходов зафиксировано 36 аномальных состояний среди 225 наблюдений (хотя, строго говоря, их всё-таки 35, потому что в 97 момент времени локация loc_1 была признана аномальной сразу с обеих точек зрения), что составляет около 15% от общего числа наблюдений.

3.8. Выводы по главе 3

В заключение этой главы кратко сформулируем её основные результаты и отметим некоторые важные, с нашей точки зрения, моменты.

Прежде всего, нужно понимать, что наличие развёрнутого экспертного мнения относительно какого-либо реального процесса — это, как правило, роскошь, недоступная аналитику данных. Квалифицированный эксперт в рамках реального процесса (например, производственного, но то же самое можно сказать и о сервисном или маркетинговом процессе) занимает в этом процессе

ключевую позицию: это главный инженер, ведущий инженер-конструктор, директор предприятия или подразделения и т. д. Таких специалистов всегда немного, и все они настолько востребованы в рамках самого процесса, что использовать их для первичной разметки данных по принципу нормы или аномалии просто невозможно, а заменить их на менее квалифицированных и, следовательно, более доступных сотрудников того же процесса нельзя, так как их экспертиза окажется заведомо некачественной, а, значит — вообще бесполезной (или даже вредной). В этом состоит одна из причин, по которой мы включили в настоящую диссертационную работу главу с синтетическими данными: нам были нужны эксперты, пусть даже и синтетические, чтобы продемонстрировать работоспособность алгоритма на всех этапах его работы, включая сравнение алгоритмической детекции аномалий с экспертной оценкой.

По итогам исследования, проведённого в этой главе, отметим:

1. Алгоритм детекции аномалий в нестабильных процессах, основанный на изложенной выше (см. выше главу 2) методике МССР (Modeling, Challenge, Calibration, Production) корректно отработал на всех этапах, включая этап сопоставления с экспертным мнением.
2. При работе с реальными (а не с синтетическими) данными мы будем использовать другой принцип поиска критических величин отклонений Δ_{cr} и Θ_{cr} , основанный не на экспертном мнении, которое, как правило, недоступно, а на анализе калибровочной кривой (см. далее, главы 4–6).

Таким образом, изложенная в этой главе (и реализованная в виде кодов в приложении С) алгоритмика показала свою работоспособность на всех этапах методики МССР, что говорит о её применимости к реальным данным о производственных, а также сервисных и маркетинговых процессах одного из ведущих отечественных автопроизводителей, к чему мы и переходим.

ГЛАВА 4. ДЕТЕКЦИЯ АНОМАЛИЙ В ПРОЦЕССЕ АВТОМОБИЛЬНОГО ПРОИЗВОДСТВА

Описанный в предыдущих главах настоящей диссертационной работы алгоритм детекции аномалий в нестабильных процессах, основанный на методике МССР (Modeling, Challenge, Calibration, Production), способен обрабатывать данные о процессах произвольной природы и диагностировать состояние процесса как аномальное вне зависимости от того, какой именно процесс подаётся ему на вход. Разумеется, он не является панацеей, и, разумеется, нельзя следовать ему слепо, игнорируя специфику процесса и его семантическое содержание. Так, например, если применить методику МССР (Modeling, Challenge, Calibration, Production) к технологическим процессам производства, в которых процесс, по определению, должен быть подвержен минимальной вариативности, а в идеале — вообще принимать исключительно константные значения, то мы получим лишь жалкое подобие хорошо известной [6, 9, 38, 56, 154, 225] и отлично себя зарекомендовавшей схемы Шухарта, с тою лишь разницей, что на внедрение и эксплуатацию новой методики потребуются время и немалые средства (а результат при этом вряд ли превзойдёт те, что получаются с применением стандартных статистических инструментов управления качеством).

Вместе с тем, современные представления о качестве и организации производства основываются на том, что на данном этапе развития предел технологического совершенства уже достигнут, и дальнейший рост качества возможен лишь тех в областях, которые сопутствуют технологиям: логистике, менеджменте, квалификации персонала, его мотивированности и т. д. Очевидно, что развитие систем управления качеством демонстрирует эволюцию парадигмы от преимущественно технократического подхода к социотехническому, при этом отчетливо прослеживаются этапы эволюции и смена парадигм [26,

58, 85].

Поэтому прежде, чем переходить к содержательной части этой главы, мы предварительно дадим беглый обзор этой эволюции.

Период технического детерминизма (до середины XX века). На начальных этапах эволюции системы управления качеством превалировала инженерно-ориентированная парадигма [26, 27, 58, 59]. Качество рассматривалось преимущественно как атрибут материального объекта, обеспечиваемый посредством стандартизации, метрологического контроля, статистического контроля процессов (Statistical Process Control, SPC) и управления дефектами. Человеческий ресурс трактовался как переменная, подверженная минимизации дисперсии посредством стандартизированных операционных процедур (Standard Operating Procedures, SOP). Ключевые показатели эффективности (Key Performance Indicators, KPI) фокусировались на показателях производительности, коэффициенте выхода годной продукции и стоимости несоответствий (Cost of Non-Conformance, CONC).

Интеграция концепции TQM (Total Quality Management) (1950-е – 1980-е годы). Возникновение TQM ознаменовало переход к социотехнической парадигме управления качеством [26, 77]. Теория всеобщего управления качеством признала персонал как ключевой ресурс, способствующий непрерывному улучшению (Kaizen) [58, 72, 85]. Внедрение инструментов вовлечения персонала, командной работы и обучения способствовало делегированию полномочий и расширению ответственности. Дополнительно к производственным KPI, стали использоваться метрики удовлетворённости потребителей.

Системы Lean Manufacturing и Six Sigma (1980-е годы – настоящее время). Данные концепции углубляют социотехнический подход. Lean Manufacturing ориентируется на устранение всех видов потерь, включая недоиспользование потенциала персонала [26, 58, 72, 73, 87]. Six Sigma, преследуя снижение вариативности процессов, акцентирует внимание на тщательный анализ причин

отклонений (Root Cause Analysis, RCA) и использование статистических методов для принятия управленческих решений [58, 63, 87]. Интеграция принципов бережливого производства и шести сигм привела к росту значимости обучения, multifunctionality сотрудников и более комплексному подходу к управлению процессами.

Современный этап (с начала XXI века). На текущем этапе развития всё большее значение придается организационной культуре качества, управлению знаниями (Knowledge Management), созданию самообучающейся организации (Learning Organization), и внедрению принципов клиентоориентированности (Customer Centricity). В дополнение к финансовым и производственным KPI, внедряются показатели вовлеченности персонала (Employee Engagement), уровня стресса, индекса лояльности сотрудников (Employee Net Promoter Score, eNPS), и текучести кадров. Наблюдается интеграция психологии труда, организационной психологии и социологии в практику управления качеством [26, 58, 88, 89, 152].

Признание значимости социальных факторов в управлении качеством является эволюционным шагом, позволяющим перейти от технократического подхода к социотехническому. Данный переход способствует повышению эффективности, улучшению качества, снижению издержек, стимулированию инноваций и обеспечению устойчивого развития организации. В этой главе мы рассмотрим одну из технологических операций, которая осуществляется на производственных предприятиях одного из ведущих отечественных автопроизводителей, не с точки зрения исполнения технологии, а как некий показатель эффективности того или иного работника, то есть — с социальной точки зрения (хотя, социальность, безусловно, понимается нами в исключительно узком, корпоративном смысле этого слова).

4.1. Производство как социальный процесс

Производство, в своей глубинной сущности, является сложным социальным

процессом, в котором человеческие взаимоотношения и организационные структуры значат гораздо больше, чем просто внедрение технологий:

1. Взаимодействие между работниками и менеджментом [27, 36, 58, 59, 63, 64] формирует фундамент успешного предприятия, определяя уровень мотивации, вовлеченности и, как следствие, производительности труда.
2. Организация труда, выстроенная с учётом человеческого фактора [36, 58, 72, 73, 78, 85], эргономики и психологического комфорта, способствует снижению стресса и повышению эффективности работы.
3. Обучение персонала [27, 63, 78], направленное не только на освоение технических навыков, но и на развитие коммуникативных компетенций и командного духа, является инвестицией в будущее компании.
4. Социальные факторы [36, 85, 86, 88], такие как культура организации, система мотивации, возможности для карьерного роста, оказывают непосредственное влияние на лояльность сотрудников и их стремление к самосовершенствованию.
5. Здоровая атмосфера в коллективе, основанная на взаимоуважении, поддержке и признании достижений, способствует развитию творческого потенциала и генерации инновационных идей [85, 86, 88].

Безусловно, новые технологии играют важную роль в повышении эффективности производства [27, 59, 63, 90], но именно социальные факторы определяют, насколько эффективно эти технологии будут использоваться [85, 88]. Пренебрежение социальными аспектами производства может привести к снижению производительности, увеличению текучести кадров, ухудшению качества продукции и, в конечном итоге, к финансовым потерям [152, 153]. Компании, осознающие важность социальных факторов и инвестирующие в развитие человеческого капитала, получают конкурентное преимущество, обеспечивающее устойчивый рост и процветание, так как в конечном итоге, именно люди, а не машины, создают ценность и двигают производство.

4.1.1. Роль организационных аспектов процесса производства

В современных теориях управления качеством организационные аспекты процесса производства играют ключевую роль, являясь фундаментом для достижения устойчивых результатов. Эффективное управление качеством сегодня не сводится к контролю выходной продукции, а подразумевает комплексный подход, интегрированный в каждый этап создания ценности. Особое внимание уделяется обучению персонала и организации рабочего места [36, 88, 90], поскольку именно эти факторы напрямую влияют на производительность, мотивацию и, как следствие, на качество выпускаемой продукции или предоставляемых услуг.

Обучение персонала в современных условиях рассматривается как стратегическая инвестиция в человеческий капитал [223, 233]. Непрерывное повышение квалификации, обучение вновь принятых исполнителей [239], освоение новых технологий и методов работы позволяют сотрудникам адаптироваться к быстро меняющимся требованиям рынка [58, 63, 78] и производственным задачам, а грамотно организованные программы обучения, учитывающие как специфику деятельности компании, так и индивидуальные потребности сотрудников, способствуют повышению их компетентности [88], снижению ошибок и улучшению качества работы.

Помимо этого, организация рабочего места, в свою очередь, оказывает непосредственное влияние на производительность труда [27, 36, 58,] и общее впечатление о компании [39, 42, 44, 106], позволяет создать эргономичное и безопасное рабочее пространство, минимизировать потери как времени, так и ресурсов, улучшить условия труда и повысить мотивацию сотрудников.

В совокупности, обучение персонала и оптимизация рабочего места представляют собой мощный инструмент повышения качества. Компании, уделяющие достаточно внимания этим организационным аспектам, получают конкурентное преимущество, обеспечивая высокое качество продукции или услуг, удовлетворённость клиентов и устойчивый рост.

4.1.2. Обучение персонала как один из факторов качества производства

Обучение персонала играет фундаментальную роль в обеспечении высокого качества производственного процесса, так как компетентность сотрудников напрямую влияет как на эффективность работы оборудования, так и на точность выполняемых операций и, в конечном итоге, на качество выпускаемой продукции. Программы обучения должны охватывать все уровни персонала, от рабочих на производственной линии до руководителей подразделений, отделов и целых производств, но в первую очередь это относится к вновь принятым исполнителям (см. ниже, раздел 4.2.1), причём, вне зависимости от должности исполнителя: это может быть как рабочий на сборочном конвейере, так и ответственное лицо на высокой позиции [27, 36, 58, 63]. Необходимо обучать сотрудников современным технологиям, новым методам работы, и не менее важно уделять внимание развитию навыков командной работы и коммуникации. Регулярное обучение помогает сотрудникам адаптироваться к изменениям в производственном процессе и осваивать новые технологии [88], а это, в свою очередь, позволяет избежать ошибок, снизить брак и повысить производительность, и, кроме того, обученные сотрудники более мотивированы и лояльны к компании.

Обучение персонала должно быть непрерывным процессом [86, 88], включающим в себя как теоретические занятия, так и практическое обучение на рабочем месте. Несомненно, эффективная система обучения персонала — это залог стабильного качества производственного процесса и успешного развития предприятия в целом, а отсутствие должного внимания к обучению неизбежно приведёт к снижению качества продукции и увеличению брака.

4.1.3. Роль организации рабочего места в управлении качеством

Не меньшую роль играет и эффективная организация рабочего места, которая является важным аспектом управления качеством производственного процесса [58, 59], так как грамотно спланированное пространство повышает

производительность, снижает риски ошибок и нарушений технологических требований, способствует созданию благоприятной рабочей атмосферы [88]. Очевидно (и, к сожалению, это очевидное утверждение требует постоянных усилий, направленных на его чёткое исполнение), что, в первую очередь, необходимо обеспечить доступность инструментов и материалов, необходимых для выполнения производственных задач, так как чёткая маркировка и систематизация оборудования позволяют сотрудникам быстро находить нужные инструменты, экономя как время, так и силы. Оптимальная эргономика рабочего места играет ключевую роль в поддержании высокой работоспособности персонала. Отметим также, что организация рабочего места — это не просто вопрос эстетики, а стратегически важный элемент, влияющий на качество продукции и производительность труда

4.1.4. Мотивация сотрудников в контексте управления качеством производственного процесса

Мотивация сотрудников является еще одним важным элементом [85, 86] в управлении качеством производственного процесса, оказывая прямое влияние на каждый аспект деятельности компании. Когда все сотрудники хорошо мотивированы, они демонстрируют повышенную вовлеченность и ответственность, что, в свою очередь, положительно сказывается на эффективности их работы.

Высокая мотивация побуждает сотрудников к более внимательному отношению к деталям, снижая вероятность ошибок на производственной линии. Они более склонны к соблюдению технических регламентов, поскольку осознают свою роль в достижении общих целей компании. Мотивированные сотрудники проявляют инициативу в улучшении процессов, предлагая инновационные решения и оптимизируя рабочие процедуры. В условиях высокой мотивации сотрудники охотно берут на себя ответственность за результаты работы, чувствуя личную заинтересованность в успехе предприятия. Они тщательнее контролируют качество продукции на каждом этапе производства,

оперативно выявляя и устраняя возможные отклонения от нормы. Мотивированные сотрудники готовы к обучению и повышению квалификации, что позволяет им осваивать новые технологии и улучшать свои профессиональные навыки.

Заметим, что в дополнение к сказанному выше, взаимодействие и сотрудничество между мотивированными сотрудниками становятся более эффективными, что способствует обмену опытом и знаниями. Они охотно делятся информацией о проблемах и предлагают свои решения, создавая атмосферу постоянного улучшения и совершенствования [86]. Мотивация способствует формированию позитивного рабочего климата, где сотрудники чувствуют поддержку и признание со стороны руководства и коллег. Когда сотрудники чувствуют, что их вклад ценится, они более лояльны к компании и готовы прилагать дополнительные усилия для достижения результатов, поэтому мотивация снижает текучесть кадров, сохраняя ценный опыт и знания внутри организации.

4.1.5. Социальные процессы как важная часть процесса производства

В заключение раздела о важности понимания производства как не только технологического, но и социального процесса, подчеркнём, что эффективное управление качеством всегда требует комплексного подхода, учитывающего взаимодействие 1) человеческого фактора и 2) технических аспектов. Мы убедились, что игнорирование социальной составляющей производственного процесса неизбежно приводит к снижению качества и эффективности работы.

В дальнейшем применение принципов социального инжиниринга в управлении производством позволит не только повысить качество продукции, но и создать устойчивое конкурентное преимущество компании. Понимание взаимосвязи между технологическими с одной стороны и социальными аспектами производства с другой стороны открывает новые возможности как для оптимизации процессов, так и для повышения общей эффективности и достижения долгосрочного успеха [88]. Игнорирование этой взаимосвязи, напротив,

обрекает любую компанию на стагнацию и полную потерю конкурентоспособности.

4.2. Данные об исполнении операции установки уплотнителей на стекла автомобиля

Хорошо известно, что корректный анализ данных возможен лишь при условии их однородности [3, 7, 13, 48, 172]. Было бы некорректно сравнивать производительность, используя для этого разные операции с принципиально различной трудоёмкостью (например, тестирование электроники автомобиля некорректно сравнивать с установкой зеркал, так как время, необходимое для исполнения этих операций отличается на порядок), но также некорректно сравнивать производительность сотрудников, обладающих принципиально различным опытом работы (не стоит сравнивать стажера с мастером).

4.2.1. Когортный анализ производительности

Мы проводим когортный анализ [7] производительности труда работников сборочного конвейера, принятых на работу в один период (конкретно — в течение декабря 2022 года) и одновременно проходящих адаптацию к процессу производства. Этот метод позволяет отслеживать динамику производительности группы работников (когорты) на протяжении определённого времени, выявляя закономерности и факторы, влияющие на их эффективность. Он позволяет гомогенизировать данные за счёт того, что все сотрудники находятся в равных условиях: они были приняты на работу в один и тот же период времени и одновременно проходили обучение и адаптировались.

Ниже приведены временные ряды их производительности⁴¹ (см. рисунок 4.1), на которых отчётливо видны периоды 1) адаптации (левее вертикальной

⁴¹ Мы называем сотрудников локациями, так как в рамках нашей парадигмы исследования процессов они действительно являются локациями. Аббревиатура из двух латинских букв означает краткие инициалы (имя и фамилию, без отчества) их имен на латинице.

прямой) и 2) регулярной работы (правее вертикальной прямой), причем, если левее вертикали тренд процесса всегда положительный, то этого нельзя сказать о поведении процесса после прохождения этапа адаптации.

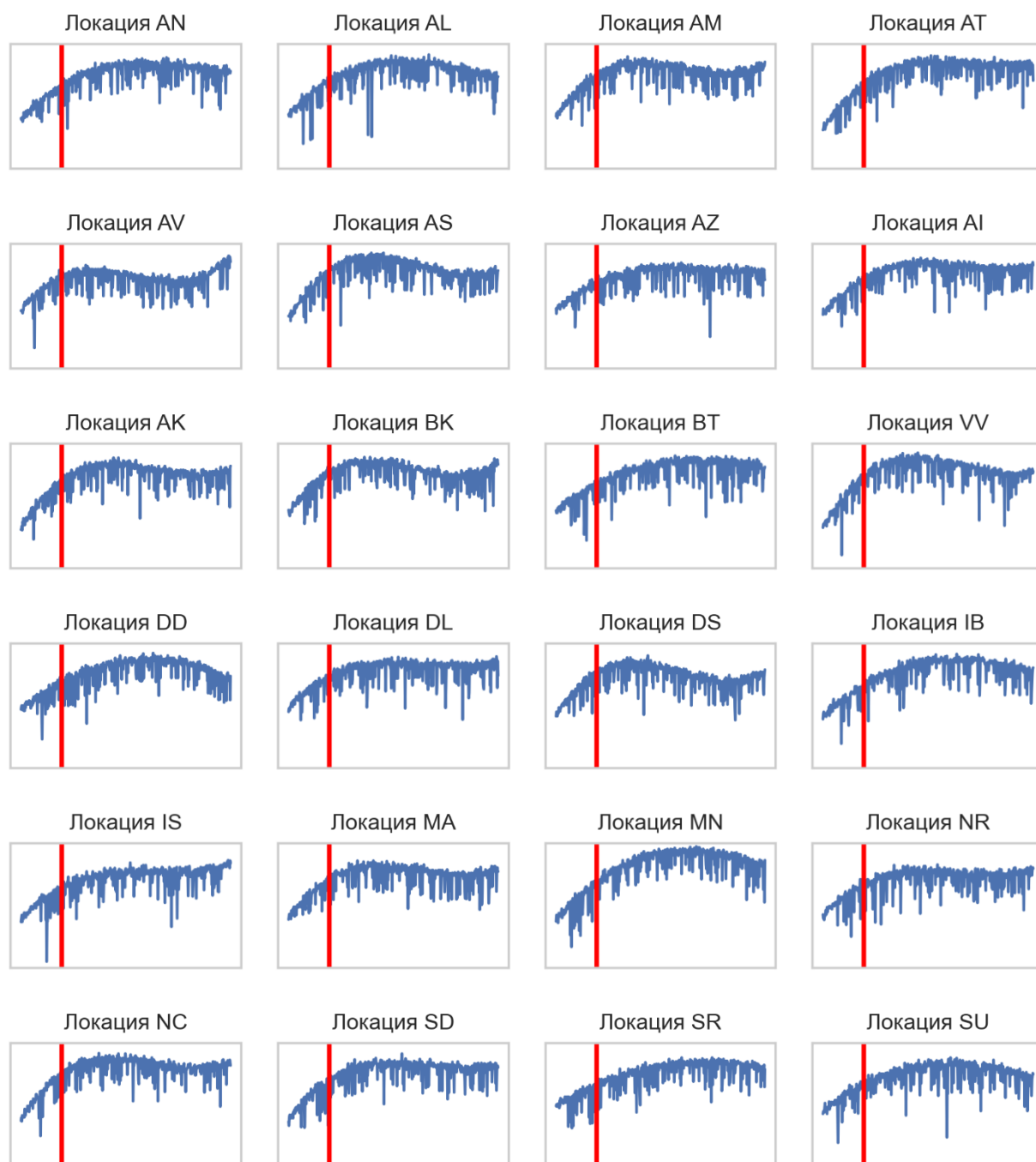


Рисунок 4.1 — Временные ряды исполнения одной из технологических операций

Заметим, что анализ данных по производительности каждой когорты позволяет выявить «узкие места» в процессе обучения, а также факторы, спо-

собствующие или, наоборот, препятствующие быстрой адаптации новых сотрудников (например, можно оценить влияние различных методов обучения, наставничества или организации рабочего места на скорость освоения навыков и достижение целевых показателей).

Сбор и анализ данных о производительности труда новых сотрудников требует чётких метрик и показателей, а также организации эффективного сбора и обработки информации, и кроме того, очень важно отметить, что когортный анализ требует постоянного мониторинга и анализа данных, так как регулярное обновление информации и анализ тенденций позволяют оперативно выявлять проблемы и принимать корректирующие меры [23, 210].

Однако, даже поверхностный взгляд на приведённые выше временные ряды позволяет сделать сразу несколько важных выводов:

1. Во-первых, обучение работников и их выход на регулярные показатели производительности происходит примерно в одинаковые сроки, независимо от уровня начальной подготовки (эти сроки на данной когорте составили 110 ± 10 наблюдений, что в целом соответствует обычному периоду адаптации работника на позиции данной технологической операции).
2. Обращает на себя внимание высокая вариативность рядов, что, видимо, говорит о недостатках в организации труда (см. ниже раздел 4.5 с гипотетической оценкой причин аномальных состояний).

4.2.2. Описание данных

В качестве когорты использовалась группа из 24 сотрудников, принятых для исполнения одной из технологических операций на конвейерной линии одного из ведущих отечественных автопроизводителей (конкретно — речь идёт об операции установки уплотнителей на стекла автомобиля), которые наблюдались в течение 515 рабочих дней. Данные представлены в формате ML, в качестве моментов времени используются рабочие дни, в качестве локаций —

сотрудники, принятые в штат в декабре 2022 года (см. таблицу 4.1).

Таблица 4.1 — Данные о числе выполненных операций в формате ML

	Локация 1	Локация 2	Локация 3	...	Локация 24
Момент 1	98	106	102	...	105
Момент 2	95	102	108	...	110
Момент 4	97	103	107	...	111
Момент 4	100	110	98	...	108
...	...	...	...	...	...
Момент 514	192	182	206	...	188
Момент 515	194	191	213	...	186

Наблюдаемая характеристика процесса — число операций, выполненное тем или иным работником в течение смены. Транспонирование переводит данные из формата ML в формат LM (см. таблицу 4.2).

Таблица 4.2 — Данные о числе выполненных операций в формате LM

	Момент 1	Момент 2	Момент 2	...	Момент 515
Локация 1	98	95	97	...	194
Локация 2	106	102	103	...	191
Локация 3	102	108	107	...	213
Локация 4	113	114	112	...	232
...	...	...	...	...	...
Локация 23	121	124	120	...	196
Локация 24	105	110	111	...	186

Кроме того, мы используем две вспомогательные таблицы, связывающие наименование момента (в качестве которого выступает дата) с его кодом и наименование локации (в качестве которого выступают инициалы работника) с её кодом (см. таблицы 4.3 и 4.4).

Таблица 4.3 — Коды моментов в процессе производства

	Код	Момент
1	mom_0	2023-01-10
2	mom_1	2023-01-11
3	mom_2	2023-01-12
...	...	...

	Код	Момент
515	mom_515	2024-12-30

Таблица 4.4 — Коды локаций в процессе производства

	Код	Локация
1	loc_0	DL
2	loc_1	AT
3	loc_2	AZ
...	...	...
24	loc_23	SD

Под период исторических наблюдений мы отводим первые 100 наблюдений, для этапа обучения, на котором будет производиться обучение моделей и калибровка отклонений — наблюдения со 101 по 150. Последние 416 наблюдений мы отводим на пилотную эксплуатацию детектора аномалий.

4.3. Детекция аномалий по времени в процессе производства

В этом разделе мы проводим поэтапную реализацию алгоритма детекции аномалий в нестабильном процессе по отношению к исполнению одной из технологических операций: начиная с моделирования этого процесса на данных, представленных выше, и заканчивая пилотной эксплуатацией детектора, который возвращает аномальные состояния в каждый момент времени и для каждой локации.

С учётом того, что вся алгоритмика нашей методики уже была нами подробно рассмотрена выше (см. главы 2 и 3), здесь мы ограничимся лишь самыми общими моментами алгоритма, уделяя основное внимание полученным результатам и идеям, которые к ним приводят, но не самим вычислениям.

4.3.1. Моделирование по времени процесса производства

Для моделирования процесса производства мы используем пул из семи моделей, относящихся к разным типам (от простых линейных моделей до сложных

моделей градиентного бустинга и нейросетевых регрессий):

1. `LinearRegression` (линейная регрессия);
2. `DecisionTreeRegressor` (регрессия на решающих деревьях);
3. `RandomForestRegressor` (регрессия на случайных лесах);
4. `GradientBoostingRegressor` (градиентный бустинг над случайными лесами);
5. `SVR` (регрессия на методе опорных векторов);
6. `KNeighborsRegressor` (регрессия на ближайших соседях);
7. `MLPRegressor` (нейронная сеть многослойного перцептрона).

Мы применяем специально разработанную для настоящего диссертационного исследования функция моделирования по времени `modeling_by_moments`, которая подробно описана нами в приложении С.1.1. Используя её в цикле по приведенному выше списку моделей с максимальным погружением авторегрессии, равным 10, получаем следующие результаты, демонстрирующее эффективность той или иной схемы моделирования (см. рисунок 4.2).

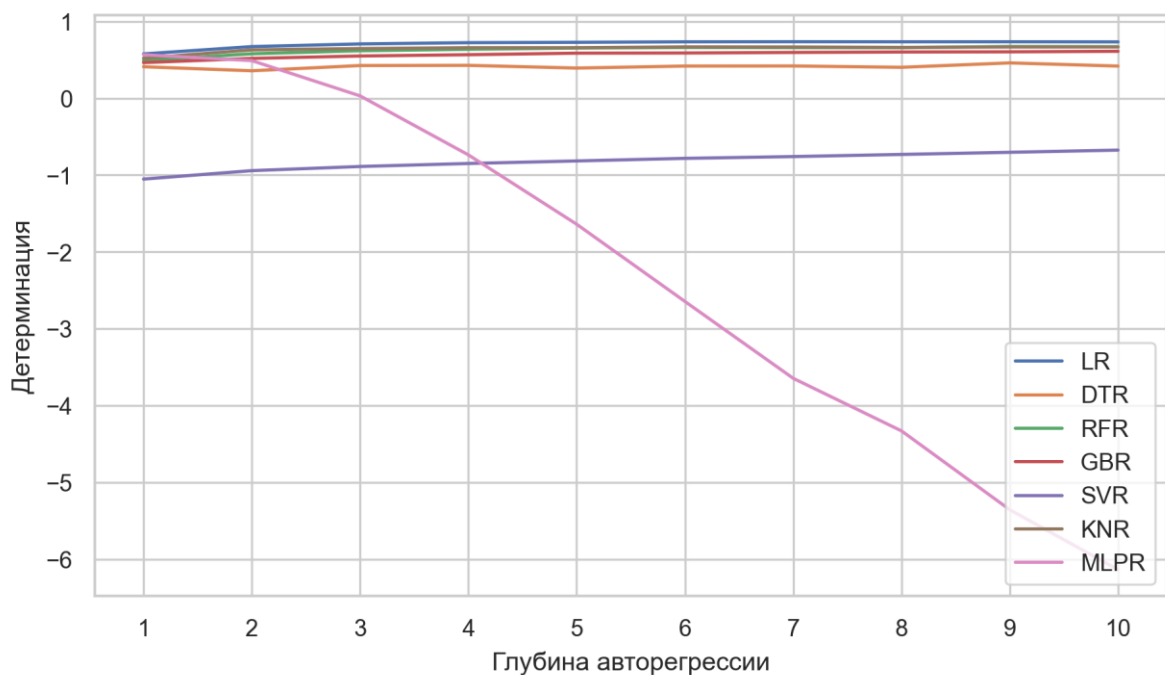


Рисунок 4.2 — Анализ схем моделирования по времени в процессе производства

Вызывает удивление поведение многослойного перцептрона (модель

MLPR), однако в наши планы не входит изучение этого феномена⁴², наша задача состоит в том, что нам нужно определить, какая именно из наших схем моделирования обладает наилучшей прогнозирующей способностью. Однако, масштаб изображения таков, что он не позволяет нам достоверно определить различия между тем или иным видом регрессионной модели, и для преодоления этого препятствия мы просто детализируем изображение, ограничив его вертикальным промежутком от 0 до 1 (см. рисунок 4.3), после чего график модели MLPR перестаёт оказывать влияние на изображение, и становится понятно, что наилучшую схему моделирования демонстрирует модель **LinearRegression** с погружением по авторегрессии на глубину 6. И хотя, не теряя точности предсказаний, можно было бы взять глубину, равную 7 или 8, мы выбираем меньшую из глубин, чтобы использовать как можно меньшее число предикторов регрессии.

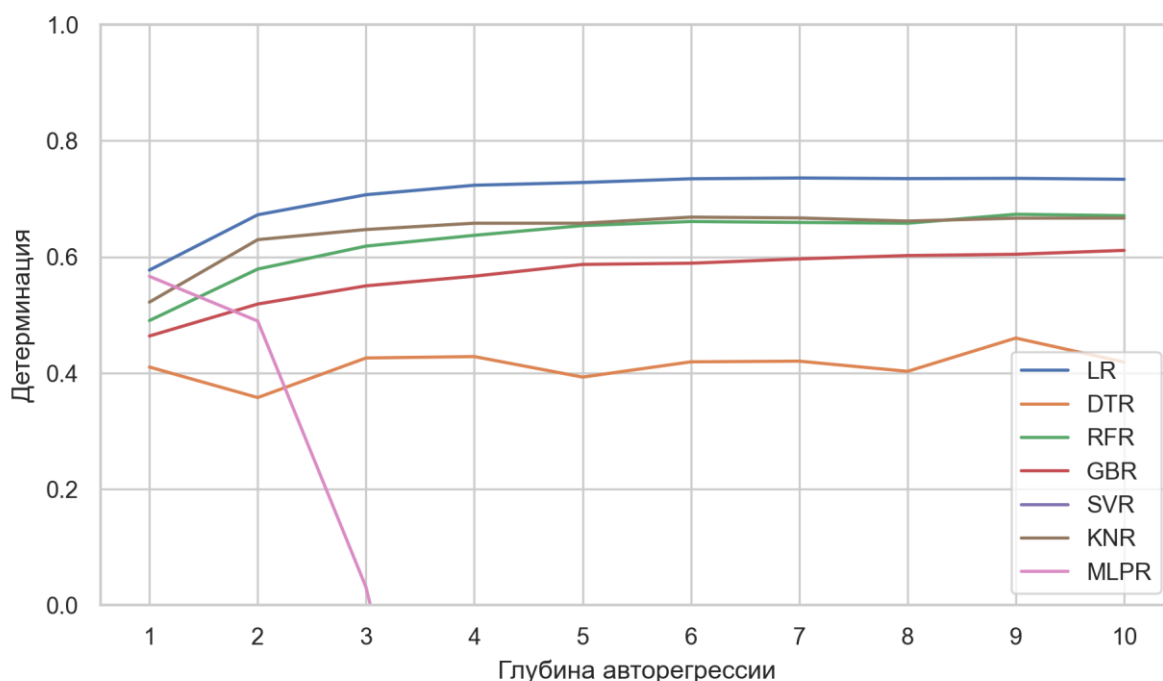


Рисунок 4.3 — Укрупнённый скоринг моделей по времени в процессе производства

⁴² Скорее всего, нейронная сеть, будучи сложной моделью по своей природе, просто переобучается на исторических данных и именно поэтому даёт такие плохие предсказания, хотя, это только наше предположение

Итак, по итогам этапа моделирования процесса оказывается, что наилучшая схема моделирования состоит в том, чтобы использовать модель линейной регрессии `LinerRegression` с глубиной авторегрессии, равной 6.

4.3.2. Калибровка отклонений по времени в процессе производства

Как отмечалось выше (см. главу 3, раздел 3.8), при работе с реальными (а не с синтетическими) данными мнение экспертного сообщества, как правило, оказывается недоступным, и к ним неприменимы методы сопоставления. Именно в такой ситуации сейчас находимся и мы: все, что мы знаем о процессе исполнения одной из технологических операций на сборке автомобилей — это данные, сосредоточенные в четырёх таблицах 4.1, 4.2, 4.3 и 4.4, но никакой экспертной разметки, относящих то или иное состояние процесса к категории нормы или аномалии, у нас нет. Тем не менее, перед нами сейчас стоит та же самая задача, что и во время анализа синтетических данных: нам нужно каким-то образом определить, какую именно величину отклонения прогнозируемого значения от истинного мы будем считать аномальной, и до тех пор, пока мы её не решим, наш прогноз, полученный нами ранее на этапе моделирования процесса будет нам совершенно бесполезен.

Для того чтобы справиться с этой задачей, мы применим следующие соображения. Наша лучшая прогнозирующая схема прогнозирует процесс не идеально, в любой момент времени и в любой локации есть какие-то отклонения реальных значений от того, что она прогнозирует. Отклонения могут быть как в положительную, так и в отрицательную сторону, но у каждого из них есть выраженность, которая определяется абсолютной величиной отклонения: чем больше абсолютная величина, отклонения, тем ярче выражена аномальность соответствующего состояния нашего процесса. Давайте выпишем абсолютные величины отклонений, которые получились во все моменты времени во всех локациях на этапе обучения, в виде списка. Этап обучения продолжался в течение 50 моментов времени в 24 локациях, следовательно, мы получим список, в котором будет $50 \cdot 24 = 1\,200$ элементов. Упорядочим список по

возрастанию и визуализируем его (см. рисунок 4.4) в виде кривой (которую здесь и везде далее будем называть калибровочной кривой).

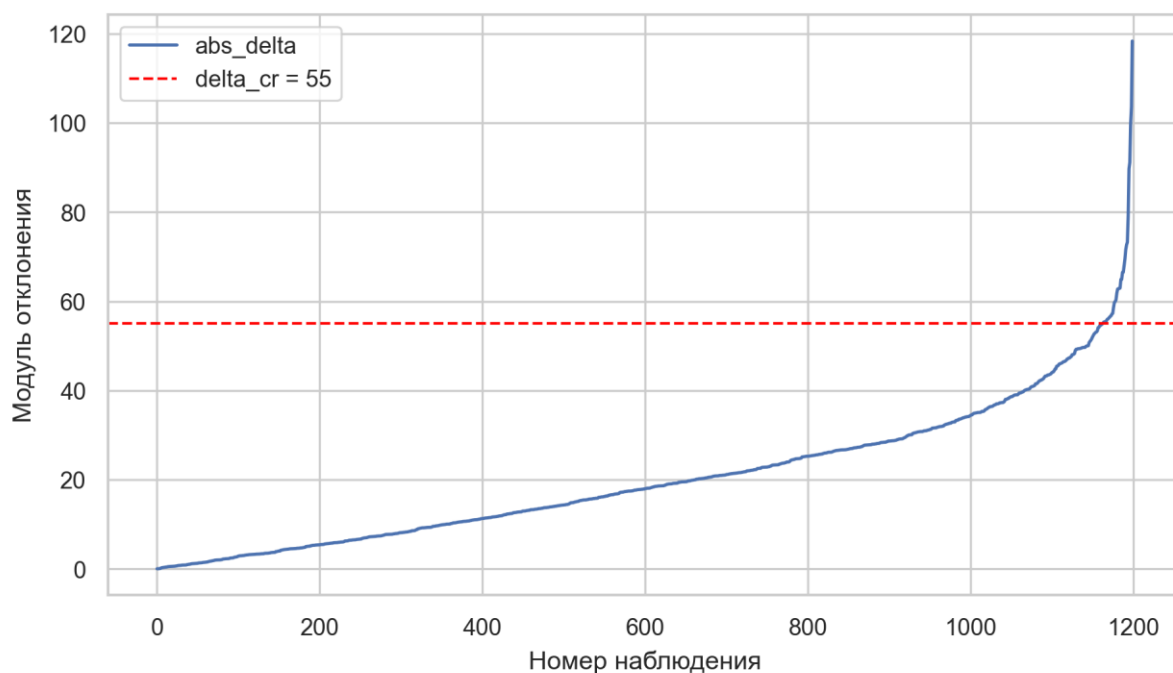


Рисунок 4.4 — Калибровочная кривая отклонений по времени в процессе производства

Анализируя калибровочную кривую, видим, что в её поведении есть две тенденции: сначала отклонения возрастают медленно, а начиная с некоторой позиции — быстро. Точка перехода от одной тенденции к другой называется локтевой точкой [191], а ординату локтевой точкой мы будем считать той границей, начиная с которой отклонения прогноза от реальности становятся аномально большими, и, таким образом, анализ калибровочной кривой показывает, что критическое отклонение по времени $\Delta_{cr} = 55$.

4.3.3. Пилотная эксплуатация детектора аномалий по времени в процессе производства

По завершении предварительных этапов в нашем распоряжении есть все сведения, необходимые для пилотной эксплуатации детектора. Во-первых, теперь мы знаем, в чем состоит наилучшая схема моделирования процесса: нам нужно проводить линейную регрессию на глубине авторегрессии, равной шести наблюдениям.

```
best_model = LinearRegression
autoregression_dive = 6
```

А во-вторых, мы знаем величину критического отклонения:

```
delt_cr = 55
```

Теперь мы применяем алгоритм пилотной эксплуатации, изложенный нами в приложении С.4, передав ему найденные нами значения 1) лучшей модели, 2) лучшей глубины авторегрессии и 3) критической величины отклонения. Для этапа пилотной эксплуатации мы используем этап из последних 415 наблюдений, каждое из которых в каждой из наших локаций детектор диагностирует как нормальное или аномальное. Например, наблюдая за локацией АК, обнаруживаем 9 аномалий⁴³ (см. рисунок 4.5 и 4.6).

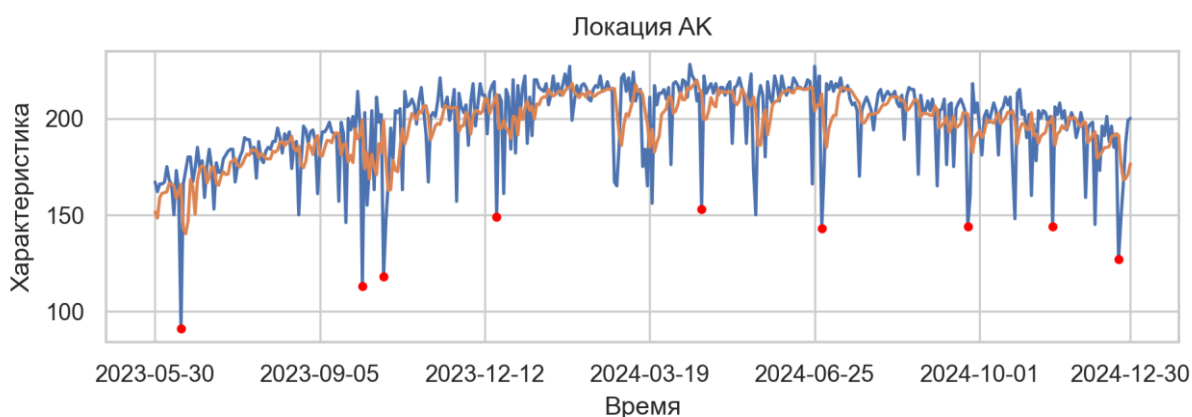


Рисунок 4.5 — Модель по времени в локации АК.



Рисунок 4.6 — Отклонения от модели по времени в локации АК.

⁴³ Много это или мало — это другой вопрос. Аналитик данных вообще никогда не дает никаких оценок, он констатирует факт. Подробнее об этом см. ниже, раздел 4.5.

На рисунке 4.5 оранжевая кривая означает спрогнозированные значения наблюдаемой характеристики (напомним, что в данном случае речь идёт о числе исполнений данным работником одной из технологических операций), а синяя кривая — истинные значения характеристики, зарегистрированные в реальности. В те моменты, когда синяя кривая слишком отклоняется от оранжевой (если быть точным, в те моменты, когда это отклонение превышает $\text{dela\_cr} = 55$), детектор фиксирует аномалию. Ниже, на рисунке 4.6 мы синей кривой изобразили собственно сами отклонения, а оранжевыми пунктиром — допустимые границы отклонений для того, чтобы подчеркнуть принцип, по которому фиксируется аномальное состояние процесса.

Точно так же мы поступаем с локацией AZ (см. рисунок 4.7, 4.8), где фиксируем 3 аномалии⁴⁴.

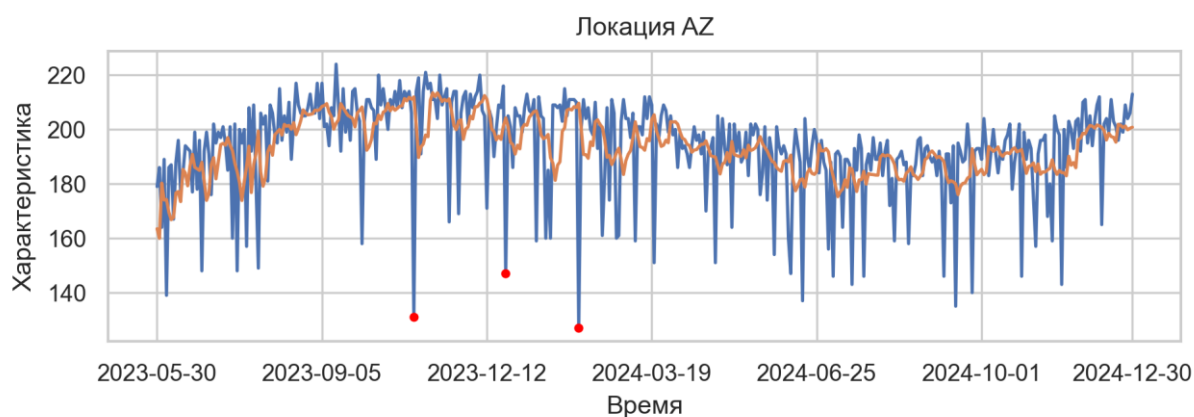


Рисунок 4.7 — Модель по времени в локации AZ

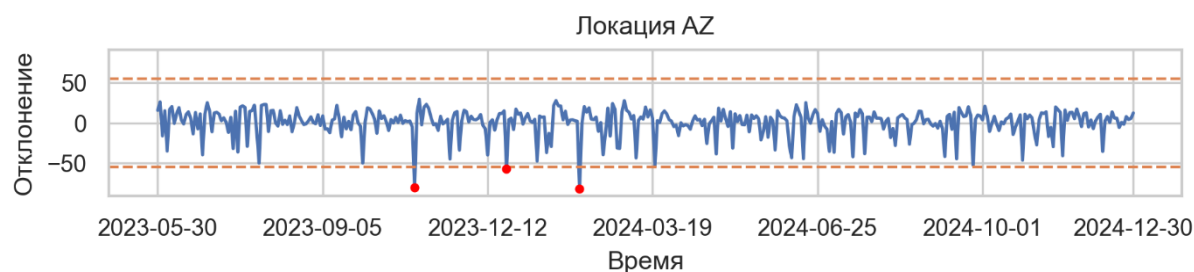


Рисунок 4.8 — Отклонения от модели по времени в локации AZ

⁴⁴ И опять-таки: мы не делаем из этого никаких выводов, мы не говорим, что работник AZ работает лучше или хуже работника АК, мы только констатируем наблюдаемый факт, который состоит в том, что на этапе пилотной эксплуатации зафиксировано три аномалии.

И так далее. Мы анализируем все локации.

Но с другой стороны, нашей целью является не оценка той или иной локации (в контексте данного раздела — того или иного работника) с позиций её высокой или низкой склонности к аномалиям. Нам хотелось бы иметь инструмент, который позволял бы нам своевременно реагировать на наступление аномального состояния в процессе, который мы хотим контролировать, и относительно которого несём ответственность за качество его протекания. В этой связи, мы генерируем сопряжённую схему: если выше (см. рисунки 4.5–4.8) мы в каждой отдельной локации помечали все аномальные моменты времени, то теперь, наоборот, в каждый отдельный момент времени будем помечать все аномальные локации. Такой подход к анализу аномальных состояний процесса мы называем календарём аномалий. Вот, например, рейтинг аномалий⁴⁵, детектированных по времени в процессе производства в момент времени 2024-04-04 (см. рисунок 4.9).

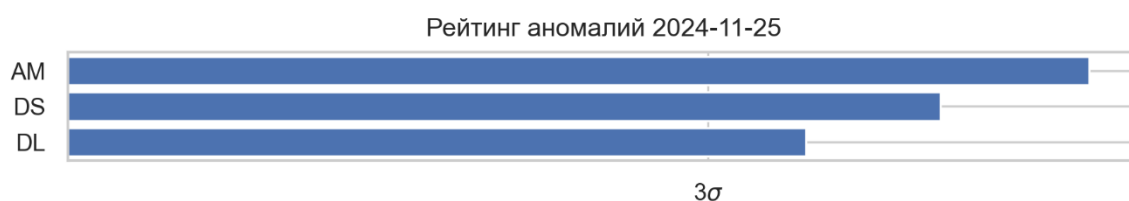


Рисунок 4.9 — Рейтинг аномальных по времени локаций в процессе производства в момент времени 2024-04-04

Здесь мы визуализируем аномалии относительно психологически значимой величины 3σ , где σ — это среднее квадратичное отклонение прогнозируемых значений наблюдаемой характеристики от её реальных значений, зафиксированное на этапе обучения. Отметим, что правило 3σ мы вообще нигде не используем, мы детектируем аномальные отклонения в силу совсем других

⁴⁵ Мы называем это рейтингом, потому что все детектированные аномалии упорядочены нами по убыванию их выраженности.

принципов⁴⁶, и в нашем прочтении аномальными могут оказаться (и часто оказываются) отклонения, гораздо меньшие, нежели 3σ . Отсечка по 3σ нужна нам лишь для того, чтобы иметь какой-то безразмерный абстрактный ориентир, при помощи которого мы могли бы сравнивать результаты детекции в процессах различной природы (производства в главе 4, сервиса в главе 5, маркетинга в главе 6 и т. д.).

В другой момент времени, мы получаем другой рейтинг аномалий с другими локациями, аномалии в которых другие и выражены по-другому (см. рисунок 4.10), но они тоже соотнесены с 3σ . Но кроме того, и это гораздо важнее с практической точки зрения, все наши аномалии упорядочены по убыванию выраженности. Значит и реакция на аномалии должна быть упорядочена, то есть, в первую очередь разбираться с причинами аномалий нужно в локациях, которые занимают «ведущие» позиции в своих рейтингах (см., например, рисунок 4.9: сначала нужно проверить локацию AM и лишь затем переходить к локациям DS и DL).

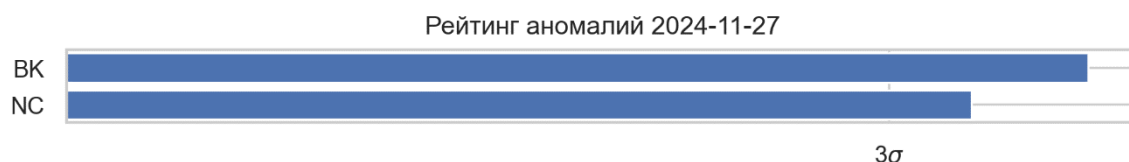


Рисунок 4.10 — Рейтинг аномальных по времени локаций в процессе производства в момент времени 2024-09-16

В какие-то из моментов наблюдения аномалии не фиксируются вообще, в другие — наоборот, их может оказаться достаточно много, но в любом случае, упорядоченность списков аномалий позволяет аналитику расставлять приоритеты в аудите подконтрольных ему локаций: в первую очередь аудиту должны подвергаться именно те локации, где выраженность аномалий выше.

⁴⁶ А именно: допустимый порог отклонений мы определяем, исходя из локтевой точки на калибровочной кривой.

Календарь аномалий по времени в процессе производства в течение последних десяти наблюдений представлен в приложении D.1.1.

4.4. Детекция аномалий по локациям в процессе производства

Напомним, что, наряду с детекцией аномалий по времени, методика МССР предполагает ещё один тип детекции аномалий, а именно: детекцию аномалий по локациям, когда значение наблюдаемой нами характеристики сравнивается не с предыдущими её значениями в этой же локации, а с её значениями в других локациях в тот же момент времени. Этапы алгоритма детекции по локациям те же, что и этапы алгоритма детекции по времени: моделирование, выбор лучшей моделирующей схемы, калибровка и, наконец, пилотная эксплуатация. При этом используются данные в формате LM (см. таблицу 4.2), с тем же разбиением на периоды: под период исторических наблюдений отводятся первые 100 наблюдений, под период обучения моделей и калибровки — следующие 50 наблюдений, а под период пилотной эксплуатации — последние 415 наблюдений.

4.4.1. Моделирование по локациям процесса производства

Так же как и выше, для моделирования процесса производства по локациям мы используем пул из тех же семи моделей (см. раздел 4.3.1). Функция моделирования по локациям описана в приложении C.2 (мы, так же как и выше, используем именно локальный вариант алгоритма, не апеллируя к ближайшим предикторам). Используем максимальное погружение в историю наблюдений, равное 10, и получаем следующие результаты (см. рисунок 4.11). Снова многослойный перцептрон MLPR демонстрирует выдающееся из общего ряда падение детерминации, однако, на этот раз, падение демонстрирует и линейная регрессия LR, и наши выводы по этому этапу таковы, что наилучшей схемой моделирования, которую мы будем использовать далее на этапе пилотной эксплуатации детектора аномальных состояний, оказывается регрессия на ближайших соседях KNR, `histiry_dive = 10`.

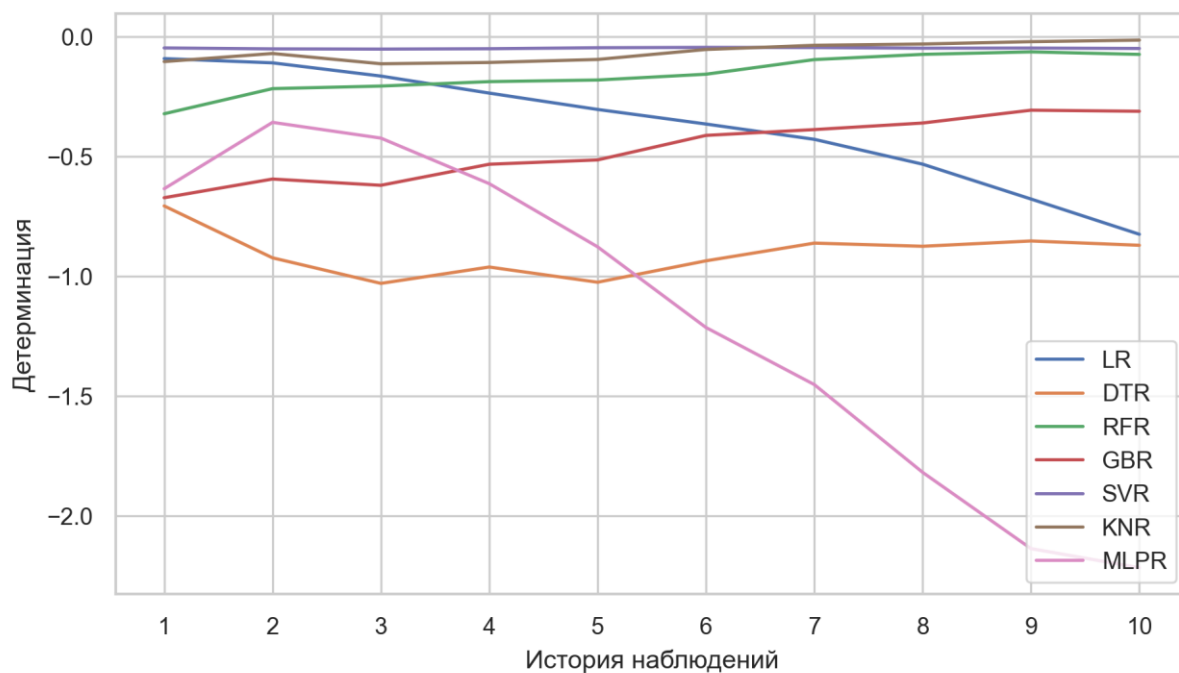


Рисунок 4.11 — Сравнительный анализ схем моделирования моделей по локациям в процессе производства

Следовательно:

```
best_model = KNeighborsRegressor
history_dive = 10
```

4.4.2. Калибровка отклонений по локациям в процессе производства

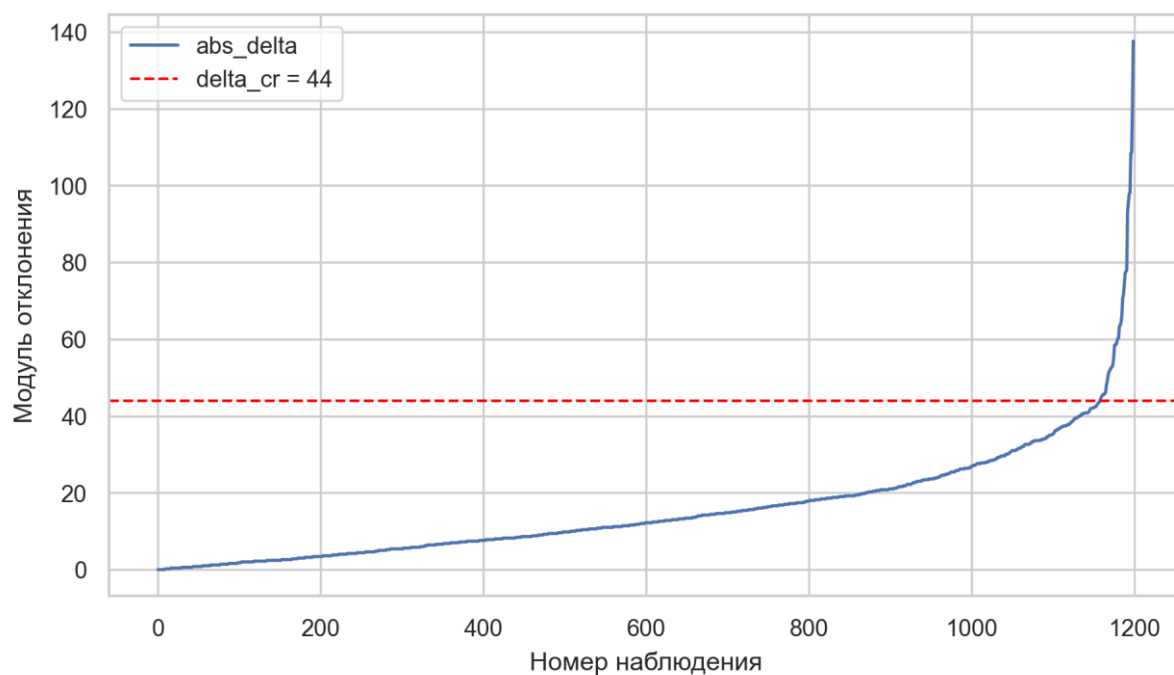


Рисунок 4.12 — Калибровочная кривая отклонений по локациям в процессе производства

Калибровку детектора аномалий проводим по локтевой точке абсолютных величин отклонений (см. рисунок 4.12). Видим, что локтевая точки приходится на величину отклонения, равную примерно 44. Следовательно:

```
delt_cr = 44
```

4.4.3. Пилотная эксплуатация детектора аномалий по локациям в процессе производства

Теперь у нас есть все необходимые сведения для пилотной эксплуатации детектора. Во-первых, мы знаем, в чем именно состоит наилучшая схема моделирования процесса: нам нужно проводить регрессию на ближайших соседях, используя при этом глубину истории наблюдений, равную десяти наблюдениям.

```
best_model = KNeighborsRegressor  
history_divide = 10
```

А во-вторых, мы знаем величину критического отклонения:

```
delt_cr = 44
```

Для этапа пилотной эксплуатации мы используем этап из последних 415 наблюдений, каждое из которых в каждой из локаций детектор диагностирует как нормальное или аномальное. Так, наблюдая за локацией AZ, обнаруживаем 12 аномалий (см. рисунок 4.13 и 4.14).

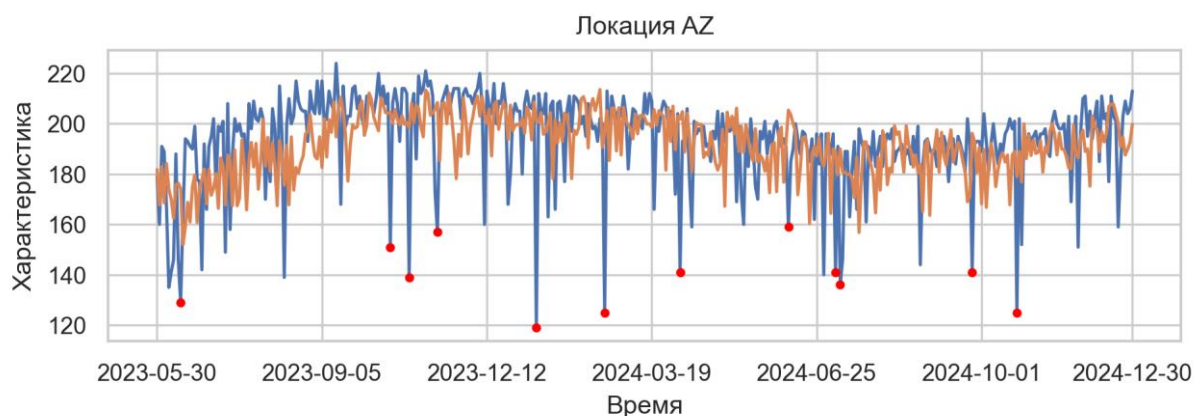


Рисунок 4.13 — Модель по локациям в локации AZ

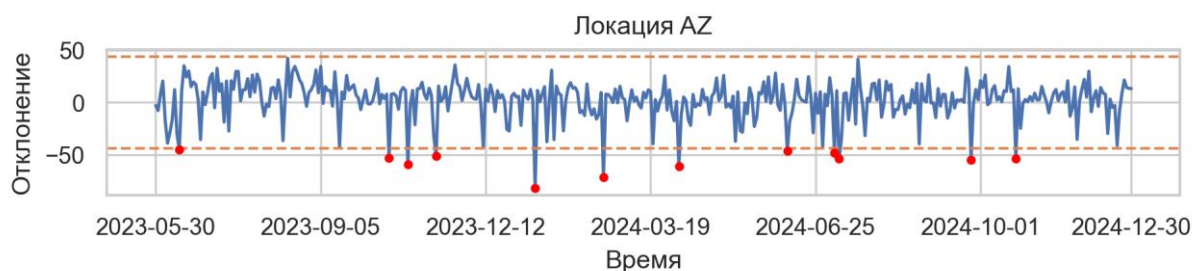


Рисунок 4.14 — Отклонения от модели по локациям в локации AZ

Отметим, что все аномалии, которые отмечены в локации AZ, являются односторонними: отклонение характеристики всегда происходит в сторону, меньшую, нежели это нами прогнозировалось, что, безусловно, свидетельствует о том, что в какие-то моменты времени AZ теряет производительность⁴⁷. Но это не всегда так. Рассмотрим, например, локацию BT (см. рисунки 4.15 и 4.16).

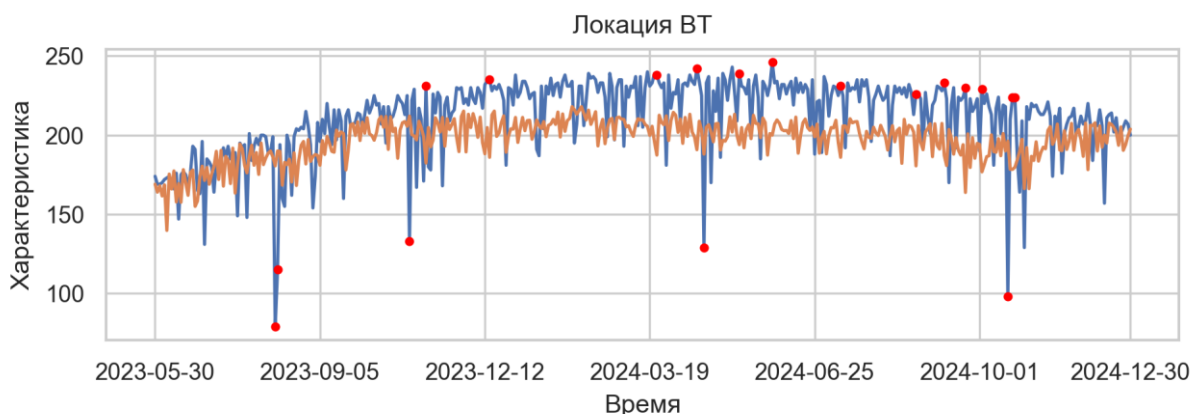


Рисунок 4.15 — Модель по локациям в локации BT



Рисунок 4.16 — Отклонения от модели по локациям в локации BT

⁴⁷ Но это ни в коем случае не указывает на причину падения. Мы ещё будем это обсуждать подробно в разделе, посвящённом интерпретации результатов.

Мы видим 17 аномальных состояний, из которых отклонения в меньшую сторону составляют 5 случаев, а отклонения в большую сторону — 12 случаев. Это, безусловно, свидетельствует о том, что ВТ работает лучше других сотрудников из своей когорты⁴⁸. Об этом же говорит и сам характер прогнозирующей кривой (оранжевая кривая на рисунке 4.15), которая почти всегда лежит ниже кривой реальных результатов.

Однако для наших целей важны не общие рассуждения, а конкретные знания об отклонениях от нормы, которые нам необходимо получать в каждый момент времени для того, чтобы оперативно на них реагировать. Поэтому мы строим ещё один календарь аномалий⁴⁹, на этот раз это будет календарь аномалий по локациям.

В каждый момент наблюдений мы формируем рейтинг аномальных локаций, указывая как абсолютную величину отклонения, так и его величину относительно 3σ . Например, в момент времени 2023-06-12 мы наблюдали аномалии в трёх локациях: АМ, MN и ВК (см. рисунок 4.17).

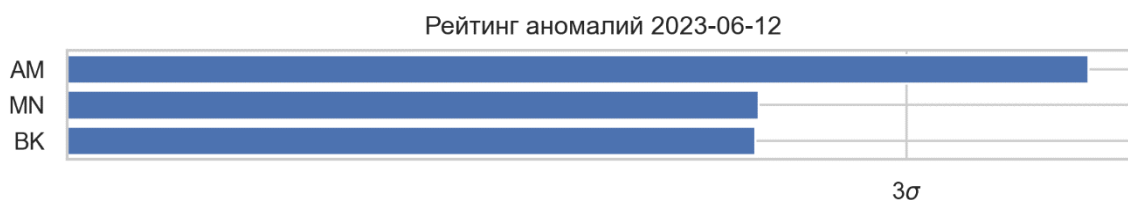


Рисунок 4.17 — Рейтинг аномальных по локациям локаций в процессе производства в момент времени 2023-06-12

Причём, выраженность аномалий такова, что, реагируя на них, мы в первую очередь должны выяснить причины несоответствия процесса его норме в локациях с наиболее выраженной аномалией: в локациях АМ, и лишь после этого мы должны перейти к локациям MN, и лишь после этого — к локациям ВК.

⁴⁸ Хотя, факт, который мы можем вполне уверенно констатировать, так же как и выше, не раскрывает причины этого явления.

⁴⁹ Напомним, что выше мы уже построили один календарь аномалий, это был календарь аномалий по времени в процессе производства.

В момент времени 23-08-09 мы видим четыре аномалии: BT, DS, DD и AI (см. рисунок 4.18).



Рисунок 4.18 — Рейтинг аномальных по локациям локаций в процессе производства в момент времени 2023-08-09

Так же как и выше, рейтинг аномалий подсказывает нам, не только сами локации, в которых следует проводить аудит, но и порядок его проведения⁵⁰: начинать следует с наиболее выраженной аномалии в локации BT, и заканчивать наименее выраженной аномалией в локации AI.

Календарь аномалий по локациям в процессе производства в течение последних десяти наблюдений представлен в приложении D.1.2.

4.5. Об интерпретации аномалий, выявленных в процессе производства

Своевременное выявление аномальных состояний в производственном процессе играет важную роль в обеспечении стабильности и эффективности работы предприятия, так как аномалии, будь то отклонения от заданных параметров или неожиданные сбои, могут указывать на скрытые проблемы, которые, если их не обнаружить и не устранить вовремя, способны привести к серьёзным последствиям [6, 19, 38, 154, 56, 162, 206, 222, 232, 245]. В этой связи, первостепенная задача — это минимизация рисков [221]. Оперативное выявление аномальных или нетипичных состояний процесса производства позво-

⁵⁰ Именно для этого и нужен рейтинг аномалий: чтобы выстраивать работу аудита наиболее оптимальным способом и подвергать проверкам наиболее проблемные локации, а те локации, в которых не было выявлено никаких несоответствий, вообще не беспокоить.

ляет предотвратить развитие критических ситуаций и избежать тем самым значительных финансовых потерь, и чем быстрее будет установлена причина той или иной аномалии, тем быстрее можно будет разработать и реализовать эффективные меры по её устранению.

В конечном счёте, своевременное выявление аномалий и оперативное принятие мер — это инвестиция в устойчивость и конкурентоспособность предприятия. Оно позволяет не только избежать негативных последствий, но и повысить эффективность производства, улучшить качество продукции и укрепить репутацию компании. Чёткий протокол реагирования на аномалии, встроенный в систему управления производством, является надёжным инструментом для поддержания стабильности и достижения поставленных целей, а постоянный мониторинг и анализ данных, в свою очередь, позволяют выявлять даже незначительные отклонения и предотвращать развитие серьёзных проблем [6, 19, 162, 225, 237].

Вместе с тем в мире современного производства, где автоматизация и сбор данных стали неотъемлемой частью производственного процесса, соблазн полагаться исключительно на формальные показатели для выявления аномалий процессов необычайно велик. Следует иметь в виду, что слепое следование этим показателям, без учёта специфики контекста конкретного производства (или даже конкретного рабочего места), превращается в опасную профанацию, способную не только не улучшить, но и ухудшить качество производственного процесса.

Автоматизированные системы мониторинга, безусловно, полезны в обнаружении отклонений от заданных параметров. Но интерпретация этих сигналов требует глубокого понимания процесса, знания его нюансов и возможных факторов, влияющих на его ход, а игнорирование специфики ситуации порождает массу проблем. Это приводит к ложным тревогам, необоснованным корректировкам, ухудшению стабильности процесса и, в конечном итоге, к снижению качества продукции. Работники, видя, что автоматизированная си-

стема часто ошибается, теряют к ней доверие и начинают игнорировать её сигналы, что увеличивает риск пропуска действительно серьёзной аномалии.

Таким образом, реагирование на аномалии производственного процесса должно быть основано не только на формальном анализе данных, но и на глубоко понимании процесса, опыте специалистов и контексте конкретной ситуации. Только сочетание этих факторов может привести к действительно эффективному повышению качества производственного процесса, тогда как профанация, основанная на слепом следовании алгоритмам, является прямой дорогой к ухудшению результатов и потере конкурентоспособности.

Давайте вспомним, что, анализируя данные об исполнении одной из технологических операций в процессе сборки на одном из крупнейших отечественных автопредприятий (см. выше, разделы 4.3.3 и 4.4.3), мы обнаружили частые и значительные отклонения производительности в отрицательную сторону: производительность работника снижается, тогда как его предыдущие результаты никак не предвещали такого снижения. В чем причина таких отклонений? Мы можем выдвинуть целый ряд гипотез.

Гипотеза 1. Все дело в состоянии работника. Возможно, он утратил форму в связи с недомоганием, или в связи с личными проблемами и душевным состоянием.

Гипотеза 2. Причина в том, что работник в те или иные дни не располагает достаточным количеством комплектующих. Он просто простаивает, потому что ему не с чем работать.

Гипотеза 3. Работников перемещают с одного рабочего места на другое, и оказавшись в новой, не организованной «под себя» обстановке, работник существенно снижает свою производительность.

Но справедлива ли та или иная из наших гипотез? Или она ошибочна? Мы этого не знаем, и опасность состоит даже не в том, что мы можем ошибиться (в той или иной степени ошибаются все и всегда). Опасность состоит в самой постановке вопроса: «В чем причина аномального отклонения?»

Интерпретация результатов анализа данных о производственном процессе — задача, требующая не только статистической грамотности, но и глубокого понимания специфики самого производства. Аналитик данных, сосредоточенный прежде всего на обработке и моделировании, может упустить (и наверняка упустит) нюансы или какие-то дела, критически важные для правильной интерпретации, в тот момент как профильный представитель производственной области, будь то технолог, инженер или начальник смены, обладает интуитивным знанием о взаимосвязях, не всегда очевидных из данных. Он понимает контекст: условия работы оборудования, качество сырья, человеческий фактор или колебания спроса. Он видит за цифрами реальные процессы, проблемы и потенциальные возможности. Кроме того, интерпретация данных — это не просто констатация фактов, а выработка чётких рекомендаций по улучшению процесса. Производственник, обладая большим практическим опытом, предлагает решения, реалистичные и эффективные в конкретных условиях производства. Аналитик же, оперируя лишь данными, может предложить (и наверняка предложит) решения, не учитывающие технологические ограничения или целесообразность.

Все это означает, что изложенный в этой главе алгоритм детекции аномалий в производственном процессе является не более чем инструментом, помогающим специалистам по управлению качеством, но не заменяющим их. Детектор сигнализирует о наличии аномалии, и, подчеркнём: это требует совсем нетривиальных решений (см. приложение А, а также — главы 1, 2 и 3 настоящей диссертационной работы), но указание её причин и включение протокола реагирования выходят за рамки его функционала.

4.6. Выводы по главе 4

Подводя краткие итоги нашей работы по исследованию процесса исполнения одной из технологических операций и выявлению в нем аномалий, отметим следующее.

1. Технологические операции в процессе производства, по определению, должны обладать стабильностью. Нарушение стабильности означает нарушение технологии, и об этом мы неоднократно говорили выше.
2. Мы рассматриваем одну из операций производственного процесса не в смысле её технологического исполнения, а в социальном смысле: как некую метрику эффективности того или иного работника.
3. В этом смысле процесс оказывается существенно нестабильным, так как зависит от множества внешних факторов: состояния работника, его подготовленности, организации рабочего мета и т. д., вплоть до цепочки взаимосвязей между подразделениями предприятия.
4. Мы формализуем сведения о процессе в виде данных нужного нам формата ML/ LM и далее работаем с данными о процессе, не делая разницы между процессом и данными, которые его описывают.
5. Применение методики МССР (Modeling, Challenge, Calibration, Production) позволяет нам сформировать на этапе пилотной эксплуатации рейтинги аномалий, которые указывают чёткую последовательность проверок для выяснения причин несоответствий и, тем самым, существенно повышают эффективность процедур аудита данного конкретного процесса.
6. Установление причин алгоритмически выявленных несоответствий и включение соответствующих протоколов реагирования на них должно осуществляться представителями предметной области и выходит за рамки компетенций аналитика данных.

Таким образом, проведённая в этой главе процедура детекции аномалий демонстрирует применимость методики МССР (Modeling, Challenge, Calibration, Production) к отдельным процессам в группе процессов производства, при условии, что эти процессы рассматриваются не с технологической, а с социальной точки зрения. В практическом смысле выявление таких несоответствий позволяет оперативно реагировать на нежелательные состояния в процессах производства и своевременно их корректировать.

ГЛАВА 5. ДЕТЕКЦИЯ АНОМАЛИЙ В ПРОЦЕССЕ СЕРВИСНОГО ОБСЛУЖИВАНИЯ АВТОМОБИЛЕЙ

Мы рассматриваем продукт как сложную комплексную сущность, которая помимо собственно изделия, со своими техническими и эксплуатационными характеристиками, включает в себя также и его сервисное сопровождение на протяжении всего его жизненного цикла (см. главу 1, раздел 1.3.1).

В современном мире конкуренция на рынке товаров и услуг невероятно высока, и именно сервисное обслуживание приобретает решающее значение для формирования положительного восприятия качества любого продукта со стороны потребителя. Изначально, качество продукта ассоциируется с его функциональностью и надёжностью, однако долгосрочное взаимодействие с ним (а обладание автомобилем — это и есть пример весьма долгосрочного взаимодействия с продуктом), а также соответствующая ему поддержка со стороны производителя формируют гораздо более глубокое и устойчивое впечатление [33, 34, 96, 97, 99]. Сервисное обслуживание включает в себя широкий спектр услуг, начиная от консультаций при выборе и заканчивая послепродажной поддержкой, ремонтом и обновлением [34, 102, 105, 107]. Качественное сервисное обслуживание демонстрирует заботу производителя о своих клиентах и готовность решать возникающие проблемы [113, 117], что значительно повышает лояльность к бренду [122]. Кроме того, оперативное устранение неисправностей укрепляет уверенность потребителя в продукте и компетентности производителя.

Игнорирование сервисного обслуживания вполне может привести к негативным последствиям, таким как потеря клиентов, ухудшение репутации и снижение продаж [34, 39, 42, 44]. В эпоху социальных сетей, негативные от-

зывы о продукте или сервисе могут быстро распространиться и нанести существенный ущерб бренду. В то же время, позитивный опыт сервисного обслуживания может стать мощным инструментом маркетинга, привлекая новых клиентов и укрепляя позиции на рынке. Более того, качественное сервисное обслуживание продлевает жизненный цикл продукта, позволяя потребителю максимально эффективно использовать его возможности [46]. Регулярное техническое обслуживание, а также своевременное обновление программного обеспечения и доступность запасных частей гарантируют долговечность и бесперебойную работу продукта. В результате, потребитель получает максимальную отдачу от своих затрат и остаётся доволен своим выбором.

В современных условиях сервисное обслуживание является критически важным элементом в интегрированной системе управления качеством, оказывая существенное влияние на достижение стратегических целей организации, связанных с удовлетворением требований потребителей и обеспечением устойчивого конкурентного преимущества [58, 63, 77, 151, 168].

Отметим, что сервисные процессы в контексте системы управления качеством выполняют целый ряд ключевых функций. Прежде всего, они выступают как канал обратной связи и источник данных для улучшения качества [31, 77, 151]: сервисное обслуживание генерирует ценную информацию о восприятии потребителем характеристик продукции/услуги, выявленных несоответствиях, функциональных отказах и эксплуатационных ограничениях. Эти данные, документируемые в рамках системы сбора и анализа данных системы управления качеством, служат основой для реализации корректирующих и предупреждающих действий, направленных на устранение первопричин проблем и предотвращение их повторного возникновения в производственных процессах и разработке новой продукции.

Кроме того, сервисные процессы обеспечивают минимизацию негативного воздействия на лояльность потребителей [27, 80], так как оперативное и эффективное устранение проблем, возникающих у потребителей в процессе потребления продукции/услуги, является ключевым фактором поддержания

лояльности и предотвращения оттока клиентской базы. Эффективная система управления рекламациями и претензиями, интегрированная в систему управления качеством, позволяет трансформировать потенциально негативный потребительский опыт в фактор удержания клиентов и повышения их удовлетворенности.

Наконец, сервисное обслуживание само по себе является индикатором результативности системы управления качеством: сервисные процессы выступают в качестве диагностического инструмента для оценки эффективности функционирования системы управления качеством. Ключевые показатели эффективности (KPI) сервисного обслуживания являются объективными индикаторами, отражающими уровень качества продукции/услуги и эффективность производственных процессов [59, 89, 151].

Можно констатировать, что значимость сервисных процессов с точки зрения управления качеством сопоставима со значимостью производственных процессов. В условиях современной экономики, характеризующейся высоким уровнем конкуренции и возрастающими требованиями потребителей, эффективная организация сервисного обслуживания является необходимым условием для обеспечения устойчивого конкурентного преимущества и достижения стратегических целей организации в области качества.

В этой главе мы проанализируем данные о процессе сервисного обслуживания автомобилей одного из известных отечественных брендов и покажем, как методика МССР (Modeling, Challenge, Calibration, Production) может использоваться для анализа процессов сервисной группы. Кроме того, для того чтобы проиллюстрировать идею данных в формате CML/ CLM, описывающих процесс в терминах нескольких характеристик (см. главу 1, раздел 1.4), в этой главе мы рассмотрим в качестве показателей сервисного процесса две его характеристики: 1) число обработанных за день заявок на гарантийное обслуживание и 2) суммарную стоимость гарантийных услуг, оказанных тем или иным сервисным центром.

Однако, прежде чем переходить к алгоритмической обработке данных,

отметим специфику процесса сервисного обслуживания, отличающую этот процесс от рассмотренного выше процесса производства (см. главу 4).

5.1. Сервисное обслуживание как процесс потребления услуги

Инициатором, активизирующим деятельность автомобильных сервисных центров, является не производитель, а потребитель, то есть, не владелец автомобильного бренда, а сообщество владельцев автомобилей. Именно они, сталкиваясь с определёнными проблемами эксплуатации, ищут решения, делятся опытом и, в конечном итоге, формируют спрос на конкретные услуги, прописанные в гарантийном регламенте [34, 96, 105]. Этот спрос, в свою очередь, подталкивает сервисные центры к развитию, внедрению новых технологий и повышению квалификации персонала.

Сообщества владельцев автомобилей, как правило, аккумулируют всю информацию о типичных неисправностях, слабых местах моделей и способах их устранения. Эта информация становится важным ресурсом как для самих владельцев, так и для сервисных центров, которые, анализируя данные, полученные от сообществ, могут эффективно диагностировать проблемы и предлагать оптимальные решения, тем самым повышая лояльность клиентов. Более того, влияние сообществ распространяется не только на техническое обслуживание, но и на выбор запчастей и расходных материалов. Владельцы автомобилей, объединённые общими интересами, обмениваются мнениями о качестве различных брендов, проводят собственные тесты и исследования, и эта информация помогает им делать осознанный выбор и, в конечном итоге, оказывает влияние на ценовую политику и ассортимент сервисных центров. Именно потребитель, в лице сообщества владельцев автомобилей, является основной движущей силой развития автомобильного сервиса, определяя его направления и стандарты качества [34, 96, 102, 105].

5.1.1. Специфика потребления сервисных услуг

Сервисное обслуживание автомобиля — это целый комплекс мероприятий,

направленных на поддержание работоспособности и безопасности, причём, в отличие от производства, где процессы чётко регламентированы, сервисное обслуживание, как потребление услуги, обладает гораздо большей гибкостью и адаптивностью. Автомобильный бренд устанавливает общие стандарты и рекомендации, но фактическое предоставление услуги зависит от множества факторов [33, 34, 99, 102, 105, 107, 113]. К ним относятся индивидуальные особенности автомобиля, условия эксплуатации, квалификация персонала и даже личные предпочтения клиента. Жёсткая регламентация со стороны бренда могла бы ограничить возможности адаптации к этим переменным, что негативно сказалось бы на качестве обслуживания⁵¹.

Тем не менее, отсутствие жёсткой регламентации не означает, что сервисное обслуживание должно быть стихийным процессом. Управление процессом сервисного обслуживания необходимо для обеспечения высокого уровня удовлетворённости клиентов и поддержания репутации бренда. Это включает в себя обучение и сертификацию персонала, внедрение систем контроля качества, сбор и анализ обратной связи от клиентов, а также — сбор и анализ данных о деятельности отдельных сервисных центров. Важно создать прозрачную и понятную для клиента систему, где он будет уверен в качестве предоставляемых услуг, а качественное управление сервисным обслуживанием — это достижение баланса между гибкостью, необходимой для удовлетворения потребностей, и стандартизацией, обеспечивающей стабильное качество и надёжность [33, 34, 96, 99, 102, 105, 107].

5.1.2. Временные паттерны в процессах потребления

Процессы потребления в широком смысле этого слова (в частности, это относится и к процессам потребления сервисных услуг) принципиально отличаются от процессов производства, хотя они и взаимосвязаны в экономическом

⁵¹ Например, клиент может предпочесть определённые запчасти или способы ремонта, не предусмотренные стандартными процедурами.

цикле. Производство ориентировано на создание новых благ и услуг, в то время как потребление связано с использованием и усвоением этих благ для удовлетворения потребностей. Можно сказать, что основное отличие заключается в направленности: производство направлено вовне, на создание продукта, а потребление — внутрь, на удовлетворение нужд индивида или общества. В отличие от производственных процессов, где часто акцент делается на оптимизации ресурсов и технологиях, потребление прежде всего связано с психологией, а также с социальными нормами и культурными особенностями. Производство стремится к стандартизации и массовости, а потребление подвержено индивидуальным предпочтениям и вариативности.

Особое значение при изучении процессов потребления приобретают повторяющиеся временные факторы и сезонность. Потребление многих товаров и услуг подвержено циклическим колебаниям, связанным с временем года, праздниками, отпусками и другими регулярными событиями⁵². Учёт таких сезонных колебаний позволяет предприятиям (в первую очередь, нас, безусловно, интересуют сервисные предприятия, так как именно они являются главными действующими лицами настоящей главы) адаптировать свои производственные и маркетинговые стратегии для более эффективного удовлетворения потребительского спроса [34, 96]. Игнорирование временных факторов и сезонности может привести к неэффективному распределению ресурсов и упущенной выгоде [34]. Предприятиям требуется всесторонне использовать данные компании и потребления сервисной услуги, активно задействовать анализ временных рядов и сезонных колебаний, так как именно эти инструменты позволяют прогнозировать потребительский спрос, оптимизировать запасы и планировать маркетинговые кампании с учётом особенностей конкретного временного периода [198]. Специфика процессов потребления такова, что

⁵² Например, спрос на тёплую одежду возрастает в зимние месяцы, а на прохладительные напитки — летом. Потребление подарочной продукции резко возрастает в период новогодних и рождественских праздников. И так далее.

она требует комплексного подхода, учитывающего не только экономические, но и социальные, психологические и временные факторы, что делает их изучение критически важным для эффективного функционирования экономики предприятия.

5.1.3. Роль сервисного обслуживания в общей оценке качества продукта

Сервисное обслуживание играет ключевую роль в общей оценке качества автомобиля, формируя у потребителя восприятие не только самого продукта, но и бренда в целом [34, 96, 105]. Современный покупатель рассматривает автомобиль не просто как средство передвижения, а как сложный механизм, требующий регулярного ухода и профессиональной поддержки, и именно качество сервиса становится определяющим фактором удовлетворённости владельца.

Первое впечатление о бренде, полученное в дилерском центре, часто переносится на восприятие автомобиля, поэтому внимательное отношение, профессионализм персонала, оперативность решения проблем — все это создаёт положительный образ компании. И напротив: некомпетентность и игнорирование потребностей клиента вполне могут испортить даже самое благоприятное впечатление от самого автомобиля. Более того, качественный сервис повышает лояльность клиентов к бренду [34, 96, 97, 99]. Владельцы, уверенные в том, что им будет оказана квалифицированная помощь, склонны оставаться приверженцами марки и рекомендовать её своим знакомым, и можно заключить, что сервисное обслуживание — это неотъемлемая часть процесса покупки и эксплуатации автомобиля, а качество сервиса напрямую влияет на формирование отношения потребителей к бренду, определяет долговечность и надёжность автомобиля, и способствует существенному повышению лояльности клиентов.

5.2. Данные о работе сети сервисных центров

В мире автомобильного бизнеса, где конкуренция так высока, а ожидания клиентов постоянно растут, своевременное отслеживание работы сервисных центров становится критически важным для успеха и устойчивости бренда. Когда речь идёт о сети, насчитывающей сотни сервисных центров (а в рамках настоящей диссертационной работы мы рассматриваем сервисную сеть одного из ведущих отечественных автопроизводителей, состоящую из 336 сервисных центров), задача мониторинга усложняется многократно, однако от этого её значимость только возрастает [104]. Централизованный контроль позволяет автомобильному бренду поддерживать качество обслуживания во всех точках контакта с клиентом [242]. Это создаёт положительный имидж компании, а также обеспечивает лояльность и способствует привлечению новых клиентов. Регулярный мониторинг помогает выявлять отклонения, оперативно реагировать на проблемы и предотвращать их эскалацию.

Несмотря на сложность задачи, современные технологии предлагают эффективные решения для мониторинга работы сервисных центров, в первую очередь — инструменты анализа данных [2–5, 7–9, 12, 17, 49, 52, 65, 173], которые позволяют оперативно получать информацию о ключевых показателях работы каждого сервисного центра, что, в свою очередь, даёт возможность принимать обоснованные управленческие решения и повышать эффективность работы всей сети.

5.2.1. Предварительное преобразование данных к формату ML/ LM

В отличие от предыдущей главы, где в нашем распоряжении с самого начала были данные о производительности работников, представленные в формате ML, на этот раз в качестве исходных сведений о работе сервисных центров мы имеем «сырые», неподготовленные данные (см. таблицу 5.1).

Таблица 5.1 — Исходные данные о работе сервисных центров

	service_id	service_name	model	defect_id	...	causer
0	44360	Самара	11183	2915004024000	...	51000

	service_id	service_name	model	defect_id	...	causer
1	44360	Самара	11183	6106082083000	...	30000
2	44360	Самара	11183	5402334000000	...	0
3	44360	Самара	11183	1001157086000	...	30000
...	...	...	...	...	...	...
1230473	44588	Тольятти	21703	8118020000000	...	51000
1230474	44588	Тольятти	21703	1703055004000	...	30000

Эти данные содержат 1230475 записей о заявках, поступивших в сервисные центры одного из ведущих отечественных автопроизводителей, за период с 2023-01-01 по 2024-12-30, которые описаны при помощи 18 признаков:

1. service_id — уникальный идентификатор заявки,
2. service_name — населённый пункт, в котором расположен сервисный центр (не сам сервисный центр),
3. model — модель автомобиля,
4. component — компонент автомобиля, который требует сервисного обслуживания по гарантии,
5. product_id — уникальный идентификатор услуги,
6. date_in — дата поступления заявки,
7. date_out — дата исполнения заявки,
8. mileage — суммарный пробег автомобиля на момент подачи заявки на обслуживание,
9. warranty — идентификатор гарантийного случая,
10. labor_cost — оплата труда по устранению дефекта,
11. operating_time — нормативное время, необходимое на устранение дефекта,
12. service_price — стоимость обслуживания,
13. part_price — стоимость запчастей, использованных для устранения дефекта,
14. materials_price — стоимость расходных материалов, использованных для устранения дефекта,
15. general_price — общая стоимость устранения дефекта,

- 16. `defect_id` — уникальный идентификатор дефекта,
- 17. `defect_name` — описание дефекта,
- 18. `causer` — подразделение, ответственное за возникновение дефекта.

Для наших целей (а мы хотим научиться отслеживать аномальные состояния сервисных центров для проведения их оперативного аудита) эти сведения являются избыточными⁵³. Мы производим агрегацию данных, используя для этого лишь две интересующие нас характеристики: 1) общее количество заявок на гарантийное обслуживание, принятых тем или иным сервисным центром в течение рабочего дня, и 2) общую стоимость заявок на гарантийное обслуживание, принятых тем или иным сервисным центром в течение рабочего дня.

Начнём с такой характеристики сервисного процесса как общее число заявок. Если мы произведём локализацию имеющегося у нас датафрейма по двум условиям, а именно: потребуем, чтобы ID сервисного центра принимал какое-то конкретное значение, и чтобы дата поступления заявки принимала какое-то конкретное значение, то мы получим датафрейм, в котором будет ровно столько записей, сколько принял в этот день этот сервисный центр. А если мы проделаем эту процедуру со всеми возможными парами сервисных центров и дат, то мы получим полные сведения о работе всех сервисных центров в течение всего промежутка наблюдений, и это будет именно та характеристика сервисного процесса, которая нам нужна.

При помощи метода `DafaFrame` библиотеки `pandas` (см. приложение С, раздел С.1) мы формируем пустой датафрейм `ML`, и заполняем его в двойном цикле по локациям (в качестве которых выступают сервисные центры из сырого набора данных, представленного выше, см. таблицу 5.1) и по моментам времени (в качестве которых выступают даты поступления заявок из сырого

⁵³ Что, разумеется, никак не уменьшает их значимости. Для решения каких-то других задач могут потребоваться данные о конкретной модели автомобиля или о распространённости конкретного вида дефектов. Эти данные являются избыточными для нашей задачи.

набора данных).

```
ML_order = pd.DataFrame()
for loc in df['service_id'].unique():
    comdition_loc = df['service_id'] = loc
    n_order_list = []
    for mom in df['date_in'].unique():
        comdition_mom = df['date_in'] = mom
        n_order = len(df.loc[condition_mom & condition_loc])
        n_order_list.append(n_order)
    ML_order['loc'] = n_order_list
```

Последующий перевод данных из формата ML в формат LM осуществляется при помощи операции транспонирования. Он не представляет труда, и мы не будем его здесь описывать.

Далее — перейдём ко второй интересующей нас характеристике: к суммарной дневной стоимости работ по гарантийному обслуживанию. Если мы производим локализацию исходного датафрейма на конкретную дату и конкретный сервисный центр, то получаем редуцированные сведения, среди которых интерес для нас представляет один столбец: **general_price**. Сумма значений в этом столбце даст общую сумму гарантийных затрат, которые понёс конкретный сервисный центр в течение конкретного рабочего дня.

```
ML_price = pd.DataFrame()
for loc in df['service_id'].unique():
    comdition_loc = df['service_id'] = loc
    price_list = []
    for mom in df['date_in'].unique():
        comdition_mom = df['date_in'] = mom
        D = df.loc[condition_mom & condition_loc]
        prise = D['general_price'].sum()
        price_list.append(prise)
    ML_price['loc'] = price_list
```

Теперь у нас есть два датафрейма в формате ML и два датафрейма в фор-

мате LM, которые мы выводим в соответствующие внешние файлы при помощи метода `to_csv` библиотеки `pandas` (см. приложение С, раздел С.1), и, на первый взгляд, у нас все готово для того, чтобы анализировать эти данные при помощи методики МССР (Modeling, Challenge, Calibration, Production), но это только на первый взгляд. Дело в том, что локации, сведения о которых предоставил нам владелец процесса, настолько вариативны, что требуют ещё одного предварительного анализа, который позволит нам избавиться от белого шума в данных и сосредоточить усилия на анализе центров, действительно значимых для процесса сервисного обслуживания.

5.2.2. Парето-анализ сервисных центров

Напомним, что принцип Парето, известный также как правило 80/20, является универсальным наблюдением, утверждающим, что примерно 80% результатов происходят от 20% усилий или причин [154]. Этот принцип, названный в честь итальянского экономиста Вильфредо Парето, впервые заметившего, что 80% земли в Италии принадлежат 20% населения, применяется в самых разных областях. В бизнесе принцип Парето указывает на то, что 80% прибыли компании часто генерируются 20% её клиентов или продуктов, и это понимание позволяет компаниям сосредоточить свои ресурсы и усилия на наиболее прибыльных областях, оптимизируя свою деятельность и повышая эффективность (например, выявление ключевых клиентов и удовлетворение их потребностей может значительно увеличить прибыль). В управлении временем принцип Парето подчёркивает, что 80% нашей продуктивности достигается за 20% времени, потраченного на работу, а это, в свою очередь, означает, что сосредоточение на наиболее важных задачах и устранение отвлекающих факторов может значительно повысить эффективность использования времени. И так далее — принцип Парето является универсальным принципом, который проявляет себя в самых разнообразных сферах человеческой деятельности.

Разумеется, принцип Парето не является строгим законом, и точные со-

отношения могут варьироваться, однако его основная идея всегда остаётся актуальной: небольшое количество факторов часто оказывает на процесс непропорционально большое влияние. В нашем случае под процессом мы понимаем процесс сервисного обслуживания автомобилей, и даже беглый взгляд на полученные выше временные ряды в формате ML показывает существенную неравномерность в распределении результатов различных сервисных центров (см. рисунок 5.1).

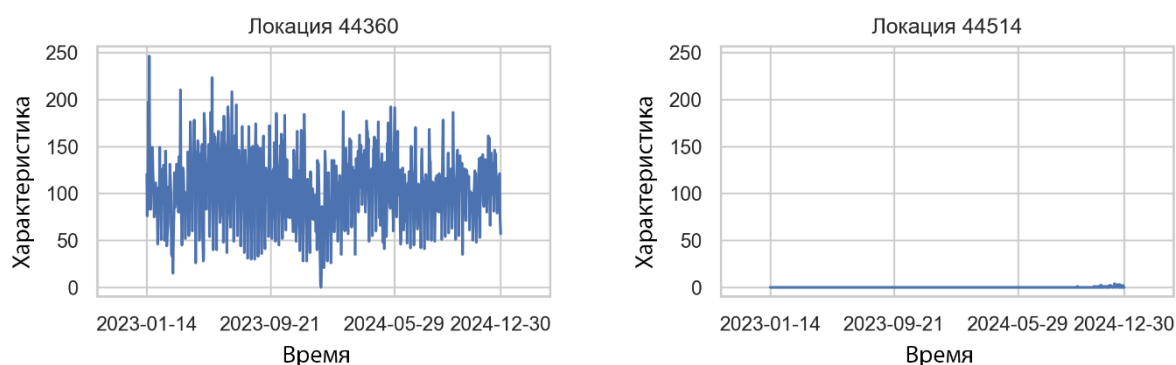


Рисунок 5.1 — Неравномерность загрузки сервисных центров

Здесь по горизонтали изменяется момент наблюдения в той или иной локации, а по вертикали — число полученных в соответствующий момент времени заявок на устранение того или иного дефекта. При этом локация 44360 демонстрирует высокую загрузку (на уровне примерно 100 заявок в день), а в локации 44514 загрузка практически отсутствует (можно считать, что это технический ноль). Много ли локаций со столь низкой загрузкой в имеющемся у нас наборе из 336 сервисных центров? Ответ — да, их много, правильнее сказать, их большинство, и совершенно очевидно, что если попытаться строить прогнозирующие модели на локациях, которые, по сути, возвращают белый шум, а не содержательные данные, то эффективность таких моделей будет крайне низка⁵⁴.

⁵⁴ Здесь мы уже не в первый раз в рамках этой диссертационной работы сталкиваемся с одним из неформальных правил машинного обучения: «Мусор на входе — мусор на выходе» (англ. «Trash in — trash out»).

В анализе данных и машинном обучении одним из ключевых этапов является предобработка данных [2–5, 7–9, 65, 210, 218, 231], причём, это критически важный процесс, поскольку качество и пригодность данных напрямую влияют на результаты моделирования и прогнозирования. Данные, содержащие только белый шум⁵⁵, не несут полезной информации и могут серьёзно исказить результаты анализа [4, 5, 7, 35, 65]. С другой стороны, данные, отражающие истинное поведение изучаемого процесса, позволяют модели выявлять значимые зависимости и закономерности. Эти данные содержат информацию о причинно-следственных связях, трендах и циклах, которые характеризуют процесс, и концентрация на таких данных позволяет построить более точные и надёжные модели, способные делать более качественные прогнозы и принимать обоснованные решения. Удаление белого шума позволяет упростить модель, снизить вычислительные затраты и повысить её обобщающую способность [2–5, 7–9, 13, 60, 65]. Очищенные данные позволяют алгоритмам машинного обучения быстрее сходиться к оптимальному решению и улучшают интерпретируемость модели [49], а фильтрация данных от белого шума и фокусировка на данных, отражающих суть явления, является фундаментальным принципом успешного анализа данных и машинного обучения [23, 48, 49]. Это позволяет создавать более надёжные, точные и полезные модели, способные приносить реальную пользу.

Извлечение из всего спектра локаций тех из них, которые в сумме обеспечивают 80% заявок на гарантийное обслуживание, не составляет большого труда и реализуется за счёт следующего алгоритма. Сначала считываем в датафрейм `d` данные в формате `ML` о числе полученных сервисными центрами заявок. Затем, при помощи суммирования и упорядочивания сумм, получаем серию `pandas`, в которой в качестве индексов выступают локации, а в качестве

⁵⁵ Белый шум — это случайные данные, не имеющие какой-либо закономерности или корреляции с изучаемым процессом. Включение такого шума в обучающий набор данных может привести к переобучению модели, когда она начинает запоминать случайные флуктуации, а не выявлять реальные закономерности.

значений — суммарное количество заявок, полученных в соответствующей локации за весь период наблюдений:

```
s = d.sum().sort_values(ascending=False)
```

Проводим несложный цикл `while`, для получения накопленной суммы до достижения условия выхода (оно состоит в том, что накопленная сумма окажется больше 80% от общего числа заявок за изучаемый период):

```
S = 0
k = 0
while S < 0.8*s.sum():
    S = S + s.iloc[k]
    k = k + 1
```

И, наконец, выводим в датафрейм `d_reduce` столбцы датафрейма `d`, обеспечивающие нужные нам 80% заявок:

```
d_reduce = d[s.index[:k]]
```

Любопытно, что в результате исполнения этого алгоритма оказывается, что 80% заявок обеспечивают 69 локаций, причём 69 — это 20.54% от общего числа локаций, число которых, напомним, равно 336. Удивительным образом, пропорция Парето проявилась и в наших процессах, лишний раз подчёркивая свой универсальный характер [154].

Так как мы изучаем процесс сервисного обслуживания при помощи двух характеристик: 1) числа заявок и 2) суммы затрат, — то мы могли бы провести аналогичный Парето-анализ данных и по второй характеристике, но мы поступим по-другому. Мы просто редуцируем данные о суммарных затратах по помощи того же набора локаций, который мы получили при Парето-анализе первой характеристики. Это позволит нам в дальнейшем оперировать полностью согласованными данными (тем более, что эти две характеристики все равно сильно коррелируют друг с другом).

5.2.3. Описание данных

Итак, теперь наши данные представлены в формате ML, в качестве моментов времени используются рабочие дни, в качестве локаций — сервисные центры одного из ведущих отечественных автопроизводителей (см. таблицу 5.2 и 5.3). В качестве изучаемых характеристик сервисного процесса выступают 1) число оформленных в течение рабочего дня заявок на гарантийное обслуживание автомобиля, 2) суммарные затраты на гарантийное обслуживание в течение рабочего дня.

Таблица 5.2 — Данные о числе заявок в формате ML

	Локация 1	Локация 2	Локация 3	...	Локация 69
Момент 1	120	86	56	...	0
Момент 2	76	48	36	...	0
Момент 3	106	63	64	...	14
Момент 4	97	43	63	...	13
...	...	...	...	...	...
Момент 714	78	10	13	...	0
Момент 715	57	25	0	...	0

Таблица 5.3 — Данные о суммарных затратах в формате ML

	Локация 1	Локация 2	Локация 3	...	Локация 69
Момент 1	119472	70610	66718	...	0
Момент 2	73799	34424	34068	...	0
Момент 3	147732	41715	93987	...	18152
Момент 4	275025	40197	60506	...	7914
...	...	...	...	...	...
Момент 714	132139	8773	17372	...	0
Момент 715	66191	28694	0	...	0

Транспонирование таблиц 5. 2 и 5.3 переводит данные из формата ML в формат LM (см. таблицы 5.4 и 5.5).

Таблица 5.4 — Данные о числе заявок в формате LM

	Момент 1	Момент 2	Момент 3	...	Момент 715
Локация 1	120	76	106	...	57
Локация 2	86	48	63	...	25

	Момент 1	Момент 2	Момент 3	...	Момент 715
Локация 3	56	36	64	...	0
Локация 4	41	36	39	...	4
...	...	...	...	...	...
Локация 68	6	13	9	...	0
Локация 69	0	0	14	...	0

Таблица 5.5 — Данные о суммарных затратах в формате LM

	Момент 1	Момент 2	Момент 3	...	Момент 715
Локация 1	119472	73799	147732		66191
Локация 2	70610	34424	41715		28694
Локация 3	66718	34068	93987		0
Локация 4	27116	27615	25252		2690
...					
Локация 68	4410	19485	8142		0
Локация 69	0	0	18152		0

Кроме того, мы используем две вспомогательные таблицы, связывающие наименование момента (в качестве которого выступает дата) с его кодом и наименование локации (в качестве которого выступает пятизначный номер сервисного центра) с её кодом (см. таблицы 5.6 и 5.7). Эти таблицы используются для интерпретации результатов, полученных на заключительном этапе методики МССР (Modeling, Challenge, Calibration, Production), когда формируются календари и рейтинги аномалий (см. приложение D, раздел D.2) для того, чтобы соотнести значение переменной с реальным значением момента или локации.

Таблица 5.6 — Коды моментов в процессе сервисного обслуживания

	Код	Момент
1	mom_1	2023-01-14
2	mom_2	2023-01-15
3	mom_3	2023-01-16
...	...	...
715	mom_715	2024-12-30

Таблица 5.7 — Коды локаций в процессе сервисного обслуживания

	Код	Локация
1	loc_1	44360
2	loc_2	11785
3	loc_3	44560
...	...	
69	loc_69	46064

5.3. Детекция аномалий по времени в процессе сервисного обслуживания для числа оформленных заявок

В этом разделе мы проведём провести весь цикл методики МССР (Modeling, Challenge, Calibration, Production), начиная с моделирования процесса по времени, заканчивая построением календаря аномалий относительно первой из заявленных нами к изучению характеристик.

5.3.1. Моделирование по времени процесса сервисного обслуживания для числа оформленных заявок

Так же как и выше (см. главу 4, разделы 4.3, 4.4), для моделирования мы используем пул из семи моделей:

8. `LinearRegression` (линейная регрессия);
9. `DecisionTreeRegressor` (регрессия на решающих деревьях);
10. `RandomForestRegressor` (регрессия на случайных лесах);
11. `GradientBoostingRegressor` (градиентный бустинг над случайными лесами);
12. `SVR` (регрессия на методе опорных векторов);
13. `KNeighborsRegressor` (регрессия на ближайших соседях);
14. `MLPRegressor` (нейронная сеть многослойного перцептрона).

Функция моделирования по времени описана нами в приложении С.1. Используя её с максимальным погружением по авторегрессии, равным 10, получаем следующие результаты (см. рисунок 5.2).

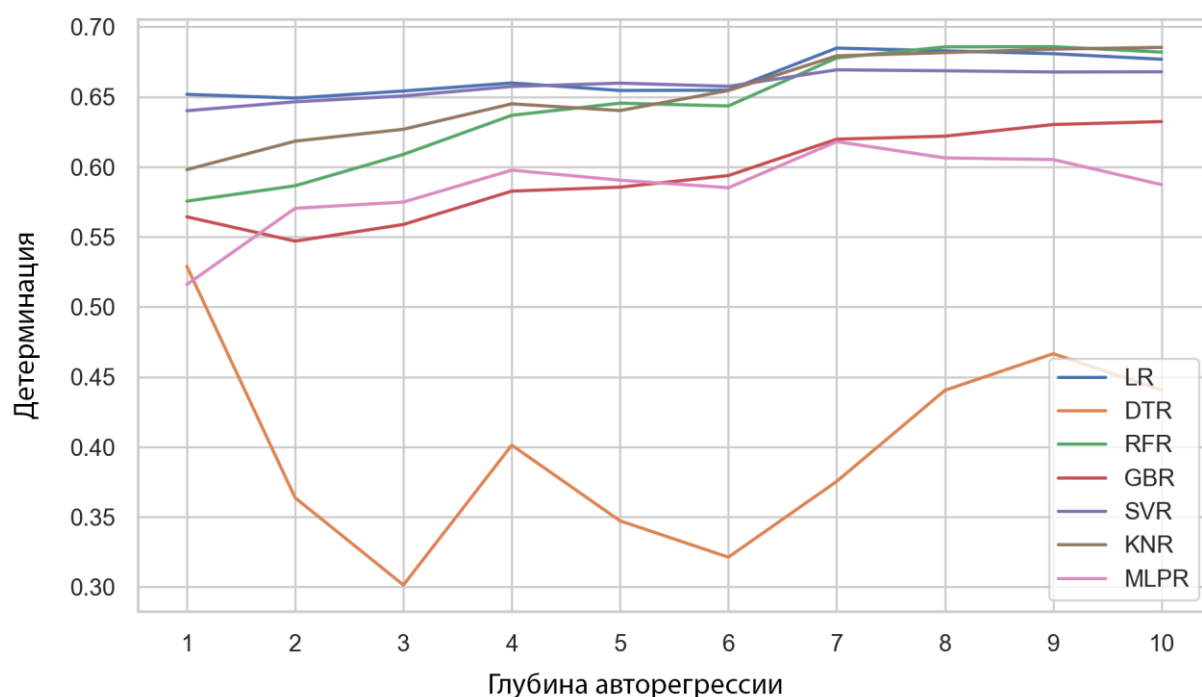


Рисунок 5.2 — Сравнительный анализ схем моделирования по времени в процессе сервисного обслуживания для числа обработанных заявок

Наилучший результат демонстрирует модель `LinearRegression` с глубиной авторегрессии 7. Именно её мы будем использовать далее, для получения калибровочной границы возможных отклонений и проведения этапа пилотной эксплуатации.

5.3.2. Калибровка отклонений по времени в процессе сервисного обслуживания для числа оформленных заявок

Так же, как и при исследовании процесса производства (см. выше, главу 4, раздел 4.3.2), при исследовании сервисных процессов мы, к сожалению, не располагаем экспертным мнением относительно нормы или аномалии того или иного состояния⁵⁶. Поэтому для определения критической величины отклонения реального значения характеристики от прогноза мы используем метод локтевой точки (см. рисунок 5.3).

⁵⁶ Это неудивительно. Наши сырые данные содержат более миллиона записей, и ни один эксперт физически не справится с таким объёмом информации.

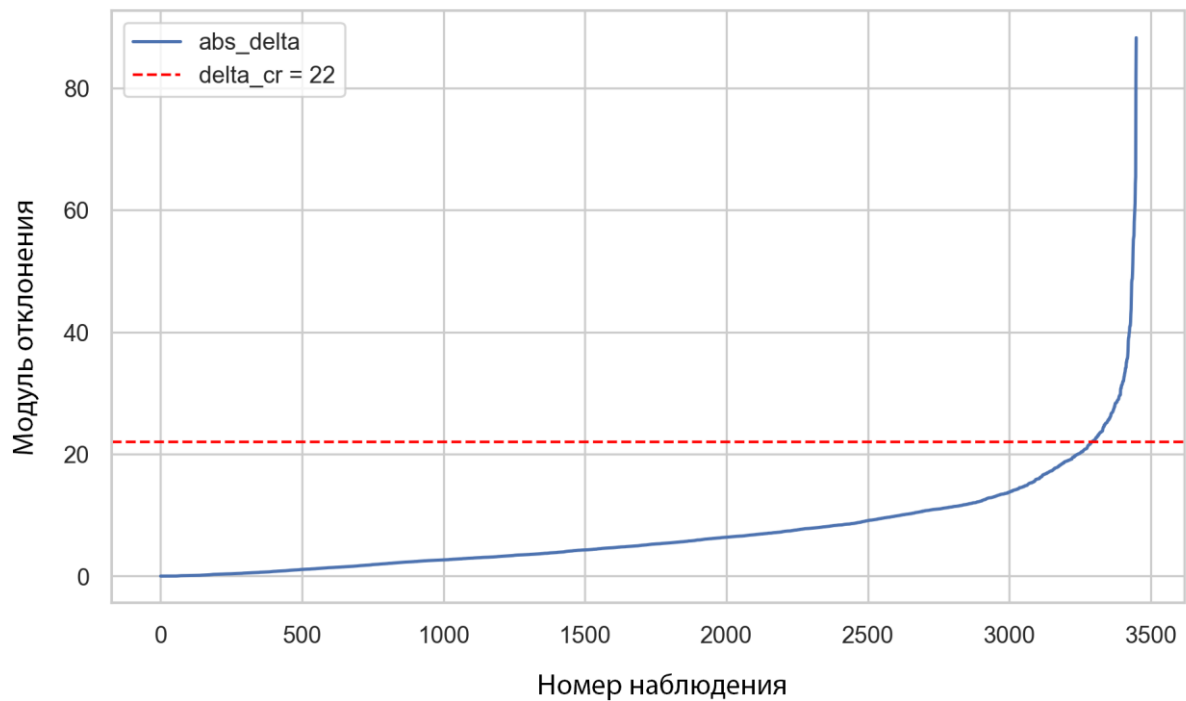


Рисунок 5.3 — Калибровочная кривая отклонений по времени в процессе сервисного обслуживания для числа обработанных заявок

Локтевая точка калибровочной кривой приходится на величину 22. Именно её мы будем использовать в дальнейшем на этапе плотной эксплуатации, считая, что любое отклонение модельного значения от реального, которое превышает $\text{delta\_cr} = 22$, сигнализирует об аномальности.

5.3.3. Пилотная эксплуатация детектора аномалий по времени в процессе сервисного обслуживания для числа оформленных заявок

Пилотная эксплуатация детектора аномалий осуществляется на последних 415 моментах времени. Мы применяем функцию `production_by_moment` (см. Приложение, раздел С.4), передавая ей все необходимые сведения. Прежде всего, описываем наилучшую схему моделирования:

```
best_model = LinearRegression
autoregression_dive = 7
```

А кроме того — критическую величину отклонения Δ_{cr} :

```
delta_cr = 22
```

Наш детектор прогнозирует значения наблюдаемой характеристики, затем вычисляет дельту между спрогнозированным и истинным значением, и, если это какое-то из отклонений оказывается больше критического, он объявляет соответствующее состояние процесса аномальным. Например, в локации 12954 детектированы 15 аномалий (см. рисунок 5.4).



Рисунок 5.4 — Модель для числа заявок по времени в локации 12954



Рисунок 5.5 — Отклонение модели для числа заявок по времени в локации 12954

В локации 42663 аномалий ещё меньше: всего 3 (см. рисунок 5.6).



Рисунок 5.6 — Модель для числа заявок по времени в локации 42663

Отклонения модели от реальности в этой локации проиллюстрированы ниже (см. рисунок 5.7), и видно, что на этот раз они выражены крайне слабо, едва преодолевая нижнюю калибровочную границу.

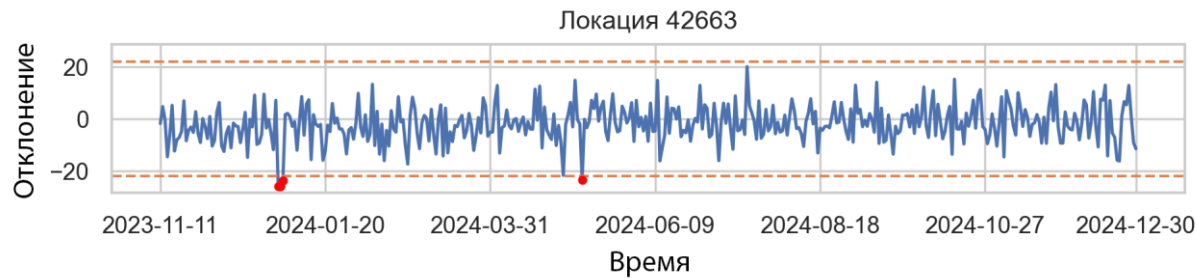


Рисунок 5.7 — Отклонение модели для числа заявок по времени в локации 42663

И так далее. Аналогичный анализ наш детектор проводит во всех имеющихся локациях. Но в контексте практического использования нас интересуют не даты аномалий в конкретной фиксированной локации, а, наоборот: локации аномалий в конкретный фиксированный момент времени. Имея такие сведения, лицо, ответственное за процесс сервисного обслуживания, сможет проводить своевременный аудит соответствующих локаций (и только их, не отвлекаясь на локации, детектированные как нормальные), причём в порядке убывания выраженности аномалий, которая характеризуется величиной отклонения. Сведения об аномалиях, зафиксированных в тот или иной момент времени мы называем календарём аномалий. Такой календарь мы формируем для всего этапа пилотной эксплуатации детектора (напомним, что он содержит 415 наблюдений), например, страница календаря за 2023-11-15 имеет следующий вид (см. рисунок 5.8):



Рисунок 5.8 — Рейтинг аномальных по времени локаций в процессе сервисного обслуживания для числа заявок в момент времени 2023-11-15

Для визуализации рейтинга мы используем психологически значимую величину 3σ , что позволяет судить о выраженности аномалий в процессах разной природы⁵⁷ (например, сравнивать календари процесса сервисного обслуживания с календарями рассмотренного выше, в главе 4, процесса производства, или с календарями маркетинговых процессов, которые мы будем рассматривать ниже, в главе 6). Представленный выше рейтинг говорит о том, что в момент времени 2023-11-15 аудит локаций следует проводить в последовательности от 4436 к 45070, начиная с наиболее выраженной аномалии, и заканчивая наименее выраженной.

А такова ситуация на 2024-03-08 (см. рисунок 5.9):

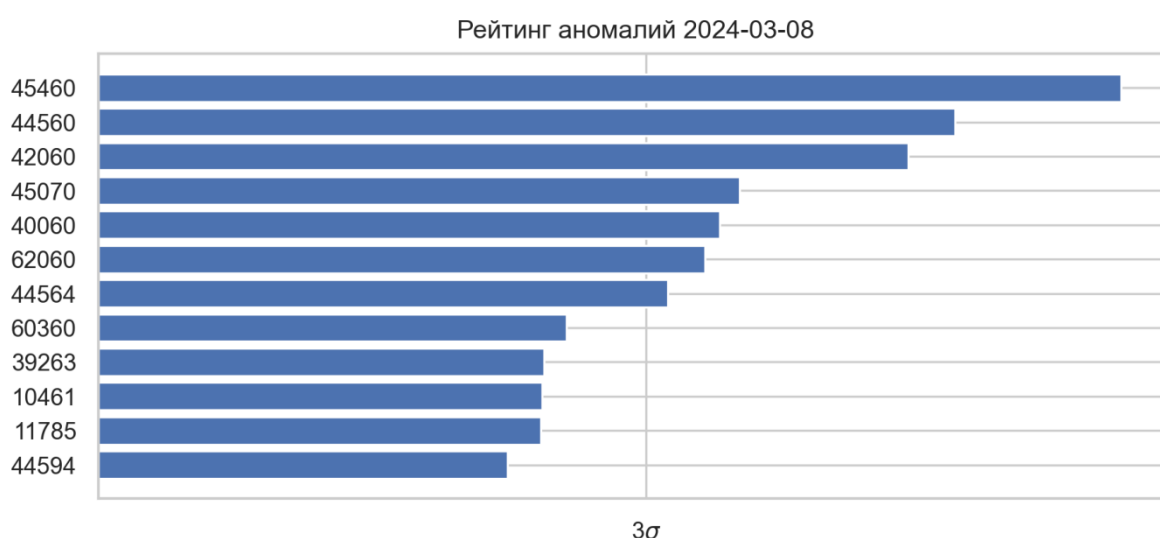


Рисунок 5.9 — Рейтинг аномальных по времени локаций в процессе сервисного обслуживания для числа заявок в момент времени 2024-03-08

Не стоит удивляться столь высокому числу аномалий. 8 марта — это вообще аномальная дата, а поскольку процесс сервисного обслуживания является, по сути, процессом потребления услуги, то и поведение потребителей в этот день оказывается соответствующим (видимо, мало кто из них занимается в этот день сервисным обслуживанием своего автомобиля).

Календарь аномалий по времени в процессе сервисного обслуживания

⁵⁷ Еще раз отметим, что для детекции аномалий мы используем правило локтевой точки (см. раздел 5.3.2) и никак не апеллируем к правилу 3σ . Поэтому в наших исследованиях аномальными могут признаваться состояния, отклонения которых гораздо меньше, чем 3σ .

по отношению к числу заявок, зафиксированных в течение последних десяти наблюдений, представлен в приложении D.2.1.

5.4. Детекция аномалий по локациям в процессе сервисного обслуживания для числа оформленных заявок

В этом разделе мы так же проводим полный цикл методики МССР (Modeling, Challenge, Calibration, Production), начиная с моделирования процесса по локациям, и заканчивая построением календаря аномалий.

5.4.1. Моделирование по локациям процесса сервисного обслуживания для числа оформленных заявок

Так же как и выше, когда мы моделировали процесс по времени (см. раздел 5.3.1), для моделирования по локациям мы используем пул из семи моделей:

1. `LinearRegression` (линейная регрессия);
2. `DecisionTreeRegressor` (регрессия на решающих деревьях);
3. `RandomForestRegressor` (регрессия на случайных лесах);
4. `GradientBoostingRegressor` (градиентный бустинг над случайными лесами);
5. `SVR` (регрессия на методе опорных векторов);
6. `KNeighborsRegressor` (регрессия на ближайших соседях);
7. `MLPRegressor` (нейронная сеть многослойного перцептрона).

Функция моделирования по локациям описана в приложении C.2. Используя её с максимальным погружением в историю наблюдений, равным 10, получаем следующие результаты (см. рисунок 5.10). Наилучший результат демонстрирует модель `LinearRegression` с глубиной истории, равной 7 (хотя, в отличие от предыдущей ситуации неожиданно хорошие результаты показывает модель многослойного перцептрона). Именно линейную регрессию с глубиной 7 мы будем использовать далее, для получения калибровочной границы возможных отклонений и проведения этапа пилотной эксплуатации.

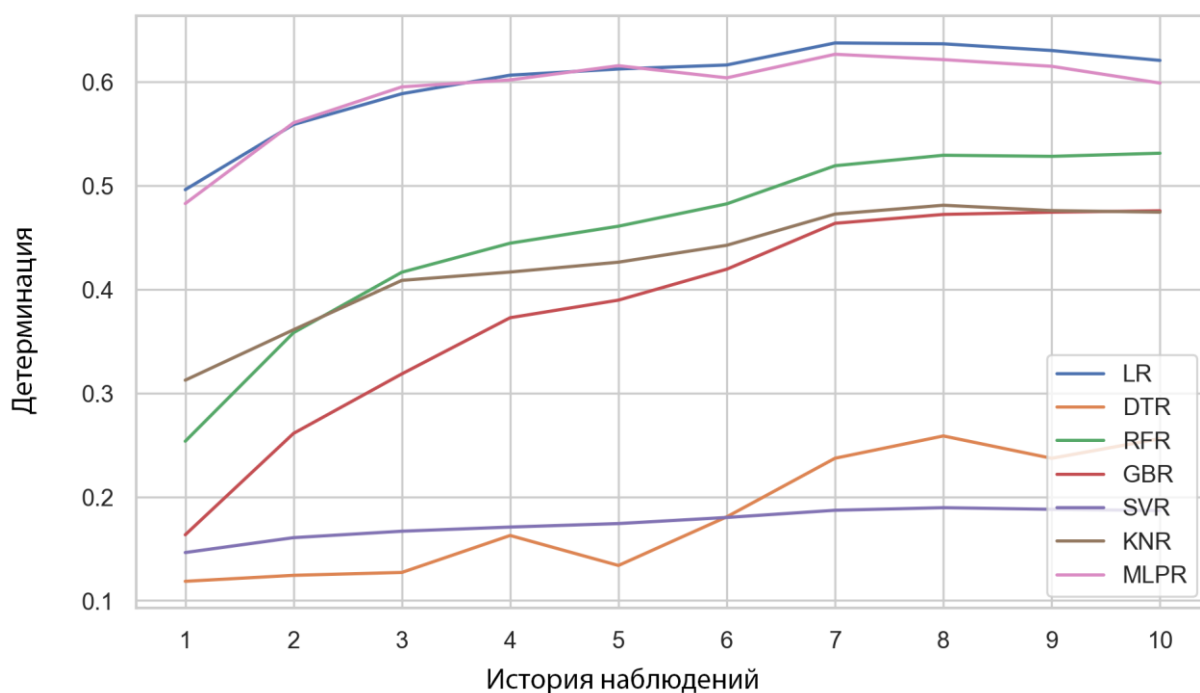


Рисунок 5.10 — Сравнительный анализ схем моделирования по локациям в процессе сервисного обслуживания для числа обработанных заявок

5.4.2. Калибровка отклонений по локациям в процессе сервисного обслуживания для числа оформленных заявок

В силу отсутствия экспертного мнения, мы применяем метод локтевой точки (см. выше, главу 4, раздел 4.3.2). Берём все отклонения модельных значений от реальных, которые получаются при наилучшей схеме моделирования, и упорядочиваем их абсолютные величины по возрастанию (см. рисунок 5.11). В результате возникает возрастающая кривая, уторая явно демонстрирует два возрастающих тренда: в начале она растет медленно, а в конце — быстро. Точка, которая отвечает смене тренда, является локтевой. Есть несколько алгоритмических методов, позволяющих вычислить положение локтевой точки: метод наибольшего удаления, метод наибольшего ускорения (метод второй производной), метод сплит-моделирования и т. д. В настоящей работе мы использовали метод сплит-моделирования (но для того, чтобы не отвлекаться от

основного сюжета диссертации, мы не излагаем его подробно, а лишь констатируем получающийся результат).

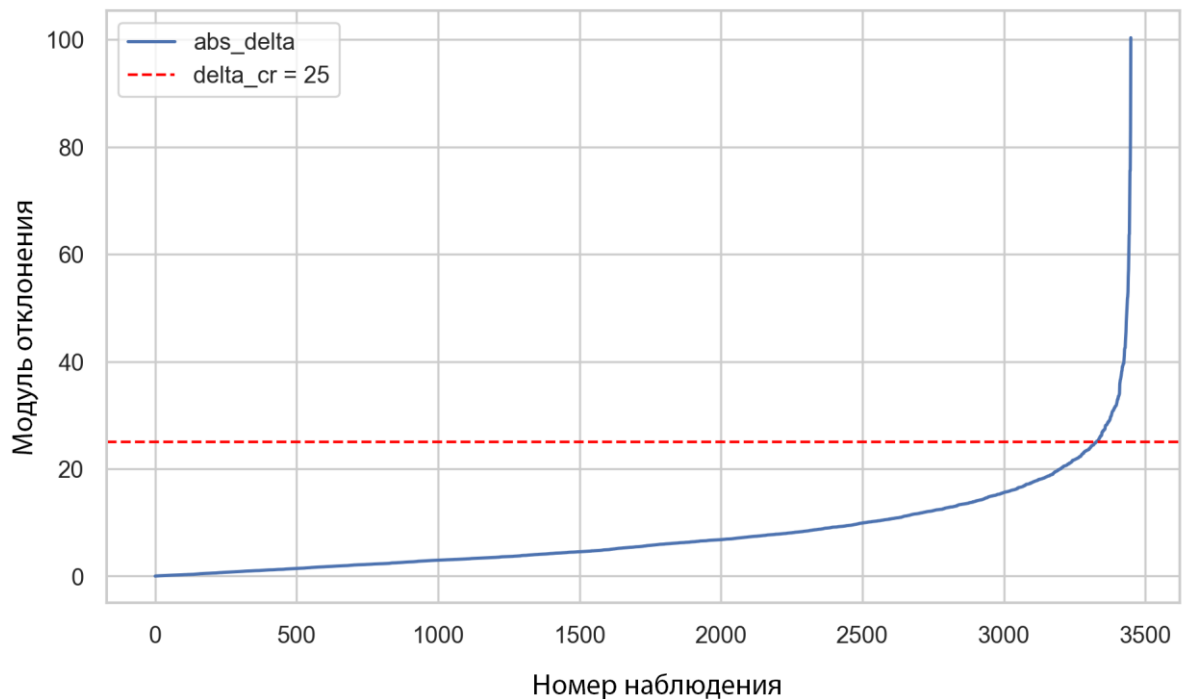


Рисунок 5.11 — Калибровочная кривая отклонений по локациям в процессе сервисного обслуживания для числа обработанных заявок

Локтевая точка калибровочной кривой приходится на величину 25. Именно её мы будем использовать в дальнейшем на этапе плотной эксплуатации, считая, что любое отклонение модельного значения от реального, которое превышает $\text{delta\_cr} = 25$, сигнализирует об аномальности.

5.4.3. Пилотная эксплуатация детектора аномалий по локациям в процессе сервисного обслуживания для числа оформленных заявок

Для этапа пилотной эксплуатации детектора аномалий используем последние 415 моментов времени. Мы применяем функцию детекции по локациям `production_by_location` (см. приложение С, раздел С.5) передавая ей все необходимые сведения. Лучшая схема моделирования:

```
best_model = LinearRegression
autoregression_dive = 7
```

Критическая величина отклонения Θ_{cr} :

teta_cr = 25

Детектор прогнозирует значения наблюдаемой характеристики, исходя из значений характеристики в других локациях. Чтобы посмотреть, насколько детекция по локациям, которую мы применяем сейчас, соответствует детекции по времени, которую мы применяли выше (см. раздел 5.3.3), мы в качестве примеров рассмотрим те же локации, что и в предыдущем разделе. В локации 12954 мы видим 18 аномалий (что близко к 15 аномалиям, зафиксированным детектором по времени). Но часть из них совпадает с аномалиями по времени (см. раздел 5.3.3), а другая часть — нет.

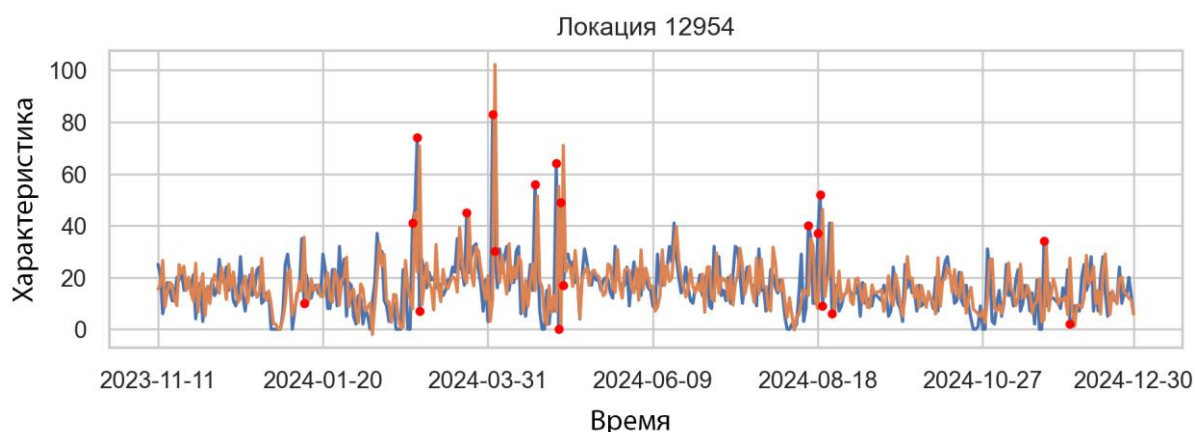


Рисунок 5.12 — Модель для числа заявок по локациям в локации 12954

Отклонения модельных значений от реально зафиксированных значений в этой локации таковы (см. рисунок 5.13).

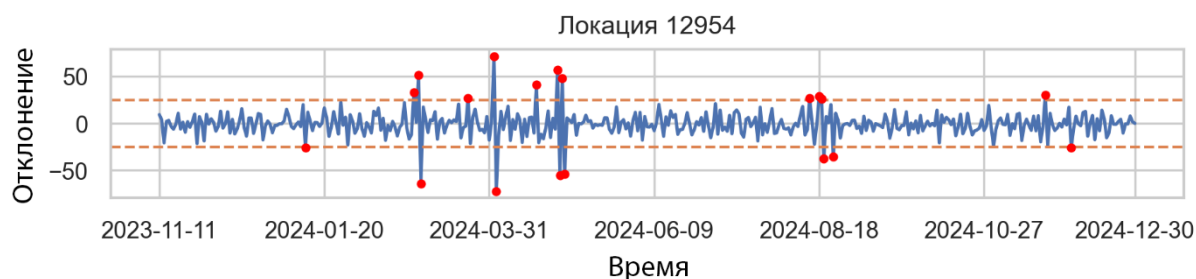


Рисунок 5.13 — Отклонение модели для числа заявок по локациям в локации 12954

В локации 42663 аномалий гораздо меньше: всего 5 (см. рисунок 5.14), что вполне согласуется с небольшим числом аномалий по времени, которые

были выявлены нам ранее (напомним, что в этой локации их было 3).



Рисунок 5.14 — Модель для числа заявок по локациям в локации 42663

Отклонения модели от реальности в этой локации проиллюстрированы ниже (см. рисунок 5.15).

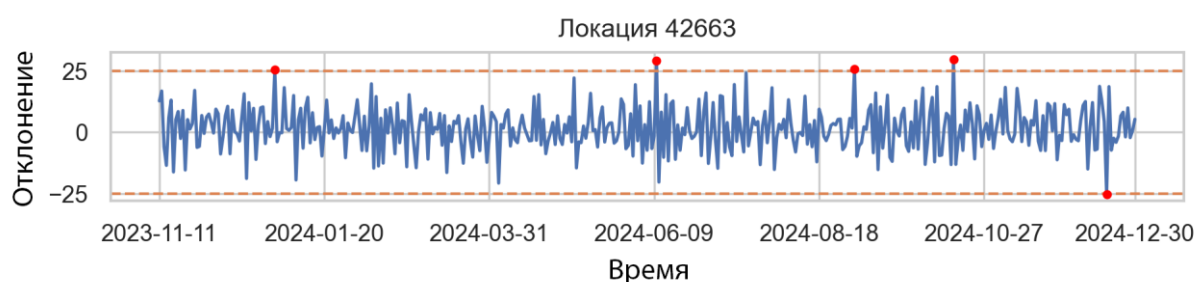


Рисунок 5.15 — Отклонение модели для числа заявок по локациям в локации 42663

Аналогичный анализ и соответствующую детекцию наш детектор проводит во всех имеющихся локациях. Для эффективного мониторинга локаций мы формируем календарь аномалий, то есть списки аномальных локаций для каждого момента времени, причём, эти списки упорядочены по убыванию, то есть, образуют рейтинг аномалий, которые фиксируются относительно наблюдаемой величины. Такой календарь мы формируем для всего этапа пилотной эксплуатации детектора (напомним, что он содержит 415 наблюдений, и, разумеется, мы не станем приводить его полностью), но, например, страница календаря за 2023-11-15 имеет следующий вид (см. рисунок 5.16):



Рисунок 5.16 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для числа заявок в момент времени 2023-11-15

Сравнивая эту ситуацию с предыдущими наблюдениями (см. рисунок 5.8), видим, что из 5 локаций, аномальных по локациям, 3 локации аномальны также и по времени, а именно — это локации 44360, 30563 и 10901, то есть, списки аномалий по времени и по локациям согласованы между собой. Но, тем не менее, они не являются совпадающими.

А такова ситуация на 2024-03-08 (см. рисунок 5.17):



Рисунок 5.17 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для числа заявок в момент времени 2024-03-08

Возникает естественный вопрос: почему аномалий по локациям 8 марта оказалось так мало? Ведь когда мы детектировали эту дату по времени, число зафиксированных аномалий было рекордно велико (см. для сравнения рисунок 5.9)? Причина именно в том, что для детекции по времени и по локациям мы используем совершенно разные принципы: в случае детекции по времени мы сравниваем состояния процесса с предыдущими состояниями той же локации, а в случае детекции по локациям — с другими локациями в тот же момент времени. Ответ на вопрос о столь разительном несоответствии результатов для двух типов детекции в том, что 8 марта все ведут себя примерно одинаково, и в эту дату локации мало отличаются друг от друга, что позволяет делать доста-

точно уверенный прогноз, ориентируясь на значения наблюдаемой характеристики (напомним, что в данном случае мы наблюдаем число заявок, обработанных в тот или иной момент времени в той или иной локации), которые были зафиксированы в других локациях. При этом, если говорить о прогнозе по времени, то поведение потребителей сервисной услуги 8 марта сильно отличается от их же поведения в предыдущие моменты времени, оно для них нетипично, и поэтому детекция по времени даёт нам множество аномалий.

Календарь аномалий по локациям в процессе сервисного обслуживания по отношению к числу заявок, зафиксированных в течение последних десяти наблюдений, представлен в приложении D.2.2.

5.5. Детекция аномалий по времени в процессе сервисного обслуживания для суммарной стоимости гарантийных работ

В этом разделе мы рассмотрим вторую из заявленных выше характеристик процесса сервисного обслуживания: накопленную за рабочий день сумму, которую необходимо потратить компании, предоставляющей сервисные услуги, на устранение дефектов по договору о гарантийном обслуживании. Данные, необходимые для исследования берём из таблицы 5.3. Это данные в формате ML: строки соответствуют моментам наблюдения, столбцы — локациям (то есть, центрам сервисного обслуживания), а в ячейках хранятся накопленные в течение рабочего дня суммы⁵⁸. Мы снова проводим весь цикл методики МССР (Modeling, Challenge, Calibration, Production), начиная с моделирования процесса по времени, и заканчивая построением календаря аномалий.

5.5.1. Моделирование по времени процесса сервисного обслуживания для

⁵⁸ Понятно, что эта характеристика коррелирует с числом оформленных в течение рабочего дня заявок, которые мы рассматривали выше (см. разделы 5.3 и 5.4): чем больше оформлено заявок, тем в большую сумму они обойдутся компании. Тем не менее, во-первых, это разные характеристики, а во-вторых, нас интересует, как связаны между собой аномалии, детектированные по этим, коррелирующим, но разным характеристикам.

суммарной стоимости работ

Так же как и выше, когда мы исследовали другие характеристики, для моделирования процесса сервисного обслуживания в контексте суммарной стоимости работ мы используем пул из семи моделей:

1. `LinearRegression` (линейная регрессия);
2. `DecisionTreeRegressor` (регрессия на решающих деревьях);
3. `RandomForestRegressor` (регрессия на случайных лесах);
4. `GradientBoostingRegressor` (градиентный бустинг над случайными лесами);
5. `SVR` (регрессия на методе опорных векторов);
6. `KNeighborsRegressor` (регрессия на ближайших соседях);
7. `MLPRegressor` (нейронная сеть многослойного перцептрона).

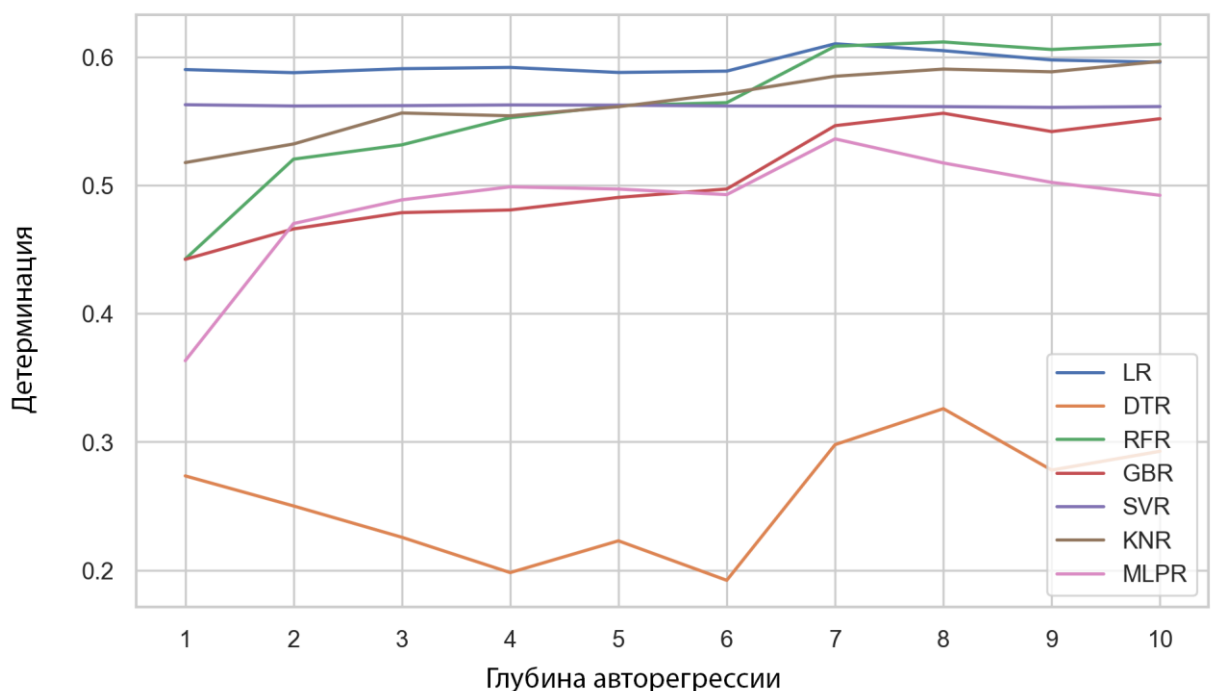


Рисунок 5.18 — Сравнительный анализ схем моделирования по времени в процессе сервисного обслуживания по отношению к суммарной стоимости работ

На рисунке 5.18 представлены результаты работы функции моделирования по времени `modeling_by_moment`, которая описана в приложении С.1. Видим,

что наилучший результат демонстрирует модель LinearRegression с глубиной авторегрессии 7. Именно её мы будем использовать далее, для получения калибровочной границы возможных отклонений, а затем для этапа пилотной эксплуатации.

5.5.2. Калибровка отклонений по времени в процессе сервисного обслуживания для суммарной стоимости работ

Для калибровки отклонений применяем метод локтевой точки.

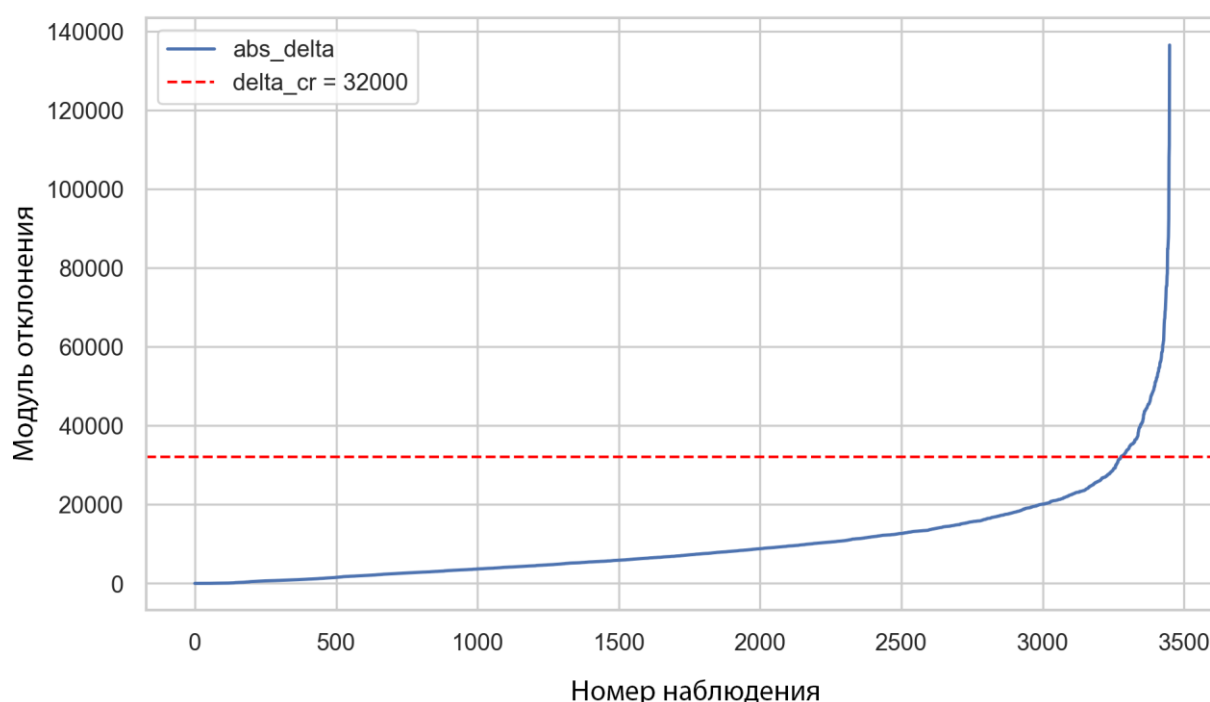


Рисунок 5.19 — Калибровочная кривая отклонений по времени в процессе сервисного обслуживания по отношению к суммарной стоимости работ

Видим, что локтевая точка калибровочной кривой (см. выше, главу 4, раздел 4.3.2) приходится на величину 32000. Именно её⁵⁹ мы будем использовать в дальнейшем на этапе плотной эксплуатации, считая, что любое отклонение модельного значения от реального, которое превышает `delta_cr =`

⁵⁹ Обратите внимание, сколь разительно отличается найденная величина от того, что получилось у нас в предыдущих разделах: на три порядка. Тем не менее, мы можем сравнивать между собой календари и рейтинги аномалий за счёт того, что при их построении мы используем масштаб относительно среднего квадратичного отклонения σ .

32000, сигнализирует об аномальности.

5.5.3. Пилотная эксплуатация детектора аномалий по времени в процессе сервисного обслуживания для суммарной стоимости работ

Этапа пилотной эксплуатации детектора аномалий, так же как и выше, мы проводим на последних 415 моментах времени. Мы применяем функцию `production_by_moment` (см. приложение С.4) передавая ей все необходимые сведения. Лучшая схема моделирования:

```
best_model = LinearRegression
autoregression_dive = 7
```

Критическая величина отклонения Δ_{cr} :

```
delta_cr = 32000
```

Выше мы применяли детектор аномалий по времени к другой характеристике и в качестве примера рассматривали локации 12954 и 42663. Рассмотрим их же для того, чтобы провести сравнение результатов. На этот раз в локации 12954 детектированы 8 аномалий (см. рисунок 5.20). В предыдущих разделах (см. рисунки 5.4 и 5.12), когда анализировалось число оформленных гарантийных заявок, было выявлено 15 и 18 аномалий (по времени и по локациям соответственно), что неплохо соотносится с нашим новым результатом.

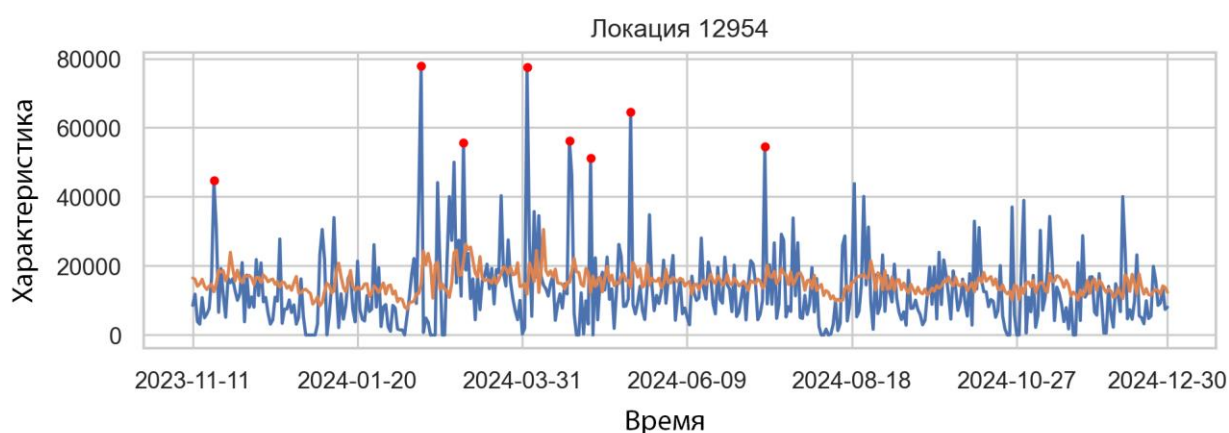


Рисунок 5.20 — Модель по времени для суммарной стоимости работ в локации 12954

Отклонения модели от реальности в этой локации проиллюстрированы ниже

(см. рисунок 5.21), причём, видно, что часть аномалий значительно превышает критически допустимый порог $\text{delta\_cr} = 32000$, хотя, если отдельно рассмотреть финальный этап периода пилотной эксплуатации, то обнаружится, что на этом этапе аномалии вообще прекращаются.

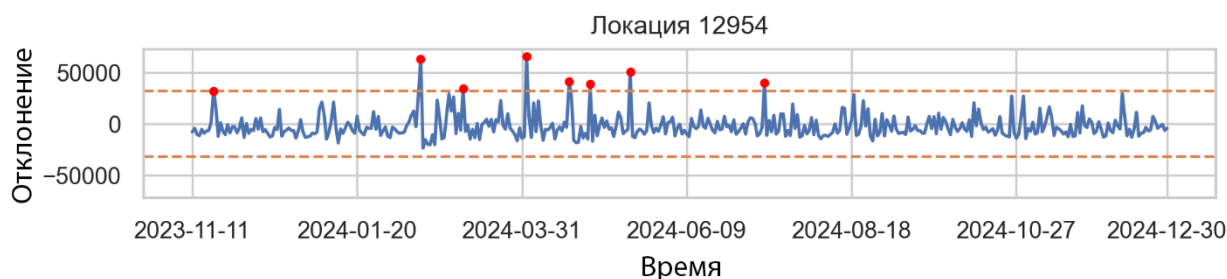


Рисунок 5.21 — Отклонение модели по времени для суммарной стоимости работ в локации 12954

В локации 42663 аномалий совсем мало: всего 3 (см. рисунок 5.22), то есть ровно столько же, сколько мы зафиксировали, анализируя аномалии по времени для числа оформленных заявок (хотя, при этом следует отметить, что они наблюдаются в другие моменты времени, что хорошо видно, если сравнить рисунки 5.22 и 5.6).

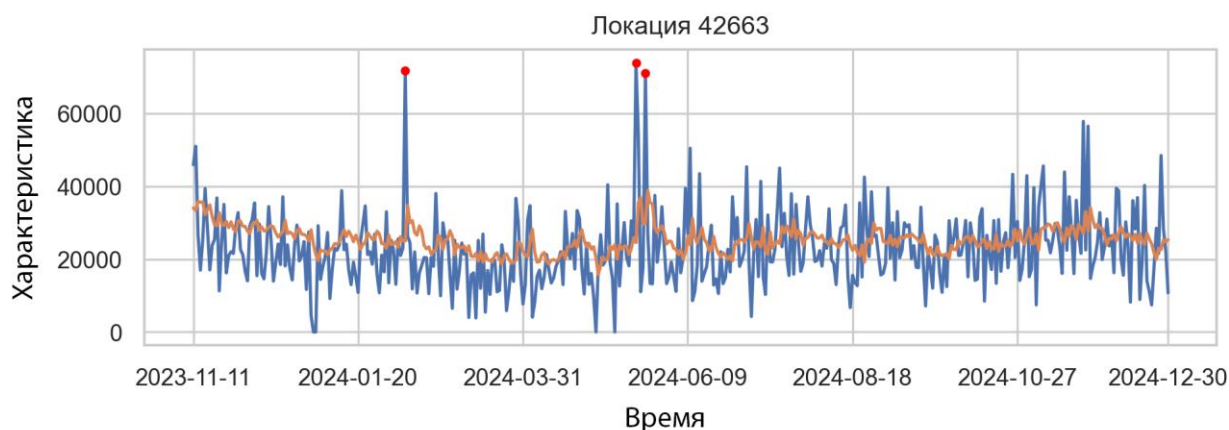


Рисунок 5.22 — Модель по времени для суммарной стоимости работ в локации 42663

Также можно вывести и посмотреть в виде отдельной иллюстрации совокупность всех отклонения прогнозируемых значений наблюдаемой характеристики от её реальных значений (см. рисунок 5.23). Мы видим, что выраженность отклонений чуть выше, нежели в случае с прогнозом по времени для

числа оформленных заявок, но что интересно — если при детекции аномалий по заявкам все отклонения уходили в отрицательную зону, то теперь, когда мы проводим детекцию аномалий по суммарным затратам, все отклонения приходятся на положительную зону⁶⁰.



Рисунок 5.23 — Отклонение модели по времени для суммарной стоимости работ в локации 42663

Аналогичный анализ и детекцию аномальных состояний мы проводим во всех имеющихся локациях (а после Парето-анализа мы исследуем процесс сервисного обслуживания в 69 локациях). С практической точки зрения нас интересует ситуации, связанные не с одной отдельно взятой локацией во все моменты времени, а наоборот: ситуации, связанные с одним моментом времени (в случае реальной, а не пилотной эксплуатации речь идёт о текущем моменте) во всех локациях сразу. Нас интересует, в каких локациях в данный момент наблюдаются аномалии, насколько они выражены, и какова упорядоченность локаций по выраженности аномалий, то есть то, что мы называем календарём аномалий.

Рассмотри для примера те же даты, которые мы использовали при детекции аномалий по времени для числа оформленных заявок (см. раздел 5.3.3). Это были две даты: будний день 2023-11-15 и выходной 2024-03-08. В момент 2023-11-15 для числа заявок тогда наблюдалось немного аномалий по времени

⁶⁰ Это может свидетельствовать о том, что в данном сервисном центре (код которого — 42663) сотрудники в силу каких-то известных только им обстоятельств навязывают клиентам дорогостоящие услуги. Но может и не свидетельствовать об этом. Вопросы интерпретации аномалий мы подробно обсуждаем в одном из следующих разделов (см. раздел 5.7).

(если быть точным, аномалии тогда были отмечены в четырёх локациях). Теперь же, когда мы анализируем аномалии для суммарных затрат на гарантийное обслуживание, ситуация и рейтинг аномальных состояний иллюстрируются следующей, аналогичной, но всё-таки другой диаграммой (см. рисунок 5.24 и, для сравнения, рисунок 5.8).

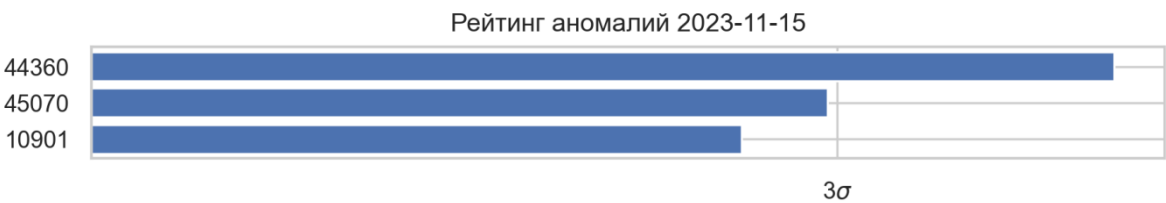


Рисунок 5.24 — Рейтинг аномальных по времени локаций в процессе сервисного обслуживания для суммарной стоимости работ в момент времени 2023-11-15

Теперь мы видим не 4, а 3 аномалии в трёх локациях, но что интересно — все эти 3 локации в момент 2023-11-15 демонстрировали также аномалии и по отношению к числу оформленных заявок. Нас как исследователей процесса сервисного обслуживания такая согласованность не может не радовать⁶¹.

Перейдём к 2024-03-08 (см. рисунок 5.25).

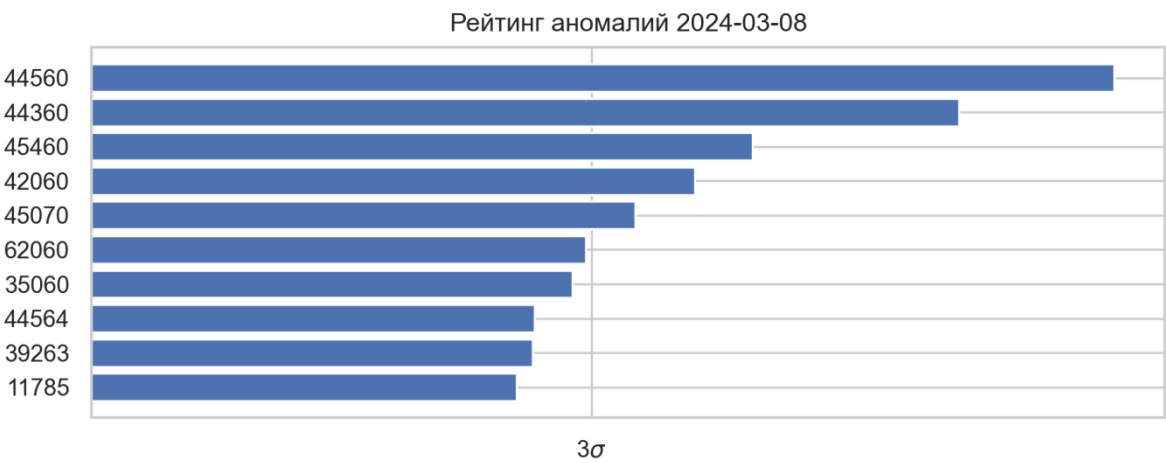


Рисунок 5.25 — Рейтинг аномальных по времени локаций в процессе сервисного обслуживания для суммарной стоимости в момент времени 2024-03-08

⁶¹ Хотя (и мы это уже отмечали выше, и ещё неоднократно отметим ниже) спешить с выводами и интерпретацией этих фактов было бы непростительной профанацией.

Как и следовало ожидать, 8 марта аномальных локаций оказалось гораздо больше — 10 локаций продемонстрировали аномальные значения суммарных затрат на гарантийное обслуживание (а если сравнивать результат с аномалиями для числа оформленных заявок, то для той характеристики их было 12). Из них 8 локаций в момент времени 2024-03-08 демонстрировали аномалии и по отношению к числу оформленных заявок, то есть мы снова вполне можем утверждать, что согласованность аномалий на этих двух характеристиках оказывается весьма высокой.

Календарь аномалий по времени в процессе сервисного обслуживания по отношению к суммарной стоимости затрат на гарантийное обслуживание, зафиксированных в течение последних десяти наблюдений, представлен в приложении D.2.3.

5.6. Детекция аномалий по локациям в процессе сервисного обслуживания для суммарной стоимости гарантийных работ

Здесь мы рассмотрим суммарную стоимость гарантийных работ с ещё одной точки зрения: с точки зрения аномальных значений этой характеристики по отношению к локациям. Данные, необходимые для исследования берём из таблицы 5.4. Это данные в формате LM: строки соответствуют локациям (то есть, центрам сервисного обслуживания), столбцы — моментам наблюдения, а в ячейках хранятся накопленные в течение рабочего дня суммы. Мы снова проводим весь цикл методики MSCP (Modeling, Challenge, Calibration, Production), начиная с моделирования процесса по времени, и заканчивая построением календаря аномалий.

5.6.1. Моделирование по локациям процесса сервисного обслуживания для суммарной стоимости работ

Так же как и выше, когда мы исследовали другие характеристики, для моделирования процесса сервисного обслуживания в контексте суммарной стоимости работ мы используем пул из семи моделей:

1. `LinearRegression` (линейная регрессия);
2. `DecisionTreeRegressor` (регрессия на решающих деревьях);
3. `RandomForestRegressor` (регрессия на случайных лесах);
4. `GradientBoostingRegressor` (градиентный бустинг над случайными лесами);
5. `SVR` (регрессия на методе опорных векторов);
6. `KNeighborsRegressor` (регрессия на ближайших соседях);
7. `MLPRegressor` (нейронная сеть многослойного перцептрона).

Применяем функцию `modeling_by_location`, которая описана нами в приложениях (см. приложение С.2) и в результате получаем сведения о прогнозирующей способности всех моделей нашего пула при изменении глубины наблюдений от 1 до 10. Визуализация сравнительного анализа моделирующих сем приведена ниже (см. рисунок 5.26). Видно, что практически неразличимые результаты демонстрируют две модели: линейная регрессия и многослойный перцептрон, но едва различимая разница все-таки есть.

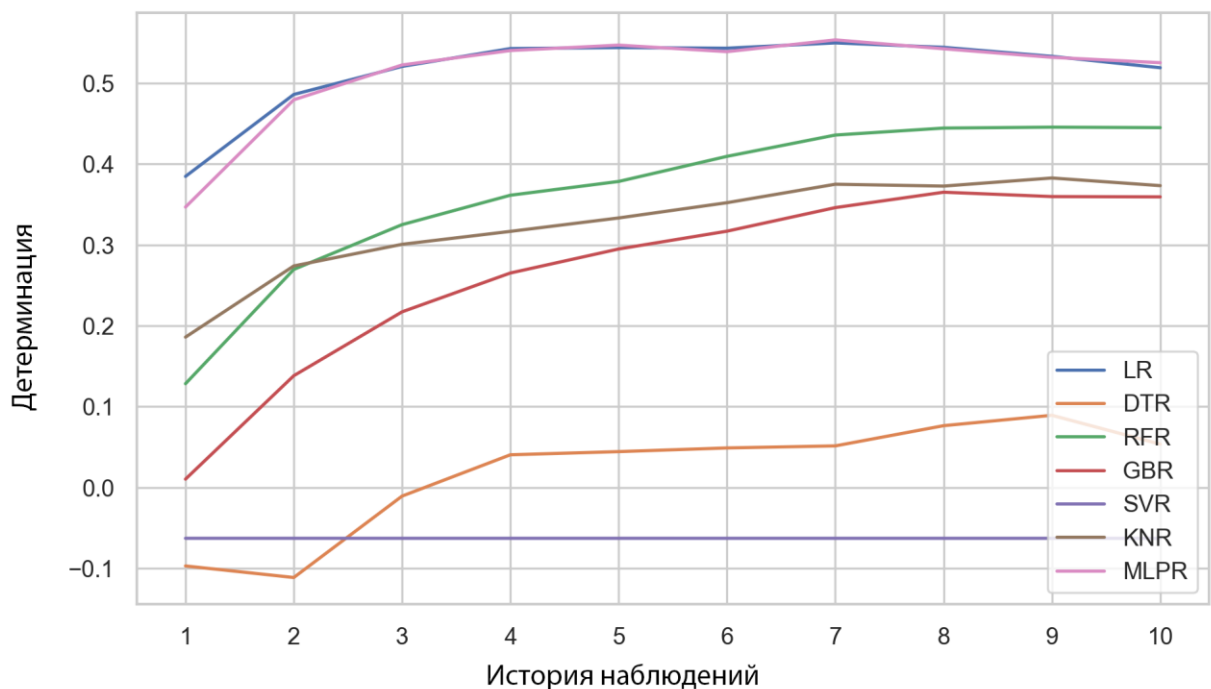


Рисунок 5.26 — Сравнительный анализ схем моделирования по локациям в процессе сервисного обслуживания для суммарной стоимости работ

Наилучший результат демонстрирует нейронная сеть многослойного перцептрона⁶² `MLPRegressor` с глубиной истории, равной 7. Именно эту схему моделирования процесса мы будем использовать далее, для получения калибровочной границы возможных отклонений и проведения этапа пилотной эксплуатации⁶³.

5.6.2. Калибровка отклонений по локациям в процессе сервисного обслуживания для суммарной стоимости работ

Снова применяем метод локтевой точки (см. Выше, главу 4, раздел 4.3.2). Берём все отклонения модельных значений от реальных, которые получаются при наилучшей схеме моделирования, и упорядочиваем их абсолютные величины по возрастанию (см. рисунок 5.27).

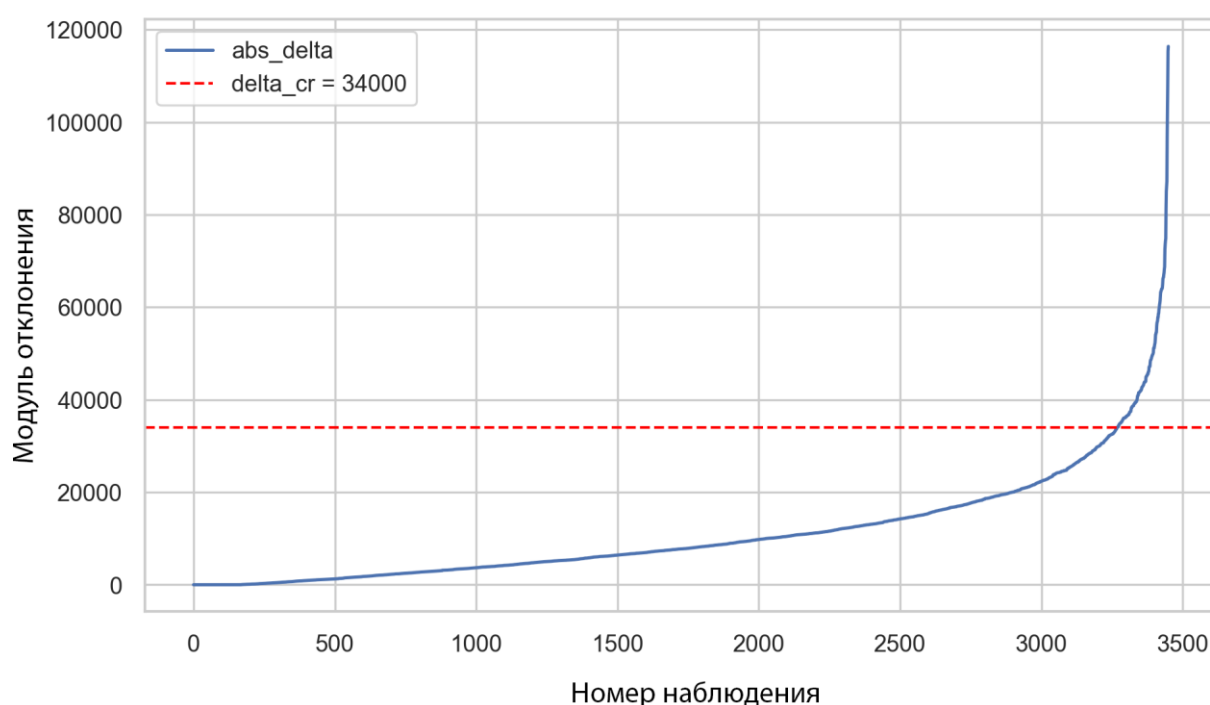


Рисунок 5.27 — Калибровочная кривая отклонений по времени в процессе сервисного обслуживания для суммарной стоимости работ

⁶² Хотя её преимущество по сравнению с линейной регрессией совсем невелико.

⁶³ Отметим еще раз, что в наши задачи не входит объяснение этого феномена. Чтобы ответить на вопрос, почему на имеющихся данных наилучшей прогнозирующей способностью обладает именно нейронная модель требует отдельного исследования, а мы лишь используем этот факт для достижения своих целей.

Локтевая точка калибровочной кривой приходится на величину 34000, и именно её мы будем использовать на этапе плотной эксплуатации, считая, что отклонение модельного значения от реального, которое превышает $\text{delta\_cr} = 34000$, сигнализирует об аномальности.

5.6.3. Пилотная эксплуатация детектора аномалий по локациям в процессе сервисного обслуживания для суммарной стоимости работ

Для этапа пилотной эксплуатации детектора аномалий мы, так же как и выше, используем последние 415 моментов времени. Мы применяем функцию `production_by_location` (см. приложение С.5) передавая необходимые сведения. Наилучшая схема моделирования:

```
best_model = MLPRegressor  
autoregression_dive = 7
```

Критическая величина отклонения Θ_{cr} :

```
teta_cr = 34000
```

Снова рассмотрим две локации — 12954 и 42663 — которые мы уже рассматривали трижды с трёх точек зрения (число заявок, детектированное по времени, число заявок, детектированное по локациям, и суммарная стоимость, детектированные по времени), ещё с одной точки зрения: с точки зрения детекции аномальных состояний по локациям, когда прогнозирование осуществляется не по отношению к прошлым состояниям этой локации, а по отношению к другим локациям в данный момент времени.

Начнём с локации 12954 (см. рисунок 5.28). Мы видим 12 аномалий, а с учётом того, что в предыдущих схемах детектирования выявлялось 15, 18 и 8 аномалий, то можно говорить о некоторой согласованности нашего нового результата с предыдущими, тем более, что мы видим одинаковые тенденции: аномалий по локациям оказывается больше, чем аномалий по времени (15 к 18 для числа оформленных заявок и 8 к 12 для суммарных затрат).

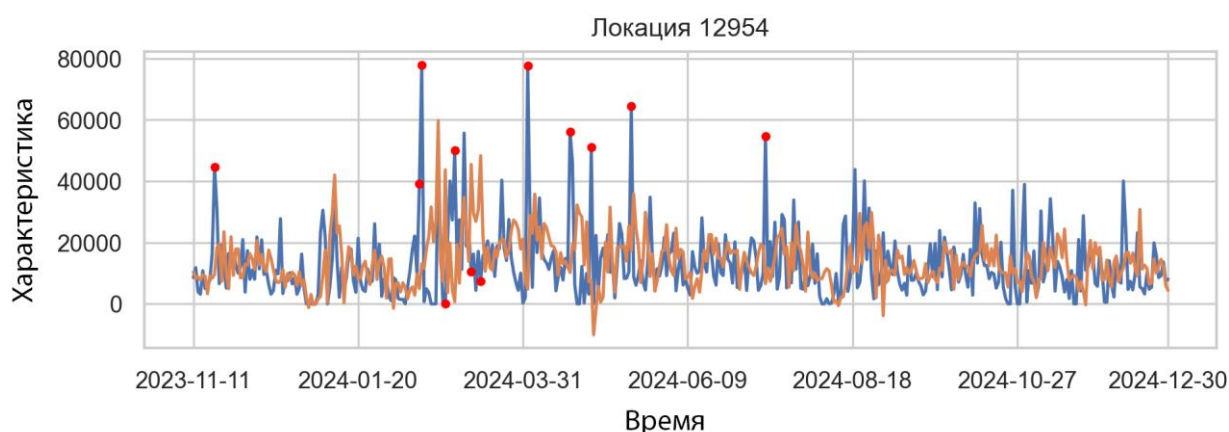


Рисунок 5.28 — Модель по локациям для суммарной стоимости работ в локации 12954

Отклонения модели от реальности в этой локации проиллюстрированы ниже (см. рисунок 5.29), причём, видно, что часть аномалий значительно превышают критически допустимый порог $\text{delta\_cr} = 34000$, а сами отклонения носят как положительный (в большую сторону), так и отрицательный (в меньшую сторону) характер, причем, значительные отклонения всегда являются положительными, тогда как отрицательные отклонения всегда выражены весьма слабо.



Рисунок 5.29 — Отклонение модели по локациям для суммарной стоимости работ в локации 12954

Во второй выбранной нами для примеров локации, а именно — в локации 42663 (см. рисунок 5.30) снова детектируется весьма незначительное число аномалий: всего 5 (напомним, что три предыдущие схемы каждый раз возвращали по 3 аномалии), и их отклонения, в отличие от локации 12954, относительно невелики (хотя тенденция, состоящая в том, что положительные отклонения выражены ярче, чем отрицательные, сохраняется и в этой локации).

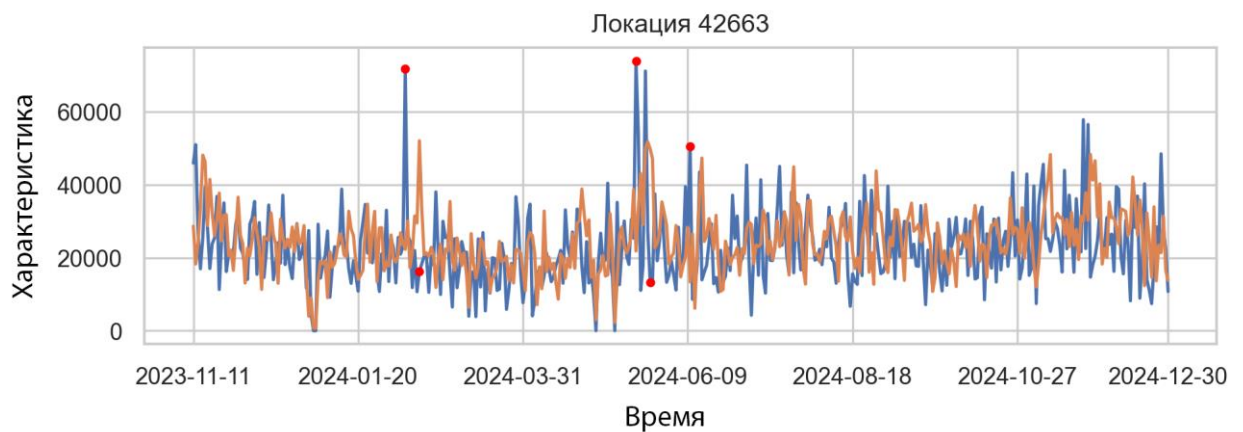


Рисунок 5.30 — Модель по локациям для суммарной стоимости работ в локации 42663

Отклонения модели от реальности представлены ниже (см. рисунок 5.31).

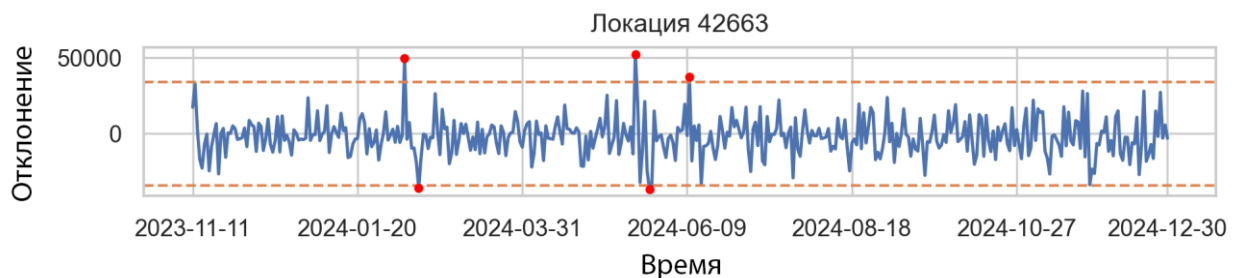


Рисунок 5.31 — Отклонение модели по локациям для суммарной стоимости работ в локации 42663

Чтобы иметь представление об аномальных состояниях процесса сервисного обслуживания в реальном времени, нам в каждый момент времени нужно знать, в каких именно локациях наблюдается аномалия, насколько она выражена, а если аномалий наблюдается несколько, то как они упорядочены по своей выраженности. Такие сведения мы называем календарём аномалий. Выше мы рассматривали страницы календарей для моментов времени 2023-11-15 и 2024-03-08. Рассмотрим эти моменты ещё раз, чтобы сравнить новые результаты с предыдущими.

В момент времени 2023-11-15 видим две аномалии (см. рисунок 5.32), причём, они обе выражены весьма слабо, а одна из них (а именно, локация 10901) была так же детектирована как аномальная и в схеме детекции аномалий по времени (см. выше, раздел 5.5.3). Заметим, что, анализируя эту дату в

смысле аномальности для числа оформленных заявок, мы получили 4 и 5 аномалий в случае детекции по времени и по локациям соответственно соответственно (см. разделы 5.3.3 и 5.4.3), а локация 10901 вошла во все эти рейтинги⁶⁴, что говорит о согласованности разных детекторов.



Рисунок 5.32 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для суммарной стоимости работ в момент времени 2023-11-15

Далее — перейдём к моменту 2024-03-08 (см. рисунок 5.33).

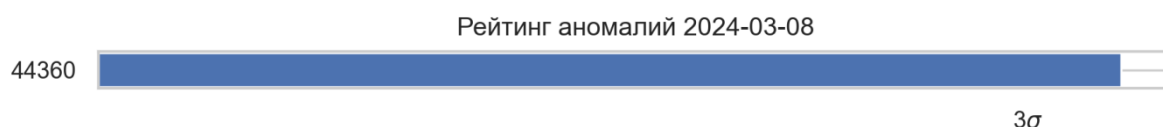


Рисунок 5.33 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для суммарной стоимости в момент времени 2024-03-08

Мы видим всего одну аномалию, а если вспомнить календарь аномалий на 2023-03-08, полученный при помощи детекции по локациям для числа оформленных заявок (см. раздел 5.4.3), а именно — тот факт, что тогда было детектировано всего 2 аномалии, — то мы имеем ещё один аспект согласованности наших алгоритмов детекции.

Календарь аномалий по локациям в процессе сервисного обслуживания по отношению к суммарной стоимости затрат на гарантийное обслуживание, зафиксированных в течение последних десяти наблюдений, представлен в приложении D.2.4.

5.7. Об интерпретации аномалий, выявленных в процессе сервисного обслуживания

⁶⁴ Отметим ещё раз, что было бы в высшей степени непрофессионально давать этому факту какую-либо содержательную интерпретацию, кроме то, что была дана только что: детекторы работают вполне согласованно.

Специфика работы сети сервисных центров крупного автомобильного бренда кроется в сложной и многоуровневой системе взаимодействия с клиентами, поставщиками запчастей и внутренними подразделениями компании. Каждый сервисный центр — это не просто мастерская, а сложный комплекс, где применяются технологии обслуживания и ремонта, разработанные производителем. Эти технологии охватывают все аспекты работы: от приёма автомобиля и диагностики его неисправностей до выполнения ремонтных работ и выдачи автомобиля клиенту.

Корректная оценка работы сервисного центра требует глубокого понимания этих процессов и умения анализировать, насколько эффективно они внедрены и соблюдаются на практике. Необходимо учитывать не только количество обслуженных автомобилей или среднее время ремонта⁶⁵, но и качество выполненных работ, удовлетворённость клиентов, квалификацию персонала и наличие необходимого оборудования.

Формальная статистика, такая как количество обращений или средний чек, может дать лишь поверхностное представление о работе сервисного центра. Она не учитывает сложность ремонтов, особенности модельного ряда, сезонность обращений и другие факторы, которые могут существенно влиять на результаты работы. Например, низкий показатель возвратов может быть достигнут за счёт сильного упрощения диагностики или использования некачественных запчастей, что в конечном итоге приведёт к недовольству клиентов и к ухудшению репутации. Только профессионал автомобильного сервиса, обладающий опытом работы в данной сфере, способен комплексно оценить деятельность сервисного центра, так как он может оценить квалификацию персонала, проверить наличие необходимого оборудования, проанализировать процесс диагностики и ремонта, а также изучить отзывы.

⁶⁵ Отметим, что в нашем исследовании фигурировало именно количество обслуженных автомобилей, скрытое за характеристикой «число оформленных заявок на гарантийное обслуживание» и сильно коррелирующей с ней характеристикой «суммарные затраты на гарантийные работы».

Кроме того, профессионал способен выявить и скрытые проблемы, которые не отображаются в формальной статистике, такие как низкая мотивация персонала, неэффективная организация работы или недостаток контроля качества⁶⁶. Он может предложить конкретные меры по улучшению работы сервисного центра и повышению удовлетворённости клиентов, а оценка, основанная на формальной статистике, без учёта перечисленных факторов, является по большому счёту профанацией и не позволяет получить объективное представление о реальном состоянии дел в сервисном центре.

Давайте рассмотрим локацию 44360 в момент времени 2023-11-15. Мы, будучи аналитиками данных, точно знаем, что в этот момент времени в этой локации нами были зафиксированы аномалии сразу по обоим наблюдаемым характеристикам (по числу оформленных заявок и по суммарной стоимости гарантийных работ), причём, обе они оказались в положительной зоне (то есть, реальные значения характеристик отклонялись от прогнозируемых значений в большую сторону). Хорошо это или плохо? На первый взгляд — хорошо, сервисный центр работает лучше, чем мы предполагали. Но в чем причина таких отклонений? — этого мы не знаем. Мы можем лишь выдвинуть гипотезы относительно происходящего.

Гипотеза 1. Жители района, в котором расположен данный сервисный центр, в результате успешной маркетинговой акции (предоставление в этот день дополнительных услуг; скидки на работы, не входящие в гарантийный перечень и т. д.) именно в этот день проявили интерес к сервисному обслуживанию своих автомобилей.

Гипотеза 2. Сотрудники данного сервисного центра неоднородны в смысле своих профессиональных качеств. Какие-то из них, обладая значительно более высокой квалификацией по сравнению с остальными, демонстрируют лучшие результаты, и в этот день была как раз их смена.

⁶⁶ Разумеется, ничего этого в нашем исследовании не было, хотя мы и предложили некоторые меры по улучшению контроля.

Гипотеза 3. Руководство данного сервисного центра, исходя из корыстных побуждений, пошло на акт корпоративного мошенничества и просто приписало к реальным результатам дополнительные величины, так как завышенные показатели усиливают их позиции в конкурентной борьбе за попадание в топ-менеджмент компании.

Какая из этих гипотез верна? Или среди них нет ни одной верной гипотезы, а причина кроется в чём-то совсем другом, о чём мы, будучи оторванными от реального положения дел, не имеем ни малейшего представления? Будем честны перед самими собой: скорее всего, в силу нашей неосведомлённости, мы не сможем дать корректную интерпретацию выявленных нами аномалий.

В условиях стремительного развития анализа данных и машинного обучения все чаще и чаще в той или иной сфере человеческой деятельности возникает закономерный вопрос: кто должен интерпретировать полученные результаты? — и ответ на него однозначен: это прерогатива профессионалов предметной области. Аналитики данных, обладая глубокими знаниями в области алгоритмов и статистики, зачастую не имеют достаточного понимания специфики конкретной отрасли, а неправильная интерпретация результатов анализа данных может привести к серьёзным последствиям. В медицине это может обернуться неправильным диагнозом и неэффективным лечением, в энергетике — к авариям и перебоям в электроснабжении, в автомобилестроении — к потере конкурентоспособности.

Специалисты в анализе данных и машинном обучении должны сосредоточиться на создании эффективных моделей и алгоритмов, а интерпретацию полученных результатов следует оставить предметникам. Только совместная работа специалистов разных областей позволит извлечь максимальную пользу из анализа данных и машинного обучения, избежав при этом дорогостоящих ошибок. Необходимо помнить, что данные — это лишь отражение реальности, а профессионалы предметной области — эксперты в самой реальности (а не в её отражении).

5.8. Выводы по главе 5

В этом разделе мы провели полную процедуру детекции аномалий ещё в одном нестабильном процессе — процессе сервисного обслуживания автомобилей. В качестве заключительных замечаний к этому разделу отметим следующее.

1. Сервисное обслуживание по сути является процессом потребления услуги, который, как и любой процесс потребления, является нестабильным, так как зависит от потребностей потребителя, а не от производителя.
2. Применяя алгоритмы методики МССР (Modeling, Challenge, Calibration, Production), мы выявляем аномальные состояния этого нестабильного по своей природе процесса во всех доступных для исследования локациях с двух точек зрения: используя в качестве характеристики процесса число обработанных за день заявок и суммарную стоимость работ по гарантии.
3. Аномалии, выявленные относительно этих двух характеристик, демонстрируют высокий уровень согласованности.
4. Полученные в результате исследования календари и рейтинги аномалий позволяют лицу, ответственному за сервисное обслуживание (или подразделению, ответственному за сервисное обслуживание) проводить проверки соответствующих локаций с большей эффективностью.
5. Установление причин алгоритмически детектированных аномалий и включение соответствующих протоколов реагирования на них должно осуществляться представителями предметной области и выходит за рамки компетенций аналитика данных.

Таким образом, проведённая в этой главе процедура детекции аномалий демонстрирует применимость методики МССР (Modeling, Challenge, Calibration, Production) к отдельным процессам в группе процессов сервисного обслуживания.

ГЛАВА 6. ДЕТЕКЦИЯ АНОМАЛИЙ В ПРОЦЕССЕ МАРКЕТИНГОВОГО СОПРОВОЖДЕНИЯ АВТОМОБИЛЬНОГО БРЕНДА

Маркетинговое сопровождение автомобильного бренда играет ключевую роль в формировании отношения потребителя к продукту [39, 42, 44, 106]. В условиях конкуренции, когда каждый бренд стремится выделиться и запомниться, именно маркетинг создаёт неповторимый образ, отражающий ценности и выгодные преимущества какого-либо конкретного автомобиля, а маркетинговая стратегия включает в себя, прежде всего, глубокое понимание целевой аудитории, её потребностей и предпочтений.

В концепции всеобщего управления качеством (TQM) обеспечение потребительской удовлетворенности и непрерывное улучшение качества рассматриваются как ключевые факторы конкурентоспособности организации, и в этой парадигме, маркетинговые процессы играют стратегически важную роль, сопоставимую по значимости с процессами производства и сервисного обслуживания, что обусловлено целым рядом факторов.

Прежде всего, это формирование атрибутов качества продукта/услуги на основе анализа требований заинтересованных сторон [31, 36, 58, 85], так как проведение систематических маркетинговых исследований (включая как количественные, так и качественные методы) позволяет выявлять латентные и манифестированные потребности целевых сегментов рынка. Полученные данные используются для определения критических для качества атрибутов продукта/услуги (CTQs, англ. — Critical to Quality) и установления целевых показателей, соответствующих ожиданиям потребителей [27, 73, 88, 89, 151, 152]. Кроме того, существенную роль играют так называемые функции качества

(QFD, англ. — Quality Function Deployment) и определение приоритетных характеристик, а соответствующая методология развертывания функции качества (QFD) позволяет трансформировать требования потребителей в технические характеристики продукта/услуги и операционные показатели процессов [63, 175].

Далее, нельзя не отметить какой фактор, как обеспечение обратной связи и стимулирование процессов улучшения качества на основе анализа данных, в рамках которого особую роль играет мониторинг удовлетворенности потребителей (CSAT, англ. — Customer Satisfaction) и сбор данных о восприятии качества, так как регулярный мониторинг удовлетворенности потребителей с использованием структурированных и неструктурированных методов (опросы, фокус-группы, аналитика социальных сетей, анализ обращений в службу поддержки и т.д.) позволяет измерять восприятие качества на различных этапах жизненного цикла продукта/услуги [64, 89, 96]. Анализ причин неудовлетворенности или незаинтересованности потребителей в продукте/ услуге и выявление отклонений от целевых показателей требует использования статистических методов анализа данных и позволяет выявлять ключевые факторы, влияющие на удовлетворенность или заинтересованность потребителей.

Наконец, в качестве отдельного очень важного фактора в контексте TQM нужно упомянуть управление репутационными рисками и поддержание лояльности потребителей [36, 87, 96, 97], а, в рамках этого направления — выделить проактивный мониторинг и оперативное реагирование на негативные проявления вовлеченности потребителей в политику бренда, так как именно такое реагирование является важным элементом управления репутацией организации, и именно оно является одной из практических целей настоящей диссертационной работы.

В рамках концепции TQM, маркетинговые процессы представляют собой не просто функцию продвижения и сбыта, а стратегический элемент системы управления качеством, обеспечивающий связь между организацией и потребителем [36, 72, 85, 86]. Интеграция маркетинговых данных и методов в

процессы управления качеством позволяет организациям адаптироваться к изменяющимся требованиям рынка, обеспечивать устойчивую конкурентоспособность и достигать высокого уровня удовлетворенности заинтересованных сторон. Пренебрежение маркетинговыми процессами в системе TQM может привести к несоответствию продукта/услуги ожиданиям потребителей, снижению лояльности и ухудшению репутации организации [85, 86].

Коммуникационная политика маркетинга всегда многоканальна и охватывает различные площадки — от традиционных СМИ до цифровых платформ и социальных сетей [37, 179, 181, 182, 183, 184, 188, 189, 190, 192, 195, 196, 197, 198, 199], при этом особое внимание уделяется нарративному контенту⁶⁷. Безусловно, помимо нарратива, важным аспектом маркетинга является формирование лояльности клиентов: всевозможные программы лояльности, специальные предложения и персонализированный сервис, которые способствуют укреплению общего доверия к бренду и стимулируют повторные покупки, — маркетинговое сопровождение автомобильного бренда является необыкновенно сложным и многогранным процессом (о чем будет сказано ниже, см. раздел 6.1), требующим высокопрофессионального подхода и разнообразных специализаций. Однако в рамках настоящей диссертационной работы мы будем рассматривать лишь взаимодействие бренда с его аудиторией на коммуникационных цифровых площадках (видеохостингах, социальных сетях и т. д.).

6.1. Маркетинговое сопровождение автомобильного бренда как процесс потребления контента

⁶⁷ В маркетинге нарратив — это инструмент, который позволяет сформировать тесную связь с аудиторией через какие-то, на первый взгляд, не имеющие никакого отношения к бренду, увлекательные и запоминающиеся истории. Он помогает бизнесу донести свою историю до клиентов, сотрудников и инвесторов, а также продемонстрировать ценности компании.

Маркетинговое сопровождение автомобильного бренда — это процесс, в котором нарративный контент играет одну из ведущих ролей. Потребители уже давно перестали видеть в автомобиле просто средство передвижения: это, скорее, символ его статуса, отражение его индивидуальности и часть личной истории, а маркетологи, понимая все это, активно используют нарратив для формирования эмоциональной связи между брендом и целевой аудиторией.

Создание привлекательной истории бренда состоит в определении его ключевых ценностей и транслировании их через различные каналы коммуникации. В частности, видеоролики, рассказывающие об инновациях, экологической ответственности и т. д. — это мощный инструмент воздействия на потребителя, а социальные сети, с их интерактивными возможностями, позволяют вовлечь аудиторию в создание собственной истории, связанной с брендом.

Нарративный контент в автомобильном маркетинге охватывает широкий спектр форматов: от тест-драйвов, представленных в виде авантюрных приключений, до историй реальных владельцев, делящихся своим опытом. Важно, чтобы контент был аутентичным и релевантным для целевой аудитории, так как успешные маркетинговые кампании строятся на глубоком понимании потребностей и желаний потребителей, а также на умении рассказывать истории, которые цепляют и запоминаются. Кроме того, инфлюенс-маркетинг⁶⁸ также играет важную роль в процессе потребления нарративного контента. Лидеры мнений, разделяющие ценности бренда, все чаще становятся эффективными трансляторами его сообщения, так как они создают контент, который выглядит органично, вызывает доверие у своей аудитории и, в конечном счёте, является проявлением «мягкой силы» бренда по отношению к его потребителям.

⁶⁸ Маркетинг влияния (англ. *influencer marketing*) — это форма маркетинга в социальных сетях, включающая рекомендации и позиционирование продукта со стороны лидеров мнений, при этом под инфлюенсером понимается человек, бренд или группа лиц, которые способны повлиять на поведение и решения своей аудитории.

6.1.1. Коммуникационные площадки для продвижения нарратива автомобильных брендов

Итак, автомобильные бренды используют различные коммуникационные площадки для продвижения своего нарратива, взаимодействия с клиентами и создания имиджа. Вот основные из них:

Социальные сети

VK — одна из самых популярных платформ для автомобильных брендов в РФ и странах ближнего зарубежья. Здесь можно делиться новостями, запускать рекламные кампании и взаимодействовать с клиентами через комментарии и сообщения.

Twitter позволяет брендам быстро реагировать на события и общаться с клиентами в реальном времени. Многие автомобильные компании используют эту платформу для анонсов и общения с поклонниками.

С ростом популярности TikTok, многие автомобильные бренды начали использовать также эту платформу для создания креативного контента, здесь они делятся короткими видео, демонстрирующими автомобили в действии.

Сайты и блоги

Важную роль играют официальные сайты брендов, которые служат основным источником информации о моделях, ценах и акциях. Они также могут содержать блоги с новостями и статьями о технологиях и инновациях.

Кроме того, бренды сотрудничают с автомобильными блогерами и журналистами для создания и распространения контента, который может привлечь внимание к их продуктам. Это могут быть как тексты, так и видеообзоры.

Видеоплатформы

Безусловно, YouTube является важной платформой для автомобильных брендов, здесь они могут публиковать рекламные ролики, тест-драйвы и обучающие видео. Бренды также могут сотрудничать с популярными видеоблогерами для расширения своей аудитории.

С ростом популярности такого явления как стриминг, некоторые автомобильные компании начали использовать Twitch для проведения живых трансляций, где они могут взаимодействовать с аудиторией и демонстрировать свои автомобили в реальном времени.

Мобильные приложения

Помимо этого, автомобильные бренды разрабатывают собственные мобильные приложения, которые позволяют пользователям получать доступ к полной информации о моделях, записываться на тест-драйвы, получать уведомления и т. д.

6.1.2. Этапы маркетинговых кампаний на цифровых площадках

Нарративная реклама на цифровых площадках — это один из наиболее эффективных инструментов маркетингового сопровождения автомобильных брендов⁶⁹, позволяющий достигать конкретных целевых аудиторий. Этот инструмент весьма сложен, и нам следует понимать его специфику, коль скоро мы собираемся анализировать аномальные состояния процесса маркетингового сопровождения. Прежде всего, следует понимать, что в нем присутствуют различные этапы. Их беглый обзор представлен ниже.

⁶⁹ Заметим, что речь может идти не только об автомобилях. Те же самые принципы используются при продвижении брендов произвольной природы, вплоть до персональных брендов артистов и блогеров. Разница лишь в размерах бюджета.

Определение целей маркетинговой кампании

Перед началом любой рекламной кампании должны быть чётко определены её цели⁷⁰. Напримет, это могут быть: 1) увеличение продаж, 2) повышение узнаваемости бренда, 3) привлечение трафика на сайт, 4) генерация лидов⁷¹, 5) и т. д. Чёткое понимание целей кампании помогает в дальнейшем точнее формировать её стратегию и выбирать наиболее подходящие инструменты.

Исследование целевой аудитории

Знание своей целевой аудитории — это ещё одно необходимое условие успешной таргетированной рекламы. Прежде чем начинать кампанию, необходимо провести исследование, чтобы понять: 1) демографические характеристики (возраст, пол, социальный статус и т. д.), 2) интересы и поведение аудитории, 3) её проблемы и потребности, 4) и т. д. Эти данные помогут создать персонализированные рекламные сообщения.

Выбор платформы для размещения рекламы

Существует множество платформ для таргетированной рекламы, каждая из которых имеет свои особенности. Наиболее популярные из них: 1) YouTube (отлично подходит для визуального контента), 2) Google Ads (для текстовой рекламы и поиска), 3) LinkedIn (подходит для B2B-сегмента и профессиональной аудитории). В отечественном сегменте коммуникационных платформ ведущую роль сейчас играют 5) Яндекс Ди-

⁷⁰ Так как в противном случае, у нас не будет никаких критериев, которые позволяли бы на судить об эффективности маркетинговой кампании и, в конечном, счёте оценивать её как успешную или, наоборот, неудачную.

⁷¹ Лидогенерация (генерация лидов) — это маркетинговая тактика, направленная на поиск потенциальных клиентов и сбор контактной информации (лидов) для их дальнейшей конвертации в реальных покупателей.

рект (система контекстной рекламы на страницах Яндекса) и 6) Вконтакте. Выбор конкретной платформы (или нескольких платформ) зависит от целевой аудитории и целей кампании.

Создание рекламного контента

Качественный контент — это основа успешной рекламной кампании. Он должен быть как минимум: 1) привлекательным и запоминающимся, 2) соответствующим всем основным интересам целевой аудитории, 3) чётким и понятным, с ясным призывом к действию, не допускающим никаких разночтений, 4) и т. д., так как, помимо этого, есть ещё огромное количество требований, которым должен удовлетворять контент, но это уже зависит от конкретной кампании и набор таких требования не является универсальным. Рекламный контент должен подвергаться постоянному тестированию в различных форматах и с разнообразными креативными решениями, для того чтобы определить, какой именно вариант работает лучше с точки зрения достижения поставленных целей (см выше, о целях кампании).

Настройка таргетинга

В маркетинге таргетинг — это процесс выбора аудитории, которой будет адресована вся маркетинговая кампания. Сама по себе эта процедура включает целый ряд разных этапов: 1) определение целевой аудитории, 2) настройка множества параметров таргетинга, 3) создание рекламного обращения, 4) настройка бюджета и расписания, 5) анализ и оптимизация результатов. После запуска кампании нужно регулярно отслеживать её результаты: количество показов, кликов, конверсий. На основе этих данных вносятся изменения в рекламную кампанию с целью достижения её наибольшей эффективности.

6.1.3. Таргетинг рекламной кампании в интернет-маркетинге

Выше мы особо отметили таргетинг маркетинговой кампании как один из её важных этапов, но он настолько важен, что мы остановимся подробнее именно на нем и посвятить несколько слов тому, как именно настраивается таргетинг на различных коммуникационных площадках с учётом их специфических особенностей. Мы рассмотрим, как именно осуществляется настройка таргета на наиболее популярных площадках, таких как YouTube, Google Search, Yandex, VKontakte и Telegram, так как данные именно с этих площадок мы будем использовать в дальнейшем (см. ниже, разделы 6.2–6.4), применяя методы детекции аномалий.

YouTube

На YouTube'е таргетинг осуществляется через выбор демографических характеристик, интересов, предпочтений и поведения пользователей. Здесь рекламодатели могут настраивать показы рекламы в зависимости от возраста, пола и местоположения, а также доступна возможность таргетинга по интересам, что позволяет показывать рекламу тем, кто уже проявил ранее интерес к схожему контенту (другими словами — на YouTube'е работают рекомендательные системы).

Google Search

Google Search предлагает весьма мощные инструменты для настройки таргетинга, и в первую очередь, здесь можно использовать ключевые слова, которые соответствуют запросам пользователей. Это позволяет показывать рекламу именно тем, кто ищет товары или услуги, которые предлагает автор рекламного обращения. Кроме того, в Google Search можно настроить также географический таргетинг, чтобы реклама показывалась только в определённых регионах.

Yandex

Yandex также предоставляет возможности для таргетинга, включая контекстную рекламу и т. д., и во многом его функционал подобен функционалу Google Search. Рекламодатели могут использовать ключевые слова и фразы, чтобы достичь целевой аудитории. Кроме того, Yandex предлагает таргетинг по интересам и поведению пользователей, что позволяет более точно настраивать рекламные кампании.

ВКонтакте

ВКонтакте является одной из самых популярных социальных сетей в России, и ее инструменты таргетинга позволяют детально настраивать аудиторию. Рекламодатели могут выбирать аудиторию по возрасту, полу, интересам и даже по поведению пользователей в сети.

Telegram

Telegram, хотя и менее традиционная платформа для рекламы, также предлагает свои возможности для таргетинга. Рекламодатели могут использовать каналы и группы, чтобы достичь своей аудитории, а настройка таргетинга в Telegram может включать выбор тематики канала и интересов подписчиков, что позволяет точно нацеливаться на нужных пользователей.

Следует помнить, что успешная настройка таргетинга требует постоянного анализа и оптимизации. Поэтому алгоритмическая детекция аномалий маркетингового процесса, о которой речь пойдет ниже, и которая позволяет отслеживать состояния процесса в режиме реального времени, является необходимым инструментом успешной рекламной кампании и играет очень важную роль в маркетинговом сопровождении автомобильного бренда.

6.2. Данные о переходах на канал автомобильного бренда

Выше мы отмечали, что маркетинговое сопровождение автомобильного

бренда является сложным процессом, содержащим множество составляющих, каждая из которых заслуживает отдельного исследования. Но наша задача состоит не в том, чтобы провести исчерпывающий анализ и дать оценку качества маркетинга. В рамках настоящей диссертационной работы нам нужно продемонстрировать, как наша методика МССР (Modeling, Challenge, Calibration, Production) способна обрабатывать данные о маркетинговом процессе, поэтому для иллюстрации наших идей мы рассмотрим лишь частный фрагмент маркетингового сопровождения⁷²: данные о переходах на официальный канал известного отечественного автомобильного бренда, а если быть точным — данные о том, из каких источников происходят эти переходы. Другими словами, это будут данные об источниках траффика.

Официальный канал бренда расположен на YouTube'е и предоставляет обширную статистику, в частности, статистику переходов за весьма долгий период существования канала, начиная с 2014 года. Число разнообразных источников перехода на канал огромно, какие-то источники дают локальные по времени вспышки активности, а потом исчезают, какие-то — наполняют траффик в микроскопических дозах, а какие-то на протяжении многих лет стабильно приводят львиную долю просмотров.

С учётом столь высокой неоднородности источников, нет никакой необходимости включать в перечень источников траффика абсолютно все зафиксированные в истории канала переходы. Достаточно рассмотреть лишь те из них, которые обеспечивают основную долю просмотров⁷³, и таких источников совсем немного, если быть точным — их пять: YouTube⁷⁴, Google Search,

⁷² Так же, как и в главе 4, где мы рассматривали исполнение всего лишь одной технологической операции в процессе производства автомобиля, тогда как сам процесс производства насчитывает тысячи таких операций.

⁷³ В этом ситуация весьма напоминает ту, с которой мы столкнулись в главе 5, когда анализировали работу сервисных центров по гарантийному обслуживанию автомобилей и обнаружили, что, в точности в соответствии с классической пропорцией Парето, 80% обслуженных заявок приходится на 20% сервисных центров.

⁷⁴ Имеется в виду поиск на YouTube'е.

Yandex, VKontakte и Telegram, а на остальные источники траффика можно вообще не обращать внимания, так как их доля в общем числе просмотров ничтожна.

Данные о переходах на официальный канал автомобильного бренда из этих источников представлены в формате ML (см. таблицу 6.1).

Таблица 6.1 — Данные о переходах на канал в формате ML

	Локация 1	Локация 2	Локация 3	Локация 4	Локация 5
Момент 1	0	0	0	0	0
Момент 2	0	0	0	0	0
Момент 3	1	2	0	0	0
Момент 4	25	7	0	75	0
...	...	...	...	...	...
Момент 554	674	685	601	50	1916
Момент 555	149	113	160	0	767

Здесь в качестве локаций рассматриваются указанные выше пять источников траффика, а в качестве моментов времени — агрегированные по неделям сведения о числе переходов, зафиксированных на канале с того или иного источника траффика, начиная с 2014-08-10 и заканчивая 2025-03-23 (всего таких моментов наблюдения получается как раз 555).

Транспонируя таблицу 6.1, получаем данные в формате LM (см. таблицу 6.2).

Таблица 6.2 — Данные о переходах на канал в формате LM

	Момент 1	Момент 2	Момент 3	...	Момент 555
Локация 1	0	0	1	...	140
Локация 2	0	0	2	...	113
Локация 3	0	0	0	...	160
Локация 4	0	0	0	...	0
Локация 5	0	0	0	...	767

Кроме того, в нашем распоряжении есть ещё две вспомогательные таблицы, которые связывают коды локаций и моментов с их реальными значениями в контексте данного исследования аномальных состояний (см. таблицы

6.3 и 6.4).

Таблица 6.3 — Коды моментов в маркетинговом процессе

	Код	Момент
1	mom_1	2014-08-10
2	mom_2	2014-08-17
3	mom_3	2014-08-24
...	...	...
555	mom_555	2025-03-27

Таблица 6.4 — Коды локаций в маркетинговом процессе

	code	location
1	loc_1	YT
2	loc_2	GGL
3	loc_3	YND
4	loc_4	VK
5	loc_5	TG

Под период исторических наблюдений мы отводим первые 100 наблюдений, под период обучения — следующие 50 наблюдений, и на последних 415 наблюдениях мы проведём пилотную эксплуатацию нашего детектора аномалий.

6.3. Детекция аномалий по времени в процессе маркетингового сопровождения

В этом разделе мы, так же как и выше при анализе производственных и сервисных процессов (см. главы 4 и 5), проведём детекцию аномальных состояний в процессе маркетингового сопровождения, последовательно исполняя этапы методики МССР (Modeling, Calibration, Challenge, Production): начиная с моделирования временных рядов и выбора наилучшей схемы моделирования, и заканчивая календарями аномалий и рекомендациями по проведению мероприятий аудита.

6.3.1. Моделирование по времени процесса производства

Для моделирования процесса маркетингового сопровождения мы используем тот же пул из семи моделей, что и выше (см. главы 4 и 5):

- 15. `LinearRegression` (линейная регрессия);
- 16. `DecisionTreeRegressor` (регрессия на решающих деревьях);
- 17. `RandomForestRegressor` (регрессия на случайных лесах);
- 18. `GradientBoostingRegressor` (градиентный бустинг над случайными лесами);
- 19. `SVR` (регрессия на методе опорных векторов);
- 20. `KNeighborsRegressor` (регрессия на ближайших соседях);
- 21. `MLPRegressor` (нейронная сеть многослойного перцептрона).

Функция моделирования по времени описана в приложении С.1, и, используя её с максимальным погружением авторегрессии, равным 10, мы получаем следующие результаты (см. рисунок 6.1).

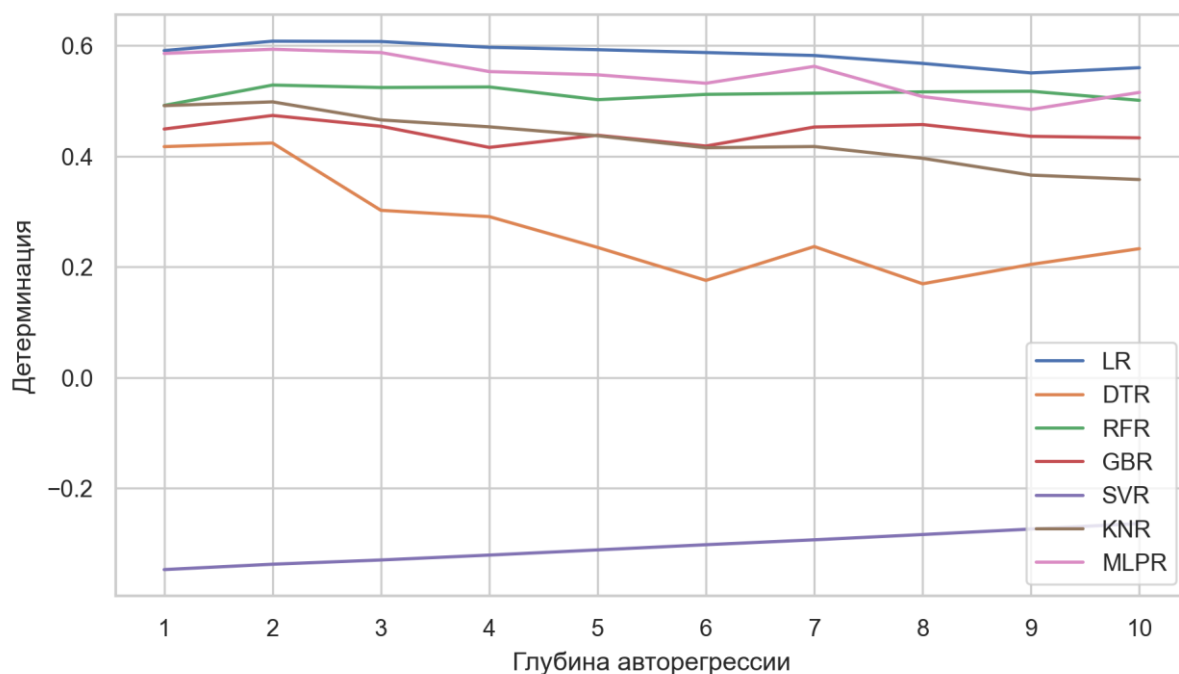


Рисунок 6.1 — Сравнительный анализ схем моделирования по времени в процессе маркетинга

Изменяя масштаб изображения в сторону его увеличения, получаем

уточнение результатов (см. рисунок 6.2).

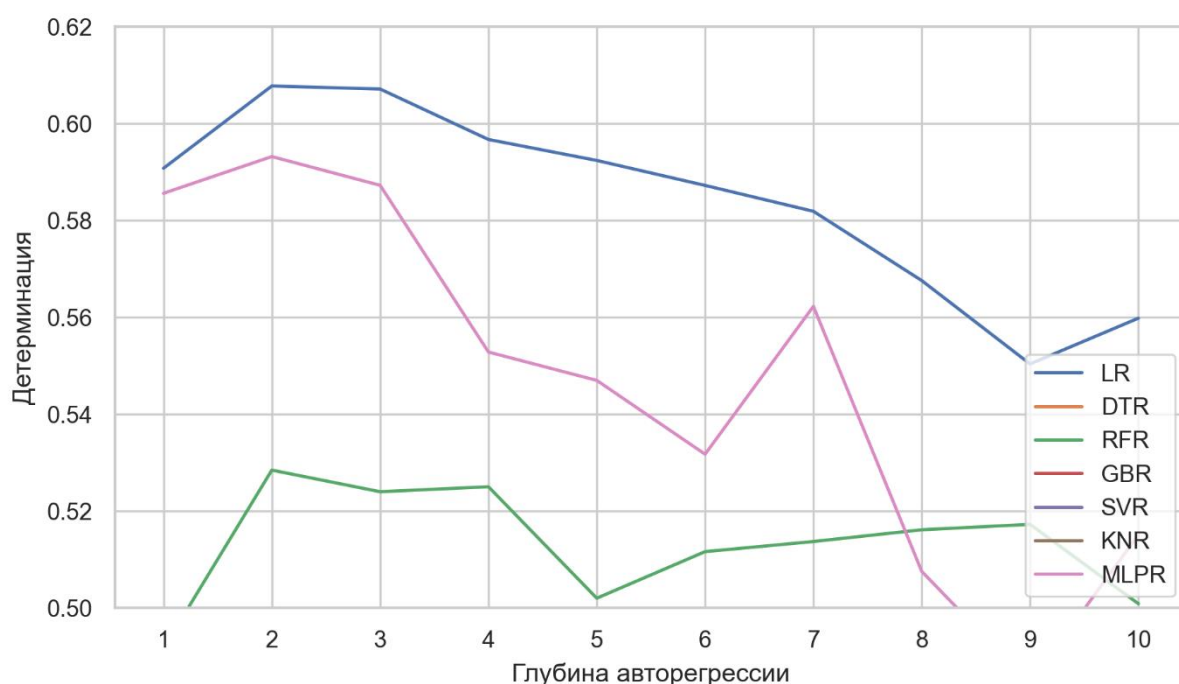


Рисунок 6.2 — Приближенный скоринг моделей по времени в процессе маркетинга

Теперь мы хорошо видим, что наилучший результат⁷⁵ показывает модель линейной регрессии `LinearRegression` с глубиной авторегрессии, равной 2. Именно эту схему моделирования мы будем использовать дальше, на этапах калибровки и пилотной эксплуатации нашего алгоритма (см. ниже, разделы 6.3.2 и 6.3.3).

6.3.2. Калибровка отклонений по времени в маркетинговом процессе

Для определения критической величины отклонения прогноза от реального значения характеристики мы проводим калибровку детектора аномалий по локтевой точке абсолютных величин отклонений (см. главу 4, раздел 4.3.2, где мы подробно описали эту идею). Калибровочная кривая в данном случае имеет следующий вид (см. рисунок 6.3).

⁷⁵ Можно, конечно, сказать, что прогнозирование, которое имеет коэффициент детерминации на уровне чуть выше 0.6, это весьма некачественное прогнозирование. Но, как будет показано ниже, причина кроется в поведении одной отдельно взятой локации.

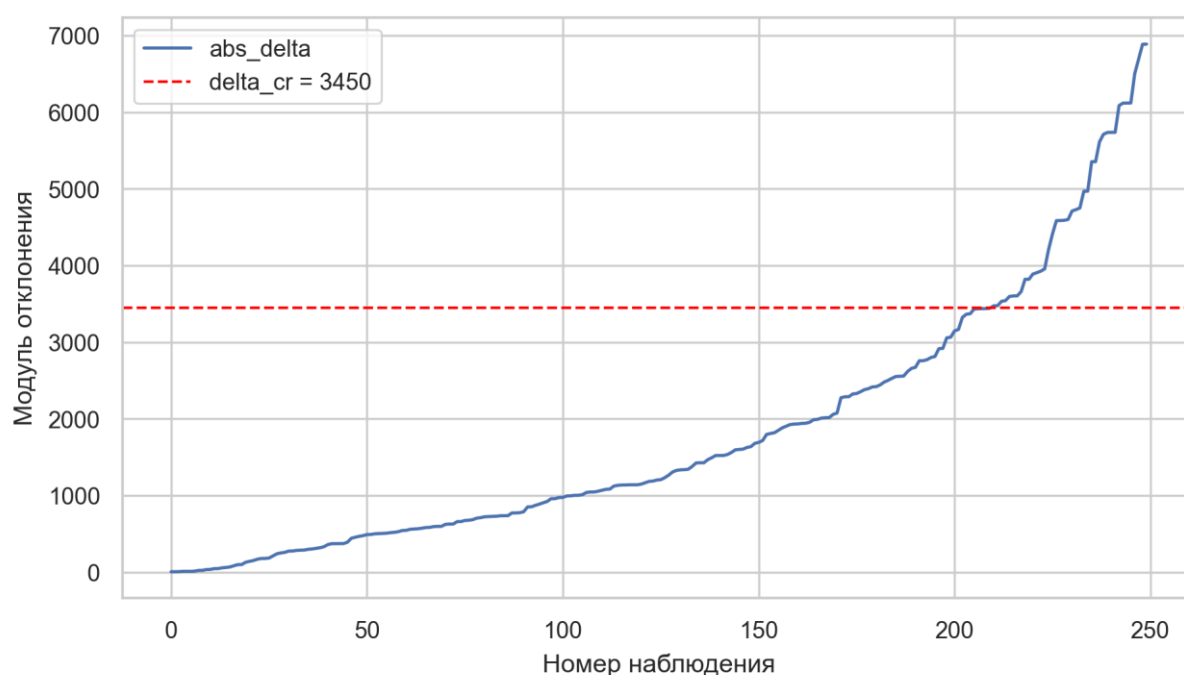


Рисунок 6.3 — Калибровочная кривая отклонений по времени в процессе маркетинга

Локтевая точка приходится на величину отклонения, равную примерно 3400. Это значение в дальнейшем мы будем считать критическим, а его превышение будет свидетельствовать об аномальном состоянии процесса.

6.3.3. Пилотная эксплуатация детектора аномалий по времени в процессе маркетинга

На этапе пилотной эксплуатации мы сводим все сведения, полученные выше, на этапах моделирования и калибровки. Наилучшая схема моделирования:

```
best_model = LinearRegression
autoregression_dive = 2
```

Критическая величина отклонения Δ_{cr} :

```
delta_cr = 3400
```

Мы применяем алгоритм пилотной эксплуатации, изложенный нами в приложении С.4, передав ему найденные нами значения 1) лучшей модели, 2) лучшей глубины авторегрессии и 3) критической величины отклонения. Для этапа пилотной эксплуатации мы используем набор из последних 415 наблюдений,

каждое из которых в каждой из локаций детектор диагностирует как нормальное или аномальное. Например, в локации GGL, обнаруживаем лишь одну аномалию за весь период пилотной эксплуатации⁷⁶ (см. рисунки 6.4 и 6.5).

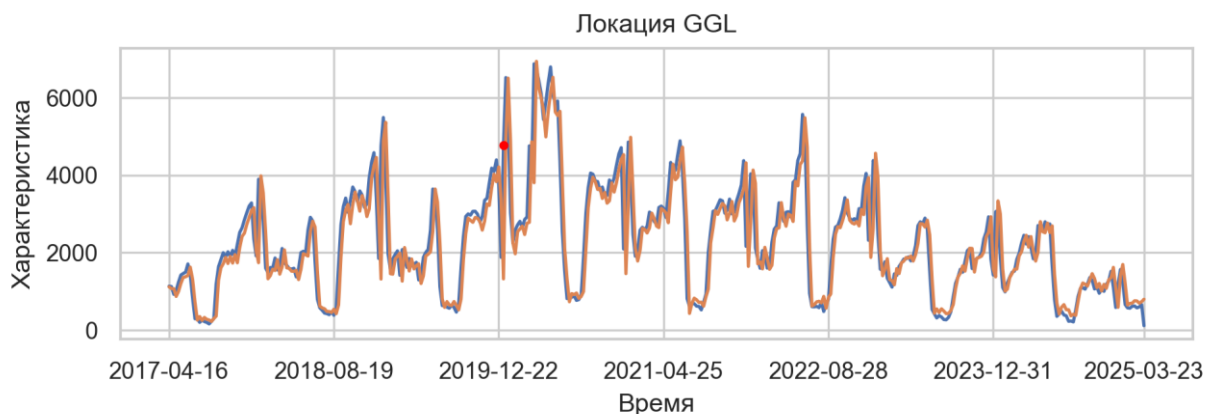


Рисунок 6.4 — Модель по времени в локации GGL



Рисунок 6.5 — Отклонения от модели по времени в локации GGL

Первое что бросается в глаза при взгляде на временной ряд локации GGL — ярко выраженные временные паттерны при общей нелинейности его тренда (см. рисунок 6.4). А кроме того, мы видим, что схема регрессионного моделирования, выбранная нами в качестве наилучшей, превосходно справляется с этим временным рядом, демонстрируя лишь одно аномальное отклонение. И даже это единственное аномальное отклонение оказывается едва выраженным, лишь совсем незначительно превышая критическое значение (см. рисунок 6.5).

Казалось бы — наша схема моделирования почти идеальна, но тогда по-

⁷⁶ Это говорит о высокой предсказуемости процесса.

чему у неё столь низкая метрика прогнозирующей способности (см. выше, раздел 6.3.1, где мы уже обращали на это внимание)? Ответ в том, что мы изучаем не одну локацию, а сразу несколько, среди которых есть плохо предсказуемая локация TG⁷⁷ (см. рисунки 6.6 и 6.7).

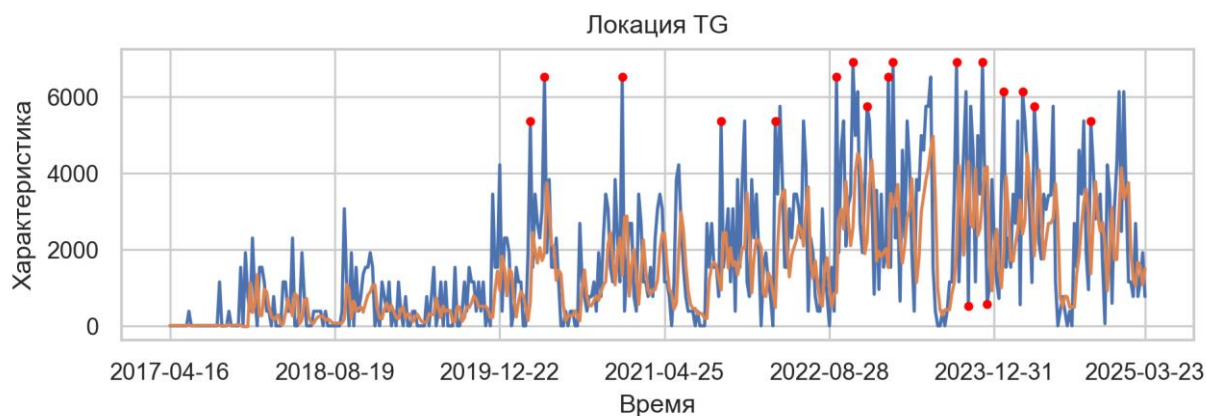


Рисунок 6.6 — Модель по времени в локации TG

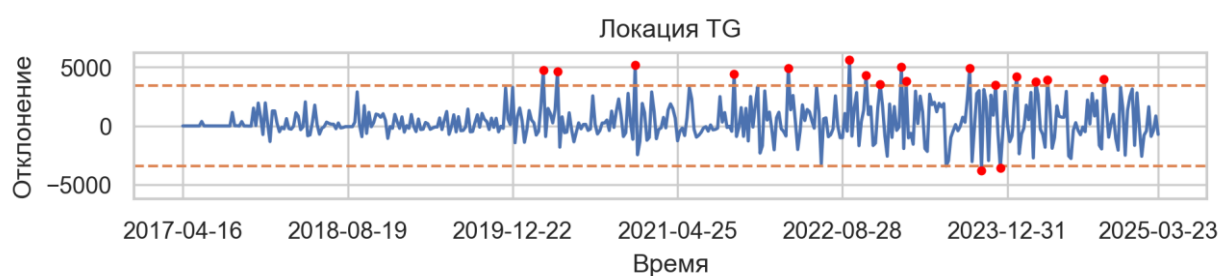


Рисунок 6.7 — Отклонения от модели по времени в локации TG

В локации Telegram мы наблюдаем сразу 18 аномалий, причём, они характеризуются отклонениями как в положительную, так и в отрицательную зону, и часть из них обладает весьма значительной выраженностью. Но самое главное — вариативность временного ряда в локации Telegram, начиная примерно с декабря 2019 года, резко возрастает, что и приводит к высокой степени непредсказуемости. Сочетание локаций с высокой (YouTube, Google) и низкой (VK и, в гораздо большей степени, Telegram) предсказуемостью приводит к тому, что на множестве всех изучаемых локаций наша схема регрессионного

⁷⁷ Причины, по которым эта локация обладает низкой предсказуемостью, требуют отдельного изучения и не входят в число наших задач.

моделирования даёт весьма посредственные результаты в смысле своей прогнозирующей эффективности.

Однако, нашей основной задачей является не создание идеальных моделей прогнозирования, а принятие эффективных управленческих решений в области мониторинга процесса маркетингового сопровождения автомобильного бренда. Нам нужно реагировать на аномальные состояния локаций в каждый момент наших наблюдений (напомним, что на этот раз моментами наблюдений выступают недели, а не дни, и поэтому у нас есть достаточное количество времени на реакцию). Поэтому в каждый момент мы формируем рейтинг аномалий, совокупность которых, распространённая на весь период пилотной эксплуатации даёт нам так называемый календарь аномалий. Например, в момент времени 2020-12-20 были зафиксированы аномалии в двух локациях, причём, аномалия в локации TG была более выраженной, нежели аномалия в локации VK (см. рисунок 6.8), а это означает, что при проведении аудита для выяснения причин зафиксированных аномалий нам нужно начинать именно с локации TG и лишь после этого переходить к локации VK (а оставшиеся три локации вообще не подвергать аудиту).

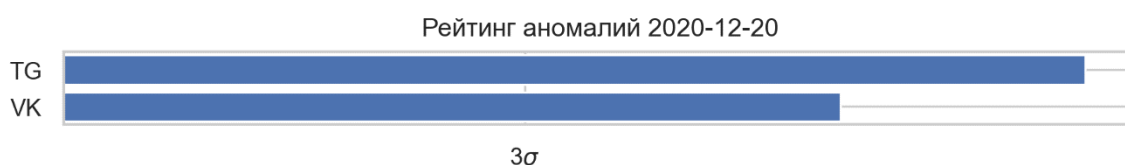


Рисунок 6.8 — Рейтинг аномальных по времени локаций в маркетинговом процессе в момент времени 2020-12-20

Отметим, что в целом в нашем процессе маркетингового сопровождения автомобильного бренда аномалии фиксируются не очень часто, большинство моментов наблюдения вообще обходится без аномалий, а если же аномалии фиксируются, то как правило, это бывает одна аномалия, из которой при всем желании не удастся получить полноценного рейтинга, и которая, как правило, приходится на локацию TG. Это объясняется тем, что, в отличие от предыдущих рассмотренных процессов (см. главу 4, где мы рассматривали процесс

производства, и главу 5, где анализировался процесс сервисного обслуживания), процесс маркетингового сопровождения бренда изучается нами только в пяти локациях (тогда как в процессе производства их было 24, а в сервисном процессе — вообще 69). Столь малое число локаций даже чисто теоретически не может демонстрировать высокое число аномалий, поэтому типичная страница календаря аномалий нашего процесса или оказывается пустой, или, в крайнем случае, содержит запись об одной зафиксированной в этот момент наблюдений аномалии (см. рисунок 6.9).

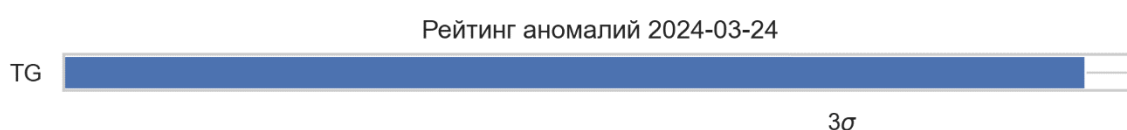


Рисунок 6.9 — Рейтинг аномальных по времени локаций в маркетинговом процессе в момент времени 2024-03-24

Календарь всех аномалий по времени, зафиксированных в течение последних десяти результативных наблюдений за процессом маркетингового сопровождения автомобильного бренда, представлен в приложении D.3.1.

6.4. Детекция аномалий по локациям в процессе маркетингового сопровождения

В этом разделе мы проведём процедуру детекции аномалий по локациям для процесса маркетингового сопровождения автомобильного бренда. При этом мы воспользуемся данными в формате LM (см. таблицу 6.2) и алгоритмами, изложенными в приложении С (см. приложения С.2, С.5).

6.4.1. Моделирование по локациям процесса маркетингового сопровождения

Для моделирования процесса маркетингового сопровождения мы используем пул из тех же семи моделей (см. раздел 6.3.1). Функция моделирования по времени описана нами в приложении С.2, используя ее, мы получаем следующие результаты (см. рисунок 6.10).

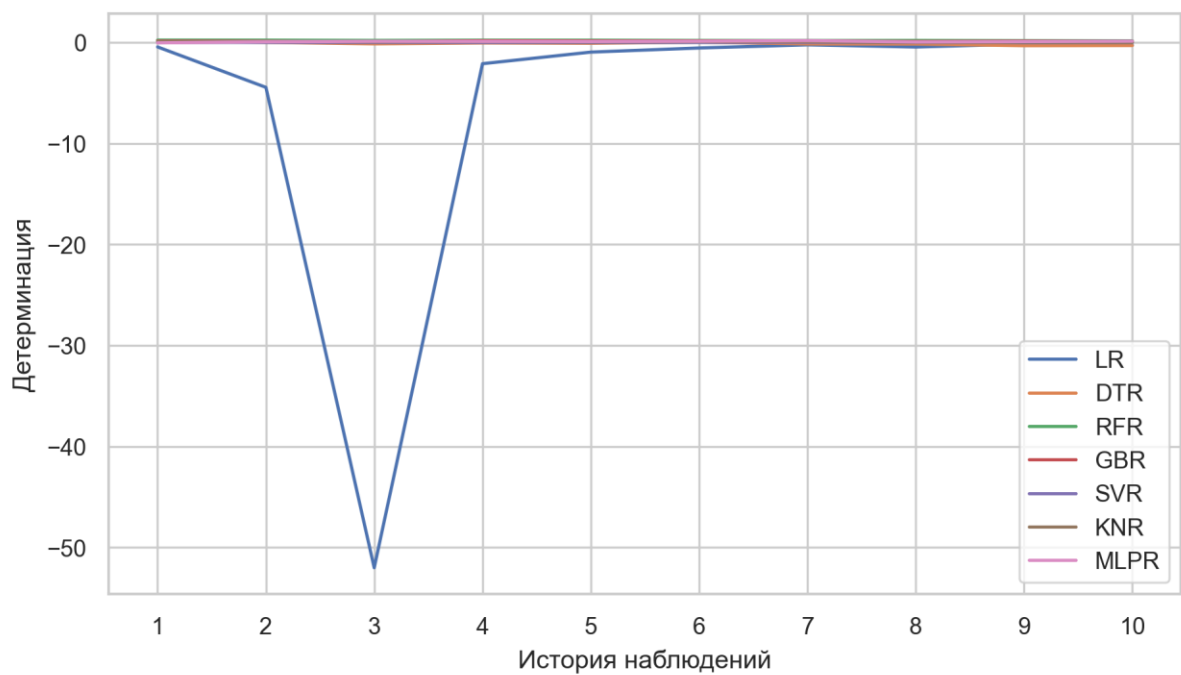


Рисунок 6.10 — Сравнительный анализ схем моделирования по локациям в процессе маркетингового сопровождения

Масштабируем иллюстрацию и в её укрупненном варианте лучшая схема моделирования становится очевидной (см. рисунок 6.11).

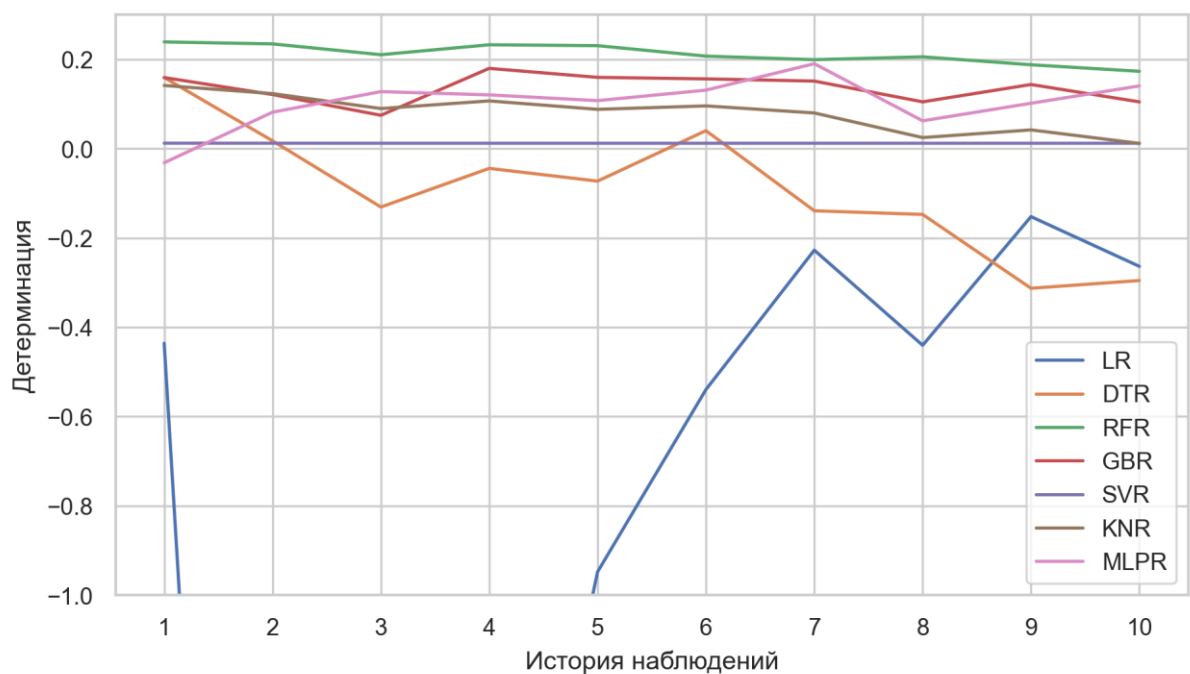


Рисунок 6.11 — Приближенный скоринг моделей по локациям в процессе маркетингового сопровождения

Наилучший результат демонстрирует древовидная модель случайного леса `RandomForestRegressor`, глубина истории наблюдений равна 1. Именно эту схему моделирования мы будем использовать дальше, на этапах калибровки и пилотной эксплуатации нашего детектора аномалий.

6.4.2. Калибровка отклонений по локациям в процессе маркетингового сопровождения

Экспертное мнение о нормальности или аномальности состояний процесса маркетингового сопровождения нам, к сожалению, недоступно (так же как и во всех предыдущих наших исследованиях).

Поэтому мы упорядочиваем абсолютные величины всех отклонений прогнозов, данных в соответствии с наилучшей прогнозирующей схемой, от истинных значений характеристики по возрастанию и проводим калибровку детектора аномалий по локтевой точке на получающейся в результате этого калибровочной кривой (напомним, что она определяется как точка смены тренда, подробнее об этом см. выше, главу 4, раздел 4.3.2). Сама кривая представлена на иллюстрации ниже (см. рисунок 6.12).

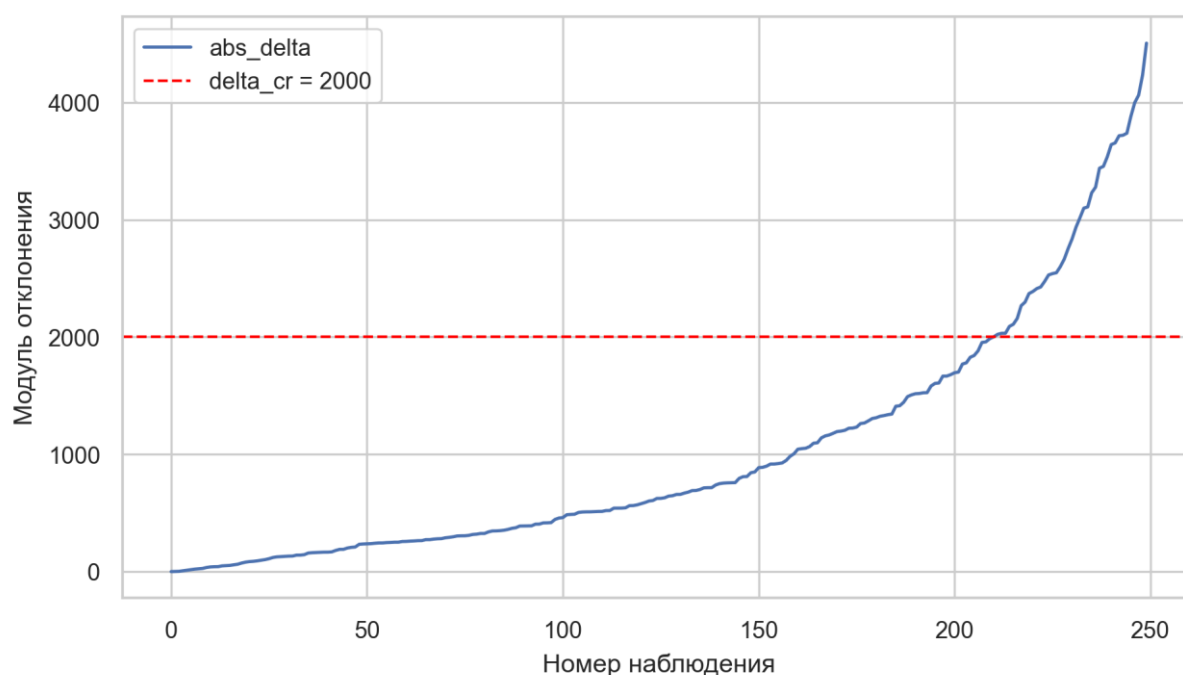


Рисунок 6.12 — Калибровочная кривая отклонений по локациям в процессе маркетингового сопровождения

Локтевая точка приходится на величину отклонения, равную примерно 2000. Тем самым, мы нашли значение, которое в дальнейшем будем считать критическим, а его превышение станет маркером аномального состояния изучаемого нами процесса.

6.4.3. Пилотная эксплуатация детектора аномалий по локациям в процессе маркетингового сопровождения

Теперь у нас есть все необходимые сведения для пилотной эксплуатации детектора. Во-первых, мы знаем, в чем именно состоит наилучшая схема моделирования процесса — нам нужно проводить регрессию при помощи случайного леса, используя глубину истории наблюдений, равную одному наблюдению.

```
best_model = RandomForestRegressor  
history_dive = 1
```

А кроме того, мы знаем величину критического отклонения Θ_{cr} :

```
teta_cr = 2000
```

Мы применяем алгоритм пилотной эксплуатации детектора по локациям, изложенный нами в приложении С.5, передав ему найденные нами значения 1) лучшей модели `best_model`, 2) лучшей глубины истории наблюдений `history_dive` и 3) критической величины отклонения `teta_cr`. Для этапа пилотной эксплуатации мы используем набор из последних 415 наблюдений, каждое из которых в каждой из локаций детектор диагностирует как нормальное или аномальное.

Выше (см. раздел 6.3) мы в качестве примера для иллюстрации работы детектора по времени использовали локации GGL и TG. Теперь мы рассмотрим те же локации с точки зрения детектора по локациям. Начнём с локации GGL (см. рисунки 6.13 и 6.14).

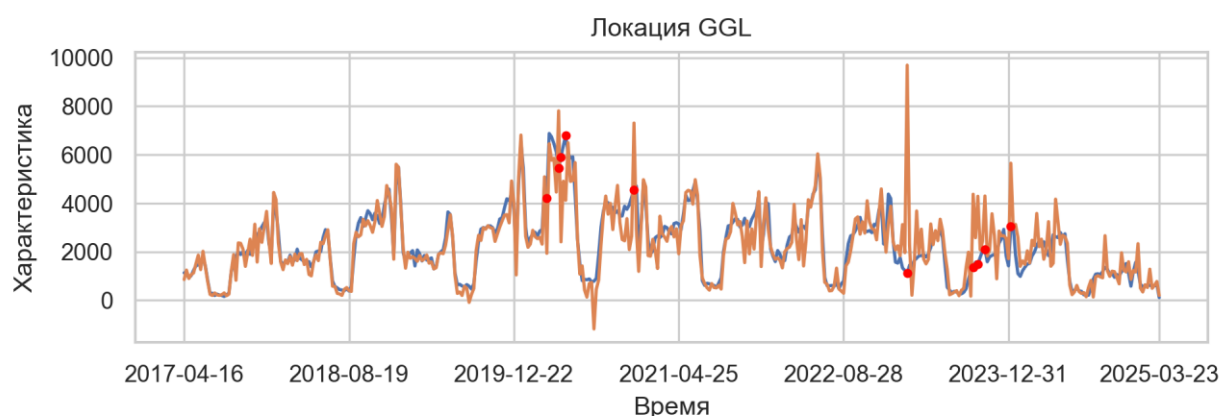


Рисунок 6.13 — Модель по локациям в локации GGL

Видим, что модель по локациям по-прежнему неплохо справляется со своей задачей: она вполне уверенно предсказывает значения наблюдаемой характеристики, хотя её качество заметно уступает модели по времени (см. для сравнения рисунок 6.4). Кроме того, мы видим ещё одно отличие: при моделировании по времени детектор вернул только одну аномалию (см. рисунки 6.4 и 6.5), теперь же, при проведении моделирования по локациям аномальными оказались сразу десять моментов наблюдения.

Выраженность отклонений представлена на иллюстрации ниже (см. рисунок 6.14), причём, она гораздо выше, чем это было при моделировании по времени⁷⁸.



Рисунок 6.14 — Отклонения от модели по локациям в локации GGL

Но все ещё хуже во второй локации, которую мы разбирали выше (см. раздел 6.3) — в локации TG (см. рисунки 6.15 и 6.16).

⁷⁸ См. для сравнения рисунок 6.5.

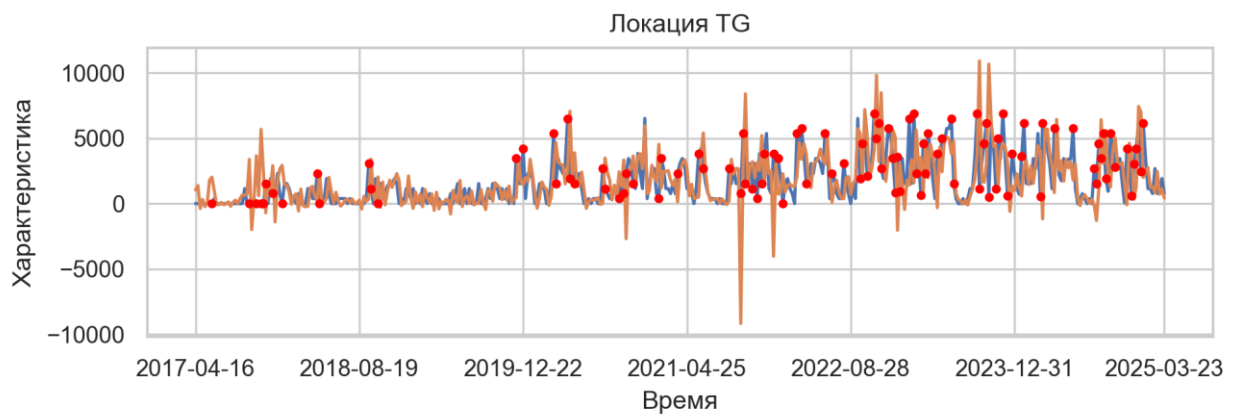


Рисунок 6.15 — Модель по локациям в локации TG



Рисунок 6.16 — Отклонения от модели по локациям в локации TG

Число зафиксированных детектором по локациям аномальных состояний настолько велико, что даже с учётом всех пяти исследуемых нами локаций общая доля аномальных состояний среди всех наблюдений составляет величину порядка 9% (для сравнения при исследовании процессов производства эта доля составляла порядка 1.5%–3%, а для процессов сервисного обслуживания — около 5%). Это, безусловно, свидетельствует о том, локация TG вносит свой негативный вклад в предсказуемость протекания процесса маркетингового сопровождения⁷⁹.

Однако для наших целей важны не общие рассуждения, а конкретные знания об отклонениях от нормы, которые нам необходимо получать в режиме реального времени для того, чтобы оперативно на них реагировать. Поэтому в каждый момент наблюдений мы формируем рейтинг аномалий. Например, в момент времени 2023-01-15 мы наблюдали аномалии сразу в трёх локациях

⁷⁹ Хотя и не говорит о причинах такого положения (см. об этом ниже в разделе 6.5).

(чего не наблюдалось ни разу на всем этапе пилотной эксплуатации детектора аномалий по времени): это локации TG, VK и YT (см. рисунок 6.17).



Рисунок 6.17 — Рейтинг аномальных по локациям локаций в процессе маркетингового сопровождения в момент времени 2023-01-15

Выраженность аномалий диктует и порядок реагирования на них: в первую очередь нужно выяснять обстоятельства протекания маркетингового процесса в локации TG, аномальность которой намного превосходит психологически значимый барьер 3σ , и только после этого переходить к локациям VK и YT.

В момент времени 24-10-13 мы видим две аномалии: в локациях VK и TG (см. рисунок 6.18) и их выраженности.

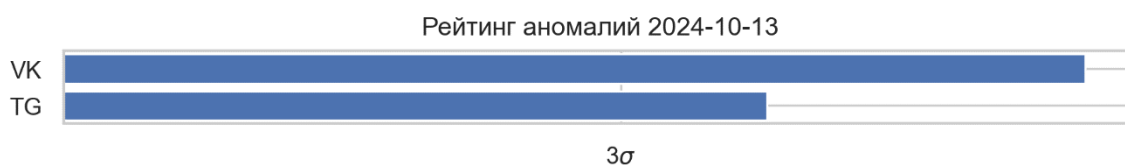


Рисунок 6.18 — Рейтинг аномальных по локациям локаций в процессе маркетингового сопровождения в момент времени 2024-10-13

Календарь всех аномалий по локациям, зафиксированных в течение последних десяти результативных наблюдений за процессом маркетингового сопровождения, представлен в приложении D.3.2.

6.5. Об интерпретации аномалий, выявленных в процессе маркетингового сопровождения

В современном мире, где конкуренция на автомобильном рынке достигает пиковых значений, своевременное выявление аномальных состояний в процессе

маркетингового сопровождения бренда становится критически важным фактором успеха. Эффективная маркетинговая стратегия — это не статичный план, а динамичный процесс, требующий постоянного мониторинга и корректировки, и любые отклонения от нормы, будь то падение трафика на сайте, снижение вовлеченности в социальных сетях или негативные отзывы в СМИ, могут сигнализировать о зарождающихся проблемах.

Своевременное обнаружение этих аномалий позволяет оперативно реагировать на изменения в потребительском поведении и рыночной конъюнктуре, так как анализ данных о продажах, запросах в поисковых системах и активности в онлайн-сообществах предоставляет ценную информацию о предпочтениях и настроениях потенциальных покупателей, а игнорирование этих сигналов может привести к упущенным возможностям, снижению лояльности клиентов и, в конечном итоге, к ухудшению финансовых показателей бренда. Более того, выявление аномалий позволяет не только предотвратить негативные последствия, но и обнаружить новые тренды и перспективные направления развития⁸⁰. В конечном счёте, своевременное выявление аномальных состояний в процессе маркетингового сопровождения автомобильного бренда — это залог его устойчивого развития и конкурентоспособности в условиях быстро меняющегося рынка.

Вместе с тем, формальная констатация аномалий в маркетинговом процессе, без глубокого анализа причин их возникновения, превращается в пустую профанацию, не приносящей реальной пользы автомобильному бренду. Простое указание на отклонения от нормы, без понимания движущих сил, породивших эти отклонения — это нечто сродни диагностике болезни без проведения анализов: такой подход не только не способствует повышению качества маркетингового сопровождения, но и потенциально вредит ему. Во-пер-

⁸⁰ Например, резкий рост интереса к определённой модели автомобиля или технологии может стать отправной точкой для разработки новых маркетинговых кампаний и продуктов.

вых, он создаёт иллюзию контроля над ситуацией, отвлекая ресурсы от действительно важных направлений. А во-вторых, он может привести к принятию ошибочных управленческих решений, основанных на поверхностных наблюдениях, а не на глубоком понимании проблемы.

Рассмотрим, например, ситуацию, когда наблюдается объективное снижение трафика на сайт автомобильного бренда. Формальная констатация этого факта, без выяснения причин, может привести к ошибочному выводу о неэффективности рекламной кампании и её срочной переработке. Однако, реальная причина может крыться в изменении алгоритмов поисковых систем или в активности конкурентов, которые запустили более агрессивную рекламную кампанию, и, вместо того, чтобы бросаться переделывать рекламу, необходимо провести анализ, выявить истинные причины падения трафика и, возможно, скорректировать SEO-оптимизацию сайта или разработать новую стратегию контент-маркетинга. Только такой подход позволит эффективно решить проблему и повысить качество маркетингового сопровождения бренда.

Игнорирование причин аномалий ведёт к весьма поверхностному реагированию на проблемы, что в долгосрочной перспективе подрывает доверие к маркетинговой команде и бренду в целом, в тот момент как комплексный подход, включающий анализ, выявление причин и разработку соответствующих решений, является единственным путём к устойчивому успеху в конкурентной автомобильной индустрии. В конечном счёте, качественное маркетинговое сопровождение требует не просто фиксации проблем, а глубокого понимания их природы и разработки эффективных стратегий для их решения.

Рассмотрим полученные нами выше результаты (см. разделы 6.3 и 6.4). В чем причина столь непредсказуемого поведения локации Telegram? Мы можем выдвинуть целый ряд предположений.

Гипотеза 1. Высокая вариативность временного ряда в этой локации объясняется несовершенством рекомендательной системы Telegram: в какие-то моменты система даёт точное попадание в целевую аудиторию, и тогда число

переходов на канал автомобильного бренда резко возрастает, в другие моменты попадания нет, и число переходов резко падает.

Гипотеза 2. Причина в том, что маркетинговая кампания проводится с использованием разных креативных решений. Те решения, которые отвечают интересам пользователей Telegram, приводят к резкому повышению числа переходов, но другие решения настолько неинтересны, что пользователи на них вообще не реагируют, и как следствие, число переходов резко падает.

Гипотеза 3. Все дело в нестабильной работе операторов мобильной связи. Так как переходы на канал бренда из локации Telegram всегда или почти всегда осуществляются именно с мобильных устройств, качество мобильной связи, ее стабильность и устойчивость оказываются основными причинами вариативности трафика.

Но все наши гипотезы являются не более чем предположениями до тех пор, пока мы их не доказали (или не опровергли). Интерпретация результатов анализа данных о маркетинговом процессе — задача, требующая глубокого погружения в нюансы отрасли и понимания психологии потребителя. Формальные методы анализа данных, безусловно, важны, однако они не могут заменить мнение специалиста: то есть, маркетолога, обладающего контекстуальными знаниями.

Маркетинговый процесс — это сложная экосистема, включающая в себя множество переменных (см., например, раздел 6.1, где мы указывали множество его различных аспектов), от креативных кампаний до каналов распространения и ценообразования. Анализ данных предоставляет лишь сухие цифры, в то время как маркетолог способен расшифровать их значение в контексте конкретной ситуации. Например, снижение конверсии рекламной кампании может быть вызвано не только неэффективным таргетингом, но и сезонными колебаниями спроса, изменениями в поведении конкурентов или даже неудачным креативным решением. Специалист по анализу данных может выявить

лишь факт снижения, в то время как маркетолог сможет определить его причину (впрочем, может и не определить). Более того, интерпретация данных требует понимания целей и стратегии компании. Маркетолог знает, какие показатели являются ключевыми для достижения бизнес-целей, и может оценить результаты анализа с этой точки зрения.

Специалисты по анализу данных, безусловно, обладают ценными навыками, но их работа всегда должна рассматриваться как инструмент, помогающий маркетологам принимать более обоснованные решения. Они могут выявлять тренды, закономерности и аномалии, но именно маркетолог должен объяснить их значение и предложить конкретные действия. В конечном счёте, успешная интерпретация данных о маркетинговом процессе требует симбиоза аналитических навыков и экспертных знаний в области маркетинга. Только в этом случае можно получить действительно ценные инсайты и разработать эффективные стратегии для достижения бизнес-целей, и маркетологи, обладающие глубоким пониманием специфики своей отрасли, являются ключевым звеном в этом процессе.

6.6. Выводы по главе 6

В этой главе мы замкнули сформулированную в самом начале настоящей диссертационной работы идею (см. главу 1, раздел 1.3.1) о том, что любой продукт можно рассматривать как совокупность процессов, относящихся к трём группам: производства (рассмотрены в главе 4), сервиса (рассмотрены в главе 5) и маркетинга (рассмотрены в настоящей главе).

Подводя краткие итоги главы, отметим следующее.

1. Процессы маркетинга весьма разнообразны, и в этом смысле они ничуть не уступают процессам производства или сервисным процессам. Мы рассматриваем лишь очень частный вариант маркетингового процесса, связанный с просмотрами нарративного контента.

2. Работа по методике МССР (Modeling, Challenge, Calibration, Production) показывает, что эта методика вполне способна справиться с данными о существенно нестабильном процессе, в котором присутствуют ярко выраженные временные паттерны.
3. Выявленные в ходе этого исследования аномальные состояния процесса маркетинга требуют серьёзного дополнительного анализа и корректной интерпретации.
4. Установление причин алгоритмически выявленных аномалий и включение соответствующих протоколов реагирования на них должно осуществляться представителями предметной области и выходит за рамки компетенций аналитика данных.

Таким образом, проведённая в этой главе процедура детекции аномалий демонстрирует применимость методики МССР (Modeling, Challenge, Calibration, Production) к отдельным процессам в группе процессов маркетингового сопровождения. С практической точки зрения выявление таких несоответствий позволяет оперативно реагировать на нежелательные состояния в маркетинговых процессах и своевременно их корректировать.

ГЛАВА 7. ИНТЕГРАЦИЯ МЕТОДОЛОГИИ И ИНСТРУМЕНТАРИЯ ДЕТЕКЦИИ АНОМАЛИЙ В ТЕОРИЮ И ПРАКТИКУ УПРАВЛЕНИЯ КАЧЕСТВОМ

Предложенная нами в настоящей диссертационной работе методика МССР (Modeling, Challenge, Calibration, Production) позволяет алгоритмически выявлять аномалии в процессах, являющихся нестабильными, таких как 1) процессы производства, рассматриваемые с социальной точки зрения (см. главу 4), 2) процессы сервисного обслуживания (см. главу 5) и 3) процессы маркетингового сопровождения (см. главу 6). Точки её приложения выходят за рамки управления качеством, в силу чего эта методика представляет самостоятельный научный интерес. Например, научные наблюдения за нестабильными процессами, в таких областях как медицина, экономика, экология, социология и т. д. — вполне допускают применение этой методики для выявления аномальных состояний и их последующего анализа с использованием инструментария соответствующих предметных областей. Более того, методика МССР (Modeling, Challenge, Calibration, Production) может быть применена к процессам, семантика которых вообще неизвестна исследователю, примером чего может служить её реализация на синтетических данных (см. главу 3), которые являются абстрактными наборами чисел, не привязаны ни к какой предметной области и, с этой точки зрения, вообще лишены какого-либо смысла. Но какое отношение все это имеет к управлению качеством как к научной и, что ещё более существенно, как к научно-практической дисциплине? И, прежде всего, — каково практическое приложение предложенных в настоящей диссертационной работе идей и методов?

Выявление аномалий в ключевых бизнес-процессах предприятия (производстве, сервисе и маркетинге) является важным фактором, определяющим результативность и эффективность аудита качества [27, 31, 72]. Аномалии, рассматриваемые как отклонения от устоявшихся паттернов функционирования процессов, представляют собой ценный источник информации для целеполагания аудиторских процедур. В контексте управления качеством, эффективный аудит базируется на ряде принципов, обеспечиваемых детекцией аномалий.

Прежде всего, это так называемый принцип RBT (Risk-Based Thinking, англ. — мышление, основанное на рисках), состоящий в том, что идентификация аномалий позволяет аудиторам применять методологию риск-ориентированного аудита, фокусируя ресурсы на процессах с наибольшим потенциалом отклонений от нормы [36, 72, 88, 89]. Это повышает оперативность и релевантность аудиторских выводов.

Кроме того, посредством выявления и анализа аномальных ситуаций реализуется цикл PDCA (Plan-Do-Check-Act, англ. — план, исполнение, проверка, действие), лежащий в основе философии непрерывного улучшения⁸¹. Аномалии инициируют анализ первопричин, разработку корректирующих и предупреждающих действий, что способствует повышению предсказуемости бизнес-процессов [36, 88, 151, 202].

Оперативная детекция аномальных состояний является критически важным элементом превентивного управления качеством и реализуется посредством:

1. Статистического управления процессами SPC (Statistical Process Control, англ. — статистическое управление процессами) [38, 56, 154, 236]. Использование статистических методов⁸² позволяет мониторить вариабельность

⁸¹ Этому будет посвящен отдельный раздел настоящей главы, (см. ниже, раздел 7.2)

⁸² Для стабильных процессов это контрольные карты. Для нестабильных процессов, являю-

процессов и выявлять отклонения, указывающие на возникновение особой причины отклонений.

2. Систем раннего предупреждения EWS (Early Warning Systems, англ. — системы раннего предупреждения) [27, 69, 151]. EWS представляют собой комплексы инструментов, предназначенных для мониторинга ключевых параметров процессов и выдачи оповещений при обнаружении⁸³.
3. Бизнес-аналитики BI (Business Intelligence, англ. — бизнес-аналитика) [69]. BI-системы интегрируют данные из различных источников и предоставляют возможности для визуализации и анализа трендов, выявления аномалий и поддержки принятия управленческих решений на основе данных⁸⁴.

Оперативное выявление аномалий играет центральную роль в обеспечении эффективного управления качеством, поскольку обеспечивает переход от реактивной стратегии исправления ошибок к проактивной стратегии предотвращения дефектов [36, 82, 88, 151, 175]. Интеграция методологий SPC, EWS и BI в систему управления качеством позволяет минимизировать риски отклонений и повысить предсказуемость процессов.

В этой главе мы рассмотрим некоторые практические и теоретические аспекты концепции детекции аномалий и основанной на ней методике МССР (Modeling, Challenge, Calibration, Production) с позиций управления качеством, разберем аспекты практического применения методологии алгоритмической детекции аномалий и опишем место, которое она занимает в общей теории и практике управления качеством.

7.1. Управление нестабильными процессами на основе выявленных

щихся предметом изучения настоящей диссертации, это методика МССР (Modeling, Challenge, Calibration, Production), рассмотренная выше (см. главы 1–2).

⁸³ В рамках настоящей диссертации EWS — это система алгоритмической детекции аномалий, которая относит любое состояние бизнес-процесса к одному из двух классов норма/аномалия.

⁸⁴ В качестве элементов BI в настоящей диссертации выступают рейтинга аномалий, которые формируются в режиме пилотной эксплуатации нашего детектора.

аномалий

Важным аспектом управления процессами (вне зависимости от того, являются ли они стабильными, или же они нестабильны) служит их аудит, проверка на предмет соответствия нормальному протеканию, когда процесс не требует операционального вмешательства, а в случае обнаруженного несоответствия — выяснения причин такой аномалии и устранения его причин [27, 36, 63, 64, 72, 87, 89, 151]. В этом разделе мы покажем, как предложенная в данной диссертационной работе концепция детекции аномалий в нестабильных процессах может быть применена в процедурах их аудита.

7.1.1. Неформальный пример

В качестве основы нашего неформального примера мы возьмём ситуацию, которую рассматривали в главе 4, где мы изучали процесс исполнения одной из технологических операций в сборке автомобиля как социальный процесс, мерой эффективности которого выступало число операций, выполненных в течение рабочего дня тем или иным сотрудником. Только пусть теперь нам необходимо отслеживать работу не 24 подчинённых, а гораздо большего их количества, например, пусть их будет 240 или даже 2 400 (для алгоритма детекции аномалий это все равно несущественно, так как алгоритм с равным успехом обработает данные как о 24, так и о 24 000 локациях процесса).

Наша задача — детектировать аномалии в работе каждого из наших сотрудников и в конечном счёте, добиться того, чтобы таких несоответствий не было вообще: тогда наша работа по управлению процессом будет считаться выполненной. И хотя, мы, разумеется, понимаем, что полного устранения аномалий нам добиться не удастся, мы стремимся к тому, что существенно минимизировать их число. Решением этой управленческой задачи служит «разбор полётов», состоящий в том, что по результатам очередного наблюдения (которое в данной ситуации фиксируется в конце каждой рабочей смены) мы под-

водим итоги работы нашего коллектива, и выясняем, в чем причины аномального результата у того или иного сотрудника.

Рассмотрим несколько сценариев решения нашей задачи.

Сценарий 1. Полный контроль процесса. Действуя по этому сценарию, мы анализируем результаты деятельности каждого нашего сотрудника, сколько бы их у нас ни было. Если их 240, то, уделяя каждому из них 1 минуту, мы потратим на это 4 часа, а уделяя каждому 30 секунд — 2 часа, и это все равно слишком много, потому что в таком случае для проведения своего «разбора полётов» мы вынужденно увеличиваем продолжительность рабочего дня на 2 часа, а это является грубейшим нарушением трудового кодекса. Таким образом, сценарий полного контроля является в высшей степени затратным, и даже если закрыть глаза на его несоответствие ТК РФ, его исполнение требует слишком больших ресурсов: временных, кадровых, а если всё-таки не нарушать законодательство и осуществлять выплаты своим сотрудникам за переработки, — то ещё и финансовых. А что будет, если число наших сотрудников не 24 и даже не 240, а хотя бы 2 400? Ответ — тогда мы вообще не сможем осуществить этот сценарий, другими словами, печальная истина состоит в том, что, полный контроль процесса мы сможем осуществить только в небольшом трудовом коллективе.

Итак, минус полного контроля — недопустимо высокие затраты. Но у него есть и важный плюс: осуществляя полный контроль, мы действительно полностью контролируем процесс, за который несём ответственность.

Сценарий 2. Выборочный контроль процесса. Отказавшись от полного контроля, мы решаем пойти другим путём и использовать для контроля нашего процесса старый добрый выборочный метод. Мы, как и выше, проводим «разбор полётов», только на этот раз мы оставляем на разбор не весь коллектив, а небольшую случайную выборку, допустим, 5 человек. В чем преимущество такого подхода? — в том, что мы сразу же нивелируем основной недостаток полного контроля, состоящий в неоправданно высоких затратах: на 5 человек

мы потратим 5 минут⁸⁵. А в чем его недостаток? — в том, что мы сразу же теряем единственное преимущество, которое давал нам полный контроль: теперь мы уже не можем утверждать, что выявили все аномалии в процессе, который должны контролировать, потому что мы выявили лишь те аномалии, которые случайно попали в нашу выборку. Итак, плюс выборочного контроля состоит в том, что он не требует слишком высоких затрат. Но у него есть и существенный минус: осуществляя выборочный контроль процесса, мы не можем быть уверенными в том, что процесс протекает нормально во всех подконтрольных локациях.

Сценарий 3. Контроль на основании детектированных аномалий.

Теперь предположим, что мы уже внедрили в наш процесс детектор аномалий, который по итогам каждого наблюдения (в нашем случае — по итогам каждого рабочего дня) возвращает нам список локаций (в нашем случае — список сотрудников), в которых были детектированы аномалии, и, более того, упорядочивает локации в соответствии с выраженностью аномалии от наиболее выраженных к наименее выраженным. Другими словами, наш детектор в конце каждого рабочего дня предоставляет нам рейтинг выявленных аномалий, например, такой, что изображён на иллюстрации ниже⁸⁶ (см. рисунок 7.1).

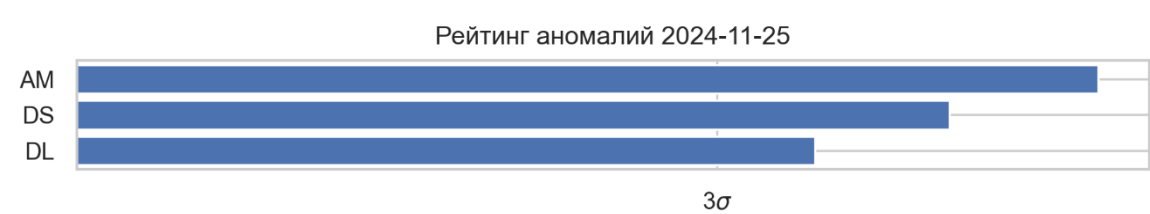


Рисунок 7.1 — 2024-04-04. Рейтинг аномальных локаций

⁸⁵ Исходя из норматива 1 минута на разбор 1 сотрудника, который, разумеется ничем не обоснован. Мы используем его здесь всего лишь для примера, чтобы подчеркнуть разницу между полным и выборочным контролем.

⁸⁶ Заметим, что мы используем иллюстрацию, которую уже использовали в главе 4 (это буквальное воспроизведение рисунка 4.9), когда изучали протекание процесса в 24 локациях. Если же число локаций больше (а в рамках этого примера мы предполагаем, что это именно так), то число выявленных аномалий тоже увеличится кратно.

Будучи приверженцами «разбора полётов», мы продолжаем нашу политику, вдумчиво и неторопливо беседуя со своими сотрудниками, только, в отличие от сценария полного контроля не со всеми, и, в отличие от сценария выборочного контроля не со случайно отобранными, а конкретно с операторами AM, DS и DL, про которых мы заранее, ещё до разговора знаем, что сегодня, 4 апреля 2024 года их работа не вполне соответствовала нормальному режиму её исполнения. Таким образом, на наш «разбор полётов» в конце смены мы оставляем только трёх человек, и, более того, сам разбор мы проводим в чётко определённой последовательности: сначала мы разбираем ситуацию сотрудника AM (так как аномальность его работы наиболее ярко выражена, о чем свидетельствует рейтинг на рисунке 7.1), затем — DS, и только после этого — DL (так как именно в таком порядке они следуют в рейтинге).

Преимущество такого подхода состоит в том, что он одновременно нивелирует недостатки как полного контроля — он не является затратным, так как не требует разбора деятельности абсолютно всех работников, так и выборочного контроля — он не содержит пропущенных аномалий, так как все аномалии уже выявлены помимо наших усилий, они известны заранее, благодаря работе алгоритма детекции аномалий.

В рамках этого примера можно отчётливо проследить, какие роли мы выделяем в качестве основных ролей применительно к нашей концепции детекции аномалий в нестабильных процессах, а именно: мы рассматриваем контроль нестабильных процессов с точки зрения ответственности за его нормальное протекание (без эксцессов, или, в нашей терминологии — без аномалий). С этой точки зрения, есть некая позиция А (это может быть конкретное лицо, либо подразделение, либо, выражаясь в более общих категориях, есть некая функция), ответственная за протекание процесса, и кроме того есть некие позиции В (и это тоже какие-то лица или подразделения, а в терминологии настоящей диссертационной работы — это локации, в которых протекает наш процесс), являющиеся исполнителями процесса, и они несут ответственность перед позицией А.

Контроль всего процесса со стороны А состоит в том, что применяется алгоритмическая процедура детекции аномалий, а управление со стороны А состоит в том, что по отношению к тем позициям В, в которых детектор выявил какие-то аномалии, А принимает меры по устранению причин аномалии, если она нежелательна, или, наоборот, меры по распространению причин выявленной аномалии на все остальные локации, если аномалия желательна⁸⁷. Таким образом, в нашем представлении об управлении нестабильными процессами ключевым является понятие ответственности: наблюдатель процесса несёт общую ответственность за весь процесс в целом; локации, в свою очередь, несут ответственность перед наблюдателем; а наша задача состоит в том, чтобы действия наблюдателя были наиболее эффективными и при этом наименее затратными.

7.1.2. Расчёт прироста эффективности аудита

Наиболее эффективными действия наблюдателя будут в случае, когда он выявляет максимально возможное количество аномалий (заметим, что не максимальное вообще, а максимально возможное в наблюдаемой им ситуации, так как иначе, он начнёт объявлять аномалиями всё подряд, не заботясь об истине, а заботясь лишь о своём KPI). А наименее затратными — когда для достижения этого результата он будет исполнять минимальное число проверок. В этом разделе мы покажем, как можно вычислить эффект от внедрения в некий подотчётный нам нестабильный процесс процедур детекции аномалий.

Напомним, что отрицательное гипергеометрическое распределение —

⁸⁷ Заметим, что такое тоже вполне возможно. Например, если в процессе маркетингового сопровождения (см. главу 6) мы обнаружим аномально высокий показатель просмотров в какой-либо из наших локаций (это и есть желательная аномалия), то мы, разумеется, предпримем все меры для того, чтобы распространить это явление и на другие источники трафика. Сможем мы это сделать, или нет — это другой вопрос. Но такая попытка с нашей стороны обязательна, так как именно в этом и состоит наша главная функция: мы должны добиться того, чтобы процесс, за протекание которого мы несём ответственность, протекал как можно более эффективно.

это распределение дискретной случайной величины X , которое определяется следующим образом:

$$P\{X = i\} = \frac{C_{i+m-1}^i C_{N-m-i}^{M-m}}{C_N^M}, \quad 0 \leq i \leq N - M.$$

параметры N, M, m — целые неотрицательные числа, удовлетворяющие условию $m \leq M \leq N$ [6, 20, 83, 169]. Отрицательное гипергеометрическое распределение обычно возникает в схеме выбора без возвращения. Если в некоторой генеральной совокупности объёма N имеется M «отмеченных» и $N - M$ «неотмеченных» элементов, и если последовательный случайный выбор (без возвращения) производится до тех пор, пока число «отмеченных» элементов в выборке не достигает фиксированного числа m , то случайная величина X — число «неотмеченных» элементов в такой выборке — подчиняется отрицательному гипергеометрическому распределению. Отрицательное гипергеометрическое распределение несимметрично. Его математическое ожидание и дисперсия вычисляются по формулам:

$$\mathbb{E}(X) = m \frac{N - M}{M + 1},$$

$$\mathbb{D}(X) = m \frac{(N + 1)(N - M)}{(M + 1)(M + 2)} \left(1 - \frac{m}{M + 1} \right).$$

Отметим, что случайная величина $X' = X + m$, то есть, объём выборки — также имеет отрицательное гипергеометрическое распределение, причём, поскольку X' получается из X смещением на константу, новая случайная величина подчиняется тому же закону распределения, хотя он и распространяется на другие, смещённые относительно исходного ряда значения:

$$P\{X' = i\} = \frac{C_{i+m-1}^i C_{N-m-i}^{M-m}}{C_N^M}, \quad m \leq i \leq N - M + m.$$

Математическое ожидание смещения равно смещению математического ожидания, поэтому:

$$\mathbb{E}(X') = m \frac{N - M}{M + 1} + m = m \frac{N + 1}{M + 1},$$

а дисперсия под действием смещения не изменяется:

$$\mathbb{D}(X') = \mathbb{D}(X).$$

Теперь, пусть у нас имеется N наблюдений за состояниями какого-либо нестабильного процесса, среди которых M состояний являются аномальными, пусть мы применили к наблюдаемым данным наш детектор, и пусть он пометил как аномальные n состояний из N , причём, когда мы подвергли их проверке, среди них обнаружилось m аномалий. Таким образом, на то, чтобы обнаружить m аномалий мы потратили n проверок. Зададимся вопросом: сколько нам пришлось бы потратить проверок для того, чтобы обнаружить то же число аномалий в исходной совокупности наблюдений? Ответ состоит в том, что число необходимых для этого проверок — это математическое ожидание рассмотренной нами выше случайной величины X' , и мы знаем, как его вычислить:

$$\mathbb{E}(X') = m \frac{N + 1}{M + 1}.$$

Коэффициент повышения эффективности теперь можно рассчитать как простое отношение⁸⁸:

⁸⁸ Мы, конечно, сделали упрощение, в какой-то степени огрубляющее итоговый результат. Но с учётом того, что число наблюдений мы предполагаем достаточно большим, дополнительные единицы в числителе и в знаменателе влияют на конечный результат несущественно, и мы избавились от них, не забыв, впрочем, указать, что получающееся равенство является всё-таки не точным, а приближенным.

$$\frac{\mathbb{E}(X')}{n} = \frac{m}{n} \cdot \frac{N+1}{M+1} \approx \frac{m}{n} \cdot \frac{N}{M} = \frac{m}{n} \left(\frac{M}{N} \right)^{-1}.$$

Заметим, что отношение m к n есть не что иное, как доля действительно аномальных состояний, среди тех, которые наш детектор вернул в качестве аномальных⁸⁹, а отношение M к N — это реальная доля аномалий среди всех состояний, которые мы наблюдаем. Поэтому мы даже переформулируем результат в виде правила:

$$\text{ImproveEffect} = \frac{\text{DetectedRatio}}{\text{GeneralRatio}},$$

— берём долю аномалий среди наблюдений, помеченных детектором в качестве аномальных, делим на долю аномалий в исходном наборе наблюдений, и то, что получается в результате, даёт нам коэффициент повышения эффективности: во столько раз меньше проверок нам придётся провести для достижения того же результата. Например, если наши данные таковы, что среди всех наблюдений 2% являются аномальными, а наш детектор таков, что среди результатов его работы 50% являются аномальными, то выигрыш от его применения составляет

$$\text{ImproveEffect} = \frac{0.5}{0.02} = 25.$$

7.2. Детекция аномалий в нестабильных процессах и принцип непрерывного улучшения

Принцип непрерывного улучшения в управлении качеством весьма часто понимается как лозунг: «Работать сегодня — лучше чем вчера, а завтра — лучше чем сегодня!», а менеджеры, апеллирующие к нему, напоминают комиссаров

⁸⁹ Кстати, эта величина называется полнотой бинарного классификатора, о чем мы отдельно упоминаем в разделе, посвящённом теории стохастического прогнозирования (см. приложение А, раздел А.7), и там же отмечаем, что среди других метрик классификации эта метрика играет особую роль. Свою роль она сыграла, например, сейчас (хотя, это и не единственное проявление её важности в процедурах детекции аномалий).

времен гражданской войны и военного коммунизма. Возможно, этот лозунг является своеобразным мотиватором трудового коллектива (хотя, это сомнительно), но всё же это скорее декларация о намерениях, а вовсе не их реализация. С нашей точки зрения, принцип непрерывного улучшения должен быть не лозунгом, а одной из внутренних функций концепции управления качеством, проявляющей себя вне зависимости от идейной стойкости менеджеров: он должен исполняться просто потому, является одним из обязательных этапов этой концепции.

В этом разделе мы покажем, как концепция детекции аномалий позволяет реализовать принцип непрерывного улучшения алгоритмически и операционально. Но сначала — несколько слов о самом принципе.

7.2.1 Философия Kaizen Масааки Имаи

Принцип непрерывного улучшения, также известный как Kaizen (改善 — яп., улучшение), является одним из ключевых элементов философии управления качеством [26, 58, 59, 72, 86–89, 151], который предполагает постоянный поиск и реализацию небольших, постепенных улучшений во всех аспектах деятельности организации. Если совсем кратко, то суть принципа состоит в нескольких основных положениях, среди которых можно выделить:

1. Непрерывность: улучшения должны происходить постоянно, а не быть разовыми акциями.
2. Небольшие шаги: вместо радикальных изменений, фокус должен делаться на небольших, постепенных улучшениях.
3. Фокус на процесс: улучшения должны быть направлены на оптимизацию процессов, а не только на конечный результат.

Масааки Имаи (Masaaki Imai) считается человеком, который не только популяризировал, но и систематизировал этот подход в своей книге «Kaizen: The Key to Japan's Competitive Success» (1986 г.), хотя сама по себе концепция непрерывного улучшения реализовывалась и ранее, ещё до публикации Имаи

(в частности, в компании Toyota, как часть их производственной системы).

Постоянное стремление к улучшениям делает организацию более гибкой и способной адаптироваться к изменяющимся условиям, в силу чего принцип непрерывного улучшения является мощным инструментом для повышения эффективности и конкурентоспособности [36, 89, 207].

7.2.2. Динамическая стандартизация измеримого нестабильного процесса

С другой стороны, адаптация к изменяющимся условиям, отмеченная в предыдущем разделе как важная составляющая принципа непрерывного улучшения — это неотъемлемая часть концепции детекции аномалий (см. главу 1, раздел 1.3.4), это та самая внутренняя её функция, о которой мы говорили выше. В рамках настоящей диссертации мы много раз заявляли и убедительно демонстрировали, что если измеримый процесс является существенно нестабильным, то бессмысленно подходить к нему с позиций какого-то раз и навсегда фиксированного константного стандарта (см., например, главы 4–6), потому что роль стандартного значения при изучении таких процессов исполняет не константа, а модель. Но вместе с тем, динамическим изменениям подвержены и критические отклонения, превышение которых свидетельствует о выходе нестабильного процесса из управляемого состояния и требует операционального вмешательства. Акцентируя эту специфику нестабильных процессов, подчеркнём следующее:

1. Стандартизация стабильных процессов требует указания двух чисел:
1) нормативного значения измеряемой характеристики и 2) максимально допустимого отклонения её измерений от нормативного значения.
2. Для нестабильных процессов нормативные значения характеристики являются функциями от времени и локации, причём, то же справедливо и для максимально допустимых отклонений — они тоже зависят от времени и от локации.

Если на уровне операционального вмешательства в подотчётный нам процесс мы вносим какие-либо корректирующие изменения в среду его протекания (например, устраняем одну из причин возникающих в нем аномалий), то мы, тем самым, изменяем и величину максимально допустимых отклонений: она становится ниже. А это значит, что требования, которые мы после вмешательства в процесс будем предъявлять к его предсказуемости, станут жёстче: то, что раньше не выходило за допустимые рамки теперь оказывается недопустимым и квалифицируется как аномалия.

В следующих двух разделах мы покажем, как работает принцип непрерывного улучшения в нашей концепции детекции аномалий при изучении неустойчивого процесса на синтетических данных. Причём, нам понадобится именно синтетика, так как при работе с реальными данными у нас нет возможности влиять на процесс: во-первых, для этого требуются полномочия, которых у нас нет; а во-вторых, при работе с реальными данными мы можем лишь констатировать факт аномалии, но без дополнительного изучения процесса (которое выходит за рамки настоящей диссертационной работы, и мы это неоднократно отмечали выше, см. главы 4–6) мы не можем указать её причину, тем более — устранить её. При работе с синтетическими данными у нас нет никаких ограничений (см. главу 3, в которой мы синтезировали неустойчивый процесс вместе с мнением тысячи синтетических экспертов).

7.2.3. Синтез данных

Мы синтезируем неустойчивый процесс, который протекает в 100 локациях и наблюдается в течение 1000 моментов времени подобно тому, как это было сделано в главе 3, целиком посвящённой работе на синтетических данных: используя в качестве нелинейных трендов семейство кубических парабол (см. раздел 3.1), а для значений временных рядов применяя нормальный шум с единичной дисперсией (см. там же). В результате получаем набор временных рядов в формате ML (см. таблицу 7.1).

Таблица 7.1 — Синтетические данные в формате ML до наложения аномалий

	Локация 1	Локация 2	Локация 3	...	Локация 100
Момент 1	6.533096	1.850326	2.925826	...	0.545310
Момент 2	6.963235	1.950080	3.052146	...	1.438918
Момент 3	6.904449	1.634107	2.922713	...	0.992128
Момент 4	6.040992	1.415177	3.134948	...	1.454622
...	...	...	...	...	...
Момент 999	4.996258	6.807757	5.505562	...	6.776973
Момент 1000	4.237322	6.656291	6.243252	...	6.954886

Например, в первой локации временной ряд (пока без аномальных значений, которые мы ещё не синтезировали) выглядит так, как представлено на иллюстрации ниже (см. рисунок 7.2).

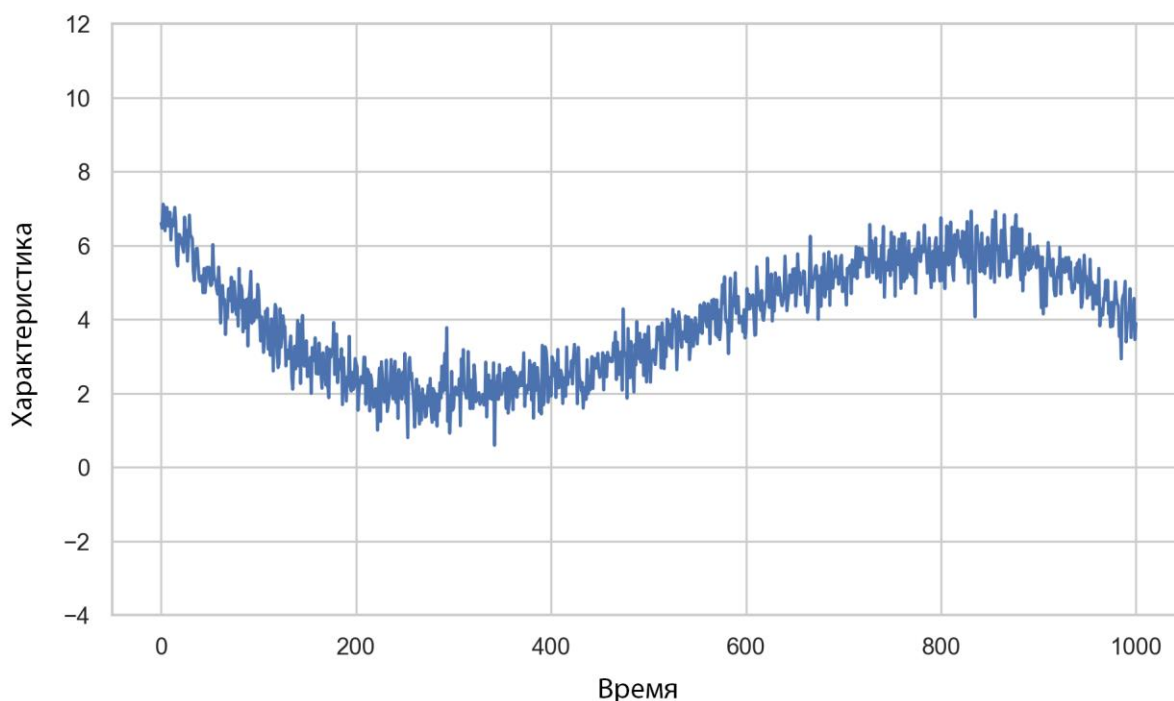


Рисунок 7.2 — Временной ряд до наложения аномалий

Далее мы сделаем целый ряд допущений. Прежде всего, допустим, что природа нашего процесса такова, что аномалии в нем возникают по двум причинам, назовём их условно — причина А и причина В, причём обе они нежелательны. Кроме того, допустим, что и по причине А, и по причине В 2% наблюдений становятся аномальными. Допустим также, что все аномалии, возникшие по причине А, являются более выраженными, нежели аномалии,

возникшие по причине В. И, наконец, мы будем считать, что все аномальные значения равномерно распределены на всем промежутке наших наблюдений от начала и до конца⁹⁰, и все наши допущения справедливы во всех локациях⁹¹.

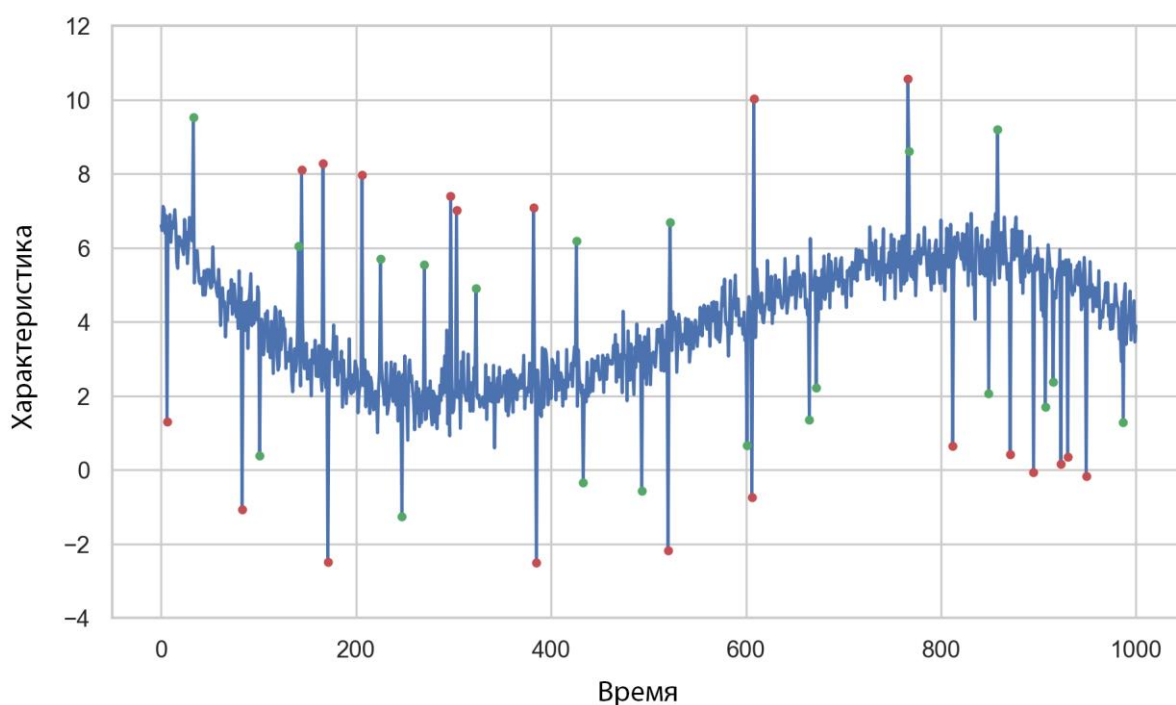


Рисунок 7.3 — Временной ряд с аномалиями двух типов

В рамках этих предположений мы, пользуясь генераторами случайных чисел, накладываем на наши данные аномальные наблюдения, и, например, в первой локации получаем временной ряд, представленный на иллюстрации ниже (см. рисунок 7.3), где красным цветом отмечены аномалии, возникающие по причине А, а зелёным — по причине В. Во всех остальных локациях ситуации аналогичны. Мы не станем проводить этап моделирования с выбором наилучшей прогнозирующей схемы из пула моделей, как мы это делали в предыдущих главах (см. главу 3, а также — главы 4–6), по двум причинам. Во-первых, мы не хотим отвлекаться от основного сюжета, который в этом раз-

⁹⁰ Понятно, что и это условие на практике будет выполняться далеко не всегда.

⁹¹ А это вообще — самое сильное допущение. В разных локациях могут действовать разные причины аномалий, не говоря уже об остальных сформулированных нами условиях.

деле состоит в том, чтобы продемонстрировать работу принципа непрерывного улучшения, а не в том, чтобы строить прогнозирующие модели. А во-вторых, мы и так знаем, какая из схем моделирования окажется наилучшей: полиномиальная регрессия третьей степени, потому что именно так мы синтезировали наши временные ряды. Для первой локации эта модель изображена ниже (см. рисунок 7.4), а с учётом сделанных нами допущений, ситуации во всех остальных локациях аналогичны.

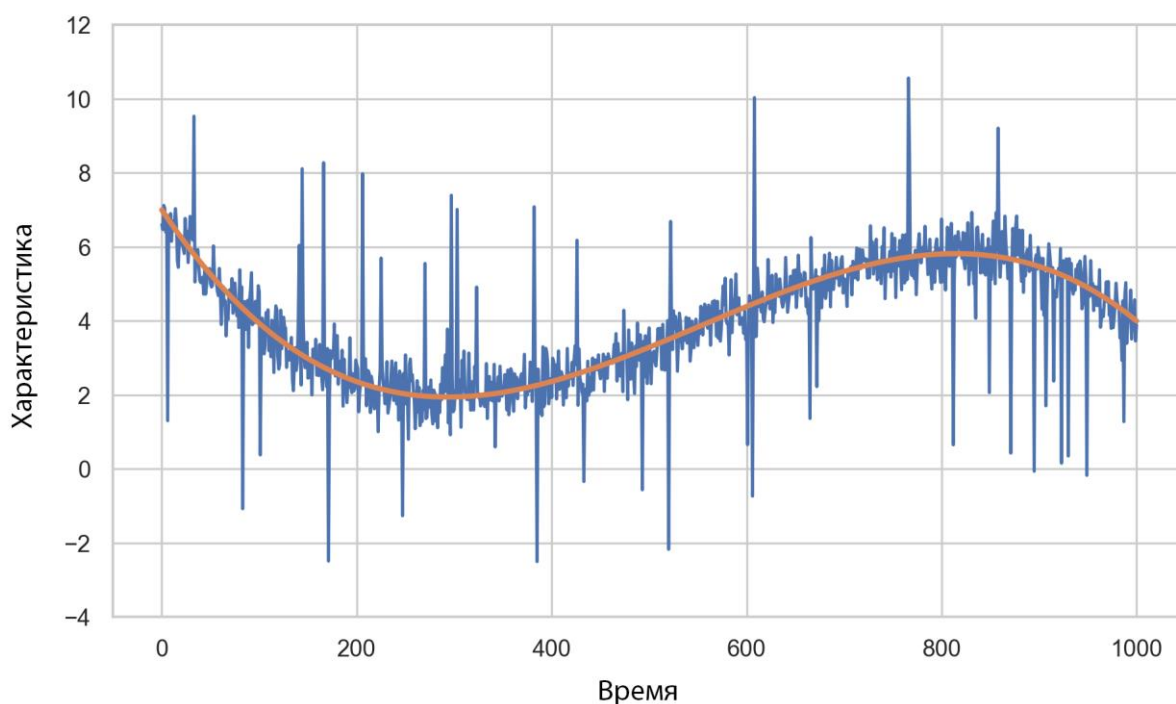


Рисунок 7.4 — Прогнозирующая модель процесса

Теперь у нас есть все, необходимое для того, чтобы найти критическую величину отклонения реального значения характеристики процесса от спрогнозированного значения Δ_{cr} , превышение которой наш детектор сочтёт аномальным. Эти процедуры мы уже много раз выполняли как на реальных (в главах 4–6), так и на синтетических (в главе 3) данных. Мы строим калибровочную кривую, после чего методом локтевой точки определяем критическую величину отклонения (см. рисунок 7.5).

Оказывается, что в условиях нашего эксперимента на синтетических данных критическое отклонение составляет величину 4.4 (понятно, что это абстрактная величина, так как синтетические данные не обладают семантикой).

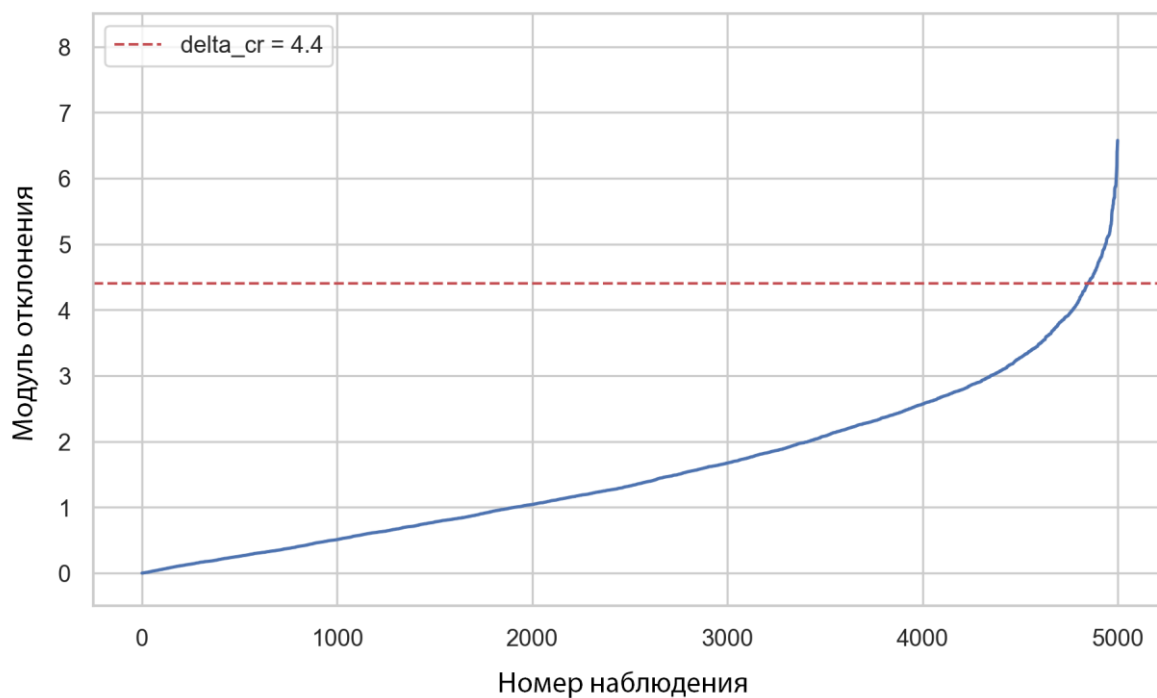


Рисунок 7.5 — Калибровка отклонений при наличии обеих причин аномальных состояний

После этого остаётся лишь отсечь наблюдения, не попадающие в допустимый коридор отклонений (см. рисунок 7.6).

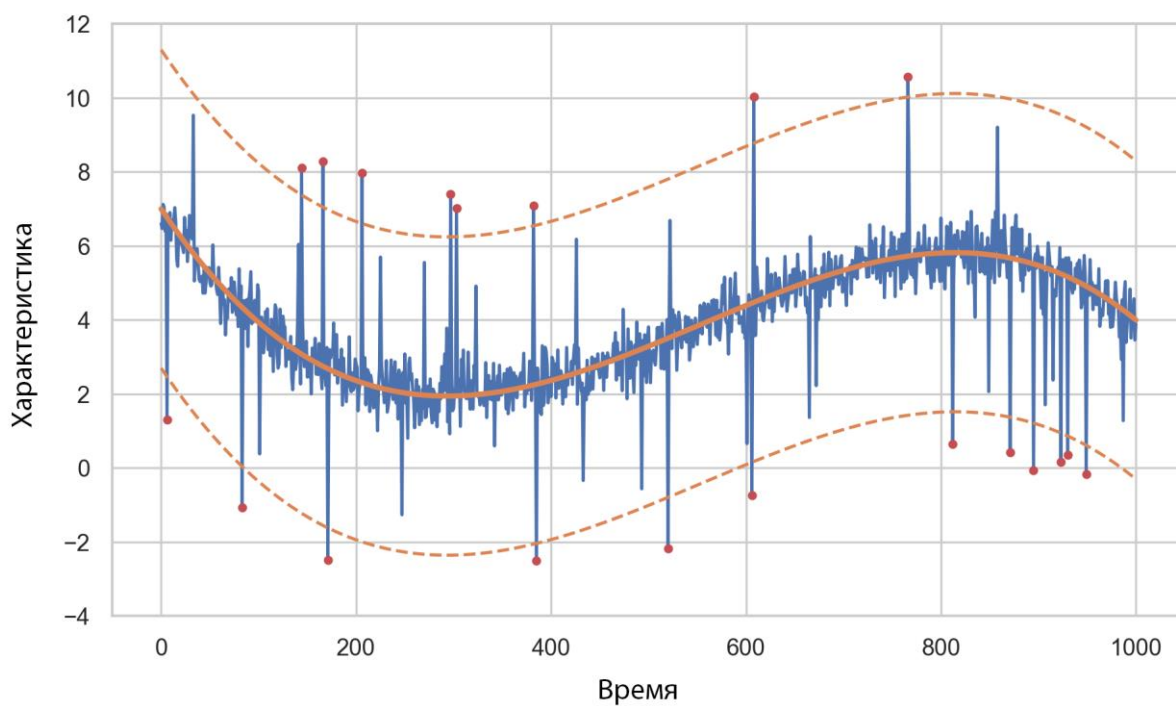


Рисунок 7.6 — Детекция аномалий до устранения причины А

Мы видим, что наш детектор хорошо справился с аномалиями, вызванными причиной А, но при этом аномалии, вызванные причиной В, он не заметил, посчитав их допустимой нормой (для сравнения см. рисунок 7.3).

7.2.4. Операциональное вмешательство в синтетический процесс и перерасчёт величины критического отклонения

Теперь допустим, что после 500 наблюдений мы выяснили, в чем состоит причина А возникновения аномальных состояний в локациях нашего процесса, и, более того — устранили эту причину, причём, во всех локациях сразу⁹². Тогда, например, в первой локации ситуация примет следующий вид (см. рисунок 7.7), и аналогично преобразуются ситуации во всех остальных локациях.

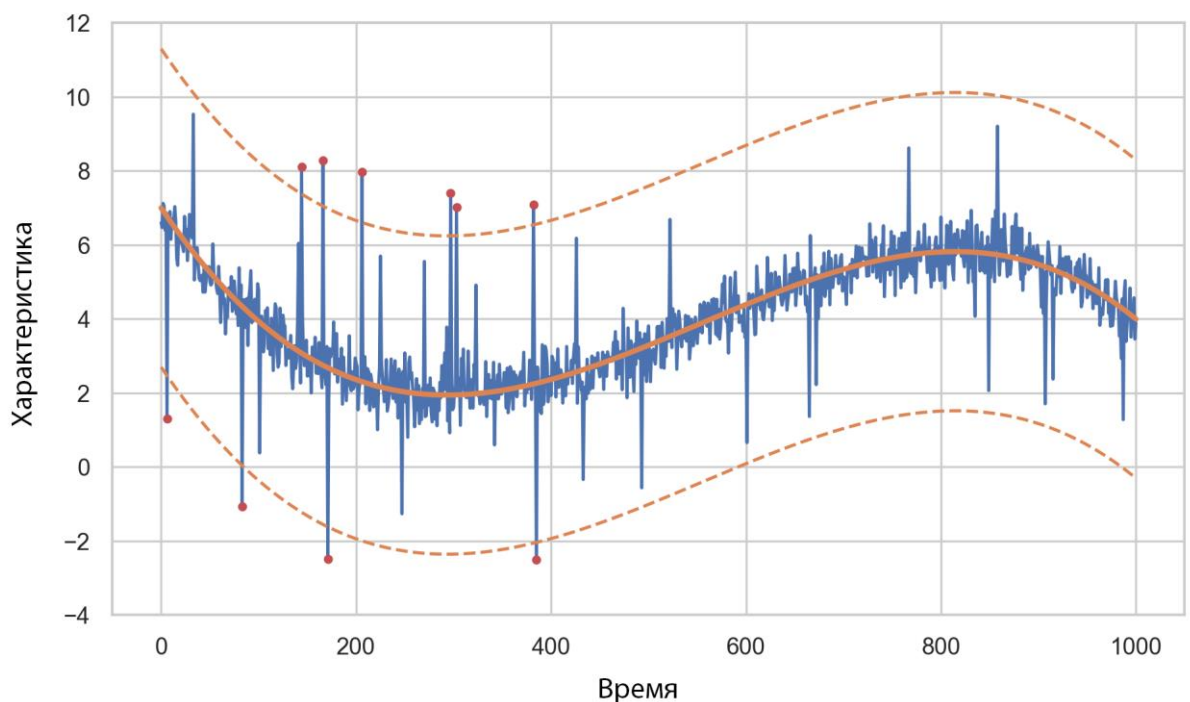


Рисунок 7.7 — Временной ряд после устранения причины А, начиная с момента 500

Здесь мы обращаем внимание на то, что, начиная с момента времени 500,

⁹² Это ещё одно допущение. Может получиться так, что в каких-то локациях причину удалось устранить, тогда как в других — это не получилось, или получилось значительно позже, а не в тот же момент и т. д. Мы, разумеется, упрощаем схему операционального вмешательства для того, чтобы подчеркнуть её суть.

наш детектор вообще перестал фиксировать какие-либо аномалии, и это при том, что мы точно знаем, что в наших наблюдениях продолжают присутствовать аномалии, продиктованные причиной В. Поэтому мы отводим период с 501-го момента по 550-й под новое обучение и вычисляем критическую величину допустимого отклонения в новых условиях, когда аномалии, вызванные причиной А отсутствуют. Для этого мы рассчитываем в каждой локации и в каждый момент времени с 501-го по 550-й все отклонения временных рядов от их трендов, организуем абсолютные величины отклонений в виде единого списка (в нем так же как и выше будет 5000 элементов) и определяем локтевую точку калибровочной кривой. В силу сразу двух причин: 1) аномалий стало меньше (не 4%, а 2% от всех наблюдений) и 2) аномалии стали менее выраженными (причина А давала более значительные отклонения) — локтевая точка должна оказаться ниже.

На иллюстрации ниже мы, для удобства восприятия, одновременно представили сразу две калибровочные кривые⁹³ (см. рисунок 7.8).

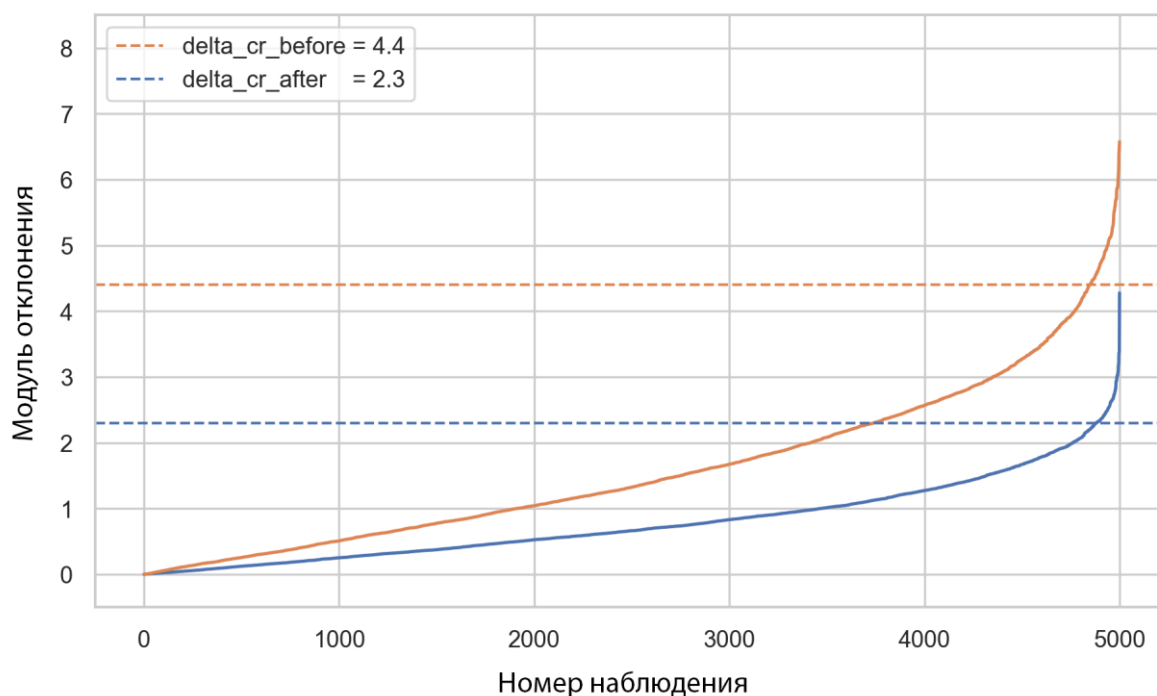


Рисунок 7.8 — Калибровочные кривые до и после устранения причины А

⁹³ Это и есть принцип непрерывного улучшения, представленный в виде двух графиков.

Таким образом, начиная с 500-го наблюдения величина допустимого порога отклонений Δ_{cr} понижается с 4.4 до 2.3 (см. рисунок 7.9), и это понижение заложено в самой алгоритмике детекции аномалий.

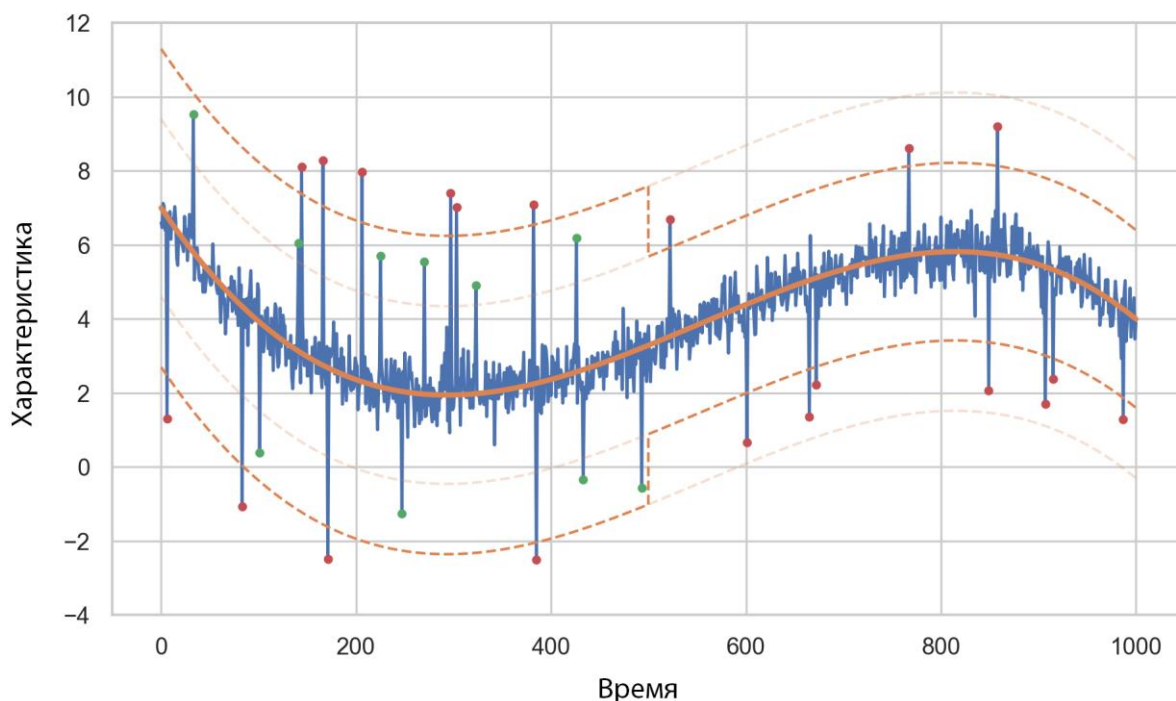
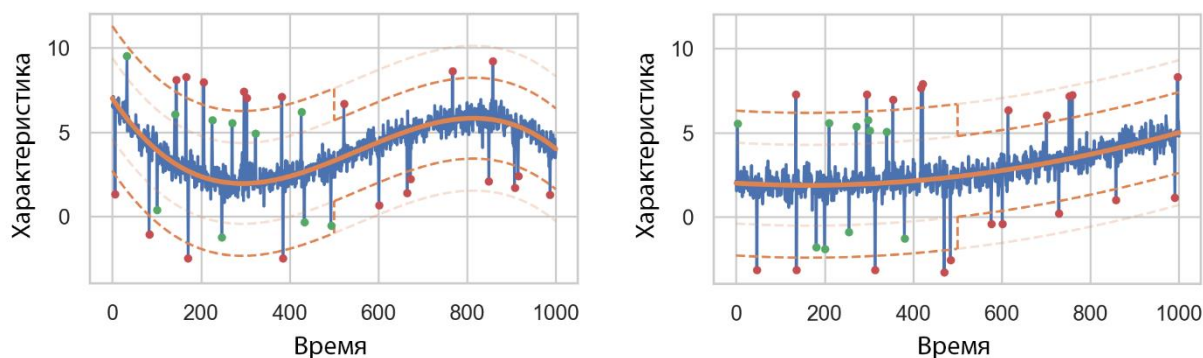


Рисунок 7.9 — Коридор допустимых отклонений до и после устранения причины А

Мы видим, что аномалии, продиктованные причиной В, на первом этапе работы детектора вообще не воспринимаются им как аномалии (зелёные точки в левой части графика), зато после повторной калибровки то, что раньше считалось допустимым, становится неприемлемым (красные точки в правой части графика): другими словами, алгоритм сам требует от процесса более плотного прилегания к тренду. Тот же эффект наблюдается и в других локациях, некоторые из них проиллюстрированы ниже (см. рисунок 7.10):



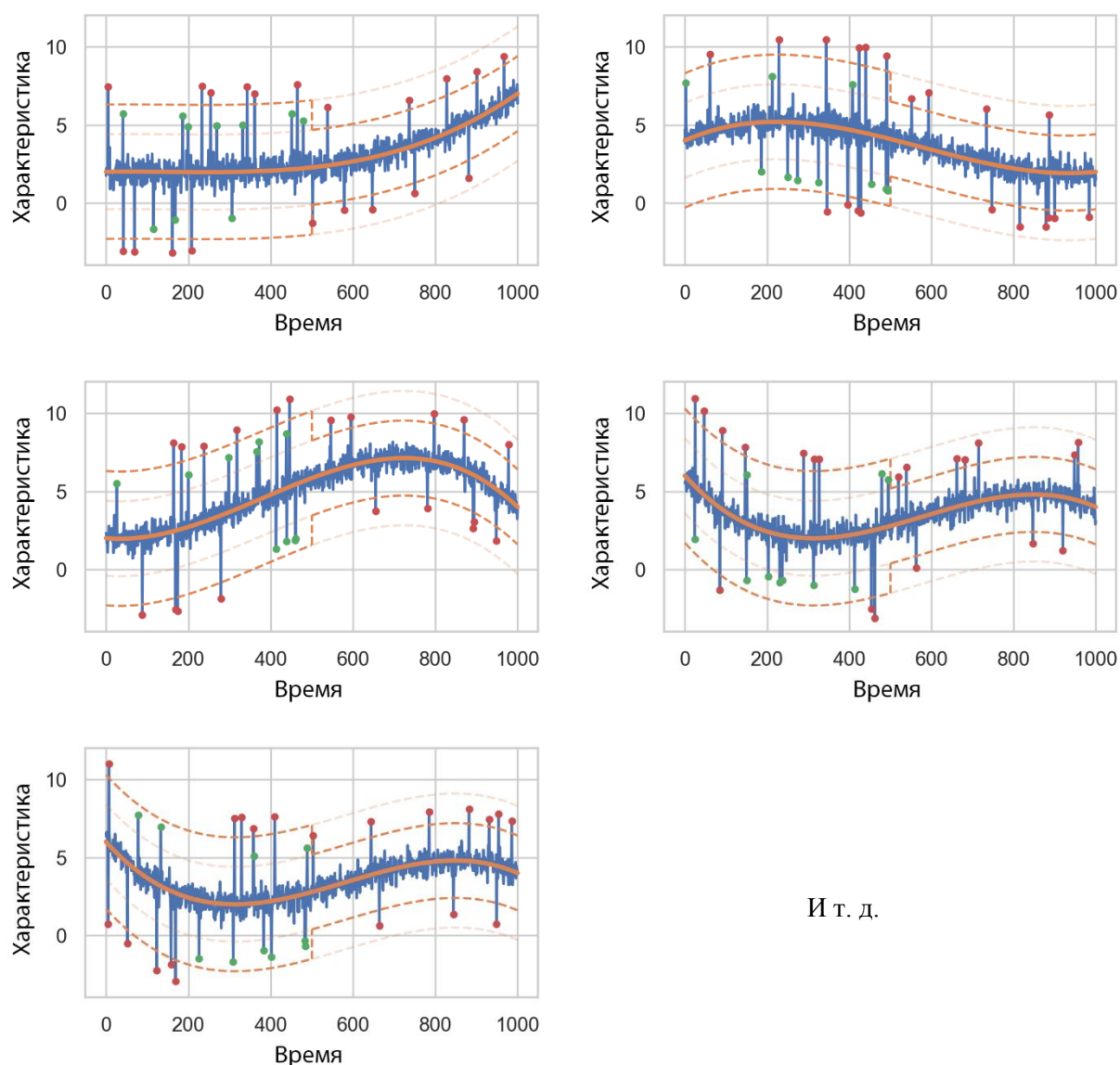


Рисунок 7.10 — Сужение коридора допустимых отклонений в различных локациях синтетического процесса в результате устранения причин аномалии А

Разумеется, мы разобрали предельно упрощённую схему повторной калибровки на предельно упрощённых синтетических данных. В реальности понижение калибровочной границы никогда не происходит одновременно — это плавный и постепенный процесс, границы никогда не понижаются одновременно сразу во всех локациях и т. д. Но это детали. Нам важно было продемонстрировать то, что принцип непрерывного улучшения заложен в самой концепции детекции аномалий, а именно — при постепенном устранении причин, возникающих в нестабильном процессе аномалий, и пересчёте калибро-

вочных границ требования к предсказуемости процесса автоматически становятся всё более жёсткими. Например, если в рассмотренном выше примере мы устраним не только причины возникновения аномалии А, но и причины аномалии В, то на следующих этапах эксплуатации детектора аномалий коридор допустимых изменений процесса станет ещё более узким, а это значит, что процесс станет более предсказуемым, и его качество станет выше.

7.3. Аспекты практической значимости концепции детекции аномалий в нестабильных процессах

На протяжении всей диссертационной работы, начиная с обсуждения её концептуальных особенностей, мы неоднократно подчёркивали связь наших идей с традиционными приёмами статистического управления качеством (см. раздел 1.3.4), которую схематично можно проиллюстрировать следующим образом (см. рисунок 7.11).

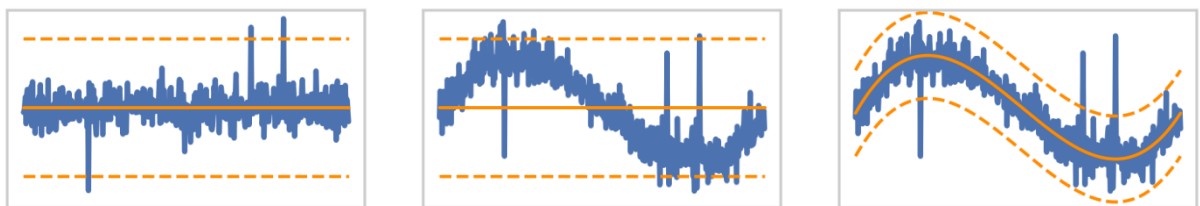


Рисунок 7.11 — а) стабильный процесс с отклонением от среднего, б) нестабильный процесс с отклонением от среднего, в) нестабильный процесс с отклонением от модели

В каком-то смысле, мы взяли хорошо известную контрольную карту Шухарта (см. рисунок 7.11 (а)) и деформировали её к нелинейному тренду нестабильного процесса (см. рисунок 7.11 (в)), так как её непосредственное применение к нестабильным процессам, как правило, приводит к некорректным результатам (см. рисунок 7.11 (б)). Практическая значимость контрольных карт неоспорима [6, 19, 38, 212], и в этом разделе мы сначала сформулируем её основные положения (см. раздел 7.3.1), а затем перенесём их на идеи и методы статистического управления качеством нестабильных процессов, которые являются основным предметом всех исследований в настоящей диссертационной работе (см. раздел 7.3.2).

7.3.1. Напоминание: практическая значимость контрольных карт Шухарта

Карта Шухарта (она же контрольная карта Шухарта) — это мощный и широко используемый инструмент статистического управления стабильными процессами [6, 19, 38, 212]. Её практическая значимость заключается в том, что она позволяет как визуализировать, так и контролировать изменчивость стабильного процесса, выявлять отклонения от нормы и предпринимать необходимые корректирующие действия для улучшения его стабильности и предсказуемости [56, 162, 226]. Укажем некоторые наиболее важные аспекты практической значимости контрольных карт так, как это обычно делают в авторитетных источниках [19, 162, 226].

- 1. Мониторинг стабильности процесса.** Основная цель карты Шухарта состоит в разделении изменчивости процесса на две составляющие:
 - а) естественная изменчивость, которая вызывается общими причинами, и
 - б) противоестественная изменчивость, которая вызывается особыми причинами. Естественной считается неизбежная случайная изменчивость, присущая любому процессу при нормальном его функционировании. Она обусловлена множеством мелких, незначительных факторов, которые весьма трудно отследить и контролировать по отдельности, при этом процесс считается находящемся в статистически управляемом состоянии, когда изменчивость обусловлена только общими причинами. Особые причины изменчивости — это неслучайные факторы, вызывающие резкие отклонения процесса от ожидаемого поведения, они могут быть связаны с какими-то внешними событиями, проблемами оборудования, ошибками персонала, изменениями в материалах и т.д. Карта Шухарта позволяет визуально обнаружить эти особые причины.
- 2. Визуализация данных во времени.** Данные о характеристиках процесса (например, размеры деталей, температура, давление) наносятся на карту в хронологическом порядке. Линия, соединяющая эти точки, показывает,

как изменяется процесс во времени. При этом на карте Шухарта присутствуют три горизонтальные линии: а) центральная линия (CL) — представляет собой среднее значение характеристики процесса за определённый период времени; б) верхний контрольный предел (UCL) — расположен выше центральной линии и показывает верхнюю границу допустимой изменчивости процесса; в) нижний контрольный предел (LCL) — расположен ниже центральной линии и показывает нижнюю границу допустимой изменчивости процесса (см. рисунок 7.11(а)). Таким образом, контрольные карты сигнализируют о выходе процесса из-под контроля: если точки данных выходят за пределы контрольных границ или образуют определённые паттерны (например, последовательность точек выше или ниже центральной линии, тренды, цикличность), это указывает на то, что в процесс вмешались особые причины и он перестал быть стабильным.

3. **Предупреждение о проблемах и своевременное принятие мер.** Карта Шухарта позволяет выявлять потенциальные проблемы ещё до того, как они приведут к браку или другим серьёзным последствиям. Анализ паттернов на карте может указывать на зарождающиеся тенденции, требующие внимания. Обнаружив признаки нестабильности, можно оперативно провести расследование, выявить коренную причину проблемы и принять корректирующие действия, чтобы вернуть процесс в управляемое состояние.
4. **Улучшение качества продукции и услуг.** Применение контрольных карт позволяет обеспечить ряд факторов повышения качества: а) сокращение брака и дефектов — устранение особых причин изменчивости приводит к уменьшению разброса параметров процесса и, как следствие, к снижению вероятности производства дефектной продукции; б) повышение соответствия требованиям — стабильный и управляемый процесс позволяет с большей уверенностью производить продукцию или предоставлять услуги, соответствующие заданным требованиям и спецификациям; в) снижение затрат — уменьшение брака, переделок и других проблем, связанных с нестабильным процессом, приводит к существенной экономии

затрат; d) улучшение удовлетворённости клиентов — более высокое качество продукции и услуг, а также стабильность процессов, приводят к повышению удовлетворённости клиентов.

5. **Принятие решений на основе данных.** Управление стабильными процессами на основании контрольных карт обеспечивает а) объективность — карта Шухарта позволяет принимать решения на основе объективных данных, а не на основе субъективных оценок или интуиции; б) измеримость — результаты применения карты Шухарта можно измерить и оценить эффективность принятых мер; в) постоянное улучшение — применение карт Шухарта является частью цикла постоянного улучшения (PDCA – Plan, Do, Check, Act), так как, после выявления и устранения особых причин изменчивости, можно пересмотреть контрольные пределы и центральную линию, чтобы отразить новое, более стабильное состояние процесса.

Общепринятое мнение специалистов в области управления качеством таково, что контрольная карта Шухарта — это не просто график, а мощный инструмент для управления процессами, улучшения качества и снижения затрат [6, 154, 162, 226, 236, 237]. Её практическая значимость заключается в предоставлении визуального, основанного на данных метода мониторинга и контроля процессов, позволяющего оперативно выявлять и устранять проблемы, а также постоянно совершенствовать процессы для достижения стабильных и предсказуемых результатов [6, 19, 154]. Она полезна в самых разных отраслях, от производства до сферы услуг, и является одним из основных инструментов в арсенале специалиста по управлению качеством.

7.3.2. Детекция аномалий как эволюция контрольных карт

Все перечисленные выше аспекты практической значимости контрольных карт Шухарта как одного из основных методов статистического управления качеством непосредственно переносятся на концепцию детекции аномалий.

На этом можно было бы закончить раздел 7.3, но мы всё-таки дадим определённые комментарии к этому утверждению. Дело в том, что при изучении процессов традиционными статистическими методами (в частности, при помощи карт Шухарта) ключевым оказывается такое понятие как стабильность процесса. Для нашей концепции, которая изначально была разработана именно для изучения нестабильных процессов, ключевым понятием является понятие его предсказуемости. При этом мы понимаем предсказуемость процесса не в том смысле, что его измеряемые характеристики всегда должны давать один и тот же результат (плюс-минус небольшие отклонения). Нет, в нашем понимании предсказуемость процесса означает, что модели машинного обучения, которые мы применяем для прогнозирования этих характеристик, обладают высокой прогнозирующей способностью: то есть, следующее значение характеристики процесса может сильно (и даже очень сильно) отличаться от её предыдущего значения, но если наш прогноз оказывается близок к реально зафиксированной величине, то процесс предсказуем.

После этого замечания мы можем перейти к более детальному разбору аспектов практической значимости, которые мы предъявили в предыдущем разделе. Начнём с самой первой позиции, где в качестве аспекта значимости контрольных карт упоминался мониторинг стабильности и просто заменим слово «стабильность» на слово «предсказуемость», после чего обнаружим, что, действительно — детекция аномалий обеспечивает **мониторинг предсказуемости нестабильных процессов**. Отметим, что то, что традиционно называется естественной изменчивостью мы называем нормальным шумом, а изменчивость, продиктованную особыми причинами мы называем аномалиями. И в результате таких несложных манипуляций с терминологией, то, что традиционно считается одним из ключевых аспектов практической значимости контрольных карт — обнаружение и визуализация особых причин нестабильности — автоматически становится аспектом практической значимости детекции аномалий: наша концепция позволяет обнаруживать аномальные со-

стояния (она для этого и была разработана), причём, не только и столько визуально (так как визуальные наблюдения за большими данными чаще всего невозможны), сколько алгоритмически (хотя соответствующие визуализации мы активно демонстрируем на протяжении всей диссертационной работы, см., например, рисунки 4.5, 4.13, 5.4, 5.12, 6.4, 6.13 и др.).

Второй часто указываемый в литературе аспект практической значимости контрольных карт — **визуализация процесса по времени** — также является неотъемлемой частью концепции детекции аномалий. Только роль центральной линии (CL) в нашем случае играет график прогнозирующей модели (см. рисунки 4.5, 4.13, 5.4, 5.12, 6.4, 6.13 и др.), а верхние и нижние границы (UCL и LCL) мы, как правило, изображаем на отдельных иллюстрациях (см., например, рисунки 4.6, 4.14, 5.5, 5.13, 6.5, 6.14 и др.), так как на реальных данных разброс наблюдаемых значений весьма высок, а характер тренда достаточно сложен, и совмещение множества линий на одном рисунке приводит к визуальной и смысловой перегруженности изображения. Если же речь идёт о визуализации синтетических данных, то мы не пренебрегаем такой возможностью, изображая верхние и нижние границы, в качестве которых выступают соответствующие смещения тренда (см. рисунки 7.6, 7.7, 7.9, 7.10 и др.).

Что касается третьего часто упоминаемого аспекта практической значимости контрольных карт Шухарта — **предупреждения о проблемах и своевременного принятия мер** — то он напрямую задействован и в нашей концепции детекции аномалий, более того, это и есть основная цель их выявлений. Мы уделяем этому действию большое внимание, когда формируем рейтинги аномалий (см. разделы 4.3.3, 4.4.3, 5.3.3, 5.4.3, 5.5.3, 5.6.3, 6.3.3, 6.4.3), более того, мы выносим предупреждения о проблемах в качестве одного из приложений настоящей диссертационной работы (см. приложение D). Собственно, сама концепция детекции аномалий, равно как и методика МССР (Modeling, Challenge, Calibration, Production) разрабатывались нами, исходя из необходимости практического применения статистических методов управления качеством к нестабильным процессам, а практика состоит не только в изучении

процесса, но и в активном вмешательстве в тех ситуациях, когда такое вмешательство необходимо.

Улучшение качества продукции и услуг, которое было упомянуто выше в качестве четвёртого аспекта практической значимости контрольных карт Шухарта, мы не станем здесь рассматривать по отношению к концепции детекции аномалий (так как такую формулировку можно применить к абсолютно любому методу управления качеством, включая и нестатистические методы), а перейдём к пятому из упомянутых выше аспектов: к **принятию решений на основе данных**. Управление нестабильными процессами на основании детектированных аномалий обеспечивает объективность (см. пункт (а) аспекта 5), так как наша концепция позволяет принимать решения на основе объективных данных, а не на основе субъективных оценок или интуиции. Мы говорили об этом, когда сравнивали два метода калибровки отклонений: калибровки на основании экспертных оценок (см. разделы 2.4, 3.6) с калибровкой по локтевой точке на калибровочной кривой (см. разделы 4.3.2, 4.4.2, 5.3.2, 5.4.2, 5.5.2, 5.6.2, 6.3.2, 6.4.2.). Кроме того, наш метод измерим (см. пункт (b) аспекта 5) — результаты его применения действительно можно измерить. И, наконец, он включает в себя принцип постоянного улучшения, этому мы даже посвятили отдельное исследование (см. раздел 7.2) и показали, как постоянное улучшение осуществляется в рамках нашей концепции детекции аномалий алгоритмически и операционально.

7.4. Выводы по главе 7

Итак, в этой главе мы рассмотрели предложенную и разработанную в настоящей диссертационной работе концепцию детекции аномалий по отношению к управлению качеством как к научной дисциплине: её место в этой научной области и связь с существующими концепциями и идеями. В качестве краткого резюме отметим следующее:

1. Наша концепция детекции аномалий является естественным продолжением идеи контрольных карт, перенесённым с группы стабильных процессов на группу нестабильных процессов.
2. Она расширяет сферу применения статистических методов управления качеством, оставаясь при этом частью всего инструментария управления качеством как научной дисциплины.
3. Все хорошо известные аспекты практической значимости контрольных карт (мониторинг, принятие решений на основе объективных данных и т. д.) переносятся на детекцию аномалий после смещения акцента со стабильности процесса на его предсказуемость.

Таким образом, проведённый в этой главе методологический анализ позволяет говорить об успешной интеграции концепции детекции аномалий в теорию и практику управления качеством.

ЗАКЛЮЧЕНИЕ

Основные результаты по работе и выводы

В ходе работы над диссертацией решена важная научно-отраслевая проблема, предложена совокупность комплексных технических решений, в результате чего достигнута цель диссертационного исследования, состоящая в разработке нового методологического инструментария статистического управления качеством при помощи детекции аномалий в процессах СМК предприятий автомобильной отрасли.

Получены следующие научно-прикладные результаты:

1. По итогам научно-прикладного анализа проблемы, связанной с анализом процессов в системе менеджмента качества автопроизводителей и необходимостью обеспечения принципа принятия решений, основанных на свидетельствах, выделено существующее в отраслевой науке и практике управления качеством противоречие, заключающееся в отсутствии адекватных статистических методов и инструментов, направленных на повышение результативности и эффективности процессов, являющихся существенно нестабильными. Также выявлено, что действующие стандарты системы менеджмента качества ISO 9001:2015, и его отраслевой аналог IATF 16949:2016 не содержат требований по применению статистических инструментов управления по отношению к нестабильным процессам. В условиях информатизации и цифровизации процессов системы менеджмента, взрывного характера роста объемов электронных данных, предназначенных для контроля, мониторинга и управления качеством, требуется опережающая разработка комплексных научно-технических решений, дающих адекватный ответ на современные вызовы. В качестве такого ответа следует рассматривать предлагаемую методологию и инструментарий детекции

аномалий в управлении качеством процессов предприятий автомобильной отрасли.

2. Сформирована научная концепция методологии детекции аномалий в процессах, входящих в СМК предприятий автомобильной отрасли. Концепция является обобщением известных концепций статистического управления качеством, таких как концепция Деминга и тесно связанная с ней концепция 6σ, и позволяет анализировать как стабильные, так и существенно нестабильные по своей природе процессы. Разработанная концепция представляет собой комплексный подход, обеспечивающий универсальное представление о различных процессах и подходах к определению нормы и аномалии на основании измеримых характеристик, и может быть использована для автоматизации управления качеством в таких процессах как маркетинг, проектирование, производство автомобильной продукции, сервисное обслуживание автомобилей в сети сервисных центров автопроизводителя, сопровождение автомобильного бренда и т. д.
3. Предложен комплекс научно-программных инструментов по обработке данных о процессах СМК автомобильной отрасли, позволяющий решать задачи по тестированию инструментов детекции аномалий (несоответствий) при мониторинге процессов. В рамках предложенного комплекса разработан инструментарий генерирования синтетических данных о протекании процессов системы менеджмента предприятий автомобильной отрасли и их экспертных оценках. В производственной практике применение предложенного комплекса обеспечивает создание виртуального инструментария объективного контроля качества мониторинга процессов СМК посредством тестирования инструментов детекции аномалий в условиях различных искусственно смоделированных экспертиз.
4. Разработаны методы моделирования состояния семейства процессов системы менеджмента, которые используются для описания производственных, сервисных и маркетинговых процессов, позволяющие одновременно

анализировать группу из множества взаимосвязанных характеристик процессов СМК, протекающих на различных участках автомобильного производства — от штамповки кузовных деталей до сборки продукта, его последующего обслуживания в сервисных центрах и маркетингового сопровождения. Эти методы особенно ценны для крупных автоконцернов, имеющих разветвлённую сеть производственных и сервисных площадок (семейство локаций). Предложенные методы предназначены для оперативной детекции аномальных состояний (несоответствий) процессов, например, отклонений в параметрах технологических характеристик, объёме выполненных операций, суммарной стоимости затрат на устранение дефектов и т. д. Использование предложенных инструментов обеспечивает значительное улучшение показателей качества на предприятиях автомобильной отрасли (в части применения к процессам статистического мониторинга качества исполнения технологических операций), укреплению репутации производителя (в части его применения к процессам сервисного обслуживания автомобилей) и повышению конкурентоспособности на рынке (в части его применения к процессам маркетингового сопровождения автомобильного бренда). А именно:

- 4.1. Для группы процессов, относящихся к эффективности выполнения технологических операций автомобильного производства, процент выявленных аномалий вырос с 4,1% до 56,4%.
- 4.2. Для группы процессов сервисного обслуживания автомобилей в фирменных сервисных центрах автопроизводителя процент выявленных аномалий вырос с 5,6% до 64,2%.
- 4.3. Для группы процессов, относящихся к маркетинговому сопровождению автомобильного бренда, процент выявленных аномалий вырос с 9,3% до 61,7%.
5. Разработан критерий детекции аномальных состояний процесса, основанный на методе локтевой точки, который позволяет автоматизировать выяв-

ление аномалий (несоответствий) процессов в системе менеджмента качества. Математический алгоритм, автоматически определяющий величину допустимых отклонений, позволяет мгновенно выявлять проблемы, такие как негативные изменения в клиентском сервисе или снижение заинтересованности потребителей в продукции автопроизводителя. Ключевое преимущество этого критерия заключается в исключении экспертной оценки, которая становится острой проблемой автомобильной отрасли в условиях ресурсных ограничений, с которыми столкнулись многие предприятия в настоящее время. Кроме того, исключение человеческого фактора повышает объективность, способствуя обоснованным управленческим и организационным решениям и как следствие — повышению качества продукции и услуг в автомобильной отрасли.

6. Разработан и реализован научно-программный комплекс моделей и алгоритмов, основанный на предложенных методологии и инструментарии детекции аномальных состояний. Комплекс, реализованный на языке Python для обеспечения гибкости и масштабируемости, включает в себя 6 ключевых алгоритмов, специально разработанных для обнаружения отклонений измеряемых характеристик процессов системы менеджмента от моделируемых трендов в режиме реального времени. Это критически важно для своевременного обнаружения потенциально опасных отклонений процессов, входящих в СМК автомобильной отрасли, от нормальных режимов их протекания. Разработанный научно-программный комплекс позволяет оперативно проводить количественный анализ выявленных аномалий (несоответствий), ранжируя их по степени выраженности, что позволяет оперативно принимать решения об уровне необходимого реагирования для каждой аномалии. Испытания комплекса на синтетических и реальных данных о процессах автомобильного производства, сервисного обслуживания автомобилей и маркетингового сопровождения автомобильного бренда под-

твердили его высокую эффективность в задачах оперативной детекции аномальных состояний, позволяя сократить время реакции на возникающие критические отклонения как минимум кратно.

7. Ключевым практическим результатом диссертации является разработка и обоснование метода принятия управленческих решений по отношению к процессам, входящим в СМК автомобильной отрасли. Метод основан на рациональном анализе данных и, в отличие от традиционного статистического выборочного контроля, часто применяемого в автомобилестроении, использует полный анализ всей доступной статистической информации о производственных, сервисных и маркетинговых процессах. Это обеспечивает более точную и объективную картину для принятия обоснованных управленческих решений, причём, основным преимуществом разработанного метода является минимизация влияния человеческого фактора и субъективных оценок, часто приводящих к ошибкам в организационных и управленческих решениях. Достоверность принимаемых решений обеспечивается за счёт использования автоматизированных алгоритмов анализа данных и чётко определённых, объективных критериев оценки. Это позволяет избежать предвзятости и обеспечивает последовательность решений, что критически важно для поддержания стабильно высокого качества выпускаемой продукции и предоставляемых услуг. Реализованный метод позволяет руководителям принимать обоснованные управленческие решения по устранению выявленных отклонений, что, в конечном итоге, приводит к снижению затрат на контроль качества.
8. Предложенные научно-технические решения нашли своё применение и внедрены в устойчивую практику таких автомобильных концернов как АО «АВТОВАЗ» и ПАО «КАМАЗ». В ПАО «КАМАЗ» на основании данных о гарантийном обслуживании реализованы решения по автоматизированному выявлению и предотвращению необоснованных ремонтов. Также внедрён комплекс для обнаружения аномалий в базах данных системы ме-

неджмента качества, снижающий риски принятия неэффективных управленческих решений и способствующий развитию системы управления качеством предприятия. В АО «АВТОВАЗ» в целях повышения эффективности контроля качества автомобилей по параметрам электромагнитной совместимости (ЭМС) внедрены методы моделирования временных рядов и состояния семейства процессов, обеспечивающие анализ параметров качества автомобилей с позиции ЭМС как для перспективных, так и для серийных моделей. Используется научно-программный инструмент для детекции аномалий (несоответствий) в управлении качеством автомобильного производства на основе анализа данных испытаний автомобилей по параметрам электромагнитной совместимости и помехоустойчивости. Кроме того, реализован метод принятия решений о качестве автомобилей по параметрам ЭМС, обеспечивающий рационализацию выводов и исключение субъективного фактора при оценке. Все предложенные инструменты прошли апробацию и доказали свою производственную эффективность. В настоящее время они используются для анализа данных о качестве автомобилей с позиции ЭМС, что позволило снизить трудоёмкость процессов исследования и анализа данных на 40% при одновременном повышении достоверности и точности полученных результатов.

Направления дальнейших исследований

В методике МССР ключевая роль отводится прогнозирующим моделям: только обладая алгоритмами с высокой прогнозирующей способностью мы сможем корректно выявлять аномальные состояния нестабильных процессов. Поэтому основные усилия в дальнейших исследованиях будут направлены на процедуры моделирования.

Мы предполагаем дальнейшую работу в следующих направлениях:

1. Построение прогнозирующих регрессионных моделей на ближайших соседях, где близость понимается в смысле предложенного в настоящей диссертации понятия близости портретов локаций и моментов (см. главу 2, раздел 2.1). Гипотетически это может повысить точность моделей.
2. Изолированная калибровка отклонений в каждой локации. При этом в качестве отклонения по времени будет использоваться не число Δ_{cr} (см. главу 2, раздел 2.5), а вектор, и аналогично — для отклонений по локациям Θ_{cr} . Гипотетически это может повысить точность детектора как бинарного классификатора.
3. Поиск аномалий для выявления несоответствий в нестабильном процессе по совокупности его характеристик. Это направление, в отличие от двух предыдущих, никак не связано с прогнозирующим моделированием и не рассматривалось в настоящей диссертации. Тем не менее, оно тоже может оказаться перспективным.

СПИСОК ЛИТЕРАТУРЫ

1. Afinogentova N. V. Analytical complex for company car-service quality assessment // N. V. Afinogentova, D. I. Panyukov, V. N. Kozlovsky. — Journal of Scientific Research and Development. 2015. Т. 2. № 11. — С. 22–30.
2. Aplaydin E. Introduction to Machine Learning / E. Aplaydin. — London: MIT Press, 2010. — 537 pp.
3. Bartz E. Hyperparameter Tuning for Machine and Deep Learning with R / E. Bartz, T. Bartz-Beielstein, M. Zaefferer, O. Mersmann. — Springer Nature Singapore Pte Ltd, 2023. — 327 pp.
4. Biehl M. The Shallow and the Deep / M. Biehl. — Groningen: University of Groningen Press, 2023. — 278 pp.
5. Bishop C. M. Pattern Recognition and Machine Learning / C. M. Bishop. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.
6. Deisenroth M. Mathematics for machine learning / M. Deisenroth, F. Faisal, C. Ong. — Cambridge University Press, 2020. — 417 pp.
7. Hand D. Principles of Data Mining / D. Hand, H. Mannila, P. Smith. — London: MIT Press, 2001. — 546 pp.
8. Harrington P. Machine Learning in Action / P. Harrington. — NY «Manning Publications Co», 2012. — 354 pp.
9. Mohri M. Foundations of Machine Learning / M. Mohri, A. Rostamizadeh, A. Talwalkar. — London: MIT Press, 2018. — 486 pp.
10. Mosin V. G. Predictive analysis of two bijectively related families of functions in l_2 , which are expressed as tuple pairs // V. G. Mosin, A. A. Abashkin — В сборнике: E3S Web of Conferences. Сер. «Topical Problems of Agriculture, Civil and Environmental Engineering, TPACEE 2020», 2020. — С. 01007.

11. Narzullaev D. Z. Transformation of types of signs for a task of the regression analysis // D. Z. Narzullaev, B. A. Abdurakhmanov, A. S. Baydullaev, S. T. Ilyasov, K. K. Shadmanov. — IOP Conference Series: Materials Science and Engineering, 2020, 862(5).
12. Shai S. Understanding Machine Learning: From Theory to Algorithms / S. Shai, B. Shai. — Cambridge University Press, 2014. — 449 pp.
13. Simeone O. A Brief Introduction to Machine Learning for Engineers / O. Simeone. — London: King's College, 2018. — 231 pp.
14. Stroganov V. I. Quality management in the automotive industry in the digital transformation era // V. I. Stroganov, V. V. Guly, V. N. Kozlovsky, D. I. Panyukov, D. I. Blagoveshchensky. — В сборнике: 2021 Systems of Signals Generating and Processing in the Field of on Board Communications, Conference Proceedings. 2021. — С. 9416116.
15. Sugiyama M. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation / Sugiyama M., Kawanabe M. — London: MIT Press, 2012. — 263 pp.
16. Абдурахманов М. А., Нарзуллаев Д. З., Фазылов Ш. Х. Развитие методов и программного обеспечения анализа данных применительно к исследованиям жизненного цикла продукции // М. А. Абдурахманов, Д. З. Нарзуллаев, Ш. Х. Фазылов. — Материалы I Международной научно-технической конференции «IPI (CALS)-2003. Информационные технологии в управлении жизненным циклом», Санкт-Петербург, 2003 — С. 42.
17. Абусев Р. А. Групповая классификация: решающие правила и их характеристики / Р. А. Абусев. — Пермь: Издательство пермского университета, 1992. — 218 с.
18. Адлер Ю. П., Шпер В. Л. Истоки статистического мышления // Ю. П. Адлер, В. Л. Шпер. — Методы менеджмента качества. 2003. № 1. — С. 34–40.
19. Адлер Ю. П. Новое направление в статистическом контроле качества — методы Тагучи / Ю. П. Адлер. — М.: Знание, 1988. — 113 с.
20. Айвазян С. А. Прикладная статистика: исследование зависимостей /

- С. А. Айвазян, И. С. Енюков, Д. Д. Мешалкин. — М.: Финансы и статистика, 1985. — 487 с.
21. Айвазян С. А. Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бушбахер, И. С. Енюков, Д. Д. Мешалкин. — М.: Финансы и статистика, 1989. — 606 с.
22. Айдаров Д. В. Воспринимаемое качество в автомобильной промышленности: Обзор международной практики // Д. В. Айдаров, В. Н. Козловский, Д. С. Самойлова. — Наука и бизнес: пути развития. 2021. № 6 (120). — С. 59–61.
23. Айзерман М. А. Метод потенциальных функций в теории обучения машин / М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр. — М.: Наука, 1970. — 384 с.
24. Алимов Ю. И. Альтернатива методу математической статистики / Ю. И. Алимов. — М: Знание, 1980. — 64 с.
25. Андерсон Т. В. Введение в многомерный статистический анализ / Т. В. Андерсон. — М: Наука, 1963. — 500 с.
26. Антипова Т. Н. Управление качеством технологических процессов / Т. Н. Антипова, Н. П. Асташева, Е. А. Жидкова и др. — Королёв: Научный консультант, 2015. — 189 с.
27. Антонова И. И. Всеобщее управление качеством. Основоположники всеобщего менеджмента качества / И. И. Антонова, В. А. Смирнов, С. А. Антонов. — Москва : РУСАЙНС, 2020. — 132 с.
28. Артюхов А. В. Реализация концепции цифровой среды поддержки управления конкурентоспособностью в автопроме // А. В. Артюхов, Д. В. Айдаров, Д. С. Гордиенко, В. Н. Козловский. — Известия Самарского научного центра Российской академии наук. 2022. Т. 24. № 6 (110). — С. 20–24.
29. Архипов В. А. Сравнительный анализ метрик качества для моделей бинарной классификации на примере кредитного скоринга // В. А. Архипов. — Вестник Алтайской академии экономики и права. 2019. № 9–2. — С. 12–15.

30. Атакищев О. И. Обнаружение аномалий и выбросов в наборах данных машинного обучения // О. И. Атакищев, И. А. Созыкин, В. Г. Грибунин, В. Д. Пастухов. — Известия Института инженерной физики. 2023. № 2 (68). — С. 81–87.
31. Афанасьев В. А. Техническое регулирование и управление качеством / В. А. Афанасьев, В. А. Лебедев, В. П. Монахова и др. — М.: КД Либроком, 2017. — 256 с.
32. Афанасьева А. А. Машинное обучение в кредитном скоринге: Роль метрик в задаче бинарной классификации // А. А. Афанасьева. — В сборнике: Управление инновационными и инвестиционными процессами и изменениями в современных условиях. Сборник материалов V Международной научно-практической конференции. Под редакцией Г.Л. Багиева, А.Г. Бездудной. Санкт-Петербург, 2022. — С. 468–472.
33. Афиногентова Н. В. Повышение качества обслуживания в сервисных компаниях через гармонизацию базовых требований стандарта ИСО 9001 // Н. В. Афиногентова, В. Н. Козловский. — Школа университетской науки: парадигма развития. 2016. № 3–4 (21–22). — С. 240–242.
34. Афиногентова Н. В. Роль качества сервисного обслуживания в конкурентоспособности организации // Н. В. Афиногентова, В. Н. Козловский. — Школа университетской науки: парадигма развития. 2016. № 3–4 (21–22). — С. 243–246.
35. Барсегян А. А. Методы и модели анализа данных: OLAP и Data Mining / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. — Спб.: БВХ-Петербург, 2004. — 336 с.
36. Басовский Л. Е. Управление качеством / Л. Е. Басовский, В. Б. Протасьев и др. — М.: Инфра-М, 2017. — 542 с.
37. Бахарева В. Б. Регрессионный анализ вовлеченности пользователя в контент // В. Б. Бахарева, В. Г. Мосин. — Тенденции развития науки и образования. 2024. № 106-11. — С. 17–21.

38. Беляев Ю. К. Вероятностные методы выборочного контроля / Ю. К. Беляев — М.: Наука, 1975. — 408 с.
39. Благовещенский Д. И. Актуализация проблемы оценки имиджа качества автобренда и основные направления ее решения // Д. И. Благовещенский, В. Н. Козловский, С. А. Васин, Д. С. Самойлова. — Известия Тульского государственного университета. Технические науки. 2021. № 6. — С. 273–279.
40. Благовещенский Д. И. Инструменты мониторинга удовлетворенности потребителей качеством автомобилей в эксплуатации // Д. И. Благовещенский, Н. А. Антонова, В. Н. Козловский, С. А. Васин. — СТИН. 2023. № 1. — С. 30–35.
41. Благовещенский Д. И. Ключевые аспекты организации службы мониторинга качества автомобилей в эксплуатации. Обобщенная группа показателей качества новых автомобилей в эксплуатации // Д. И. Благовещенский, В. Н. Козловский, С. А. Васин, Н. А. Антонова. — Известия Тульского государственного университета. Технические науки. 2022. № 7. — С. 239–247.
42. Благовещенский Д. И. Ключевые компоненты развития имиджа автобренда с точки зрения потребительской удовлетворенности качеством продукции и услуг // Д. И. Благовещенский, Н. А. Антонова, В. Н. Козловский, С. А. Васин. — СТИН. 2023. № 1. — С. 38–40.
43. Благовещенский Д. И. Комплексная программа улучшений как инструмент развития организации: Стратегический уровень организации и управления // Д. И. Благовещенский, В. Н. Козловский, В. Г. Мосин, О. И. Антипова. — Известия Тульского государственного университета. Технические науки. 2024. № 7. — С. 3–6.
44. Благовещенский Д. И. Модернизация традиционных инструментов оценки имиджа качества автомобильного бренда // Д. И. Благовещенский, В. Н. Козловский, С. А. Васин, Р. Р. Гафаров. — Известия Тульского государственного университета. Технические науки. 2021. № 6. — С. 285–294.

45. Благовещенский Д. И. Проблемы и направления развития цифровых инструментов и индикаторов мониторинга качества автомобилей в эксплуатации // Д. И. Благовещенский, В. Н. Козловский, С. А. Васин. — СТИН. 2021. № 10. — С. 21–25.
46. Благовещенский Д. И. Статистические инструменты обеспечения оперативного мониторинга качества продукции на этапах жизненного цикла // Д. И. Благовещенский, В. Н. Козловский, Д. И. Панюков. — Известия Самарского научного центра Российской академии наук. 2022. Т. 24. № 2 (106). — С. 16–26.
47. Богданов В. В. Выявление аномалий в работе информационных систем с помощью машинного обучения // В. В. Богданов, Н. А. Домуховский, Д. В. Лейчук, А. Н. Синадский, Д. Е. Комаров. — Защита информации. Инсайд. 2020. № 3 (93). — С. 31–35.
48. Бринк Х. Машинное обучение / Х. Бринк, Дж. Ричардс, М. Феверолф. — СПб.: Питер, 2017. — 336 с.
49. Броневиц А. Г. Нечеткие модели анализа данных и принятия решений / А. Г. Броневиц, А. Е. Лепский. — М.: ВШЭ, 2022. — 264 с.
50. Брюс П. Практическая статистика для специалистов Data Science / П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.
51. Бурков А. Инженерия машинного обучения / А. Бурков. — М: ДМК Пресс, 2022. — 306 с.
52. Бурков А. Машинное обучение без лишних слов / А. Бурков. — СПб: Питер, 2020. — 192 с.
53. Бутакова М. А. Обнаружение аномалий в информационных потоках и экспериментальные исследования на моделях машинного обучения // М. А. Бутакова, А. Н. Гуда, А. А. Чернов. — Вестник Ростовского государственного университета путей сообщения. 2020. № 2 (78). — С. 153–162.
54. Вапник В. Н. Восстановление зависимостей по эмпирическим данным / В. Н. Вапник. — М.: Наука, 1979. — 448 с.

55. Вапник В. Н. Теория распознавания образов / В. Н. Вапник, А. Я. Червоненкис. — М.: Наука, 1974. — 417 с.
56. Варжапетян А. Менеджмент качества. Принятие решений о качестве, управляемом заказчиком / А. Варжапетян, и др. - М.: Вузовская книга, 2007. — 360 с.
57. Воронцов К. В. Система эмпирического измерения качества алгоритмов классификации // К. В. Воронцов, А. С. Инякин, А. В. Лисица. — Всероссийская конференция «Математические методы распознавания образов-13». — М.: МАКС Пресс, 2007. — С. 577–580.
58. Вумек Дж. П., Джонс Д. Т. Бережливое производство: как избавиться от потерь и добиться процветания вашей компании / Дж П. Вумек, Д. Т. Джонс. — пер. с англ. М.: Альпина Бизнес Букс, 2004. — 473 с.
59. Вумек Дж., Джонс Д. Бережливое обеспечение: Как построить эффективные и взаимовыгодные отношения между поставщиками и потребителями / Дж П. Вумек, Д. Т. Джонс. — пер. с англ. М.: Альпина Бизнес Букс, 2006. — 264 с.
60. Вьюгин, В. В. Математические основы теории машинного обучения и прогнозирования / В. В. Вьюгин. — М.: МЦИМО. — 2013. — 387 с.
61. Гамзаев Р. Г. Применение машинного обучения для поиска аномалий сетевого трафика с целью выявления и предупреждений инцидентов информационной безопасности // Р. Г. Гамзаев, Г. К. Масков, В. М. Моргунов. — Методы и технические средства обеспечения безопасности информации. 2023. № 32. — С. 129–131.
62. Гельфанд А. М. Развитие алгоритмов машинного обучения для выявления аномалий в радиочастотном спектре // А. М. Гельфанд, С. А. Руденко. — В сборнике: Актуальные проблемы инфотелекоммуникаций в науке и образовании. Сборник научных статей XIII Международной научно-технической и научно-методической конференции в 4 т. Санкт-Петербург, 2024. — С. 238–241.
63. Гембрис С. Управление качеством / С. Гембрис, Й. Геррманн. — М.:

СмартБук, 2013. — 128 с.

64. Гличев А. В. Основы управления качеством продукции / А. В. Гличев. — М.: АМИ, 1998. — 356 с.
65. Годлевский В. Е. Применение статистических методов в автомобилестроении / В. Е. Годлевский. — Самара: ГП «Перспектива», 2003. — 196 с.
66. Головкин В. А. Нейронные сети: обучение, организация и применение / В. А. Головкин. — М.: ИПР-ЖР, 2001. — 257 с.
67. Гололобов Н. В. Сравнение эффективности выявления аномалий алгоритмами машинного обучения без учителя // Н. В. Гололобов, Е. Ю. Павленко. — Проблемы информационной безопасности. Компьютерные системы. 2022. № 2. — С. 135–147.
68. Горячкин Б. С. Анализ чувствительности метрик бинарной классификации к дисбалансу данных // Б. С. Горячкин, А. А. Чечнев. — E-Scio. 2021. № 4 (55). — С. 23–34.
69. Грозденский С. Я. Большие данные: история, перспективы, потенциал // С. Я. Грозденский, Е. А. Калачева. — Стандарты и качество. 2017. № 8. — С. 64–67.
70. Губарев А. В. Информационное обеспечение системы менеджмента качества / А. В. Губарев. — М.: ГЛТ, 2013. — 132 с.
71. Гурина А. О. Метод обнаружения аномалий на основе машинного обучения // А. О. Гурина, В. Л. Елисеев, С. А. Петренко, Д. Д. Ступин. — В сборнике: Материалы XII мультиконференции по проблемам управления (МКПУ-2019). Материалы XII мультиконференции: в 4 томах. 2019. — С. 60–65.
72. Гусев А. В. Концепция, логика, инструментарий, информативные связи при взаимодействии автопроизводителя и поставщиков автомобильных компонентов // А. В. Гусев, В. Н. Козловский, В. Г. Мосин, И. А. Беляева. — Известия самарского научного центра Российской академии наук. 2025. Т. 27, № 5. — С. 94–101.
73. Дайзенрот М. П. Математика в машинном обучении / М. П. Дайзенрот,

- А. А. Дальдо, Ч. С. Он. — Спб.: Питер, 2024. — 512 с.
74. Демкин А. С. Развитие методов мониторинга качества автомобилей в период гарантийной эксплуатации / 5 А.С. Демкин, А.С. Саксонов, В.Н. Козловский, В.Г. Мосин // Известия Самарского научного центра Российской академии наук. 2025. Т. 27, № 5. С. 166–171.
75. Джордж М. Л. Бережливое производство + шесть сигм в сфере услуг: как скорость бережливого производства и качество шести сигм помогают совершенствованию бизнеса / М. Л. Джордж. — 2-е изд. М.: Манн, Иванов и Фербер, 2017. — 447 с.
76. Джуран. Дж. Качество в истории цивилизации. Эволюция, тенденции и перспективы управления качеством: в 3 т. / под ред. Дж. Джурана. — М.: РИА «Стандарты и качество», 2004.
77. Дид Э. Методы анализа данных / Э. Дид. — М.: Финансы и статистика, 1985. — 357 с.
78. Добровидов А. В. Выбор ширины окна ядерной функции в непараметрической оценке производной плотности методом сглаженной кросс-валидации // А. В. Добровидов, И. М. Рудько. — Автоматика и телемеханика. 2010. № 2. — С. 42–58.
79. Домбровский В. В оптимальные стратегии прогнозирующего управления системами со случайными параметрами, описываемыми многомерной регрессионной моделью с марковским переключением режимов // В. В. Домбровский, Т. Ю. Пашинская. — Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2019. № 48. — С. 4–12.
80. Дремина М. А. Проектный подход к разработке и внедрению систем менеджмента качества / М. А. Дремина, В. А. Копнов, А. А. Станкин. — СПб.: Лань, 2015. — 304 с.
81. Друкер П. Ф. Задачи менеджмента в XXI веке / П. Ф. Друкер. — пер. с англ. М.: Вильямс, 2000. — 272 с.

82. Дунаева А. И. Сравнение методов машинного обучения в задачах обнаружения аномалий в больших данных // А. И. Дунаева, А. В. Истранин. — В сборнике: Вопросы управления и экономики: современное состояние актуальных проблем. Сборник статей по материалам II международной научно-практической конференции. 2017. — С. 7–18.
83. Дшхунян В. Л. Процессы и менеджмент качества в развитии экономических успехов предприятия / В. Л. Дшхунян, Т. Г. Никольская. — М.: Трек, 2005. — 144 с.
84. Загоруйко Н. Г. Алгоритмы обнаружения эмпирических закономерностей / Н. Г. Загоруйко, В. Н. Ёлкина, Г. С. Лбов. — Новосибирск: Наука, 1985. — 61 с.
85. Загоруйко Н. Г. Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. — Новосибирск: ИМ СО РАН, 1999. — 261 с.
86. Закс Ш. Теория статистических выводов / Ш. Закс. — М.: Мир, 1975. — 779 с.
87. Ивкин А. Н. Протокол обнаружения аномалий с внедренным модулем негативной селекции и правилами машинного обучения // А. Н. Ивкин. — Инфокоммуникационные технологии. 2019. Т. 17. № 2. С. — 189–193.
88. Имаи М. Гемба кайдзен. Путь к снижению затрат и повышению качества / М. Имаи. — М.: Альпина Паблишер, 2010. — 336 с.
89. Имаи М. Кайдзен. Ключ к успеху японских компаний / М. Имаи. — М.: Альпина Паблишер, 2014. — 274 с.
90. Исаев Л. К., Малинский В. Д. Обеспечение качества: стандартизация, единство измерений, оценка соответствия / Л. К. Исаев, В. Д. Малинский. — М.: Изд-во стандартов, 2001. — 274 с.
91. Исикава К. Японские методы управления качеством / К. Исикава. — сокр. пер. с англ. М.: Экономика, 1988. — 256 с.
92. Казинцев А. В. Шесть Сигм в России. Методика снижения потерь, дефектов, издержек / А. В. Казинцев. — М.: Типография «Новости», 2009. — 368 с.

93. Кане, М. М. Системы, методы и инструменты менеджмента качества / М. М. Кане. — М. ; СПб. ; Н. Новгород : Питер, 2008. — 560 с.
94. Караваев А. В. Оценка важности категориальных признаков с использованием one-hot-кодирования для модели линейной и гребневой регрессии // А. В. Караваев, В. Г. Мосин. — В сборнике: Прикладная математика и информатика: современные исследования в области естественных и технических наук. VIII Международная научно-практическая конференция (школа-семинар) молодых ученых: сборник материалов. Тольятти, 2022. — С. 174–179.
95. Карташевская Е. С. Сравнительный анализ методов машинного обучения без учителя в задачах выявления аномалий в сетях интернета вещей // Е. С. Карташевская. — Инфокоммуникационные технологии. 2022. Т. 20. № 4. — С. 73–79.
96. Клейменов С. И. Проблема скорости реагирования автопроизводителя на производственные дефекты автомобилей в эксплуатации // С. И. Клейменов, С. А. Шанин, В. Н. Козловский. — В сборнике: Управление качеством. Избранные научные труды семнадцатой Международной научно-практической конференции. 2018. — С. 212–218.
97. Козловский В. Н. Актуальные вопросы стратегического менеджмента качества автомобильных корпораций // В. Н. Козловский, Д. В. Айдаров, С. И. Клейменов, С. А. Шанин. — Автомобильная промышленность. 2019. № 3. — С. 1–7.
98. Козловский В. Н. Актуализированный комплекс показателей мониторинга качества автомобилей в эксплуатации // В. Н. Козловский, Д. В. Айдаров, Д. И. Панюков, М. В. Шакурский, С. А. Шанин. — Автомобильная промышленность. 2020. № 10. — С. 5–12.
99. Козловский В. Н. Алгоритм оценки качества производственно-сервисной системы высокотехнологичного предприятия. Часть 1 // В. Н. Козловский, Е. А. Вакулич, Н. Р. Шахов. — Методы менеджмента качества. 2021. № 7. — С. 38–45.

100. Козловский В. Н. Алгоритм оценки качества производственно-сервисной системы высокотехнологичного предприятия. Часть 2 // В. Н. Козловский, Е. А. Вакулич, Н. Р. Шахов. — Методы менеджмента качества. 2021. № 8. — С. 28–33.
101. Козловский В. Н. Аспекты создания профессиональной группы экспертов при решении проблем качества продукции автопрома // В. Н. Козловский, Д. В. Антипов, Д. И. Панюков, С. И. Клейменов. — Автомобильная промышленность. 2018. № 5. — С. 1–6.
102. Козловский В. Н. Балльная оценка качества деятельности предприятий фирменного автосервиса // В. Н. Козловский, Д. И. Благовещенский, Н. Р. Шахов. — Методы менеджмента качества. 2021. № 3. — С. 22–29.
103. Козловский В. Н. Важные аспекты организации аналитической деятельности в области управления качеством новых автомобилей в эксплуатации // В. Н. Козловский, Г. В. Смирнов, А. Н. Чекмарев, Д. В. Антипов. — Известия Самарского научного центра Российской академии наук. 2017. Т. 19. № 6. — С. 5–12.
104. Козловский В. Н. Вероятностно-статистическое моделирование в вопросах цифровизации процессов управления конкурентоспособностью // В. Н. Козловский, Д. В. Айдаров, Г. Л. Юнак, С. И. Клейменов. — Методы менеджмента качества. 2018. № 9. — С. 26–32.
105. Козловский В. Н. Вклад качества работы фирменной сети автосервиса в конкурентоспособность автопроизводителя // В. Н. Козловский, Д. И. Благовещенский, Н. Р. Шахов, М. М. Васильев. — Известия Тульского государственного университета. Технические науки. 2021. № 4. — С. 264–274.
106. Козловский В. Н. Воспринимаемая потребителями ценность качества продукции автомобильной промышленности // В. Н. Козловский, Д. В. Айдаров, Д. И. Панюков, М. М. Васильев. — Автомобильная промышленность. 2019. № 1. — С. 1–7.
107. Козловский В. Н. Дистанционный мониторинг качества деятельности предприятий автосервиса // В. Н. Козловский, Н. В. Афиногентова,

- С. А. Дементьев. — Стандарты и качество. 2016. № 3. — С. 92–95.
108. Козловский В. Н. Дистанционный мониторинг качества фирменного автосервиса // В. Н. Козловский, Д. И. Благовещенский, Н. Р. Шахов. — Методы менеджмента качества. 2021. № 2. — С. 46–52.
109. Козловский В. Н. Имидж бренда автопроизводителя с позиции качества продукции // В. Н. Козловский, Д. И. Благовещенский, А. С. Клентак. — Автомобильная промышленность. 2024. № 3. — С. 1–5.
110. Козловский В. Н. Инструменты комплексных улучшений качества работы предприятий фирменной сети автосервиса // В. Н. Козловский, Д. И. Благовещенский, Н. Р. Шахов, Л. А. Васин. — Известия Тульского государственного университета. Технические науки. 2021. № 4. — С. 286–293.
111. Козловский В. Н. Инструменты оценки качества продукции при исследовании перспективных автомобильных рынков // В. Н. Козловский, С. А. Васин, Д. И. Благовещенский, О. В. Пантюхин. — В сборнике: Отечественный и зарубежный опыт обеспечения качества в машиностроении. IV Всероссийская научно-техническая конференция с международным участием: сборник докладов. Тула, 2023. С. — 116–119.
112. Козловский В. Н. Интеллектуализация управления качеством: Проблемы и перспективы // В. Н. Козловский, Г. Л. Юнак, Д. В. Айдаров, С. И. Клейменов. — Стандарты и качество. 2018. № 2. — С. 80–85.
113. Козловский В. Н. Интеллектуальная информационная система диагностики состояния автономных транспортных объектов // В. Н. Козловский, С. В. Петровский, А. П. Новикова. — Фундаментальные исследования. 2016. № 6–1. — С. 73–77.
114. Козловский В. Н. Информационно-аналитический метод определения ключевых проблем надежности в автомобилестроении // В. Н. Козловский, А. В. Заятров, Д. И. Панюков. — Грузовик. 2016. № 3. — С. 39–45.
115. Козловский В. Н. Комплекс мониторинга качества автомобилей //

- В. Н. Козловский, В. В. Иванов, А. В. Заятров, Д. И. Панюков. — Грузовик. 2017. № 3. — С. 30–39.
116. Козловский В. Н. Методика контроля эффективности деятельности предприятий сервисной сети автопроизводителя // В. Н. Козловский, С. А. Шанин, С. А. Дементьев. — Автомобильная промышленность. 2017. № 10. — С. 1–6.
117. Козловский В. Н. Методология анализа и прогнозирования качества автомобилей в эксплуатации // В. Н. Козловский, Д. В. Антипов, А. В. Заятров. — Актуальные проблемы экономики. 2016. Т. 186. № 12. — С. 387–398.
118. Козловский В. Н. Модели мониторинга качества автомобилей в эксплуатации // В. Н. Козловский, Д. И. Панюков, Н. В. Афиногентова, А. В. Заятров. — Автомобильная промышленность. 2017. № 1. — С. 1–5.
119. Козловский В. Н. Мониторинг качества автомобилей в эксплуатации // В. Н. Козловский, Д. И. Панюков, А. В. Заятров, С. А. Шанин. — Автомобильная промышленность. 2017. № 2. — С. 1–5.
120. Козловский В. Н. Мониторинг качества работы предприятий автосервиса // В. Н. Козловский, Н. Р. Шахов. — Автомобильная промышленность. 2022. № 2. — С. 1–5.
121. Козловский В. Н. Мониторинг удовлетворенности потребителей качеством автомобилей // В. Н. Козловский, Д. В. Антипов, Д. И. Панюков. — Стандарты и качество. 2016. № 6. — С. 100–105.
122. Козловский В. Н. Нелинейные процессы оптимизации в рамках цифровой среды управления качеством / В. Н. Козловский, Д. В. Айдаров, С. И. Клейменов, М. М. Васильев // Методы менеджмента качества. — 2019. № 3. — С. 26–33.
123. Козловский В. Н. Определение целей в области качества для предприятий автомобильной промышленности // В. Н. Козловский, Д. И. Панюков, Р. Р. Гафаров. — В сборнике: Избранные научные труды двадцатой Международной научно-практической конференции «Управление качеством».

Москва, 2021. — С. 207–211.

124. Козловский В. Н. Организация и управление комплексными мероприятиями по улучшению качества продукции и услуг в автомобилестроении // В. Н. Козловский, Д. И. Благовещенский, Д. В. Айдаров, Н. Р. Шахов. — Автомобильная промышленность. 2021. № 10. — С. 1–5.
125. Козловский В. Н. Оценка качества деятельности предприятий фирменного автосервиса // В. Н. Козловский, Н. Р. Шахов. — Методы менеджмента качества. 2021. № 1. — С. 38–43.
126. Козловский В. Н. Оценка потребительской ценности качества продукции автомобилестроения // В. Н. Козловский, Д. В. Айдаров, Д. И. Благовещенский, Д. И. Панюков. — Автомобильная промышленность. 2021. № 8. — С. 1–7.
127. Козловский В. Н. Перспективные системы диагностики управления автономным транспортным объектом // В. Н. Козловский, В. В. Дебелов, О. И. Деев, А. Ф. Колбасов, С. В. Петровский, А. П. Новикова. — Грузовик. 2017. № 6. — С. 21–28.
128. Козловский В. Н. Потребительская ценность качества автомобилей // В. Н. Козловский, Г. Л. Юнак, Д. В. Айдаров, С. А. Шанин. — Стандарты и качество. 2017. № 12. — С. 76–80.
129. Козловский В. Н. Потребительская ценность качества: Апробация инструментария // В. Н. Козловский, Д. В. Айдаров, Д. И. Благовещенский. — Методы менеджмента качества. 2020. № 9. — С. 50–54.
130. Козловский В. Н. Потребительская ценность качества: Метод прогнозирования // В. Н. Козловский, Д. В. Айдаров, Д. И. Благовещенский. — Методы менеджмента качества. 2020. № 8. — С. 44–51.
131. Козловский В. Н. Потребительская ценность качества: Методология формирования и оценки // В. Н. Козловский, Д. В. Айдаров, Д. И. Благовещенский, Д. И. Панюков. — Методы менеджмента качества. 2020. № 6. — С. 34–41.

132. Козловский В. Н. Потребительская ценность качества: Первичный анализ данных // В. Н. Козловский, Д. В. Айдаров, Д. И. Благовещенский, Д. И. Панюков. — Методы менеджмента качества. 2020. № 7. — С. 34–40.
133. Козловский В. Н. Проблемы и перспективы развития системы мониторинга показателей качества автомобилей в эксплуатации // В. Н. Козловский, А. В. Седельников, А. И. Хаймович, А. Н. Чекмарев. — Известия Самарского научного центра Российской академии наук. 2017. Т. 19. № 1. — С. 31–39.
134. Козловский В. Н. Прогнозное моделирование в вопросах обеспечения гарантии новых автомобилей // В. Н. Козловский, Д. И. Панюков, Н. В. Афиногентова. — Актуальные проблемы экономики. 2015. Т. 164. № 2. — С. 420–426.
135. Козловский В. Н. Программы улучшения: Мифы и реальность // В. Н. Козловский, Г. Л. Юнак, Д. И. Благовещенский, А. С. Клентак. — Стандарты и качество. 2020. № 5. — С. 87–91.
136. Козловский В. Н. Разработка и развитие систем менеджмента качества автомобильных корпораций // В. Н. Козловский, Д. В. Айдаров, Д. И. Панюков, С. А. Шанин. — Автомобильная промышленность. 2019. № 5. — С. 1–6.
137. Козловский В. Н. Разработка и реализация аналитического комплекса по определению целей в области качества для предприятий автомобильной промышленности // В. Н. Козловский, Д. В. Айдаров, М. М. Васильев, С. А. Шанин. — Автомобильная промышленность. 2018. № 4. — С. 1–6.
138. Козловский В. Н. Разработка и реализация подхода к определению целей в области качества автопроизводителя по результатам гарантийной эксплуатации автомобилей // В. Н. Козловский, Д. В. Айдаров, С. А. Шанин. — Наука и техника. 2018. Т. 17. № 4. — С. 320–330.
139. Козловский В. Н. Разработка методики межфункциональной проектной деятельности для решения проблем качества автомобилей в эксплуатации // В. Н. Козловский, Д. В. Айдаров, У. В. Брачунова, А. Д. Муталов. —

Известия Самарского научного центра Российской академии наук. 2020. Т. 22. № 1 (93). — С. 24–30.

140. Козловский В. Н. Ранжирование проблем качества новых автомобилей // В. Н. Козловский, Д. И. Панюков, А. В. Заятров. — Автомобильная промышленность. 2016. № 12. — С. 1–5.
141. Козловский В. Н. Самонастраиваемая целевая функция в системе оценки качества // В. Н. Козловский, Г. Л. Юнак, Д. В. Айдаров, Р. Р. Гафаров. — Стандарты и качество. 2019. № 6. — С. 94–99.
142. Козловский В. Н. Система обратной связи с потребителем при измерении качества автомобилей // В. Н. Козловский, Д. В. Айдаров, Д. И. Панюков. — Методы менеджмента качества. 2019. № 5. — С. 32–36.
143. Козловский В. Н. Современный аналитический комплекс измерения качества автомобилей в эксплуатации // В. Н. Козловский, Н. В. Афиногенова, А. В. Заятров, Е. В. Полякова. — Автомобильная промышленность. 2016. № 4. — С. 1–3.
144. Козловский В. Н. Статистическое управление качеством продукции и сложных процессов // В. Н. Козловский, Г. Л. Юнак, Д. И. Благовещенский, А. Г. Сорокин. — Стандарты и качество. 2021. № 10. — С. 98–104.
145. Козловский В. Н. Стратегическое планирование конкурентоспособности с точки зрения качества // В. Н. Козловский, С. А. Шанин, Д. И. Панюков. — Стандарты и качество. 2017. № 3. — С. 76–80.
146. Козловский В. Н. Стратегия цифровизации: Новое качество производства // В. Н. Козловский, С. И. Клейменов, Д. В. Айдаров, Д. И. Панюков. — Методы менеджмента качества. 2020. № 2. — С. 46–55.
147. Козловский В. Н. Управление качеством в эпоху цифровой трансформации // В. Н. Козловский, Д. И. Благовещенский, С. И. Клейменов, Н. Р. Шахов. — Автомобильная промышленность. 2021. № 2. — С. 1–6.
148. Козловский В. Н. Цифровая среда поддержки аналитики и управления конкурентоспособностью в автомобилестроении // В. Н. Козловский,

- С. И. Строганов, Д. В. Айдаров, Р. Р. Гафаров. — Автомобильная промышленность. 2020. № 3. — С. 1–3.
149. Козловский В. Н. Цифровая среда поддержки управления конкурентоспособностью // В. Н. Козловский, Д. В. Айдаров, Д. И. Панюков, М. М. Васильев. — Стандарты и качество. 2018. № 6. — С. 86–89.
150. Козловский В. Н. Цифровизация производства: Новый формат статистических инструментов управления качеством // В. Н. Козловский, Г. Л. Юнак, С. И. Клейменов, Д. И. Благовещенский. — Стандарты и качество. 2020. № 7. — С. 102–107.
151. Козловский В. Н. Цифровое производство: Качество должно быть прогнозируемым // В. Н. Козловский, Д. В. Айдаров, С. И. Клейменов, А. В. Крицкий. — Стандарты и качество. 2020. № 3. — С. 73–77.
152. Козловский В. Н. Эффективность улучшений как отражение уровня аналитики // В. Н. Козловский, Г. Л. Юнак, Д. В. Айдаров. — Стандарты и качество. 2018. № 12. — С. 56–60.
153. Козырева Н. И. Применение машинного обучения для обнаружения аномалий в поведении пользователя // Н. И. Козырева, М. М. Шапко. — В сборнике: VI Научный форум «Телекоммуникации: теория и технологии» ТТТ-2023. Материалы XXV Международной научно-технической конференции. Казань, 2023. — С. 486–487.
154. Кондо Й. Управление качеством в масштабах компании: становление и этапы развития / Й. Кондо. — Н. Новгород: СМЦ «Приоритет», 2002. — 252 с.
155. Конти Т. Качество в XXI веке. Роль качества в обеспечении конкурентоспособности и устойчивого развития / Т. Конти, Ё. Кондо, Г. Ватсон, пер. с англ. А. Раскина. — М.: РИА «Стандарты и качество», 2005. — 280 с.
156. Коул Р. Блистательный Agile. Гибкое управление проектами с помощью Agile, Scrum и Kanban / Р. Коул, Э. Скотчер. — СПб: Питер, 2018. — 302 с.
157. Кох Р. Принцип 80/20 / Р. Кох. — М.: Эксмо, 2012 — 448 с.
158. Кочедыков Д. А. Применение логических алгоритмов классификации в

- задачах кредитного скоринга и управления риском кредитного портфеля банка // Д. А. Кочедыков, А. А. Ивахненко, К. В. Воронцов. — Всероссийская конференция «Математические методы распознавания образов-13», М.: МАКС Пресс, 2007. — С. 484–488.
159. Кочедыков Д. А. Система кредитного скоринга на основе логических алгоритмов классификации // Д. А. Кочедыков, А. А. Ивахненко, К. В. Воронцов. — Всероссийская конференция «Математические методы распознавания образов-12», М.: МАКС Пресс, 2005. — С. 349–353.
160. Кощеев М. И. Машинное обучение при определении аномалий в работе электроэнергетического объекта для обеспечения информационной безопасности // М. И. Кощеев, М. В. Никандров. — В сборнике: Проблемы и перспективы развития энергетики, электротехники и энергоэффективности. Материалы II Международной научно-технической конференции. 2018. — С. 221–226.
161. Коэн М. И. Прикладная линейная алгебра для исследователей данных / М. И. Коэн. — М.: ДМК Пресс, 2023. — 328 с.
162. Крянев А. В. Метрический анализ и обработка данных / А. В. Крянев, Г. В. Лукин, Д. К. Удумян. — М.: Физматлит, 2012. — 308 с.
163. Лагутин М. Б. Наглядная математическая статистика / М. Б. Лагутин. — М.: П-центр, 2003 — 475 с.
164. Лapidус В. А. Всеобщее качество (TQM) в российских компаниях / В. А. Лapidус. — Н. Новгород: ООО СМЦ «Приоритет», 2008. — 432 с.
165. Лапко А. В. Обучающиеся системы обработки информации и принятия решений. Непараметрический подход / А. В. Лапко, С. В. Ченцов, С. И. Крохов, Л. А. Фельдман. — Новосибирск: Наука, 1996. — 295 с.
166. Лбов Г. С. Методы обработки разнотипных экспериментальных данных / Г. С. Лбов. — Новосибирск: Наука, 1981. — 160 с.
167. Лбов Г. С. Логические решающие функции и вопросы статистической устойчивости решений / Г. С. Лбов, Н. Г. Старцева. — Новосибирск: Изд-во Института математики СО РАН, 1999. — 211 с.

168. Лесковец Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджамаран, Д. Ульман. — М.: ДМК Пресс, 2016. — 498 с.
169. Литвак Б. Г. Экспертные оценки и принятие решений / Б. Г. Литвак. — М.: Патент, 1996. — 271 с.
170. Магомедов Ш. Г. Обнаружение аномалий с помощью машинного обучения и анализа графов в задачах борьбы с цифровым мошенничеством // Ш. Г. Магомедов, С. А. Павельев, А. С. Добротворский. — В сборнике: Информационные технологии в экономике и управлении. Материалы III Всероссийской научно-практической конференции. 2018. — С. 50–55.
171. Мазур И. И. Управление качеством / И. И. Мазур, В. Д. Шапиро. — М.: Высшая школа, 2004. — 336 с.
172. Мандель И. Д. Кластерный анализ / И. Д. Мандель. — М.: Финансы и статистика, 1988. — 177 с.
173. Маслов Д. В. От качества к совершенству. Полезная модель EFQM / Д. В. Маслов. — М.: Стандарты и качество, 2008. — 150 с.
174. Матушкин В. Д. Применение кроссвалидации для оценки качества распознавания аварий на аэс с применением искусственных нейронных сетей // В. Д. Матушкин, Ю. Б. Воробьев, М. К. Седов. — В книге: Безопасность АЭС и подготовка кадров. Тезисы докладов. Обнинск, 2023. — С. 106–107.
175. Меджедович Д. Алгоритмы и структуры для массивных наборов данных / Д. Меджедович, Э. Тахирович. — М.: ДМК Пресс, 2024. — 340 с.
176. Мерков А. Б. Распознавание образов. Построение и обучение вероятностных моделей / А. Б. Мерков. — М.: URSS, 2014. — 238 с.
177. Мерков А. Б. Распознавание образов. Введение в методы статистического обучения / А. Б. Мерков. — М.: URSS, 2011. — 256 с.
178. Минько Э. В. Качество и конкурентоспособность / Э. В. Минько, М. Л. Кричевский. — СПб. : Питер, 2004. — 267 с.
179. Мосин В. Г. Обоснование методологии детекции аномалий в качестве нового цифрового инструмента управления качеством // В. Г. Мосин,

В. Н. Козловский. — Известия Самарского научного центра Российской академии наук, т. 26, № 3, 2024, С. 16-22.

180. Мосин В. Г. Кодирование категориальных данных в задачах регрессионного анализа // В. Г. Мосин, И. П. Егорова. — В сборнике: Инновационное развитие современной науки: Теория, методология, практика. Сборник статей II Всероссийской научно-практической конференции. Петрозаводск, 2021. — С. 163–169.
181. Мосин В. Г. Линеаризация целевой функции в регрессионных задачах методом сингулярных разложений // В. Г. Мосин. — В книге: Математическое моделирование. Тезисы II Международной конференции. Москва, 2021. — С. 66–67.
182. Мосин В. Г. Анализ качества электронного контента. Разбиение данных о потреблении контента на оптимальное число кластеров при помощи алгоритма KMEANS // В. Г. Мосин, В. Н. Козловский. — Известия Тульского государственного университета. Технические науки. 2024. № 2. — С. 604–609.
183. Мосин В. Г. Анализ стабильности процессов с сильно коррелирующими признаками // В. Г. Мосин, В. Н. Козловский, О. И. Антипова, С. А. Васин. — Известия Тульского государственного университета. Технические науки. 2024. № 5. — С. 11–16.
184. Мосин В. Г. Бинарная модель оценки качества и востребованности электронного контента и ее вероятностный прогноз // В. Г. Мосин, В. Н. Козловский, Д. И. Благовещенский. — Известия Тульского государственного университета. Технические науки. 2023. № 12. С. — 79–84.
185. Мосин В. Г. Востребованность контента как один из факторов качества информационного канала // В. Г. Мосин, В. Н. Козловский — Качество и жизнь. 2023. № 4 (40). — С. 6–11.
186. Мосин В. Г. Детекция аномалий информационного канала на основе прогнозирующих моделей в решении задач анализа качества контента // В. Г. Мосин, В. Н. Козловский, О. В. Пантюхин. — Известия Тульского

- государственного университета. Технические науки. 2024. № 3. — С. 421–425.
187. Мосин В. Г. Детекция аномалий информационного канала на основе прогнозирующих моделей в решении задач анализа качества контента // В. Г. Мосин, В. Н. Козловский, О. В. Пантюхин. — Известия Тульского государственного университета. Технические науки. 2024. № 3. — С. 421–425.
188. Мосин В. Г. Инструменты управления качеством. Поиск калибровочных границ методами машинного обучения // В. Г. Мосин, В. Н. Козловский, О. И. Антипова, Р. Р. Гафаров. — Известия Тульского государственного университета. Технические науки. 2024. № 7. — С. 10–15.
189. Мосин В. Г. Мониторинг данных о результативности процессов системы менеджмента. Настройка параметров кластеризации DBSCAN по подходящим значениям метрики силуэта // В. Г. Мосин, В. Н. Козловский, С. А. Васин, Г. С. Стрелков. — Известия Тульского государственного университета. Технические науки. 2024. № 4. — С. 28–33.
190. Мосин В. Г. Мониторинг качества. Сравнительный анализ двух методик поиска значимых предикторов регрессионной модели // В. Г. Мосин, В. Н. Козловский, О. В. Пантюхин, Д. В. Антипов. — Известия Тульского государственного университета. Технические науки. 2024. № 5. — С. 39–45.
191. Мосин В. Г. Новые инструменты мониторинга качества процессов: Детекция аномалий в электронных данных о потреблении контента методом кластеризации dbscan // В. Г. Мосин, В. Н. Козловский, Дм. В. Ненашев. — СТИН. 2024. № 8. — С. 49–52.
192. Мосин В. Г. Новые инструменты мониторинга качества. Анализ потребления контента на разреженных данных // В. Г. Мосин, В. Н. Козловский, О. В. Пантюхин — Известия Тульского государственного университета. Технические науки. 2024. № 1. — С. 208–214.
193. Мосин В. Г. Новые инструменты мониторинга качества. Мониторинг

- данных, временные ряды потребления контента и усиление их прогнозирующих свойств ансамблированием слабых моделей // В. Г. Мосин, В. Н. Козловский, Д. И. Благовещенский — Известия Тульского государственного университета. Технические науки. 2024. № 1. — С. 237–246.
194. Мосин В. Г. Новые инструменты управления качеством. Метод поиска локтевой точки для определения приемлемых значений метрик машинного обучения // В. Г. Мосин, В. Н. Козловский, А. П. Новикова — Известия Тульского государственного университета. Технические науки. 2024. № 4. — С. 53–56.
195. Мосин В. Г. О влиянии востребованности контента на некоторые характеристики вовлеченности в его потребление // В. Г. Мосин, В. Н. Козловский, О. В. Пантюхин — Известия Тульского государственного университета. Технические науки. 2024. № 1. — С. 250–258.
196. Мосин В. Г. О некоторых особенностях определения качества продукции на примере данных одного из ведущих отечественных автопроизводителей // В. Г. Мосин, В. Н. Козловский, Н. А. Антонова. — Известия Самарского научного центра Российской академии наук. 2023. Т. 25. № 4 (114). — С. 47–55.
197. Мосин В. Г. О некоторых проблемах моделирования измеряемых социально-психологических переменных // В. Г. Мосин — В сборнике: Математическое образование в современном мире: теория и практика. Материалы Всероссийской научно-методической конференции с международным участием. Самара, 2022. — С. 165–174.
198. Мосин В. Г. Повышение производительности регрессионных моделей при оценке качества потребления электронного контента // В. Г. Мосин, В. Н. Козловский. — Известия Тульского государственного университета. Технические науки. 2024. № 2. — С. 593–599.
199. Мосин В. Г. Понижение размерности данных методом SVD на примере данных о потреблении контента // В. Г. Мосин, В. Н. Козловский. — Известия Тульского государственного университета. Технические науки. 2024.

- № 2. — С. 599–603.
200. Мосин В. Г. Применение сингулярных разложений к обнаружению выбросов в данных о потреблении информационного контента // В. Г. Мосин, В. Н. Козловский, С. А. Васин. — Известия Тульского государственного университета. Технические науки. 2024. № 3. — С. 464–470.
201. Мосин В. Г. Сезонная структура потребления контента // В. Г. Мосин, В. Н. Козловский, Д. И. Благовещенский. — Известия Тульского государственного университета. Технические науки. 2024. № 3. — С. 490–495.
202. Мосин В. Г. Сравнительный анализ двух метрик кластеризации данных о потреблении контента // В. Г. Мосин, В. Н. Козловский. — Известия Тульского государственного университета. Технические науки. 2024. № 2. — С. 609–615.
203. Мэдфи К. П. Вероятностное машинное обучение / К. П. Мэдфи. — М: ДМК Пресс, 2023. — 990 с.
204. Наумов В. Н. Методы прогнозирования временных рядов / В. Н. Наумов. — СПб.: Лань, 2024. — 196 с.
205. Немогай Н. В. Управление качеством. Менеджмент качества / Н. В. Немогай. — Минск: ТетраСистемс, 2010. — 171 с.
206. Нив Г. Организация как система: принципы построения устойчивого бизнеса Эдвардса Деминга / Г. Нив. — пер. с англ. М.: Альпина Бизнес Букс, 2016. — 368 с.
207. Никифоров А. М. Методы и алгоритмы преобразования типов признаков в задачах анализа данных / А. М. Никифоров, Ш. Х. Фазылов. — Ташкент: Фан, 1988. — 131 с.
208. Нилд Т. Математика для Data Science / Т. Нилд. — СПб.: Питер, 2025. — 352 с.
209. Орлов А. И. Нечисловая статистика / А. И. Орлов. — М.: М3-Пресс, 2004. — 613 с.
210. Панде П. Что такое «Шесть сигм»? : революционный метод управления качеством / П. Панде, Л. Холп. Москва: Альпина Бизнес Букс, 2004. —

157 с.

211. Панюков Д. И. Формирование эффективной FMEA-команды // Д. И. Панюков, В. Н. Козловский, С. А. Шанин. — Стандарты и качество. 2017. № 7. — С. 68–72.
212. Парамонов И. Ю. Методы и модели исследования сложных систем и обработки больших данных / И. Ю. Парамонов, В. А. Смагин, Н. Е. Косых, А. Д. Хомоненко. — СПб.: Лань, 2020. — 236 с.
213. Попков Ю. С. Рандомизированное машинное обучение при ограниченных объёмах данных: от эмпирической вероятности к энтропийной рандомизации / Ю. С. Попков, А. Ю. Попков, Ю. А. Дубнов. — М.: Ленанд, 2019. — 320 с.
214. Пылов П. А. Фундаментальные типы кросс-валидации для оценки качества моделей машинного и глубокого обучения // П. А. Пылов, О. А. Ивина. — В сборнике: Россия молодая. Сборник материалов XIII Всероссийской научно-практической конференции с международным участием. Редколлегия: К.С. Костиков (отв. ред.) [и др.]. Кемерово, 2021. — С. 31520.1–31520.7.
215. Растригин Л. А. Экстраполяционные методы проектирования и управления / Л. А. Растригин, Ю. П. Пономарев. — М.: Машиностроение, 1986. — 120 с.
216. Розов Б. Б. Экономико-статистические модели с переменной структурой / Б. Б. Розов, И. В. Котюков, М. А. Ягольницер. — Новосибирск: Наука, 1984. — 243 с.
217. Рязанов В. В. О некоторых моделях голосования и методах их оптимизации // В. В. Рязанов, О. В. Сенько. Распознавание, классификация, прогноз. — Т. 3, 1990. — С. 106–145.
218. Савкова Е. Н. Методы менеджмента качества. Контроль и испытания продукции / Е. Н. Савкова, П. С. Серенков, Н. А. Жагора. — Минск: Новое знание; Москва: ИНФРА-М, 2016. — 479 с.
219. Самойлов В. Е. Выявление аномалий в трафике корпоративной сети с

- использованием алгоритмов машинного обучения // В. Е. Самойлов, О. Ю. Сейченко. — Вестник современных цифровых технологий. 2023. № 16. — С. 30–40.
220. Севернов В. В. Применение алгоритмов машинного обучения для детектирования аномалий в потоках больших данных показаний датчиков // В. В. Севернов, Б. В. Севернов. — Автоматизация. Современные технологии. 2023. Т. 77. № 10. — С. 453–459.
221. Сейновски Т. Антология машинного обучения. Важнейшие исследования в области ИИ за последние 60 лет / Т. Сейновски. — М: «Бомбора», 2021. — 440 с.
222. Серенков П. С. Методы менеджмента качества : методология организационного проектирования инженерной составляющей системы менеджмента качества / П. С. Серенков. — Минск: Новое знание; Москва: ИНФРА-М, 2011. — 490 с.
223. Серенков П. С. Методы менеджмента качества. Процессный подход / П. С. Серенков, А. Г. Курьян, В. П. Волонтей. — Минск: Новое знание; Москва: ИНФРА-М, 2014. — 440 с.
224. Серенков П. С. Методы менеджмента качества. Методология управления риском стандартизации / П. С. Серенков, В. Л. Гуревич, В. М. Романчук, А. В. Янушкевич. — Минск: Новое знание; Москва : ИНФРА-М, 2014. — 255 с.
225. Симчера В. М. Методы многомерного анализа статистических данных / В. М. Симчера. — М.: Финансы и статистика, 2008. — 400 с.
226. Старцева Е. В. Управление качеством в современной инновационной среде / Т. Е. Старцева, Н. П. Асташева, Т. Н. Антипова и др. — Королёв: Научный консультант, 2018. — 338 с.
227. Стеллман Э. Постигая Agile / Э. Стеллман, Д. Грин. — М: МИФ, 2015. — 448 с.
228. Степашкина А. С. Численные методы и машинное обучение в метрологии / А. С. Степашкина. — Москва, Вологда: Инфра-Инженерия, 2024. —

144 с.

229. Уиллер Д., Чамберс Д. Статистическое управление процессами: Оптимизация бизнеса с использованием контрольных карт Шухарта / Д. Уиллер, Д. Чамберс. — пер. с англ. М.: Альпина Бизнес Букс, 2009. — 409 с.
230. Успенская Е. О. Алгоритмы обнаружения аномалий данных на основе методов машинного обучения // Е. О. Успенская. — В книге: Радиоэлектроника, электротехника и энергетика. тезисы докладов Двадцать восьмой международной научно-технической конференции студентов и аспирантов. Москва, 2022. С. 191.
231. Фадеева М. А. Применение метрик качества моделей бинарной классификации для анализа модели данных // М. А. Фадеева — Инженерные кадры - будущее инновационной экономики России. 2023. № 1. — С. 1104–1107.
232. Федоськина Л. А. Менеджмент качества и обеспечение безопасности в автомобильном бизнесе / Л. А. Федоськина. — М.: Инфра-М, 2018. — 320 с.
233. Фейгенбаум А. Контроль качества продукции / А. Фейгенбаум. — пер. с англ. М.: Экономика, 1986. — 471 с.
234. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. — М.: ДМК Пресс, 2015. — 402 с.
235. Хардле В. Прикладная непараметрическая регрессия / В. Хардле. — М.: Мир, 1993. — 352 с.
236. Харрингтон Д. Х. Управление качеством в американских корпорациях / Д. Х. Харрингтон. — сокр. пер. с англ. М.: Экономика, 2013. — 272 с.
237. Хейдт М. Изучаем Pandas / М. Хейдт. — Москва: ДМК Пресс, 2018. — 438 с.
238. Хенсен Б. Контроль качества. Теория и применение / Б. Хенсен. — М.: Прогресс, 2007. — 400 с.

239. Хованов И. В. Математические основы теории шкал измерения качества / И. В. Хованов. — Л.: ЛГУ, 1983. — 137 с.
240. Хофманн Д. Измерительно-вычислительные системы обеспечения качества / Д. Хофманн. — М.: Энергоатомиздат, 1991. — 272 с.
241. Хрисанфова К. С. Применение полиномиальных признаков в области линейного регрессионного анализа // К. С. Хрисанфова, В. Г. Мосин, В. П. Корнеев. — В сборнике: Математика и математическое моделирование. Материалы Всероссийского конкурса научно-исследовательских работ (проектов) обучающихся. Самара, 2023. — С. 67–73.
242. Хьюберт Р. К. Общее управление качеством: Личностные и организационные изменения / Р. К. Хьюберт. — М.: ЗАО «Олимп-Бизнес», 2005. — 256 с.
243. Целых А. Н. Применение временных рядов для анализа больших данных / А. Н. Целых, В. С. Васильев, Э. М. Котов. — Южный федеральный университет, 2021 г. — 83 с.
244. Шанин С. А. Параметры тревожных сигналов в области качества автомобилей в эксплуатации // С. А. Шанин, С. И. Клейменов, В. Н. Козловский. — В сборнике: Управление качеством. Избранные научные труды семнадцатой Международной научно-практической конференции. 2018. — С. 409–414.
245. Шахов Н. Р. Базовые принципы и показатели системы цифровой, дистанционной оценки качества деятельности предприятий фирменной сети автосервиса // Н. Р. Шахов, В. Н. Козловский, С. А. Васин. — СТИН. 2022. № 5. — С. 30–34.
246. Шахов Н. Р. Основные аспекты разработки методики балльной оценки качества деятельности предприятий фирменного автосервиса // Н. Р. Шахов, В. Н. Козловский, С. А. Васин. — СТИН. 2022. № 5. — С. 34–38.
247. Шестаков О. В. Центральная предельная теорема для функции обобщенной кросс-валидации при пороговой обработке вейвлет-коэффициентов // О. В. Шестаков. — Информатика и ее применения. 2013. Т. 7. № 2. —

С. 40–49.

248. Шурыгин А. М. Прикладная стохастика: робастность, оценивание, прогноз / А. М. Шурыгин. — М.: Финансы и статистика, 2000. — 223 с.
249. Язвенко П. В. Развитие системы менеджмента качества и её современное состояние: Научный обзор // П. В. Язвенко, М. П. Воронов. — Научное обозрение. Технические науки. 2016. № 4. — С. 140–148.

СПИСОК ИЛЛЮСТРАТИВНОГО МАТЕРИАЛА

Список иллюстраций

Рисунок 1.1 — Карта Шухарта. Нормальное протекание стабильного процесса.....	33
Рисунок 1.2 — Карта Шухарта. Выход стабильного процесса за контрольные границы	33
Рисунок 1.3 — Нестабильный процесс	51
Рисунок 1.4 — Контроль нестабильного процесса по среднему значению	52
Рисунок 1.5 — Контроль нестабильного процесса по модельным значениям	53
Рисунок 1.6 — Контроль нестабильного процесса по серии константных стандартов	54
Рисунок 1.7 — Нестабильный процесс в нескольких локациях	58
Рисунок 2.1 — Портрет локации	76
Рисунок 2.2 — Портрет момента.....	77
Рисунок 2.3 — Близость локаций	78
Рисунок 2.4 — Близость моментов	79
Рисунок 2.5 — Распределение отклонений	96
Рисунок 2.6 — Вариация порога аномалий.....	96
Рисунок 3.1 — Кубические параболы из одного семейства с разными реперами.....	121
Рисунок 3.2 — Семейство реперов, порождающее семейство трендов	123
Рисунок 3.3 — Кривая тренда, проходящая через базовый репер.....	125
Рисунок 3.4 — Подобные семейства трендов	127
Рисунок 3.5 — Семейство трендов.....	128
Рисунок 3.6 — Синтетические временные ряды с различной величиной шума.....	129

Рисунок 3.7 — Портреты девяти локаций синтетического процесса....	132
Рисунок 3.8 — Портреты ста моментов синтетического процесса	133
Рисунок 3.9 — Сравнительный анализ схем моделирования по времени.....	138
Рисунок 3.10 — Сравнительный анализ схем моделирования по локациям.....	141
Рисунок 3.11. — Распределение отклонений δ	145
Рисунок 3.12 — Калибровка отклонений в формате ML по отношению к сбалансированному экспертному сообществу	154
Рисунок 3.13 — Калибровка отклонений в формате ML по отношению к экспертному сообществу с преобладанием ястребов.....	155
Рисунок 3.14 — Калибровка отклонений в формате ML по отношению к экспертному сообществу с преобладанием голубей	156
Рисунок 3.15 — Калибровка отклонений в формате LM по отношению к сбалансированному экспертному сообществу	161
Рисунок 3.16 — Калибровка отклонений в формате LM по отношению к экспертному сообществу с преобладанием ястребов.....	161
Рисунок 3.17 — Калибровка отклонений в формате LM по отношению к экспертному сообществу с преобладанием голубей	162
Рисунок 4.1 — Временные ряды исполнения одной из технологических операций.....	177
Рисунок 4.2 — Анализ схем моделирования по времени в процессе производства	181
Рисунок 4.3 — Укрупнённый скоринг моделей по времени в процессе производства	182
Рисунок 4.4 — Калибровочная кривая отклонений по времени в процессе производства	184
Рисунок 4.5 — Модель по времени в локации АК.....	185
Рисунок 4.6 — Отклонения от модели по времени в локации АК.	185
Рисунок 4.7 — Модель по времени в локации AZ	186

Рисунок 4.8 — Отклонения от модели по времени в локации AZ.....	186
Рисунок 4.9 — Рейтинг аномальных по времени локаций в процессе производства в момент времени 2024-04-04	187
Рисунок 4.10 — Рейтинг аномальных по времени локаций в процессе производства в момент времени 2024-09-16	188
Рисунок 4.11 — Сравнительный анализ схем моделирования моделей по локациям в процессе производства	190
Рисунок 4.12 — Калибровочная кривая отклонений по локациям в процессе производства	190
Рисунок 4.13 — Модель по локациям в локации AZ	191
Рисунок 4.14 — Отклонения от модели по локациям в локации AZ	192
Рисунок 4.15 — Модель по локациям в локации ВТ	192
Рисунок 4.16 — Отклонения от модели по локациям в локации ВТ.....	192
Рисунок 4.17 — Рейтинг аномальных по локациям локаций в процессе производства в момент времени 2023-06-12	193
Рисунок 4.18 — Рейтинг аномальных по локациям локаций в процессе производства в момент времени 2023-08-09	194
Рисунок 5.1 — Неравномерность загруженности сервисных центров..	211
Рисунок 5.2 — Сравнительный анализ схем моделирования по времени в процессе сервисного обслуживания для числа обработанных заявок ..	217
Рисунок 5.3 — Калибровочная кривая отклонений по времени в процессе сервисного обслуживания для числа обработанных заявок ..	218
Рисунок 5.4 — Модель для числа заявок по времени в локации 12954	219
Рисунок 5.5 — Отклонение модели для числа заявок по времени в локации 12954.....	219
Рисунок 5.6 — Модель для числа заявок по времени в локации 42663.....	219
Рисунок 5.7 — Отклонение модели для числа заявок по времени в локации 42663.....	220
Рисунок 5.8 — Рейтинг аномальных по времени локаций в процессе	

сервисного обслуживания для числа заявок в момент времени 2023-11-15	220
Рисунок 5.9 — Рейтинг аномальных по времени локаций в процессе сервисного обслуживания для числа заявок в момент времени 2024-03-08	221
Рисунок 5.10 — Сравнительный анализ схем моделирования по локациям в процессе сервисного обслуживания для числа обработанных заявок	223
Рисунок 5.11 — Калибровочная кривая отклонений по локациям в процессе сервисного обслуживания для числа обработанных заявок ..	224
Рисунок 5.12 — Модель для числа заявок по локациям в локации 12954	225
Рисунок 5.13 — Отклонение модели для числа заявок по локациям в локации 12954	225
Рисунок 5.14 — Модель для числа заявок по локациям в локации 42663	226
Рисунок 5.15 — Отклонение модели для числа заявок по локациям в локации 42663	226
Рисунок 5.16 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для числа заявок в момент времени 2023-11-15	227
Рисунок 5.17 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для числа заявок в момент времени 2024-03-08	227
Рисунок 5.18 — Сравнительный анализ схем моделирования по времени в процессе сервисного обслуживания по отношению к суммарной стоимости работ	229
Рисунок 5.19 — Калибровочная кривая отклонений по времени в процессе сервисного обслуживания по отношению к суммарной стоимости работ	230

Рисунок 5.20 — Модель по времени для суммарной стоимости работ в локации 12954.....	231
Рисунок 5.21 — Отклонение модели по времени для суммарной стоимости работ в локации 12954	232
Рисунок 5.22 — Модель по времени для суммарной стоимости работ в локации 42663.....	232
Рисунок 5.23 — Отклонение модели по времени для суммарной стоимости работ в локации 42663	233
Рисунок 5.24 — Рейтинг аномальных по времени локаций в процессе сервисного обслуживания для суммарной стоимости работ в момент времени 2023-11-15.....	234
Рисунок 5.25 — Рейтинг аномальных по времени локаций в процессе сервисного обслуживания для суммарной стоимости в момент времени 2024-03-08	234
Рисунок 5.26 — Сравнительный анализ схем моделирования по локациям в процессе сервисного обслуживания для суммарной стоимости работ	236
Рисунок 5.27 — Калибровочная кривая отклонений по времени в процессе сервисного обслуживания для суммарной стоимости работ .	237
Рисунок 5.28 — Модель по локациям для суммарной стоимости работ в локации 12954.....	239
Рисунок 5.29 — Отклонение модели по локациям для суммарной стоимости работ в локации 12954	239
Рисунок 5.30 — Модель по локациям для суммарной стоимости работ в локации 42663.....	240
Рисунок 5.31 — Отклонение модели по локациям для суммарной стоимости работ в локации 42663	240
Рисунок 5.32 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для суммарной стоимости работ в момент времени 2023-11-15.....	241

Рисунок 5.33 — Рейтинг аномальных по локациям локаций в процессе сервисного обслуживания для суммарной стоимости в момент времени 2024-03-08	241
Рисунок 6.1 — Сравнительный анализ схем моделирования по времени в процессе маркетинга	259
Рисунок 6.2 — Приближенный скоринг моделей по времени в процессе маркетинга	260
Рисунок 6.3 — Калибровочная кривая отклонений по времени в процессе маркетинга	261
Рисунок 6.4 — Модель по времени в локациях GGL	262
Рисунок 6.5 — Отклонения от модели по времени в локациях GGL.....	262
Рисунок 6.6 — Модель по времени в локациях TG	263
Рисунок 6.7 — Отклонения от модели по времени в локациях TG.....	263
Рисунок 6.8 — Рейтинг аномальных по времени локаций в маркетинговом процессе в момент времени 2020-12-20	264
Рисунок 6.9 — Рейтинг аномальных по времени локаций в маркетинговом процессе в момент времени 2024-03-24	265
Рисунок 6.10 — Сравнительный анализ схем моделирования по локациям в процессе маркетингового сопровождения	266
Рисунок 6.11 — Приближенный скоринг моделей по локациям в процессе маркетингового сопровождения.....	266
Рисунок 6.12 — Калибровочная кривая отклонений по локациям в процессе маркетингового сопровождения.....	267
Рисунок 6.13 — Модель по локациям в локациях GGL	269
Рисунок 6.14 — Отклонения от модели по локациям в локациях GGL..	269
Рисунок 6.15 — Модель по локациям в локациях TG	270
Рисунок 6.16 — Отклонения от модели по локациям в локациях TG	270
Рисунок 6.17 — Рейтинг аномальных по локациям локаций в процессе маркетингового сопровождения в момент времени 2023-01-15	271
Рисунок 6.18 — Рейтинг аномальных по локациям локаций в процессе	

маркетингового сопровождения в момент времени 2024-10-13	271
Рисунок 7.1 — 2024-04-04. Рейтинг аномальных локаций.....	282
Рисунок 7.2 — Временной ряд до наложения аномалий	291
Рисунок 7.3 — Временной ряд с аномалиями двух типов.....	292
Рисунок 7.4 — Прогнозирующая модель процесса.....	293
Рисунок 7.5 — Калибровка отклонений при наличии обеих причин аномальных состояний	294
Рисунок 7.6 — Детекция аномалий до устранения причины А	294
Рисунок 7.7 — Временной ряд после устранения причины А, начиная с момента 500	295
Рисунок 7.8 — Калибровочные кривые до и после устранения причины А.....	296
Рисунок 7.9 — Коридор допустимых отклонений до и после устранения причины А.....	297
Рисунок 7.10 — Сужение коридора допустимых отклонений в различных локациях синтетического процесса в результате устранения причин аномалии А.....	298
Рисунок 7.11 — а) стабильный процесс с отклонением от среднего, б) нестабильный процесс с отклонением от среднего, с) нестабильный процесс с отклонением от модели	299
Рисунок А.1 — Регрессия временного ряда с линейным трендом	367
Рисунок А.2 — Линейная регрессия временного ряда с нелинейным трендом.....	368
Рисунок А.3 — Полиномиальная кубическая регрессия временного ряда с нелинейным трендом	369
Рисунок А.4 — Полиномиальная регрессия 15 степени временного ряда с нелинейным трендом	369
Рисунок А.5 — Интуитивная иллюстрация метрических методов	371
Рисунок А.5 — Аппроксимация нелинейного тренда решающим деревом глубины 1	385

Рисунок А.6 — Аппроксимация нелинейного тренда решающим деревом глубины 3	386
Рисунок А.7 — Аппроксимация нелинейного тренда решающим деревом глубины 5	386
Рисунок А.8 — Две случайные выборки из исходного временного ряда	388
Рисунок А.9 — Два решающих дерева на двух случайных выборках ..	388
Рисунок А.10 — Соотношение метрик precision и recall при изменении параметров классификатора.....	402
Рисунок А.11 — ROC-кривые и метрики ROC AUC	405
Рисунок А.12 — Классическая кросс-валидация (а) и кросс-валидация временного ряда (b).....	413

Список таблиц

Таблица 1.1 — Процесс в формате ML (Moment, Location)	59
Таблица 1.2 — Процесс в формате LM (Location, Moment)	60
Таблица 1.3 — Процесс в формате ML с дополнительными признаками.....	62
Таблица 1.4 — Процесс в формате LM с дополнительными признаками.....	62
Таблица 1.5 — Процесс в формате MLC (Moment, Location, Characteristic)	64
Таблица 1.6 — Процесс в формате LMC (Location, Moment, Characteristic)	64
Таблица 2.1 — Формат ML: нуль-мерные данные	70
Таблица 2.2 — Формат ML: авторегрессия со сдвигом на одну позицию.....	71
Таблица 2.3 — Формат ML: авторегрессия со сдвигом на две позиции .	71
Таблица 2.4 — Формат ML: авторегрессия со сдвигом на k позиций.....	72
Таблица 2.5 — Формат LM: нуль-мерные данные	73
Таблица 2.6 — Формат ML: регрессия с добавлением одного исторического момента	74
Таблица 2.7 — Формат ML: регрессия с добавлением двух исторических моментов	75
Таблица 2.8 — Формат ML: регрессия с добавлением k исторических моментов	75
Таблица 2.9 — Разбиение данных в формате ML на этапы истории и обучения.....	80
Таблица 2.10 — Авторегрессия на локальных данных.....	83
Таблица 2.11 — Прогноз по времени для первого момента периода обучения.....	83

Таблица 2.12 — Вторая итерация разбиения данных в формате ML на периоды исторических наблюдений и обучения	84
Таблица 2.13 — Прогноз по времени для первых двух моментов периода обучения	84
Таблица 2.14 — Полный прогноз по времени для периода обучения.....	84
Таблица 2.15 — Авторегрессия на глобальных данных	86
Таблица 2.16 — Авторегрессия на ближайших локациях	87
Таблица 2.17 — Разбиение данных в формате LM на периоды истории и обучения	89
Таблица 2.18 — Регрессия на локальных данных	90
Таблица 2.19 — Полный прогноз по локациям для периода обучения...	90
Таблица 2.20 — Регрессия на глобальных данных	91
Таблица 2.21 — Регрессия на ближайших моментах	93
Таблица 2.22 — Истинные значения наблюдаемой характеристики для периода обучения	94
Таблица 2.23 — Прогноз по времени наблюдаемой характеристики для периода обучения	95
Таблица 2.24 — Абсолютные отклонения реальных значений от прогноза по времени	95
Таблица 2.25 — Индикатор аномалий по времени для периода обучения	97
Таблица 2.26 — Транспонированные истинные значения для периода обучения	98
Таблица 2.27 — Прогноз по локациям для периода обучения.....	99
Таблица 2.28 — Отклонения реальных значений от прогноза по локациям.....	99
Таблица 2.29 — Индикатор аномалий по локациям для периода обучения	100
Таблица 2.30 — Совокупный индикатор аномалий на этапе обучения	101
Таблица 2.31 — Сопоставление алгоритмической и экспертной	

оценки.....	103
Таблица 2.32 — Сопоставление алгоритмической оценки и оценки экспертной группы.....	105
Таблица 2.33 — Данные в формате ML в первый момент эксплуатации.....	109
Таблица 2.34 — Прогноз по времени на второй момент эксплуатации, полученный в первый момент эксплуатации	109
Таблица 2.35 — Данные в формате ML во второй момент эксплуатации.....	110
Таблица 2.36 — Абсолютные отклонения прогнозов по времени во второй момент эксплуатации	110
Таблица 2.37 — Индикаторы аномалий по времени во второй момент эксплуатации.....	110
Таблица 2.38 — Прогноз по локациям на второй момент эксплуатации, полученный во второй момент эксплуатации.....	111
Таблица 2.39 — Абсолютные отклонения по локациям для наблюдаемой характеристики во второй момент эксплуатации	111
Таблица 2.40 — Индикаторы аномалий по локациям во второй момент эксплуатации.....	111
Таблица 2.41 — Суммарные индикаторы аномалий во второй момент эксплуатации.....	112
Таблица 2.42 — Данные в формате ML во второй момент эксплуатации.....	112
Таблица 2.43 — Прогноз по времени на третий момент эксплуатации, полученный во второй момент эксплуатации.....	113
Таблица 2.44 — Структура отдельной записи в отчёте детектора аномалий	114
Таблица 3.1 — Данные для формирования кривой тренда по базовому реперу	124
Таблица 3.2 — Сформированные данные кривой тренда по базовому	

реперу	124
Таблица 3.3 — Синтетические тренды в формате ML.....	126
Таблица 3.4 — Синтетические данные в формате ML с дисперсией нормального шума $\sigma = 0.5$	130
Таблица 3.5 — Синтетические данные в формате LM с дисперсией нормального шума $\sigma = 0.5$	130
Таблица 3.5 — Истинные значения периода обучения в формате ML .	144
Таблица 3.6 — Предсказанные значения для периода обучения в формате ML	144
Таблица 3.7 — Отклонения предсказанных значений от истинных в формате ML	144
Таблица 3.8 — Синтетическое экспертное сообщество	150
Таблица 3.9 — Совокупное мнение о наличии или отсутствии аномалий синтетического экспертного сообщества, ориентированного на прогноз по времени.....	151
Таблица 3.10 — Истинные значения на этапе обучения в формате LM	159
Таблица 3.11 — Предсказанные значения на этапе обучения в формате LM.....	159
Таблица 3.12 — Отклонения предсказанных значений от истинных в формате LM	160
Таблица 3.13 — Список аномалий по времени на синтетических данных на этапе пилотной эксплуатации	163
Таблица 3.14 — Список аномалий по локациям на синтетических данных на этапе пилотной эксплуатации	165
Таблица 3.15 — Общий список аномалий на синтетических данных на этапе пилотной эксплуатации	165
Таблица 4.1 — Данные о числе выполненных операций в формате ML	179
Таблица 4.2 — Данные о числе выполненных операций в формате LM	179

Таблица 4.3 — Коды моментов в процессе производства	179
Таблица 4.4 — Коды локаций в процессе производства	180
Таблица 5.1 — Исходные данные о работе сервисных центров	206
Таблица 5.2 — Данные о числе заявок в формате ML	214
Таблица 5.3 — Данные о суммарных затратах в формате ML	214
Таблица 5.4 — Данные о числе заявок в формате LM	214
Таблица 5.5 — Данные о суммарных затратах в формате LM	215
Таблица 5.6 — Коды моментов в процессе сервисного обслуживания	215
Таблица 5.7 — Коды локаций в процессе сервисного обслуживания...	216
Таблица 6.1 — Данные о переходах на канал в формате ML	257
Таблица 6.2 — Данные о переходах на канал в формате LM	257
Таблица 6.3 — Коды моментов в маркетинговом процессе	258
Таблица 6.4 — Коды локаций в маркетинговом процессе	258
Таблица 7.1 — Синтетические данные в формате ML до наложения аномалий	291

ПРИЛОЖЕНИЕ А. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СТОХАСТИЧЕСКОГО ПРОГНОЗИРОВАНИЯ

Идея детекции аномалий по отношению к модели нестабильного процесса, предложенная нами в первой главе (см. раздел 1.3, в частности, рисунок 1.5), является вполне очевидной, однако её реализация не так проста в силу известного принципа, хорошо знакомого всем, кто занимается инженерной разработкой: дьявол кроется в деталях. Все наши рассуждения относительно алгоритмической детекции строятся некой на предпосылке и следствии вида: «Пусть у нас есть модель процесса, обладающая высокой прогнозирующей способностью. Тогда ...», — но кто сказал, что она у нас есть? Написать такую модель — это совсем не тривиальная задача [2–5, 7–9, 12, 17, 23, 48, 65, 172], а провести её корректную валидацию и обосновать предиктивную эффективность — это ещё одна задача, и она, в некотором смысле, сложнее предыдущей [3, 5, 7, 8, 23, 49, 218].

Когда мы говорили о хорошо известных традиционных инструментах управления качеством [6, 19, 56, 154, 212], мы упомянули четыре наивных инструмента элементарной статистики (стратификация, гистограмма, анализ Парето и диаграмма рассеивания), один наивный инструмент сбора данных (контрольный листок) и один наивный инструмент мониторинга (карта Шухарта). Теперь, когда перед нами стоит задача методологического обоснования концепции динамического выявления несоответствий, нам понадобятся другие инструменты, причём, их будет гораздо больше и они будут гораздо сложнее. Используя метафору для сравнения масштабов сложности, можно сказать, что сложность необходимого нам инструментария соотносится со сложностью рассмотренных выше инструментов примерно как реляционная база данных

соотносится с контрольным листком. И, хотя эта метафора (как и все другие метафоры) не отличается высокой точностью, она, тем не менее, даёт некоторое представление о том, с чем нам предстоит работать.

Нам понадобятся прогнозирующие модели, то есть, алгоритмы, способные по некоторому имеющемуся на входе массиву данных предъявлять на выходе значения нужного нам показателя для новых данных, то есть, нового, не имевшегося у нас ранее массива той же структуры [2–5, 7–9, 12, 15, 48, 51, 172]. При изучении различных процессов, формирующих продукт (производственных, сервисных или маркетинговых), речь идёт о том, что на входе мы располагаем данными обо всех предыдущих состояниях процесса в наборе локаций, а на выходе мы хотели бы получить значения изучаемой нами характеристики в каждой из локаций в следующий момент времени, тогда мы могли бы утверждать, что нам удалось смоделировать процесс, и на основании вычисляемых модельных данных делать выводы о нормальности или, наоборот, аномальности протекания подконтрольного нам процесса.

Решение таких задач обладает своей спецификой [3, 5, 9, 23, 48, 65], и для того, чтобы оперировать далее нужными нам понятиями и приёмами (и просто чтобы использовать общепринятую в данной области терминологию), в этом приложении мы совершим небольшой экскурс в область анализа больших данных и машинного обучения. Этот раздел ни в коем случае не претендует на хотя бы какую-нибудь полноту изложения и широкий охват различными тем, связанных с машинным обучением: мы не затронем вопросов, связанных с предварительной обработкой данных и их нормализацией [13, 17, 172, 204], не станем рассматривать методы регуляризации моделей с целью повышения их прогнозирующей эффективности [2, 8, 48, 173], не коснёмся алгоритмов мультиклассовой классификации [3, 4, 51, 73, 172, 173] и т. д. Мы рассмотрим лишь те аспекты машинного обучения, которые используются нами в реализации алгоритмов методики МССР (Modeling, Challenge, Calibration, Production).

А.1. Обучение по прецедентам

Пусть имеется множество объектов $X \in \mathbb{R}^n$, множество допустимых ответов Y (по-другому называемых откликами), и пусть существует отображение $\mathcal{F} : X \rightarrow Y$, значения которого $y_i = \mathcal{F}(x_i)$ известны лишь на некотором конечном подмножестве объектов $\{x_1, x_2, \dots, x_\ell\} \subset X$. Пары объектов и соответствующих им откликов (x_i, y_i) называются прецедентами, а вся совокупность таких пар $X^\ell = (x_i, y_i)_{i=0}^\ell$ называется обучающей выборкой (англ. — training sample). Задача обучения по прецедентам состоит в том, чтобы по обучающей выборке восстановить решающий алгоритм $a : X \rightarrow Y$, который максимально точно приближал бы отображение $\mathcal{F}(x)$, причём, не только на обучающей выборке X^ℓ , но и на всем множестве X .

Признаком (англ. — feature) f любого объекта называется результат измерения какой-либо его характеристики. В зависимости от того, что из себя представляет множество D_f всех допустимых значений признака f , признаки делятся на несколько типов:

если $D_f = \{0, 1\}$, то f — это бинарный признак;

если — это конечное множество, то f — это номинальный признак;

если D_f — это конечное упорядоченное множество, то f — это порядковый признак;

если $D_f = \mathbb{R}$, то f — это числовой (или количественный) признак.

Для набора признаков $f_1, f_2, \dots, f_n$ вектор, составленный из значений $(f_1(x), f_2(x), \dots, f_n(x))$ называется признаковым описанием объекта $x \in X$, причём, как правило, когда говорят об объектах из X , имеют в виду именно их признаковые описания. Признаковые описания всех объектов естественным образом образуют прямоугольную матрицу $X = (f_i(x_j))_{n \times \ell}$, которую называют матрицей данных или дата-фреймом (англ. — data frame).

В зависимости от природы множества допустимых ответов, задачи обучения по прецедентам делятся на два типа: классификационные и регрессионные. Если отклики образуют конечное множество, то задача обучения является задачей классификации (англ. — classification). В таком случае множество объектов разбивается на непересекающиеся классы

$$K_y = \{x \in X : \mathcal{F}(x) = y\}$$

а алгоритм $a(x)$ должен давать ответ на вопрос, какому классу принадлежит объект x (в частности, если $Y = \{0, 1\}$, то классификация оказывается бинарной). Если $Y = \mathbb{R}$, то задача обучения — это задача восстановления регрессии (англ. — regression estimation), и тогда для каждого объекта x алгоритм $a(x)$ должен возвращать значение, наиболее близкое к истинному значению y .

Моделью в теории обучения по прецедентам называется некоторое параметрическое семейство алгоритмов

$$\mathcal{A} = \{a(x, \lambda) \mid a : X \rightarrow Y\}$$

зависящее от набора параметров λ , методом обучения — отображение

$$\mu : X^\ell \times Y \rightarrow \mathcal{A}$$

которое произвольной конечной выборке $X^\ell = (x_i, y_i)_{i=0}^\ell$ ставит в соответствие некоторый алгоритм $a \in \mathcal{A}$, а процесс подбора оптимальных параметров λ называют обучением модели (англ. — fitting). В задачах обучения по прецедентам чётко различаются два этапа: на этапе обучения метод μ по выборке X^ℓ строит алгоритм $a = \mu(X^\ell)$, а на этапе применения алгоритм a применяется к новым объектам x и возвращает прогнозируемые значения откликов $y = a(x)$. Этап обучения наиболее сложен, и, как правило, состоит в поиске параметров модели, доставляющих оптимальное значение некоторому

заданному заранее функционалу качества.

Для определения оптимальности функционала качества вводится весьма важное понятие функции потерь (англ. — loss function), то есть, некоторой неотрицательной функции $\mathcal{L}(a, x)$, характеризующей величину ошибки алгоритма a на объекте x . Обобщение функции потерь на всю обучающую выборку, по определению — это функционал качества:

$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{x \in X^\ell} \mathcal{L}(a, x)$$

называемый так же функционалом средних потерь или эмпирическим риском, так как он вычисляется по эмпирическим данным $X^\ell = (x_i, y_i)_{i=0}^\ell$. При решении задач классификации применяется функция потерь вида

$$\mathcal{L}(a, x) = \begin{cases} 0, & a(x) = y \\ 1, & a(x) \neq y \end{cases}$$

которая называется индикатором ошибки, а функционал качества Q с такой функцией потерь представляет собой не что иное, как относительную частоту ошибок классификатора. При решении регрессионных задач используется квадратичная функция потерь:

$$\mathcal{L}(a, x) = (a(x) - y)^2$$

а сопряжённый функционал качества Q называется средней квадратичной ошибкой алгоритма a на выборке X^ℓ . Классический метод обучения, называемый минимизацией эмпирического риска (empirical risk minimization, ERM), заключается в том, чтобы найти в заданной модели $\mathcal{A}$ алгоритм a , доставляющий минимальное значение функционалу качества Q на заданной обучающей выборке X^ℓ :

$$\mu(X^\ell) = \arg \min_{a \in \mathcal{A}} Q(a, X^\ell)$$

Решением этой оптимизационной задачи служит алгоритм с оптимально подобранными параметрами λ . В таком случае говорят, что модель обучена, впрочем, хорошо она обучена или плохо — это другой вопрос, мы подробно рассмотрим его чуть позже (см. разделы А.7 и А.8).

А.2. Комментарии к основным понятиям обучения по прецедентам

Итак, как мы отметили выше, основным объектом обучения по прецедентам служат массивы данных, так называемые дата-фреймы, представляющие собой числовые матрицы⁹⁴, причём, часть из них описывается при помощи признаков и образует набор предикторов обучаемой модели, а другая часть (содержащая, как правило, один признак) является целевой (англ. — target), и именно её должна предсказывать обученная модель, для чего и развивается вся теория обучения по прецедентам.

Дата-фрейм, по сути, представляет собой таблицу, строки которой нумеруются, начиная с нуля (так как такова договорённость, принятая в большинстве языков программирования, и это справедливо не только для дата-фреймов, но и вообще — для любых итерируемых объектов: списков, массивов, кортежей и т. д.), а столбцы именуются именами признаков. В такой конфигурации признак — это столбец дата-фрейма, а объект — это его строка (см. таблицу А.1). Важно понимать, что прецедент — это тоже строка, но дополненная значением отклика, или, выражаясь, точнее, объект описывается как вектор:

$$(x_{i1}, x_{i2}, \dots, x_{in}) \in \mathbb{R}^n$$

а прецедент — как вектор, размерность которого больше на единицу:

⁹⁴ Строго говоря, элементами дата-фреймов не всегда служат числа, часто оказывается, что это строковые или логические выражения. Но, во-первых, нечисловые данные ещё до моделирования, на этапе предварительной подготовки переводятся в числа, для этого есть специальные методы, их много, и мы не будем здесь их касаться, а во-вторых, в рамках настоящей диссертации все данные всегда именно числовыми

$$(x_{i1}, x_{i2}, \dots, x_{in}, y_i) \in \mathbb{R}^{n+1}$$

Таблица A.1 — Дата-фрейм: предикторы и целевая функция

	x_1	x_2	x_3	x_4	x_5	...	x_{n-1}	x_n	y
0	x_{01}	x_{02}	x_{03}	x_{04}	x_{05}	...	$x_{0,n-1}$	$x_{0,n}$	y_0
1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	...	$x_{1,n-1}$	$x_{1,n}$	y_1
2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	...	$x_{2,n-1}$	$x_{2,n}$	y_2
3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	...	$x_{3,n-1}$	$x_{3,n}$	y_2
...	...	...	...	...	...	...	...	...	...
k-1	$x_{k-1,1}$	$x_{k-1,2}$	$x_{k-1,3}$	$x_{k-1,4}$	$x_{k-1,5}$	...	$x_{k-1,n-1}$	$x_{k-1,n}$	y_{k-1}
k	$x_{k,1}$	$x_{k,2}$	$x_{k,3}$	$x_{k,4}$	$x_{k,5}$	...	$x_{k,n-1}$	$x_{k,n}$	y_k

Объекты дата-фрейма по-другому называются наблюдениями, а отклики — метками, и, вообще, в стохастическом прогнозировании имеется настолько широкий спектр синонимичных выражений, что для того, чтобы не потеряться в терминах, мы приведём здесь список синонимов, которые будем использовать, при этом уже не оговаривая каждый раз их синонимичность:

1. дата-фрейм — матрица данных, таблица данных, массив данных⁹⁵;
2. столбец дата-фрейма — признак, или, на сленге используют слово фича (от англ. feature);
3. то, что требуется спрогнозировать — целевой столбец, целевой признак, целевая функция, отклик, метка, на сленге — таргет (от англ. target);
4. все столбцы дата-фрейма кроме целевого — переменные, предикторы.
5. строка дата-фрейма — объект, наблюдение.

Если каждому наблюдению сопоставлена метка, данные называются размеченными, а задачу обучения по прецедентам схематично можно пред-

⁹⁵ Не следует путать дата-фрейм с дата-сетом. Дата-фрейм — это всегда таблица, подобная таблице A.1 по своей структуре. А дата-сетом можно назвать любой набор данных: список, написанный вручную на листе бумаги, фотографии в смартфоне и т. д.

ставлять себе в виде дата-фрейма, в котором часть данных является размеченной, а у другой части разметка отсутствует, и для тех объектов, значения откликов для которых неизвестны, требуется их найти, или, если быть точным — найти их максимально точные приближения (см. таблицу А.2).

Таблица А.2 — Дата-фрейм: задача прогнозирования

	x_1	x_2	x_3	x_4	x_5	...	x_{n-1}	x_n	y
0	x_{01}	x_{02}	x_{03}	x_{04}	x_{05}	...	$x_{0,n-1}$	$x_{0,n}$	y_0
1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	...	$x_{1,n-1}$	$x_{1,n}$	y_1
2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	...	$x_{2,n-1}$	$x_{2,n}$	y_2
3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	...	$x_{3,n-1}$	$x_{3,n}$	y_2
...	...	...	...	...	...	...	...	...	...
k-1	$x_{k-1,1}$	$x_{k-1,2}$	$x_{k-1,3}$	$x_{k-1,4}$	$x_{k-1,5}$	...	$x_{k-1,n-1}$	$x_{k-1,n}$	y_{k-1}
k	$x_{k,1}$	$x_{k,2}$	$x_{k,3}$	$x_{k,4}$	$x_{k,5}$	...	$x_{k,n-1}$	$x_{k,n}$	y_k
k+1	$x_{k+1,1}$	$x_{k+1,2}$	$x_{k+1,3}$	$x_{k+1,4}$	$x_{k+1,5}$	...	$x_{k+1,n-1}$	$x_{k+1,n}$	
k+2	$x_{k+2,1}$	$x_{k+2,2}$	$x_{k+2,3}$	$x_{k+2,4}$	$x_{k+2,5}$	...	$x_{k+2,n-1}$	$x_{k+2,n}$	
...	...	...	...	...	...	...	...	...	
k+s-1	$x_{k+s-1,1}$	$x_{k+s-1,2}$	$x_{k+s-1,3}$	$x_{k+s-1,4}$	$x_{k+s-1,5}$	...	$x_{k+s-1,n-1}$	$x_{k+s-1,n}$	
k+s	$x_{k+s,1}$	$x_{k+s,2}$	$x_{k+s,3}$	$x_{k+s,4}$	$x_{k+s,5}$	...	$x_{k+s,n-1}$	$x_{k+s,n}$	

Размеченный фрагмент данных объявляется обучающей выборкой, на которой учится модель (то есть, по известным значениям меток подбираются параметры алгоритма, минимизирующие функционал качества), а после того как модель обучена, её применяют к неразмеченному фрагменту, получая на выходе недостающие значения откликов. Именно эту процедуру — заполнение незаполненных ячеек с откликами — и называют стохастическим прогнозированием. Если прогнозируется временной ряд, то под прогнозом понимают буквальное предсказание будущего значения отклика (или нескольких будущих значений), но, вообще говоря, прогноз не обязательно направлен в будущее. В случае, когда признаковые описания объектов не имеют естественного отношения порядка (в частности, если они не нумеруются моментами времени), то в задаче прогнозирования размеченные и неразмеченные данные не

обязательно следуют друг за другом (как это представлено в разметке таблицы А.2), они вполне могут поменяться местами, а, скорее всего, — вообще располагаются в совершенно произвольном порядке.

В качестве заключительного комментария отметим еще, что число признаков, как правило измеряется десятками, тогда как число объектов — сотнями тысяч или миллионами⁹⁶, поэтому матрицы данных отличаются специфической конфигурацией: они очень сильно вытянуты по вертикали.

А.3. Линейные и полиномиальные регрессии

Напомним, что задача прогнозирования называется регрессионной, если её метки относятся к числовому типу, то есть, если $Y \subset \mathbb{R}$, а, в силу того, что её все её объекты тоже являются числовыми векторами $X \subset \mathbb{R}^n$, обучение регрессионной модели можно представлять себе как поиск отображения $\mathbb{R}^n \rightarrow \mathbb{R}$, которое лучше всего приближает истинное соответствие между объектами и таргетами. Возможных отображений бесконечно много, но мы можем упростить задачу и договориться, что будем искать наилучшее только среди линейных функций вида⁹⁷:

$$y = w_1x_1 + w_2x_2 + \dots + w_nx_n + w_0$$

где y — это целевая функция (таргет), $(x_1, x_2, \dots, x_n)$ — признаковое представление объекта (вектор признаков), а $w_1, w_2, \dots, w_n$ — параметры модели. Вектор $w = (w_1, w_2, \dots, w_n)$ часто называют вектором весов, а число w_0 — свободным коэффициентом, или сдвигом (англ. — bias). Более компактно линейную модель можно записать в виде

⁹⁶ Это, разумеется, не всегда так. В аналитике данных приходится иметь дело с ситуациями, которые описываются тысячами признаков, и даже — с ситуациями, когда число объектов меньше числа признаков. Но такие ситуации, скорее, являются исключениями.

⁹⁷ Именно поэтому регрессии такого вида называют линейными.

$$y = \langle x, w \rangle + w_0$$

где угловые скобки означают скалярное произведение. Таким образом, в случае регрессионной задачи ищется не какое-то абстрактное отображение, а конкретный вектор

$$(w_0, w_1, w_2, \dots, w_n) \in \mathbb{R}^{n+1}$$

причём, признаковое описание объекта обычно дополняют одним фиктивным признаком, тождественно равным единице

$$x = (1, x_1, x_2, \dots, x_n) \in \mathbb{R}^{n+1}$$

после чего линейная модель уже можно описать при помощи скалярного произведения без сдвига:

$$y = \langle x, w \rangle$$

В качестве функции потерь в линейной регрессии используется квадрат L_2 -нормы вектора разности предсказаний модели и истинных значений целевой функции, так как такая функция потерь обладает несколькими замечательными свойствами:

1. L_2 -норма вектора — это евклидово расстояние между вектором предсказаний алгоритма и вектором истинных значений таргета, то есть, приближение задаётся при помощи самого простого, понятного и, следовательно, легко интерпретируемого расстояния;
2. с точки зрения статистики это означает, что наши данные состоят из линейного сигнала и нормально распределенного шума.

Используя определение L_2 -нормы, получаем выражение для функции потерь, связывающей алгоритм a , с прецедентом (x, y) , а с учётом того, что сам алгоритм для линейной регрессионной модели, по сути, представляет собой вектор коэффициентов w , выражение для функции потерь приобретает вид вполне

конкретной функции от переменных $w_0, w_1, \dots, w_n$:

$$\mathcal{L}(w, x) = (\langle x, w \rangle - y)^2$$

После этого функционал качества получается как средний риск по всем объектам обучающей выборки

$$Q(w, X^\ell) = \frac{1}{\ell} \sum_{x \in X^\ell} \mathcal{L}(w, x) = \frac{1}{\ell} \sum_{x \in X^\ell} (\langle x, w \rangle - y)^2$$

и обучение модели сводится к задаче оптимизации

$$Q(w, X^\ell) \rightarrow \min_w$$

где минимум берётся в смысле обычного минимума функции нескольких переменных $w_0, w_1, \dots, w_n$.

Для решения задач оптимизации имеются разные подходы: как точные аналитические, так и приближенные, причём на практике, как правило, реализуются приближенные методы, но, в принципе, если объем данных не очень высок, ничто не мешает применить и традиционные приёмы оптимизации. Классический анализ состоит в поиске локального минимума функционала качества, а с учётом того, что функционал качества, построенный на основе L_2 -нормы, является выпуклой функцией переменных $w_0, w_1, \dots, w_n$, его локальный минимум одновременно является и глобальным. Это гарантирует единственность точного аналитического решения, и было бы решающим аргументом в пользу аналитических методов, если бы не их вычислительная сложность, которая часто не позволяет реализовывать такие методы за разумное время. Что же касается градиентных методов, которые лежат в основе приближенных решений задачи выпуклой оптимизации, то они образуют отдельный, очень сложный раздел математики, и в рамках настоящей диссертационной работы мы не будем их рассматривать даже в беглом формате, чтобы не отклоняться от основного сюжета. Отметим лишь, что и в том, и в другом случае

задача в конечном счёте состоит в минимизации суммарного отклонения (см. рисунок А.1, где из соображений наглядности выделены не все отклонения, а только те, номера которых кратны пятидесяти).

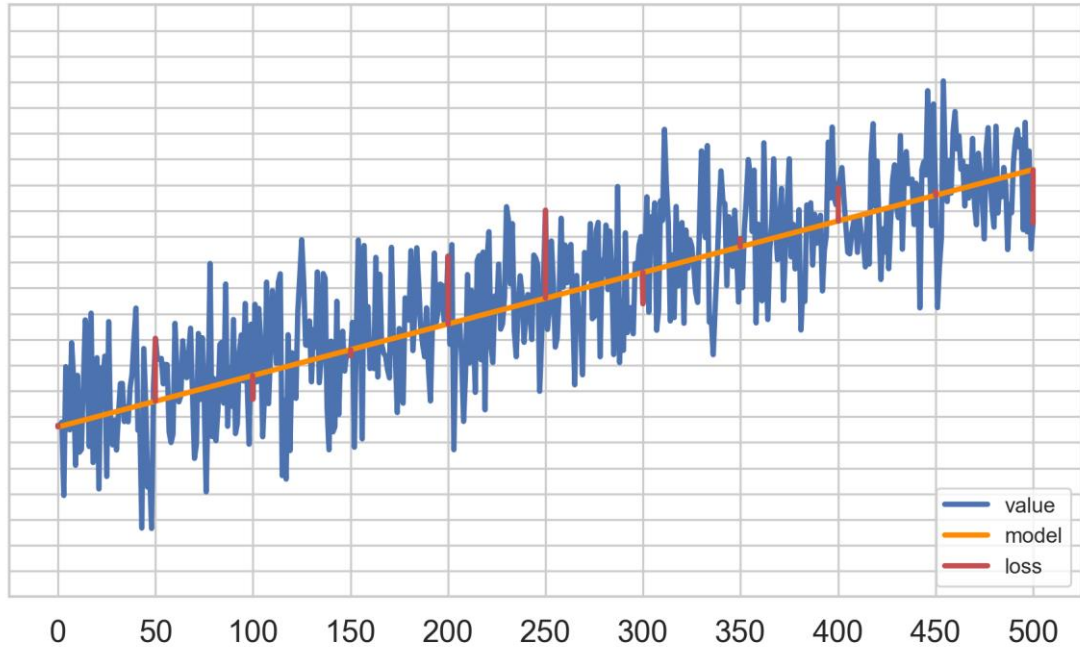


Рисунок А.1 — Регрессия временного ряда с линейным трендом

На иллюстрации выше представлена одномерная ситуация, когда в качестве единственного признака выступает номер наблюдения, а данные обладают ярко выраженным линейным трендом. Однако, как это ни парадоксально, линейная регрессионная модель способна аппроксимировать и нелинейные тренды, оставаясь при этом линейной. Это хорошо известный приём, известный под названием полиномиальной регрессии, и состоящий в генерации новых признаков, являющихся алгебраическими выражениями от старых. Например, если есть только один исходный признак x_1 , то мы можем положить

$$x_0 = x_1^0 = 1, \quad x_1 = x_1^1 = x_1, \quad x_2 = x_1^2, \quad \dots, \quad x_k = x_1^k$$

после чего искать модель в виде полинома

$$y = w_0 + w_1 x_1 + w_2 x_1^2 + \dots + w_k x_1^k = \langle x, w \rangle$$

и в то же время — по-прежнему в виде линейной комбинации вновь сгенерированных признаков⁹⁸.

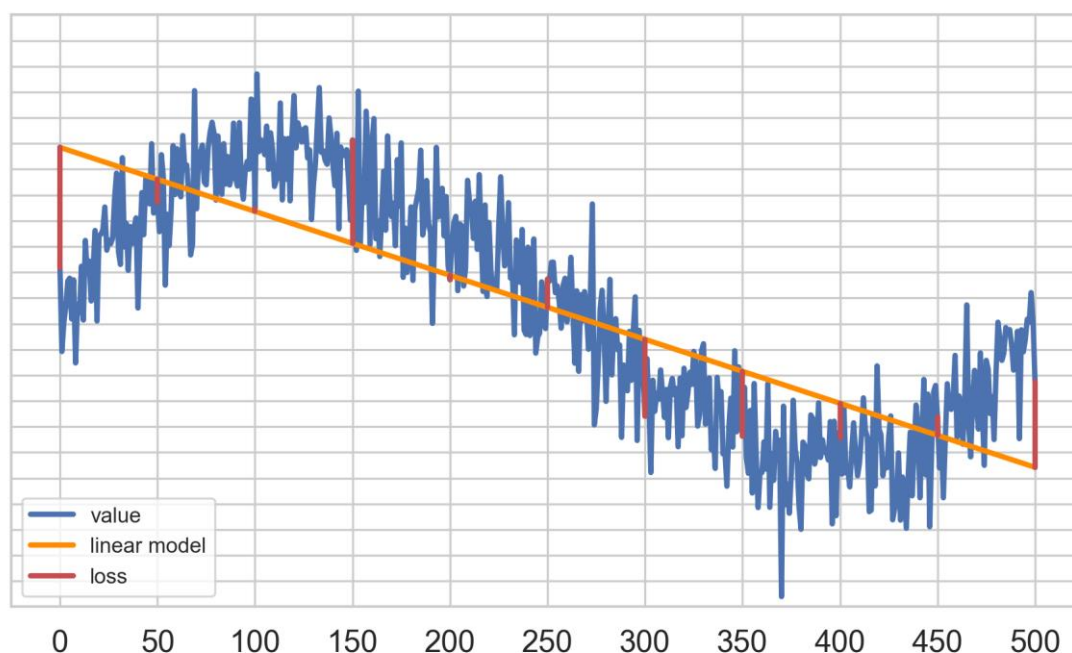


Рисунок А.2 — Линейная регрессия временного ряда с нелинейным трендом

Если применить алгоритм линейной регрессии к временному ряду с нелинейным трендом, то результат, разумеется, оставит желать лучшего (см. рисунок А.2, где мы, так же как и выше, выделили красным отклонения каждого пятидесятого наблюдения), слишком высоки будут значения функции потерь.

Вводя дополнительные признаки, равные второй и третьей степеням времени, мы можем существенно улучшить ситуацию и добиться того, что кривая тренда, став фрагментом кубической параболы, начнёт более точно воспроизводить поведение временного ряда, и, соответственно, суммарные отклонения, определяющие функционал качества, уменьшатся (см. рисунок А.3, где мы, так же как и выше, выделили красным отклонения каждого пятидесятого наблюдения, а также — ср. их с предыдущими на рисунке А.2).

⁹⁸ Вообще, к признакам можно применять не только алгебраические, но и любые другие функции: логарифмические, тригонометрические и т. д. — регрессия все равно будет оставаться линейной, потому что это будет по-прежнему линейная комбинация признаков.

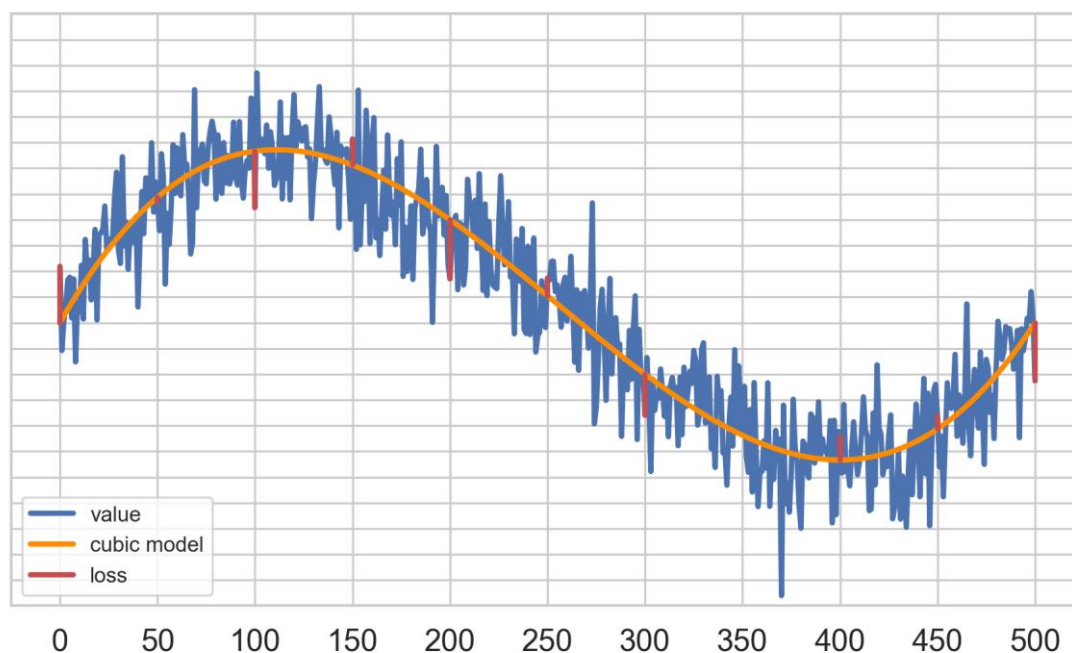


Рисунок А.3 — Полиномиальная кубическая регрессия временного ряда с нелинейным трендом

Но дальнейшее повышение степени аппроксимирующего полинома уже не улучшает результат (см. рисунок А.4).

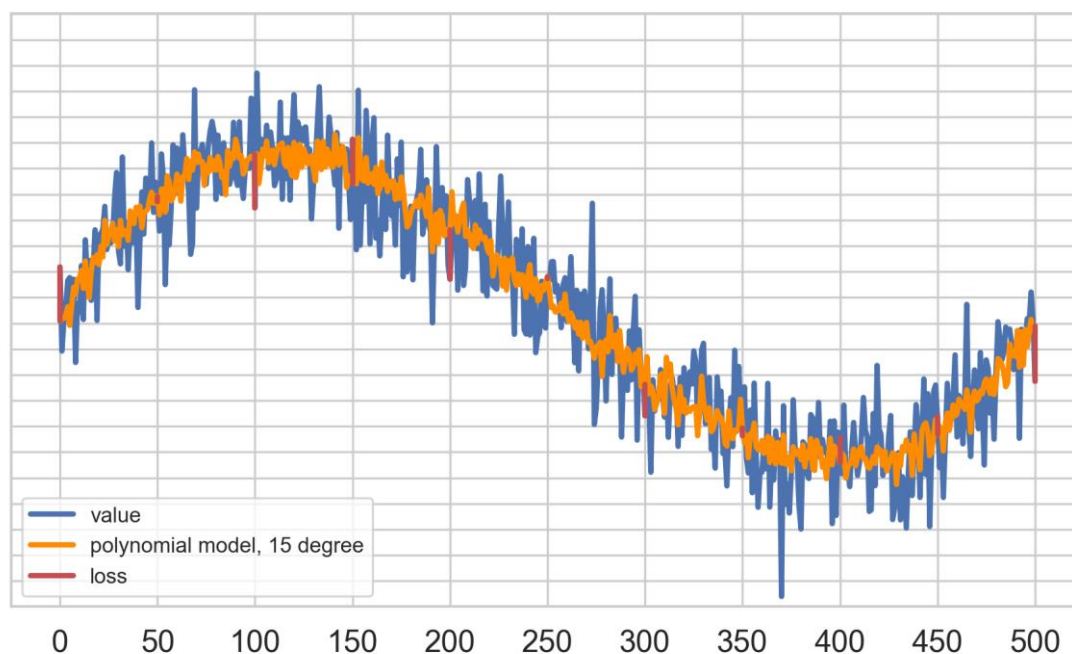


Рисунок А.4 — Полиномиальная регрессия 15 степени временного ряда с нелинейным трендом

В связи с этим, нельзя не отметить ограничения полиномиальной регрессии, которые проявляются, во-первых, в комбинаторной сложности генерации

полиномиальных признаков в случае регрессионных задач высокой размерности (при наличии единственного признака, как на приведённых выше иллюстрациях, где в качестве признака выступает номер наблюдения, повышение степени полиномиальной регрессионной модели не представляет труда, но если признаков хотя бы сто пятьдесят, генерация всех их всевозможных алгебраических комбинаций уже не осуществляется за разумное время), а во-вторых, в склонности полиномиальных моделей к переобучению при возрастании степени полинома. Поэтому при обучении линейных моделей нельзя неограниченно повышать их сложность, так как при этом снижается их обобщающая способность (аналогичные замечания см. ниже, разделы А.4 и А.5).

А.4. Метрические методы kNN

Идея метрических методов kNN (англ. — k nearest neighbors) состоит в том, что близкие или похожие объекты должны обладать близкими или похожими откликами. Алгоритмы этого класса почти не имеют фазы обучения. Вместо этого они просто запоминают всю обучающую выборку, а на этапе предсказания просто ищут объекты, похожие на целевой. Такой процесс называют *lazy learning*, потому что никакого обучения, по сути, не происходит. Метрические модели являются непараметрическими, они не делают явных допущений о глобальных законах, которым подчиняются данные⁹⁹. Метрические методы локальны: они исходят из допущения, что свойства объекта можно узнать, имея представление о его соседях. Эти соображения могут быть полезными, особенно в случае сложно устроенных данных, для которых трудно придумать глобальную модель. С другой стороны, из-за *lazy learning* алгоритм становится абсолютно неприменимым при большом количестве данных. Алгоритмы ближайших соседей очень просты, и тем не менее, они довольно точны и хорошо

⁹⁹ Например, линейная регрессия основывается на предположении о том, что изучаемая закономерность линейная (с неизвестными коэффициентами, которые восстанавливаются по выборке), линейная бинарная классификация — что существует гиперплоскость, разделяющая классы и т. д.

интерпретируемы — поэтому их часто используют в качестве бейзлайнов.

Интуитивное представление о метрических методах kNN дает иллюстрация ниже (см. рисунок А.5). Допустим, мы проводим классификацию объектов на два класса — синий или жёлтый, и нам дана некоторая обучающая выборка с целевым объектом, который отмечен серым цветом.

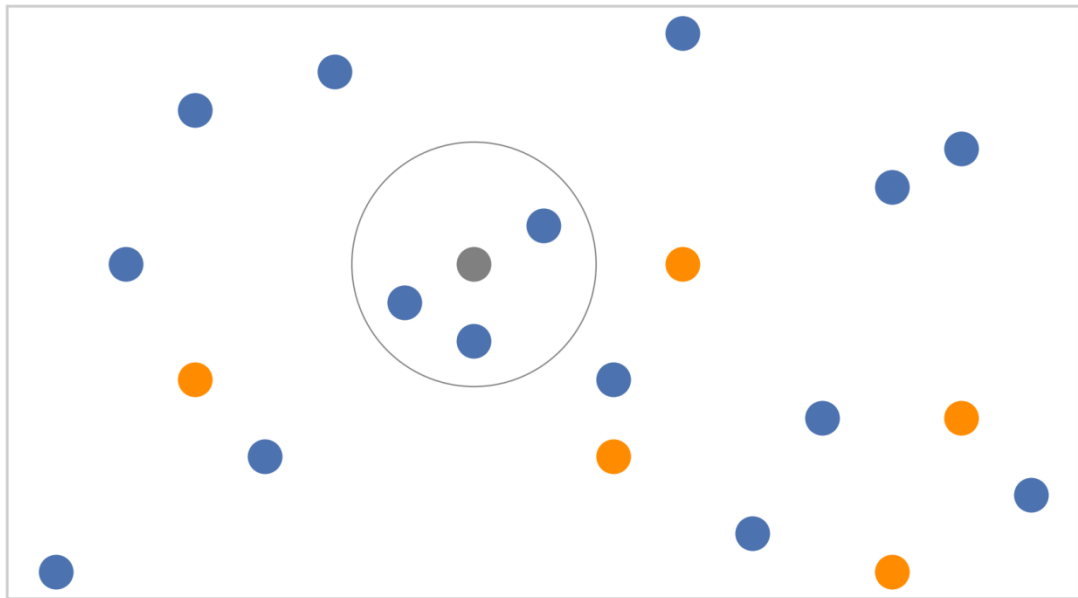


Рисунок А.5 — Интуитивная иллюстрация метрических методов

Мы хотим определить, к какому классу относится серый объект, причём, интуитивно очевидно, что он должен быть синим, потому что все его соседи синие. Эта интуиция и отражает суть метода KNN — прогнозировать отклик целевого объекта, исходя из известных нам свойств объектов обучающей выборки, которые максимально похожи на него.

Переходя к формальному описанию алгоритма, рассмотрим сначала задачу многоклассовой классификации. Пусть имеется набор прецедентов $(x_i, y_i)_{i=1}^{\ell}$, где объекты $x_i \in X^{\ell} = \mathbb{R}^n$, а отклики принадлежат конечному множеству классов $y_i \in Y^{\ell} = \{\text{class}_1, \dots, \text{class}_m\}$. Пусть также задана некоторая симметричная по своим аргументам функция расстояния $\rho : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, +\infty)$, и пусть нам требуется классифицировать новый

объект u . Для этого мы находим k объектов обучающей выборки X^ℓ , наиболее близких к u в смысле расстояния ρ : $X_k^\ell(u) = \{x_{(1)}, \dots, x_{(k)}\}$ по следующему правилу:

$$\rho(u, x_{\text{in}}) \leq \rho(u, x_{\text{out}}) \quad \forall x_{\text{in}} \in X_k^\ell(u), \quad \forall x_{\text{out}} \in X^\ell \setminus X_k^\ell(u)$$

Тогда класс нового объекта, который является результатом работы алгоритма a , естественным образом определяется как наиболее часто встречающийся класс среди объектов из $X_k^\ell(u)$:

$$a(u) = \arg \max_{y \in Y^\ell} \sum_{i=1}^k [y_{(i)} = y]$$

то есть, для каждой метки класса количество соседей с такой меткой можно посчитать, просто просуммировав по всем соседям индикаторы событий, соответствующих тому, что метка соседа равна y . Легко заметить, что этот алгоритм позволяет также оценивать вероятности каждого из классов. Для этого достаточно просто посчитать частоты классов соседей:

$$\mathbb{P}(u \sim y) = \frac{1}{k} \sum_{i=1}^k [y_{(i)} = y]$$

Несмотря на то что формально в алгоритме kNN отсутствует фаза обучения, этот алгоритм, тем не менее, может легко переобучиться. Это происходит из-за того, что параметрами алгоритма можно считать всю обучающую выборку, как правило, очень большую по размеру. Тем самым, такой алгоритм становится сложным алгоритмом с большим количеством параметров, и поэтому ему легко подстроиться под конкретные данные.

Важную роль в метрических методах машинного обучения играет выбор подходящей метрики. В большинстве случаев можно применить обычную евклидову метрику:

$$\rho(u, v) = \sqrt{\sum (u_i - v_i)^2}$$

Однако иногда евклидова метрика оказывается плохо приспособленной для описания метрических соотношений между объектами дата-фрейма, и тогда используются другие метрики, спектр которых довольно-таки широк. Это, прежде всего, метрика Минковского:

$$\rho(u, v) = \left(\sum |u_i - v_i|^p \right)^{1/p}$$

которая служит обобщением евклидовой метрики, получающейся из метрики Минковского при $p = 2$. Также используют манхеттенскую метрику:

$$\rho(u, v) = \sum |u_i - v_i|$$

Она тоже является частным случаем метрики Минковского при $p = 1$. Чаще всего манхеттенская метрика используется в пространствах очень высокой размерности из-за лучшей устойчивости к выбросам. Например, если два объекта в 1000-размерном пространстве почти идентичны, но сильно отличаются по одному из признаков, то это почти наверняка свидетельствует о выбросе в этом признаке, и объекты, скорее всего, очень близки, несмотря на этот выброс. Однако обычное евклидово расстояние усилит различие в единственном признаке и сделает их более далёкими друг от друга. Этого недостатка лишена манхэттенская метрика, так как в ней вместо квадрата используется модуль.

В ситуациях, связанных с анализом текстов применяется косинусное расстояние. Эта метрика хороша тем, что не зависит от норм векторов, и такое поведение бывает полезно в некоторых задачах, например при поиске похожих документов. В качестве признаков там часто используются частота используемых слов. При этом интуитивно очевидно, что если в тексте использовать каждое слово в два раза чаще, то тема этого текста поменяться не должна.

В этом случае нам не важна норма векторного представления объекта, и в задачах, связанных с текстами, оценивают не расстояние в буквальном смысле этого слова, а близость направлений векторных представлений объектов:

$$\rho(u, v) = 1 - \cos \theta = 1 - \frac{\langle u, v \rangle}{|u||v|}$$

И, наконец, ещё одна часто используемая метрика — это расстояние Жаккара, применяемое, когда исследуемые объекты представляют собой некоторые множества:

$$\rho(u, v) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Отметим, что, несмотря на некоторые эвристические соображения по выбору метрики¹⁰⁰, её можно считать одним из гиперпараметров модели и подбирать соответствующими способами. Часто качество модели сильно зависит от выбора метрики, а иногда выбрать правильную метрику очень тяжело. Например, если данные имеют сильно отличающийся масштаб, то выбрать подходящую метрику бывает невозможно, и нужно проводить нормализацию.

У оригинальной версии алгоритма kNN есть один большой недостаток, состоящий в том, что он никак не учитывает расстояния до соседних объектов, хотя эта информация может быть полезной. Это происходит потому, что в формуле оптимизации функционала качества все индикаторы истинности учитываются с одинаковыми коэффициентами, но было бы логично назначать индикаторам веса, которые тем больше, чем ближе объект располагается к целевому объекту в смысле какой-либо назначенной выше метрики. Такой алгоритм называется взвешенным kNN алгоритмом (англ. — weighted kNN) и задаётся следующей формулой:

¹⁰⁰ Упомянутые функции мы называем «метриками», но, конечно же, они не обязаны быть метриками в строгом математическом смысле. Они неотрицательны и симметричны, но могут не удовлетворять неравенству треугольника.

$$a(u) = \arg \max_{y \in Y^\ell} \sum_{i=1}^k w_i [y_{(i)} = y]$$

Отметим, что есть большое множество различных вариантов выбора весов для объектов, которые условно можно разделить на две большие группы. В первой группе относятся те методы, в которых веса зависят только лишь от порядкового номера объекта в отсортированном по близости к объекту u массиве объектов обучающей выборки $X_k^\ell(u)$. В этой группе затухание весов чаще всего происходит по линейному закону:

$$w_i = \frac{k + 1 - i}{k}$$

или, что происходит несколько реже, по экспоненциальному:

$$w_i = q^i, \quad 0 < q < 1$$

Однако при применении таких весовых коэффициентов мы также не используем всю информацию, которая нам доступна: используется лишь порядок соседей, порождаемый расстояниями, но не используются сами расстояния.

В результате таких наблюдений возникает вторая группа весовых методов, в которой вес — это некоторая функция K от расстояния ρ , причём, очевидно, что 1) она должна быть положительной на всей своей области определения (иначе модель будет поощрять несовпадение с некоторыми ближайшими соседями), а кроме того, 2) необходимо, чтобы она монотонно не возрастала (тогда вес близких соседей будет больше, чем вес далёких). Определённые таким образом функции $K : \mathbb{R} \rightarrow \mathbb{R}$ называются ядерными (англ. — kernel), с их помощью вычисляются веса соседей в оптимизации функционала качества:

$$a(u) = \arg \max_{y \in Y^\ell} \sum_{i=1}^k K \left(\frac{\rho(u, x_{(i)}(u))}{h} \right) [y_{(i)} = y]$$

где — некоторое положительное число, называемое шириной окна.

От выбора ядра K зависит гладкость модели, но на её качество этот выбор почти не влияет. Ширина окна, напротив, влияет как раз на качество модели: при слишком маленькой ширине модель сильно подстраивается под обучающую выборку и теряет свою обобщающую способность; при слишком большой ширине модель становится слишком простой и тоже теряет обобщающую способность. Универсальной ширины окна не существует, и для каждой задачи её приходится специально подбирать. Так же, как не существует и универсальных ядер.

Наиболее простой ядерной функцией является прямоугольное ядро, оно разрывно и кусочно-постоянно:

$$K(z) = \frac{1}{2} [|z| \leq 1]$$

Чуть более сложный вариант ядра даёт треугольная ядерная функция, она уже непрерывна, хотя, являясь кусочно-линейной, она, разумеется, все ещё не является гладкой¹⁰¹:

$$K(z) = (1 - |z|) [|z| \leq 1]$$

Ещё один интересный со многих точек зрения вариант ядра — это так называемое ядро Епанчикова:

$$K(z) = \frac{3}{4} (1 - z^2) [|z| \leq 1]$$

оно непрерывно и является гладким везде кроме двух точек $z = \pm 1$. Пример всюду гладкой ядерной функции доставляет биквадратное ядро:

¹⁰¹ Гладкой называется функция, такая, что 1) она непрерывна, 2) её производная тоже непрерывна. Гладкие многообразия, определяемые такими функциями, характеризуются тем, что касательные пространства к ним не терпят дискретных скачков.

$$K(z) = \frac{15}{16} (1 - z^2)^2 [|z| \leq 1]$$

И, наконец, одно из наиболее часто используемых ядер — это бесконечно гладкое везде в $\mathbb{R}$ ядро Гаусса:

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-2z^2}$$

На практике чаще всего используют либо прямоугольное ядро для простоты вычислений, либо, в случае, когда важна гладкость модели, ядро Гаусса, причём, гладкость играет существенную роль в регрессионных задачах, к решению которых метрическими методами kNN мы переходим.

Алгоритм kNN легко переносится на задачу регрессии. Самый простой вариант — возвращать среднее значение окликов ближайших соседей:

$$a(x) = \frac{1}{k} \sum_{i=1}^k y_{(i)}(u)$$

Чуть более сложный вариант — вычислять не обычное, а среднее взвешенное откликов при помощи весов, задаваемых каким-либо ядром:

$$a(x) = \frac{\sum_{i=1}^k K\left(\frac{\rho(u, x_{(i)}(u))}{h}\right) y_{(i)}(u)}{\sum_{i=1}^k K\left(\frac{\rho(u, x_{(i)}(u))}{h}\right)}$$

Этот метод обычно называется методом Надаря – Ватсона. Он объединяет целое семейство методов для решения регрессионных задач, называемое ядерной регрессией (англ. — kernel regression).

Однако, можно подойти к восстановлению регрессии и с ещё более глубоких позиций, применив почти те же идеи оптимизации, что и выше, при решении задач классификации, но адаптировав их к регрессионной специфике.

Дело в том, что в задачах регрессии все отклики y почти наверняка будут различны, поэтому сумма индикаторов истинности $[y_{(i)}(u) = y]$ окажется состоящей не более чем из одного слагаемого, а это значит, что максимум всегда будет достигаться на одном ближайшем соседе, и, следовательно, метод всегда будет вырождаться в 1NN. Отметим, что в задаче классификации индикатор — естественная мера близости двух объектов: если объекты совпадают, то значение 1, если различаются, то 0, и проблема в том, что в задаче регрессии объекты являются действительными числами, а для них функция, которая выдаёт отличное от нуля значение лишь в случае точного равенства — это очень плохая мера близости. В случае непрерывных значений откликов естественно использовать более гладкие функции для выражения близости, поэтому, для обобщения формулы оптимизации на задачу регрессии нам достаточно всего лишь заменить индикатор на более гладкую функцию. При этом для действительных чисел чаще всего рассматривают не близость, а расстояние между ними, то есть некоторую метрику. Например, в качестве такой метрики можно взять квадрат евклидова расстояния $(y_{(i)}(u) - y)^2$. Отметив, что максимизация близости эквивалентна минимизации расстояния, получим следующую формулу оптимизации модели:

$$a(u) = \arg \min_{y \in Y^\ell} \sum_{i=1}^k K \left(\frac{\rho(x_{(i)}(u))}{h} \right) (y_{(i)} - y)^2$$

Ядерная регрессия, основанная на метрических методах kNN, обладает как своими достоинствами, так и своими недостатками. К достоинствам можно отнести следующие факторы:

1. этот алгоритм является непараметрическим, то есть, не делает явных предположений о распределении данных;
2. он прост в объяснении и интерпретации;
3. он достаточно точен, хотя и уступает чаще всего градиентному бустингу и

случайному лесу (см. ниже, разделы А.5 и А.6);

4. наконец, он может быть использован как для регрессии, так и для классификации.

Вместе с тем, несмотря на все свои очевидные преимущества, алгоритм kNN не лишён и минусов:

1. он неэффективен по памяти, поскольку для его применения приходится хранить всю обучающую выборку;
2. он чувствителен к масштабу данных и к неинформативным признакам;
3. для применения алгоритма необходимо, чтобы метрическая близость объектов совпадала с их семантической близостью, чего на практике бывает очень не просто добиться.

А.5. Решающие деревья и случайные леса

Ещё одно семейство моделей стохастического прогнозирования образуют решающие деревья (англ. — decision trees). Решающее дерево предсказывает значение целевой переменной с помощью применения последовательности простых решающих правил, которые называются предикатами. Этот процесс согласуется с естественным для человека процессом принятия решений, и, хотя обобщающая способность моделей на решающих деревьях невысока, их предсказания вычисляются довольно просто, из-за чего решающие деревья часто используют как фрагменты для построения ансамблей, то есть, моделей, делающих предсказания на основе агрегации предсказаний других моделей.

В древесных моделях решение о том, какое значение метки будет сопоставлено текущему объекту выборки, принимается с помощью прохода от корня дерева к некоторому листу. В каждом узле дерева решений находится предикат, если предикат верен для текущего примера из выборки, осуществляется переход в правого потомка, если нет — в левого. В листьях дерева записаны прогнозы, и как только проход по дереву оказывается в листе, соответствующему объекту приписывается то значение метки, которое находится в

этом листе. Это объясняет, в частности, то, что любое дерево осуществляет кусочно-постоянную аппроксимацию целевой зависимости.

Более строго в теории прогнозирования дерево решений определяется как бинарное дерево, в котором:

1. каждой внутренней вершине v приписан предикат $B_v : X \rightarrow \{0, 1\}$, где X — это пространство объектов;
2. каждой листовой вершине v приписан прогноз $c_v \in Y$, где Y — это область значений целевой переменной.

В ходе предсказания осуществляется проход по этому дереву к некоторому листу, причём, для каждого объекта $x \in X$ движение начинается из корня, и в каждой внутренней вершине проход продолжается вправо, если $B_v(x) = 1$, и влево, если $B_v(x) = 0$. Вообще, предикат B_v может иметь, произвольную структуру, но на практике чаще используют просто сравнение с порогом $t \in \mathbb{R}$ по какому-то признаку¹⁰²:

$$B_v(x, t) = [x \leq t]$$

При проходе через узел дерева с данным предикатом объекты будут отправлены в правое поддерево, если значение признака x у них меньше либо равно t , и в левое — если больше. В дальнейшем мы всегда будем по умолчанию использовать именно такие предикаты.

Алгоритм решающего дерева неявно содержит в себе несколько вспомогательных функций, которые выбираются так, чтобы итоговое дерево было способно минимизировать функционал качества $Q(a, X^\ell)$, связанный с некоторой функцией потерь $\mathcal{L}(a, x)$: ответ алгоритма $\text{Ans}(X_v)$, функция остановки алгоритма $\text{Stop}(X_v)$ и функция ветвления дерева $\text{Branch}(X_v)$, где

¹⁰² Квадратными скобками в теории стохастического прогнозирования принято обозначать индикатор истинности: он равен нулю, если значение в скобках ложно, и единице, если оно истинно.

$X_v \subset X^\ell$ — множество объектов из обучающей выборки X^ℓ , попавших в текущий лист v .

Функцией $\text{Ans}(X_v)$, вычисляющей ответ для листа по попавшим в него объектам из обучающей выборки, может быть, например:

1. в случае задачи классификации — метка самого частого класса;
2. в случае задачи регрессии — среднее, медиана или другая статистика целевой переменной;
3. в обоих случаях — другая модель, например, листы в дереве, задающем регрессию, могут быть линейными функциями или даже полиномами, обученными на данных, попавших в лист¹⁰³.

Критерий остановки $\text{Stop}(X_v)$ — функция, которая решает, нужно ли продолжать ветвление или же пора уже остановиться. Обычно это какое-то тривиальное правило: остановиться если все объекты в листе получились достаточно однородными и их не слишком много; или если следующий сплит превысит изначально заданную глубину дерева и т. д.

Однако, самая сложная компонента алгоритма — функция ветвления $\text{Branch}(X_v)$, измеряющая, насколько хорош предлагаемый сплит, то есть, насколько улучшится функционал качества в случае, если получившиеся два листа будут терминальными, по сравнению с ситуацией, когда сама исходная вершина — это лист. Функция $\text{Branch}(X_v)$ выбирает такой сплит, который даёт наиболее существенное улучшение.

Основная идея критерия ветвления заслуживает того, чтобы уделить ей отдельное внимание и остановиться на ней более подробно. Пусть дерево возвращает ответы $c \in \mathbb{R}$, и пусть задана некоторая функция потерь $\mathcal{L}(c, x)$, ко-

¹⁰³ Везде ниже мы будем предполагать, что каждый лист просто возвращает константу, но следует понимать, что это не всегда так.

торая зависит от того, какую именно константу возвращает дерево для каждого объекта x , попавшего в текущий лист. Наша задача — найти оптимальный сплит $X_v = X_{v'} \sqcup X_{v''}$, где символ $\sqcup$ означает объединение непересекающихся подмножеств, и прежде, чем решать эту задачу, мы можем, используя все объекты из ещё неразделенного листа v , вычислить тот константный таргет c , который предсказало бы дерево, будь текущая вершина терминальной, используя тот факт, что функционал качества $Q(c, X_v)$, определяемый функцией потерь $\mathcal{L}(c, x)$, должен принимать минимальное значение. Функционал качества — это среднее значение функции потерь по всем объектам из X_v :

$$Q(c, X_v) = \frac{1}{|X_v|} \sum_{x \in X_v} \mathcal{L}(c, x)$$

Оптимальное значение функционала качества получается минимизацией по всем возможным константным ответам c на листе v :

$$H(X_v) = \min_{c \in Y} \frac{1}{|X_v|} \sum_{x \in X_v} \mathcal{L}(c, x)$$

Получившаяся величина $H(X_v)$ называется информативностью листа (англ. — impurity, т. е. загрязнённость), и чем она ниже, тем лучше объекты в листе можно приблизить константным значением¹⁰⁴. Теперь допустим, что мы провели какое-то разбиение листа $X_v = X_{v'} \sqcup X_{v''}$ на две части, причём для определённости будем считать, что лист v' является левым потомком листа v , а лист v'' — правым. Тогда функционал качества для разбиения можно представить в виде выражения:

¹⁰⁴ Хотя, и русскоязычный термин «информативность» и англоязычный термин «загрязнённость» — неточно отражают суть явления. Гораздо точнее был бы термин «неопределённость», так как чем ниже неопределённость, тем меньше отличаются друг от друга таргеты для объектов листа, тем с большей уверенностью мы можем приписать всему листу одно константное значение.

$$Q(c, X_{v'} \sqcup X_{v''}) = \frac{1}{|X_v|} \left(\sum_{x \in X_{v'}} \mathcal{L}(c, x) + \sum_{x \in X_{v''}} \mathcal{L}(c, x) \right)$$

Тождественно преобразовав его, получим:

$$\frac{1}{|X_v|} \left(|X_{v'}| \frac{1}{|X_{v'}|} \sum_{x \in X_{v'}} \mathcal{L}(c, x) + |X_{v''}| \frac{1}{|X_{v''}|} \sum_{x \in X_{v''}} \mathcal{L}(c, x) \right)$$

и, после минимизации по $c \in Y$:

$$H(X_{v'} \sqcup X_{v''}) = \frac{|X_{v'}|}{|X_v|} H(X_{v'}) + \frac{|X_{v''}|}{|X_v|} H(X_{v''})$$

Теперь для того чтобы принять решение о разделении, мы можем сравнить значение информативности исходного листа $H(X_v)$ со значением информативности сплита $H(X_{v'} \sqcup X_{v''})$. Их разность равна

$$H(X_v) - \frac{|X_{v'}|}{|X_v|} H(X_{v'}) - \frac{|X_{v''}|}{|X_v|} H(X_{v''})$$

причём, для симметрии её принято умножить на $|X_v|$, после чего и получается критерий ветвления:

$$\text{Branch}(X_v) = |X_v| \cdot H(X_v) - |X_{v'}| \cdot H(X_{v'}) - |X_{v''}| \cdot H(X_{v''})$$

Выше мы уже отмечали (см. раздел А.1.1), что в случае регрессии с минимизацией средней квадратичной ошибки функция потерь имеет вид

$$\mathcal{L}(a, x) = (a(x) - y)^2$$

для любого алгоритма a , но, поскольку решающее дерево возвращает всегда константу, то алгоритм $a(x) = c$ для любого объекта $x \in X_v$, и, значит

$$\mathcal{L}(a, x) = \mathcal{L}(c, x) = (c - y)^2$$

Подставляя эту функцию потерь в формулу для информативности, получим:

$$H(X_v) = \min_{c \in Y} \frac{1}{|X_v|} \sum_{x \in X_v} (c - y)^2$$

Вместе с тем, хорошо известно, что в случае квадратичной функции потерь минимум для константного предсказания достигается при константе, равной среднему значению множества откликов, то есть, при

$$c = \frac{1}{|X_v|} \sum_{x \in X_v} y$$

Отсюда получаем окончательное выражение для формулы информативности, которая полностью определяет критерий ветвления:

$$H(X_v) = \frac{1}{|X_v|} \sum_{x \in X_v} (c - y)^2, \quad c = \frac{1}{|X_v|} \sum_{x \in X_v} y$$

Нетрудно видеть, что информативность оказалась равной дисперсии таргетов, попавших в лист, и, после этого замечания, критерий ветвления становится настолько изящен, что допускает даже простую словесную формулировку: оценка значения в каждом листе — это среднее, а выбирать сплиты надо так, чтобы сумма дисперсий в листьях потомков была как можно меньше.

Из описанной нами выше структуры дерева решений следует сразу несколько интересных его свойств, которые мы отметим, резюмируя идею древесных алгоритмов:

1. выученная функция всегда оказывается кусочно-постоянной, из-за чего её производная равна нулю везде, где она задана.
2. в отличие от рассмотренных выше (см. раздел А.3) линейных моделей, дерево решений не сможет экстраполировать найденные зависимости за границы области значений обучающей выборки.
3. дерево решений способно идеально приблизить обучающую выборку и при этом ничего не выучить: для этого достаточно построить такое дерево,

в каждый лист которого будет попадать только один объект.

Это значит, что при обучении решающего дерева следует не просто приближать обучающую выборку как можно лучше, но и стремиться оставлять дерево как можно более простым, чтобы результат обладал хорошей обобщающей способностью (аналогичное замечание для линейных моделей см. выше, раздел А.3). А именно:

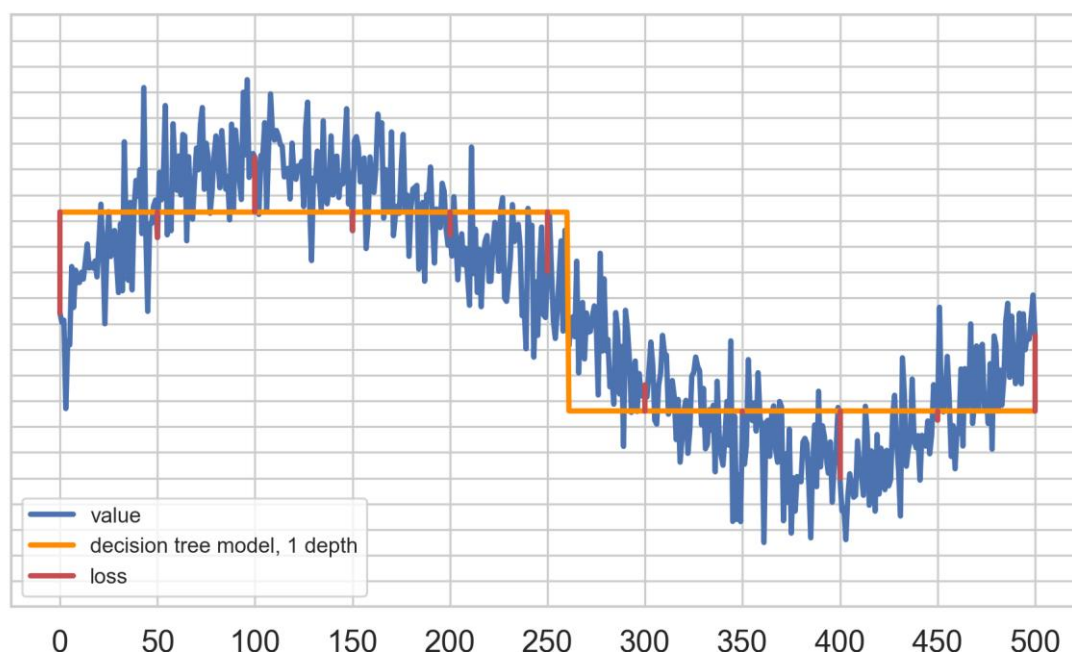


Рисунок А.5 — Аппроксимация нелинейного тренда решающим деревом глубины 1

При увеличении глубины дерева с одного (см. рисунок А.5) до трёх (см. рисунок А.6) точность модели повышается. И это вполне естественно, поскольку при этом происходит большее количество делений множества возможных значений переменной. Так, например, на пропущенном нами этапе дерева глубины два происходит ровно два деления: сначала исходный промежуток изменения переменной делится на два, а за тем — каждый из них ещё раз делится на два, в результате чего получается четыре промежутка, а на глубине, равной трём, происходит три деления и получается уже восемь промежутков.

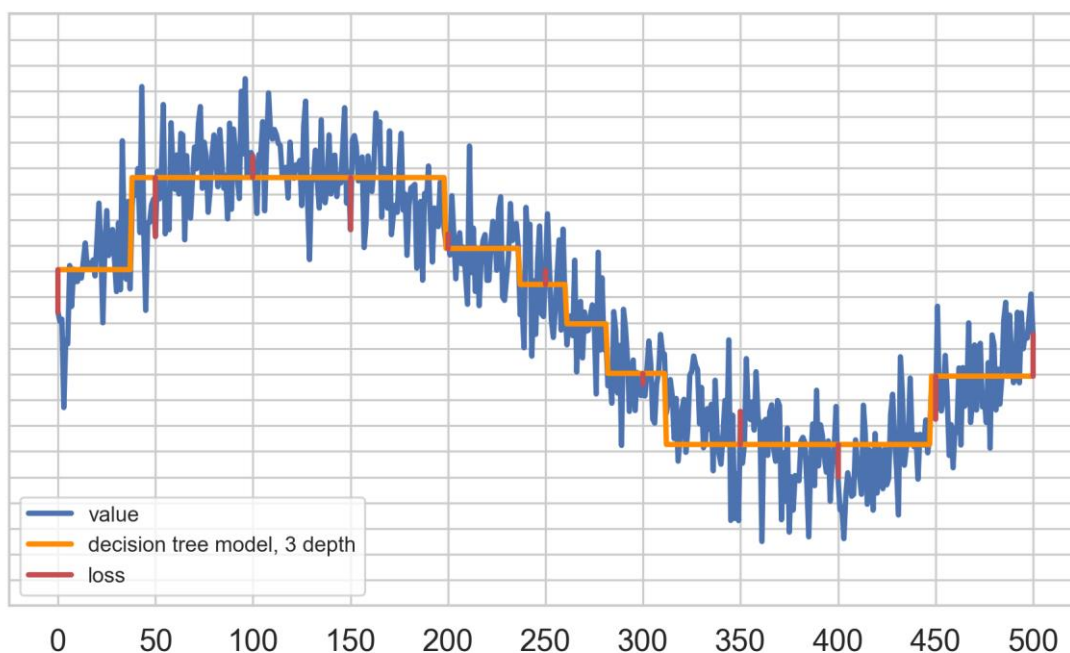


Рисунок А.6 — Аппроксимация нелинейного тренда решающим деревом глубины 3

Но уже на глубине пять мы наблюдаем целый ряд артефактов: в районе 0, в районе 230, в районе 275 и в районе 315 (см. рисунок А.7).

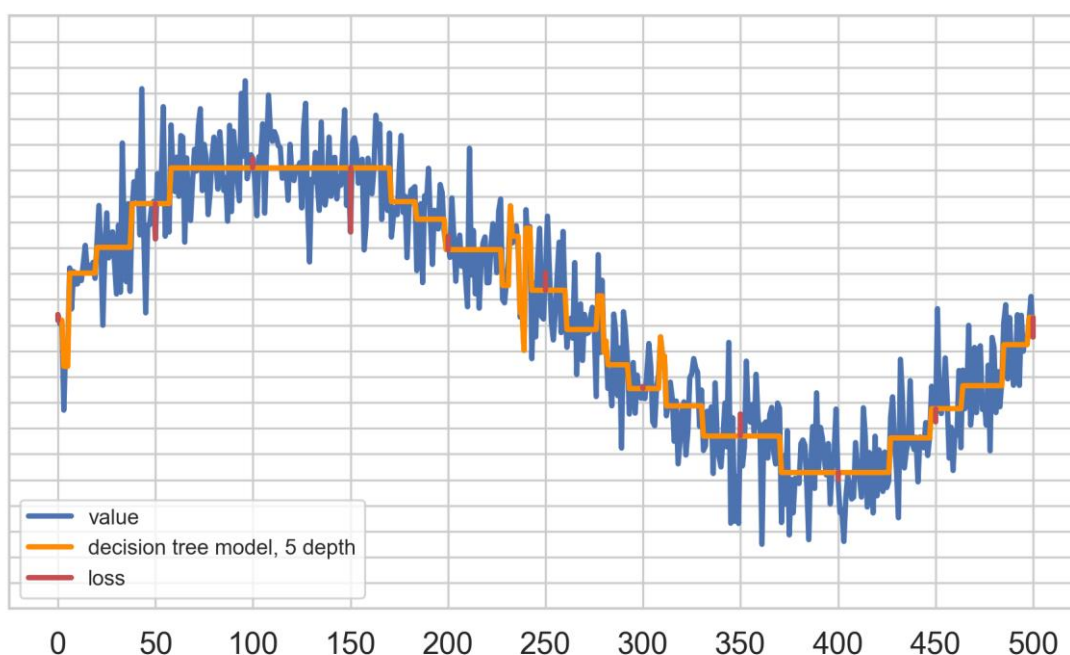


Рисунок А.7 — Аппроксимация нелинейного тренда решающим деревом глубины 5

Это означает переобучение модели, то есть, тот факт, что модель слишком хорошо подстраивается под имеющиеся обучающие данные, но при попытке воспроизвести алгоритм на тестовых данных она начнёт выдавать слишком

большое количество слишком больших ошибок.

Чтобы избежать переобучения и добиться одновременного выполнения несовместимых (на первый взгляд) требований простоты модели и её точности, используется приём, называемый бэггингом (англ. — bagging или bootstrap aggregation), который состоит в следующем. Пусть обучающая выборка X^ℓ состоит из ℓ объектов. Из этой совокупности выбираются ℓ примеров равновероятно, с возвращением. Получается новая выборка X_1^ℓ , в которой некоторых элементов исходной выборки может и не оказаться, а какие-то, наоборот, могут войти в неё несколько раз. С помощью некоторого алгоритма b на этой, вновь полученной частичной выборке обучается модель $b_1(x) = b(x, X_1^\ell)$. Затем эта процедура повторяется: формируется вторая выборка X_2^ℓ , тоже из ℓ элементов с повторением и с помощью того же алгоритма b на ней обучается вторая модель $b_2(x) = b(x, X_2^\ell)$. После k итераций, получаются k моделей, обученных на k выборках, а для того, чтобы получить одно итоговое предсказание, предсказания всех моделей усредняются:

$$a(x) = \frac{1}{k} \sum_{i=1}^k b_i(x)$$

Процесс генерации подвыборок с помощью семплирования с возвращением называется бутстрепом (англ. — bootstrap), модели $b_1(x), b_2(x), \dots, b_k(x)$ называются базовыми алгоритмами (хотя, наверное, лучше было бы назвать их базовыми моделями), а модель $a(x)$ называется ансамблем этих моделей.

В качестве примера возьмём временной ряд и, дважды выделив из него случайным образом некоторую часть наблюдений, получим два новых временных ряда¹⁰⁵ (см. рисунок А.8):

¹⁰⁵ Разумеется, в них будет меньшее число элементов. Тем не менее, оба они будут относиться к одному и тому же семейству временных рядов, так как они порождаются одним, общим для них обоим, прародителем.

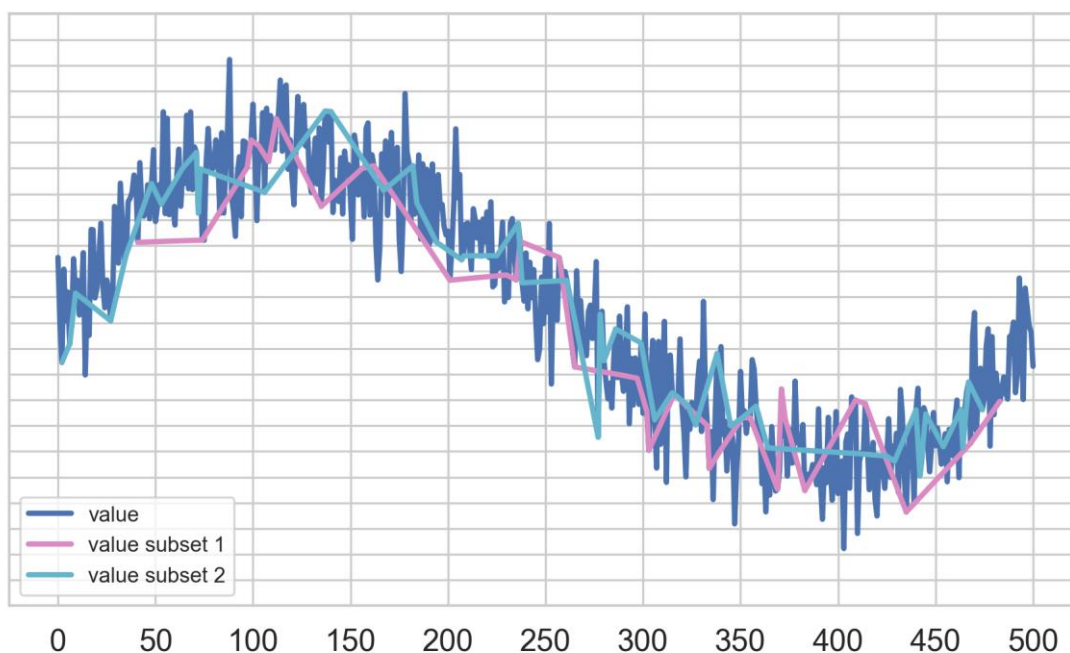


Рисунок А.8 — Две случайные выборки из исходного временного ряда

Оба раза мы можем построить решающее дерево небольшой глубины и получить, тем самым, две кусочно-постоянные модели, которые заведомо не будут переобученными. Ниже (см. рисунок А.9) представлены две такие модели с решающими деревьями глубины 2, обученные на двух случайных выборках из общего числа наблюдений.

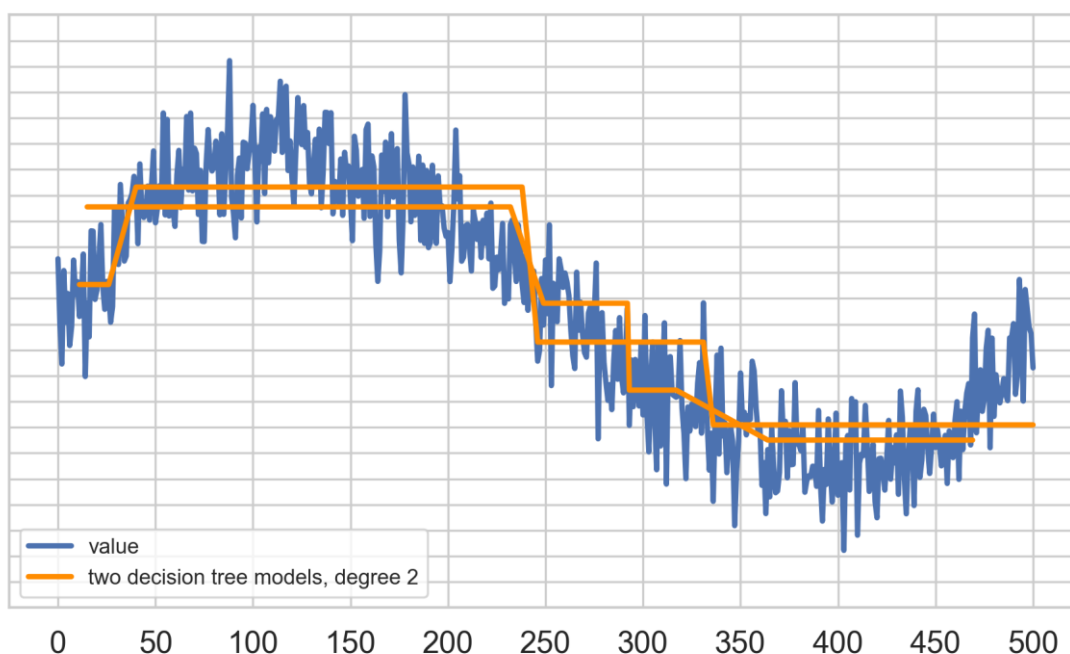


Рисунок А.9 — Два решающих дерева на двух случайных выборках

Каждая из этих двух моделей неточна и весьма грубо описывает целевую зависимость. Но их среднее арифметическое, будет более точным, так как разноимённые отклонения погасят друг друга при сложении, после чего среднее двух моделей будет представлять собой более гладкую кривую. Если же повторить эту процедуру достаточное неоднократно, каждый раз усредняя результат прогнозирования, то в итоге мы получим приемлемый результат.

Заметим, что ансамбль моделей на основе беггинга строится на неявном предположении, что базовые алгоритмы некоррелированы, и за счёт этого получается уменьшение дисперсии ансамбля относительно входящих в него базовых алгоритмов. Однако в реальности это не так: базовые алгоритмы всегда учат одну и ту же зависимость на пересекающихся выборках, поэтому корреляция, на самом деле, всегда ненулевая. На практике оказывается, что строгое выполнение предположения о некоррелированности не обязательно — достаточно, чтобы алгоритмы были просто не похожи друг на друга, и на этом строится развитие идеи бэггинга для решающих деревьев — случайный лес (анг. — random forest).

Случайный лес — это ансамбль алгоритмов, где в качестве базового алгоритма выступает решающее дерево. Его основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации или регрессии, но за счёт их большого количества получающийся в итоге усреднённый результат оказывается кратно лучшим по сравнению с каждым из его базовых алгоритмов.

Алгоритм случайного леса осуществляется по следующей схеме:

1. Для построения i -го дерева:

- а) Сначала, как и в обычном бэггинге, из обучающей выборки X^ℓ выбирается с возвращением случайная подвыборка X_i^ℓ того же размера, что и X^ℓ .

- b) Затем, в процессе обучения дерева в каждой вершине случайно выбираются $n' < n$ признаков, где n — полное число признаков (другими словами, применяется так называемый метод случайных подпространств), и среди них ищется оптимальный сплит.
2. Для получения предсказания ансамбля на тестовом объекте, усредняются отдельные ответы деревьев (в случае регрессии) или берётся самый популярный класс (в случае классификации).

При этом возникает ещё одна, дополнительная степень свободы в выборе признаков, причём, это происходит в каждом узле каждого дерева, что позволяет добиться большего разнообразия базовых алгоритмов, их меньшей коррелированности и, как следствие, большей устойчивости результирующей модели.

А.6. Градиентный бустинг над решающими деревьями

Бустинг (англ. — boosting) — это ансамблевый метод, в котором так же, как и в методах, рассмотренных выше, строится множество базовых алгоритмов из одного семейства, объединяющихся затем в более сильную модель. Отличие состоит в том, что в бэггинге и случайном лесе базовые алгоритмы учатся независимо и параллельно, а в бустинге — последовательно.

Каждый следующий базовый алгоритм в бустинге обучается так, чтобы уменьшить общую ошибку всех своих предшественников, в следствие чего, итоговая композиция будет иметь меньшее смещение, чем каждый отдельный базовый алгоритм (хотя и уменьшение разброса также может происходить). Основная цель бустинга — уменьшение смещения, поэтому в качестве базовых алгоритмов часто выбирают алгоритмы с высоким смещением и небольшим разбросом, например, если в качестве базовых классификаторов выступают деревья, то их глубина должна быть небольшой — обычно не больше 2-3 уровней. Ещё одной важной причиной для выбора моделей с высоким сме-

щением в качестве базовых является то, что такие модели, как правило, быстрее учатся. Это важно для их последовательного обучения, которое может стать очень дорогим по времени, если на каждой итерации будет учиться сложная модель. Хотя случайный лес — мощный и достаточно простой для понимания и реализации алгоритм, на практике он чаще всего сильно уступает бустингу. Поэтому бустинг сейчас — основное продакшн-решение, если работа происходит с табличными данными¹⁰⁶.

Бустинг, который использует деревья решений в качестве базовых алгоритмов, называется градиентным бустингом над решающими деревьями, (англ. — Gradient Boosting on Decision Trees, GBDT). Такой бустинг способен эффективно находить нелинейные зависимости в данных различной природы. Вообще, этим свойством обладают все алгоритмы, которые используют деревья решений, однако именно GBDT обычно выигрывает в подавляющем большинстве задач. Благодаря этому он широко применяется во многих областях машинного обучения, а так же — в задачах из индустрии: поисковом ранжировании, рекомендательных системах, таргетировании рекламы, предсказании погоды, выбора пункта назначения такси и т. д.

На уровне интуиции идея градиентного бустинга состоит в следующем. Пусть для решения регрессионной задачи выбирается алгоритм из семейства алгоритмов $\mathcal{B}$, под которым мы будем понимать множество всевозможных решающих деревьев (от англ. branch — ветвление). Если мы обучим единственное решающее дерево $b_1(x)$, то качество такой модели, скорее всего, будет низким. Однако, благодаря этому слабому и неточному решению, возникает новая информация о том, на каких объектах построенное дерево давало точные предсказания, а на каких ошибалось. Используем эту информацию и обучим ещё одну модель. Допустим, что предсказание первой модели $b_1(x)$ на объекте x отличается от истинного значения метки y на величину, равную $\delta(x)$,

¹⁰⁶ В работе нетабличными данными, такими, например, как изображения или тексты доминируют нейросети.

то есть: $b_1(x) = y + \delta(x)$. Тогда, если мы сможем обучить вторую модель $b_2(x)$, которая для объекта x возвращала бы ответ, равный $(-\delta(x))$, то сумма двух моделей будет возвращать точное значение:

$$b_1(x) + b_2(x) = (y + \delta(x)) + (-\delta(x)) = y$$

Другими словами, если вторая наша модель научится предсказывать разницу между реальным значением и ответом первой, то это позволит уменьшить ошибку композиции. В реальности вторая модель тоже не сможет обучиться идеально, поэтому мы обучим третью модель, которая будет компенсировать неточности первых двух, и будем продолжать так до тех пор, пока не построим устраивающую нас композицию из k алгоритмов.

Здесь будет уместно провести аналогию с разложением аналитической функции в ряд Тейлора. Всем известно, что бесконечно дифференцируемую функцию $f(x)$ в окрестности x_0 можно представить в виде бесконечной суммы степенных функций:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n$$

Нулевое слагаемое в этом разложении является константой, и вряд ли можно сказать, что она хоть как-то приближает функцию $f(x)$. Однако, после добавления первого слагаемого возникает линейная функция, которую уже можно назвать хоть и очень грубым, но всё-таки приближением. Если добавить ещё одно слагаемое, то получится квадратичная функция, которая чуть лучше приближает $f(x)$ и т. д. Каждая следующая элементарная функция увеличивает точность приближения, но менее заметна в общей сумме. И если нам не требуется абсолютно точное разложение, то вместо бесконечного ряда Тейлора мы можем ограничиться суммой достаточного количества его первых элементов, представив, таким образом, нужную нам функцию в виде суммы более простых функций.

Перенос этой идеи на алгоритмы машинного обучения означает, что в машинном обучении мы, прежде всего, выбираем подходящий алгоритм — подобно тому, как при разложении функции в ряд мы заранее решили, что будем приближать её степенными функциями, а не логарифмическими, экспоненциальными или какими-либо ещё. При этом мы можем выбрать сложный алгоритм, который сразу хорошо выучит истинную зависимость, а можем обучить простой, который выучит истинную зависимость весьма и весьма посредственно. Затем мы добавляем к нему ещё один простой алгоритм, чтобы уточнить предсказание первого, затем — ещё один и т. д. Продолжая этот процесс, получаем композицию простых алгоритмов, где первый алгоритм весьма грубо приближает истинную зависимость, а каждый следующий делает приближение всё более и более точным (хотя, и вклад каждого нового алгоритма в общую точность композиции будет все ниже с каждым шагом).

Применительно к нашей главной задаче восстановления регрессии со среднеквадратичной функцией потерь, построение модели состоит в том, чтобы минимизировать функционал качества

$$Q(a, Y^\ell, X^\ell) = \frac{1}{\ell} \sum_{i=0}^{\ell} (a(x_i) - y_i)^2 \rightarrow \min$$

Для решения регрессионной задачи методом градиентного бустинга строится композиция из k алгоритмов, которые берутся из семейства $\mathcal{B}$

$$a(x) = a_k(x) = b_1(x) + b_2(x) + \dots + b_k(x)$$

где в качестве множества базовых алгоритмов используется семейство $\mathcal{B}$ решающих деревьев некоторой фиксированной (как правило, небольшой) глубины. Используя известные методы построения решающих деревьев, мы обучаем алгоритм $b_1(x) \in \mathcal{B}$, который наилучшим образом приближает целевую переменную:

$$b_1 = \arg \min_{b \in \mathcal{B}} Q(b, Y^\ell, X^\ell)$$

Построенный алгоритм, скорее всего, работает не идеально¹⁰⁷. Поэтому далее мы вычисляем, насколько сильно отличаются предсказания этого дерева от истинных значений:

$$S_1^\ell = Y^\ell - b_1(X^\ell)$$

Это нужно для того, чтобы скорректировать наш алгоритм $b_1(x)$ с помощью алгоритма $b_2(x)$. Найти совершенный алгоритм, скорее всего, не получится, но по крайней мере так мы сможем выбрать из семейства наилучшего представителя для такой задачи. Таким образом, второе решающее дерево будет обучаться предсказывать разности между тем, что предсказывает первый алгоритм и истинными значениями отклика:

$$b_2 = \arg \min_{b \in \mathcal{B}} Q(b, S_1^\ell, X^\ell)$$

При этом композиция моделей $a_2(x) = b_1(x) + b_2(x)$ станет более точно предсказывать целевую переменную y .

Далее рассуждения повторяются. На k -ом шаге вычисляется разность между правильным ответом и текущим предсказанием композиции из $k - 1$ алгоритмов:

$$S_{k-1}^\ell = Y^\ell - \sum_{i=1}^{k-1} b_i(X^\ell) = Y^\ell - a_{k-1}(X^\ell)$$

Затем k -й алгоритм учится предсказывать эту разность

¹⁰⁷ Более того, если базовый алгоритм работает слишком хорошо на обучающей выборке, то, скорее всего, он переобучен.

$$b_k = \arg \min_{b \in \mathcal{B}} Q(b, S_{k-1}^\ell, X^\ell)$$

и композиция завершается по формуле:

$$a_k(x) = a_{k-1}(x) + b_k(x)$$

В заключение отметим работу градиентных методов в бустинге, а заодно — проясним название метода, который по определению является градиентным бустингом над решающими деревьями, но собственно градиент нам пока нигде не встречался. Дело в том, что прямая оптимизация нашего функционала качества — это задача, которая в общем случае решается только теоретически, а с вычислительной точки зрения она слишком сложна для решения за разумное время. Поэтому на практике минимизируют не сам функционал Q , а линейную часть его разложения в ряд Тейлора:

$$Q(a_{k-1} + b, Y^\ell, X^\ell) \approx Q(a_{k-1}, Y^\ell, X^\ell) + \langle b(X^\ell), \text{grad}(Q) \rangle$$

где угловые скобки означают скалярное произведение, а градиент берётся в точке $a_{k-1}(X^\ell, Y^\ell)$. С учётом того, что первое слагаемое $Q(a_{k-1}, Y^\ell, X^\ell)$ является константой и не влияет на аргумент минимизации, получается следующая оптимизационная задача:

$$\langle b(X^\ell), \text{grad}(Q) \rangle \rightarrow \min$$

в которой напрямую фигурирует градиент функционала Q . Поскольку наше минимизируемое выражение — это скалярное произведение двух векторов, его значение минимизируют такие значения вектора $b(x_i)$, которые похожи на антиградиент, то есть, пропорциональны отрицательным значениям градиента: $-\text{grad}_i(Q)$. Поэтому на каждой итерации базовые алгоритмы b_k обучаются предсказывать значения антиградиента функции потерь по текущим

предсказаниям композиции.

А.7. Метрики регрессии и классификации

Задачи машинного обучения по прецедентам, как правило, состоят в восстановлении зависимости между парами (признаковое описание, целевая переменная) по данным, доступным для анализа. Алгоритмы машинного обучения позволяют построить модель, аппроксимирующую эту зависимость, при этом почти наверняка модель будет ошибаться на некоторых объектах: будь она даже идеальной, шум или выбросы не позволят добиться точной аппроксимации. Разные модели будут ошибаться на разных объектах и в разной степени, поэтому одна из ключевых задач машинного обучения состоит в том, чтобы подобрать подходящий критерий, который позволит сравнивать различные модели¹⁰⁸.

Выбор критерия оценки зависит от типа решаемой задачи, которые на практике могут широко варьироваться и относиться практически к любому виду человеческой деятельности:

1. решить, сколько упаковок с конкретным видом продукта нужно привезти в конкретный магазин, чтобы предложение соответствовало спросу, и не пришлось выбрасывать излишки (сфера деятельности — ритейл);
2. увеличить счастье пользователя от работы с сервисом, чтобы он стал лояльным и приносил стабильный доход (сфера деятельности — телеком);
3. решить, нужно ли направить человека на дополнительное обследование или же имеющихся данных достаточно для постановки диагноза (сфера деятельности — медицинская диагностика);
4. и так далее.

¹⁰⁸ Отметим, что качество модели нельзя оценивать на обучающей выборке. Как минимум, это нужно делать на отложенной (тестовой) выборке, но если это позволяют время и вычислительные ресурсы, стоит прибегнуть и к более надёжным способам проверки — например, кросс-валидации (см. раздел А.8)

В каждом таком конкретном случае может возникнуть (и, как правило, возникает) целая иерархия метрик. Например, если речь идёт о стриминговом музыкальном сервисе, пользователям которого мы решили предоставить доступ к сгенерированным нейросетью трекам (не защищённым авторским правом, а потому совершенно бесплатным), то в качестве иерархии метрик можно предложить следующую упорядоченную последовательность:

1. самый верхний уровень — будущий доход сервиса, его невозможно измерить в моменте, так как он слишком сложным образом зависит от совокупности всех усилий всех подразделений сервиса;
2. медианная длина сессии, служащая оценкой удовлетворённости пользователей, которая, как мы надеемся, повлияет на их желание продолжать платить за подписку, её придётся измерять в продакшене, так как важна реакция настоящих пользователей на новшество;
3. доля удовлетворённых качеством сгенерированной музыки ассессоров, на которых она будет протестирована до выхода в продакшен;
4. функция потерь, ориентируясь на которую мы будем обучать генеративную сеть.

На этом примере видны сразу несколько общих закономерностей, характерные для различных видов оценки. Во-первых, метрики бывают offline и online. Online-метрики вычисляются по данным, собираемым с работающей системы (например, общий доход компании или медианная длина сессии — это именно online-метрики). Наоборот: offline-метрики вычисляются до введения модели в эксплуатацию, и некоторые из них могут вычисляться только например, с привлечением специальных людей, ассессоров¹⁰⁹. В данном разделе нас будут интересовать offline-метрики, которые могут быть получены вообще

¹⁰⁹ Этот приём часто применяется, когда в качестве метрики выступает реакция живого человека: например, так поступают поисковые компании, которые предлагают специально отобраннным людям оценить качество ранжирования экспериментальной системы ещё до того, как рядовые пользователи увидят эти результаты в обычном порядке.

без привлечения людей, и которые вычисляются на самых ранних этапах работы над моделью, когда ещё нет необходимости её внедрения даже в пилотном варианте.

Важно понимать разницу между функцией потерь, с которой мы неоднократно сталкивались выше, и метрикой качества, к описанию которой мы только пойдём. Функция потерь возникает в тот момент, когда мы сводим задачу построения модели к задаче оптимизации, причём, обычно требуется, чтобы она обладала хорошими свойствами (например, часто бывает нужно, чтобы она была дифференцируемой). Метрика — это внешний, объективный критерий качества, зависящий не от параметров модели, а только от предсказанных меток.

Самый простой и вместе с тем интуитивно очевидный пример метрики эффективности модели возникает в моделях бинарной классификации. Это доля верно предсказанных классов (англ. — accuracy):

$$\text{Accuracy}(a, X^\ell, Y^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

или же, наоборот: семантически сопряжённая ей метрика — доля ошибочных классификаций (англ. — error rate):

$$\text{ErrorRate}(a, X^\ell, Y^\ell) = 1 - \text{Accuracy}(a, X^\ell, Y^\ell)$$

Однако, посмотрев на метрику accuracy чуть более внимательно, легко заметить, что у неё есть весьма существенные недостатки. Во-первых, она не способна учесть дисбаланс классов. Например, в задаче диагностики редких заболеваний классификатор, предсказывающий всем пациентам отсутствие болезни будет иметь весьма высокую accuracy просто потому, что больных людей в выборке намного меньше. А во-вторых, она не учитывает цену ошибки на объектах разных классов. Для примера снова можно взять задачу медицинской диагностики: если ошибочно положительный диагноз для здорового

больного обернётся лишь ещё одним обследованием, то ошибочно отрицательный вердикт может привести к смерти пациента.

Поэтому при исследовании классификационных моделей применяются более сложные метрики, среди которых одной из основных является матрица несоответствий (англ. — *confusion matrix*). Исторически задача бинарной классификации — это задача об обнаружении чего-то редкого в большом потоке объектов, в связи с чем класс, который представляет для нас интерес, называется положительным, а оставшийся — отрицательным. Для каждого объекта в выборке возможно четыре ситуации:

1. Мы предсказали положительную метку и угадали. Такие объекты мы будем относить к группе TP истинно положительных объектов (англ. — *true positive*). Здесь мы использовали True, потому что предсказание было правильным, а *positive* — потому что предсказали положительную метку.
2. Мы предсказали положительную метку, но при этом ошиблись в своём предсказании, и на самом деле она является как раз отрицательной. Такие объекты образуют группу FP ложно положительных объектов (англ. — *false positive*). Здесь мы использовали False, потому что предсказание было неправильным.
3. Аналогично, объекты, на которых мы предсказали отрицательную метку и угадали, образуют группу TN всех истинно отрицательных объектов (англ. — *true negative*); True, потому что мы сделали верное предсказание, а *negative* — потому что метки, которые мы предсказывали, были отрицательными.
4. И наконец, объекты, на которых мы предсказали отрицательную метку, но ошиблись, и на самом деле они были положительными, образуют группу FN ложно отрицательных объектов (англ. — *false negative*); False, потому что предсказание оказалось ошибочным, а *negative* — потому что предсказывалась отрицательная метка.

Для удобства все эти числа изображают в виде таблицы, которую и называют матрицей несоответствий:

$$\text{ConfusionMatrix}(a, X^\ell, Y^\ell) = \begin{pmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{pmatrix}$$

Отметим, что если вычислить долю недиагональных элементов, то получится метрика *error rate*, о которой мы говорили выше:

$$\text{ErrorRate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

а доля диагональных элементов даёт метрику *accuracy*:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Рассмотрим ситуацию, когда положительный класс — это редкое событие, например, поисковую систему, в которой хранятся миллиарды документов, а релевантных к конкретному поисковому запросу на несколько порядков меньше, и пусть мы хотим решить задачу бинарной классификации «документ *d* релевантен по запросу *q*». Благодаря большому дисбалансу, *accuracy* тождественного классификатора, который объявляет все документы нерелевантными, будет близка к единице, так как для такого классификатора высокое значение метрики обеспечивается членом *TN*, в то время для пользователей более важен высокий *TP*. Поэтому в случае выраженной асимметрии классов, нужно использовать метрики, которые не учитывают истинно отрицательные объекты группы *TN* и ориентируются на истинно положительные, то есть, относящиеся к группе *TP*.

Рассмотрим долю правильно предсказанных положительных объектов среди всех объектов, идентифицированных классификатором в качестве положительных. Она представляет собой весьма важную для асимметричных классов метрику, которая называется точностью (англ. — *precision*):

$$\text{Precision}(a, X^\ell, Y^\ell) = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

В рамках примера с поиском документов эта метрика показывает долю релевантных документов среди всех, что найдены классификатором. Чем меньше ложно положительных срабатываний будет допускать модель, тем выше будет её метрика precision.

Если же рассмотреть долю правильно найденных положительных объектов среди всех объектов положительного класса, то получится ещё одна не менее важная метрика, которая называется полнотой (англ. — recall):

$$\text{Recall}(a, X^\ell, Y^\ell) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

В рамках нашего примера новая метрика recall показывает долю найденных документов среди всех релевантных: чем меньше ложно отрицательных срабатываний, тем выше recall модели.

Приоритет той или иной из этих двух метрик сильно зависит от контекста, в котором они будут использоваться. В каких-то случаях важна полнота, и тогда разработчики ориентируются на метрику recall, в других ситуациях, наоборот, важнее точность, и тогда применяется метрика precision. Например, задача кредитного скоринга — это, по сути, задача бинарной классификации, в которой потенциальные заёмщики делятся на два класса: класс заёмщиков, исполняющих свои кредитные обязательства, и класс потенциальных неплательщиков, причём, класс законопослушных клиентов гораздо шире класса неплательщиков. И в этом примере все зависит от кредитной политики финансовой организации. Если политика направлена на увеличение кредитной массы, то кредиты будут выдаваться при первой же возможности и с минимальным набором требований к заёмщику: в таком случае для классификатора важна будет полнота (метрика recall). Если же текущая политика организации требует снижения объёма кредитов, то кредиты будут выдаваться неохотно,

только в случае полной гарантии их возвратов: и тогда, наоборот, приоритетной будет точность классификатора (метрика precision).

Идеальный классификатор одновременно обладает высоким значением precision и высоким значением recall, однако, на практике этого не удаётся добиться практически никогда (см. рисунок A.10), и печальная истина состоит в том, что идеальных классификаторов не существует (как, впрочем, не существует и идеальных вариантов всех других моделей машинного обучения).

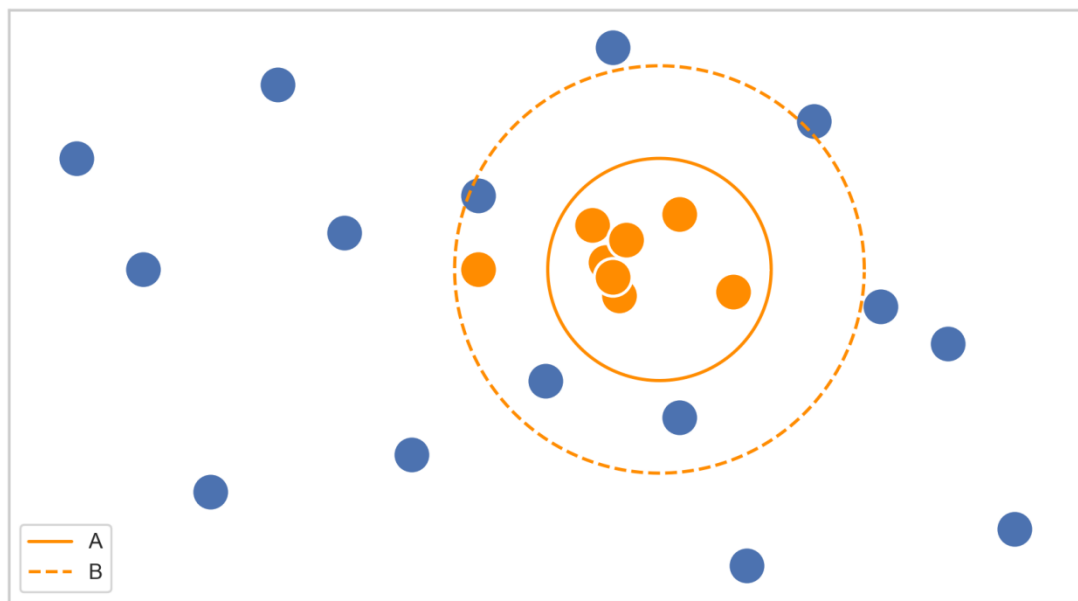


Рисунок A.10 — Соотношение метрик precision и recall при изменении параметров классификатора

Дело в том, что, повышая точность классификатора, мы, как правило, понижаем его полноту, и наоборот: повышая полноту, мы неизбежно понижаем точность. Так, на иллюстрации выше мы изобразили работу классификатора, который действует по принципу: «внутри этой окружности находятся жёлтые объекты, а все, что расположено вне неё, относится к классу синих объектов»¹¹⁰. Классификатор А обладает высокой, точнее говоря — идеальной

¹¹⁰ Мы рассматриваем только объекты, полностью попавшие внутрь окружности. Объекты, попавшие лишь какой-то своей частью, считаются находящимися во внешней зоне.

точностью: все, что этот классификатор посчитал жёлтым объектом, действительно оказалось жёлтым. Поэтому

$$\text{Precision}(A) = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{7}{7 + 0} = 1$$

Вместе с тем, его полнота не идеальна, так как один из действительно жёлтых объектов был отнесён к классу синих. Воспользовавшись формулой соответствующей метрики, получаем:

$$\text{Recall}(A) = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{7}{7 + 1} = \frac{7}{8}$$

Если мы хотим повысить полноту нашего алгоритма, то мы увеличиваем радиус нашей окружности и получаем классификатор В. Его полнота, действительно, становится выше, а если быть точным, то у классификатора В она идеальна, так как он выявил абсолютно все жёлтые объекты:

$$\text{Recall}(B) = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{8}{8 + 0} = 1 > \frac{7}{8}$$

Но в то же самое время классификатор В теряет точность, так как два синих объекта оказываются идентифицированы как жёлтые:

$$\text{Precision}(B) = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{8}{8 + 2} = \frac{8}{10} < 1$$

Как precision, так и recall вполне могут использоваться в качестве самостоятельных метрик классификации, однако нам гораздо удобнее иметь какой-то универсальный инструмент, который одновременно учитывал бы и точность, и полноту классификатора. В качестве компромиссного варианта часто используется так называемая F-мера, которая по определению является средним гармоническим мер recall и precision:

$$F(a, X^\ell, Y^\ell) = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}$$

Ещё один популярный комбинированный вариант, учитывающий одновременно и полноту, и точность классификатора, состоит в применении метрики ROC AUC (название происходит от двух английских аббревиатур: ROC — receiver operating curve, AUC — area under curve). Эта метрика рассчитана на применение бинарных классификаторов, возвращающих не класс, а вероятность принадлежности к единичному классу, а сам класс получается бинаризацией классификатора по некоторому фиксированному порогу. Если нашей целью является построение прогноза в терминах метки класса, то следует учитывать, что в зависимости от порога мы будем получать разные предсказания и, соответственно, разное качество алгоритма. Чем ниже порог отсечения, тем больше объектов модель будет относить к положительному классу, и наоборот: чем выше порог отсечения, тем меньше мы будем получать прогнозов с положительной меткой.

При уменьшении порога отсечения мы будем правильно предсказывать всё большее число положительных объектов, но также и неправильно предсказывать положительную метку на всё большем числе отрицательных объектов. В связи с этим, сначала вводятся две вспомогательные метрики TPR и FPR: где TPR (англ. — true positive rate) — это полнота, то есть, доля положительных объектов, правильно предсказанных положительными, а FPR (англ. — false positive rate) — доля отрицательных объектов, неправильно предсказанных положительными:

$$\text{TPR}(a, X^\ell, Y^\ell) = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR}(a, X^\ell, Y^\ell) = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Обе эти величины растут при уменьшении порога. Кривая в осях TPR/FPR, которая получается при изменении порога бинаризации от 0 до 1, называется ROC-кривой¹¹¹. Чем лучше классификатор разделяет два класса, тем больше

¹¹¹ Название «ROC-кривая» (англ. — receiver operating characteristic, то есть, рабочая характеристика приёмника) происходит от того, что она была разработана для операторов военных радиолокационных приёмников во время Второй мировой войны.

площадь под ROC-кривой — именно эта площадь используется в качестве метрики ROC AUC (см. рисунок A.11).

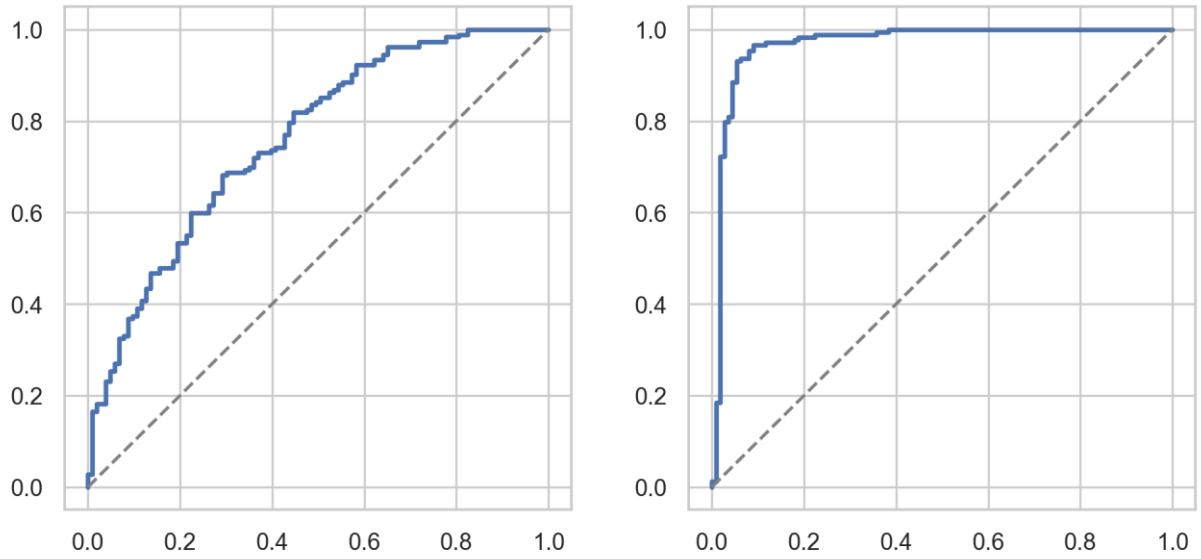


Рисунок A.11 — ROC-кривые и метрики ROC AUC

На иллюстрации выше представлены ROC-кривые двух классификаторов, причём, второй классификатор выигрывает у первого по ROC AUC.

Формально метрика ROC AUC определяется равенством:

$$\text{ROC AUC}(a, X^\ell, Y^\ell) = \frac{\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} [y_i < y_j] [a(x_i) < a(x_j)]'}{\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} [y_i < y_j]}$$

где квадратные скобки означают индикатор истинности, принимающий значение 0 в случае, если условие не выполнено, и 1, если оно выполнено, а штрихованные квадратные скобки — модифицированный индикатор истинности, который принимает дополнительное значение 0.5 в случае точного равенства.

В заключение этого раздела мы скажем несколько слов о метриках эффективности регрессионных задач. Они не так разнообразны, как в случае с классификаторами, но, тем не менее, их тоже несколько, и их тоже следует представлять.

Прежде всего отметим, что в задачах регрессии целевая метка `target` принимает потенциально бесконечное число различных значений, и природа этих

значений, как правило, связана с каким-либо процессом измерений: величины температуры в определённый момент времени на метеостанции, количества прочтений статьи на сайте, количества проданных продуктов в конкретном магазине, сети магазинов или стране и т. д. В каких-то случаях метка является целым числом, а иногда — это совершенно произвольное вещественное число, причём, обычно случаи целочисленных меток моделируются так, как если бы это были обычные вещественные числа¹¹².

В отличие от задач классификации, постановка регрессионной задачи такова, что по обучающей выборке (X^ℓ, Y^ℓ) где $y_i \in \mathbb{R}$ требуется найти оптимальный алгоритм $a : \mathbb{R}^\ell \rightarrow \mathbb{R}$, при этом величину возникающей ошибки $e_i = a(x_i) - y_i$ называют регрессионным остатком, а вся такая совокупность величин e_i , в некотором смысле, аналогична матрице несоответствий в задаче классификации. Интуитивно очевидно, что чем меньше регрессионный остаток, тем лучше модель воспроизводит реальные измерения, но для формализации этой интуиции применяют довольно-таки широкий спектр регрессионных метрик, построенных на вычислении разного рода ошибок. Наиболее распространёнными из метрик регрессии являются MSE, RMSE и MAE. Метрика MSE (англ. — mean squared error) определяется следующим равенством:

$$\text{MSE}(a, X^\ell, Y^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Иногда для того, чтобы показатель эффективности MSE имел размерность исходных данных, извлекают квадратный корень и получают показатель RMSE (англ. — root mean squared error):

¹¹² При таком подходе может оказаться, что модель А лучше модели В по некоторой метрике, но при этом предсказания модели А могут быть не целыми, что не соответствует поставленной бизнес-задаче. Если в бизнесе предполагается именно целочисленный ответ, то и оценивать нужно именно округление.

$$\text{RMSE}(a, X^\ell, Y^\ell) = \sqrt{\text{MSE}(a, X^\ell, Y^\ell)}$$

Если RMSE не подходит для решения задачи, например, если нужно сравнивать модели на выборках с большим количеством выбросов в данных, то прибегают к метрике MAE (англ. — mean absolute error):

$$\text{MAE}(a, X^\ell, Y^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

Метрики MSE и MAE относятся к категории метрик, ориентирующихся на абсолютную величину регрессионного остатка, но очень часто эта величина может быть не вполне подходящей для оценки модели. Рассмотрим, например, задачу прогнозирования продаж: пусть продукт А продаётся в количестве 100 штук, а продукт В в количестве 10 штук, и пусть базовая модель предсказывает количество продаж продукта А как 98 штук, а продукта В как 8 штук. Ошибки на этих объектах добавляют 4 штрафных единицы в MAE. Пусть также есть две модели, претендующие на улучшение базовой модели: первая предсказывает продажи товара А в количестве 99 штук, а товара В в количестве 8 штук; вторая предсказывает для товара А 98 штук, а для товара В 9 штук. Обе конкурирующие модели улучшают MAE базовой модели на 1 единицу. Однако, с точки зрения бизнеса, вторая модель предпочтительнее, так как предсказание продажи редких товаров обладает более высоким приоритетом. Один из способов учесть это требование состоит в том, чтобы рассматривать не абсолютную, а относительную ошибку на объектах.

Для этого часто используют метрику MAPE (англ. — mean absolute percentage error), которая определяется равенством:

$$\text{MAPE}(a, X^\ell, Y^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{|a(x_i) - y_i|}{|y_i|}$$

причём, если в знаменателе оказывается ноль, то за неверное предсказание

накладывается заранее фиксированный штраф, а для верного предсказания соответствующее слагаемое считается нулевым. Другой способ уйти от нуля в знаменателе состоит в применении метрики SMAPE (англ. — symmetric mean absolute percentage error):

$$\text{SMAE}(a, X^\ell, Y^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{2|a(x_i) - y_i|}{|a(x_i)| + |y_i|}$$

При применении этой метрики ноль в знаменателе возможен только в том случае, если он верно будет предсказан, и соответствующее слагаемое, по определению, полагают нулевым.

Однако, часто приходится сравнивать разные модели, обученные на данных, имеющих разную природу, и применяющихся для решения разных бизнес-задач. В таких ситуациях сравнение ошибок, интерпретированных в разных предметных областях и измеряющихся в разных единицах, не имеет никакого смысла: нельзя сравнить отклонение на два километра с ошибкой, составляющей два килограмма. Это приводит к идее универсальной метрики, которая плохо интерпретируется (или вообще не интерпретируется) в предметной области моделей, но зато применима при решении любой регрессионной задачи и, поэтому, может использоваться для сравнения любых регрессионных моделей. В качестве такой метрики, как правило используется коэффициент детерминации регрессионной модели¹¹³, который определяется при помощи следующего равенства:

$$R^2(a, X^\ell, Y^\ell) = 1 - \sum_{i=1}^{\ell} \frac{(a(x_i) - y_i)^2}{(\bar{y} - y_i)^2}$$

где $\bar{y}$ означает среднее значение целевой функции. Интерпретация значений

¹¹³ Везде в рамках настоящей диссертации в качестве метрики эффективности наших моделей мы используем именно коэффициент детерминации.

коэффициента детерминации такова:

1. $R^2 = 1$, в этом случае модель идеально предсказывает данные;
2. $R^2 = 0$, в этом случае модель работает по принципу «ленивого предсказания», то есть, прогноз модели константен и соответствует среднему арифметическому фактических целевых значений;
3. наконец, если $R^2 < 0$, то модель работает хуже, чем ленивое использование среднего.

А.8. Кросс-валидация

В машинном обучении прогнозирующие и аппроксимирующие модели выполняют различные функции, хотя обе могут использоваться для анализа данных и извлечения знаний. Аппроксимирующая модель стремится выявить закономерности в данных и создать некое смоделированное представление результата. Она подходит для анализа зависимости между переменными, где основное внимание уделяется выявлению структуры данных, а не предсказанию конкретных значений. Аппроксимирующие модели часто используются для упрощения сложных систем, чтобы сделать их более предсказуемыми, позволяя исследовать влияние различных факторов на результирующую переменную. С другой стороны, прогнозирующая модель направлена на предсказание значений на основе известных входных данных. Прогнозирующие модели могут использоваться для предсказания каких-либо характеристик изучаемого процесса, основываясь на исторических данных и различных факторах, влияющих на их величину. Они сосредотачиваются на точности предсказаний и могут применяться в ситуациях, где важна высокая степень уверенности в результатах. Важнейшей чертой прогнозирующих моделей является так называемая обобщающая способность модели, то есть, качество, позволяющее модели, обученной на одном наборе данных, делать хороший прогноз на другом наборе, который был неизвестен модели на этапе её обучения, но, тем не менее, относится к той же генеральной совокупности, что и объекты обучающей

выборки.

Отметим, что в машинном обучении под прогнозом понимают не обязательно прогноз будущего. Прогноз в машинном обучении — это значения целевого показателя на любых новых для модели данных, не обязательно связанных с хронологией событий (например, в банковском скоринге, обучив модель на одной группе клиентов, делают прогноз для другой группы). Это важно подчеркнуть, потому что, когда речь заходит о временных рядах, и при этом прогноз действительно приобретает временную специфику, эту же специфику приобретает и оценка прогнозирующей способности, о чем мы ещё будем говорить в дальнейшем (см. ниже в этом же разделе).

Для оценки прогнозирующей способности той или иной модели применяют уже рассмотренные нами метрики эффективности: в зависимости от задачи это могут быть как метрики классификации, так и регрессионные метрики (см. раздел А.7). Как правило, при этом используют процедуру кросс-валидации, которая имеет различные варианты, но все они основаны на том, что множество доступных для анализа данных разбивается на обучающую и тестовую выборки¹¹⁴, модель обучается на обучающей выборке, а метрики эффективности вычисляются на тестовой:

$$X = X_{\text{train}} \sqcup X_{\text{test}}, \quad Y = Y_{\text{train}} \sqcup Y_{\text{test}}$$

В простейшем случае разбиение производится один раз по формуле, данной выше. Однако и в этом, простейшем случае есть определённый риск получить несбалансированное разбиение. Например, может оказаться так, что данные для бинарного классификатора упорядочены по значению метки: сначала следуют одни единицы, а затем — одни нули (или наоборот). Тогда, обучаясь на первой половине, модель приучится выдавать одну и ту же метку в качестве выхода своего алгоритма, а получив на вход вторую половину, не сможет дать

¹¹⁴ Впрочем, иногда разбиение происходит на три части: обучающую, валидационную и тестовую.

ни одного правильного прогноза. Чтобы нивелировать риск, связанный с порядком следования данных, в момент разбиения данных на train и test используют случайное перемешивание.

Однако чаще в кросс-валидации применяют разбиение данных на несколько блоков, этот метод называется k -fold (англ. — fold, то есть, складка), и когда говорят о кросс-валидации, как правило, имеют в виду именно его. Множество всех данных разбивается на k блоков, которые называются фолдами:

$$X = X_1 \sqcup \dots \sqcup X_k, \quad Y = Y_1 \sqcup \dots \sqcup Y_k$$

Затем последовательно один из фолдов объявляется тестовой выборкой, а объединение всех оставшихся — обучающей:

$$(X, Y)_{\text{train}} = (X, Y) \setminus (X_i, Y_i), \quad (X, Y)_{\text{test}} = (X_i, Y_i)$$

и каждый раз происходит оценивание модели так же, как если бы разбиение было единственным, после чего оценки усредняются.

Метод k -fold реализуется как следующий алгоритм:

1. Фиксируется некоторое целое число k (обычно это число от 5 до 10), меньшее числа объектов в наборе данных.
2. Данные разбиваются на k одинаковых частей (в последней части может быть меньше семплов, чем в остальных). Эти части называются фолдами.
3. Далее происходит k итераций, во время каждой из которых один фолд выступает в роли тестового множества, а объединение остальных — в роли тренировочного. Модель учится на $k - 1$ фолде и тестируется на оставшемся.
4. Финальная оценка модели получается усреднением k получившихся тестовых результатов.

После применения кросс-валидации методом k -fold для одной модели у нас останется k экземпляров (иногда говорят — k инстансов) этой модели, обученных на разных подмножествах обучающей выборки. В связи с этим, возможные следующие варианты дальнейшего использования модели:

1. Использовать усреднение всех k инстансов;
2. из этих k инстансов выбрать тот, который набрал лучший скор на своём тестовом фолде, и применять дальше его;
3. заново обучить модель уже на всех k фолдах и делать предсказания уже этой моделью.

Указать, какой именно из этих способов следует применять, можно лишь в условиях реально реализуемого проекта в зависимости от конкретной задачи и имеющихся вычислительных возможностей.

Метод k -fold даёт более надёжную оценку качества модели, чем однократное разбиение на `train` и `test`, так как обучение и тест модели происходят на разных подмножествах исходного множества данных. Однако проведение k итераций обучения и теста может быть вычислительно затратным, и поэтому метод обычно применяют либо когда данных достаточно мало, либо при наличии большого количества вычислительных ресурсов, позволяющих проводить все k итераций параллельно.

В случае, когда данных мало, прибегают к методу LOO (англ. — `leave one out`), который является частным случаем метода k -fold, в нём каждый фолд состоит ровно из одного семпла:

$$X = x_1 \sqcup \dots \sqcup x_k, \quad Y = y_1 \sqcup \dots \sqcup y_k$$

Для валидации на каждой итерации методу LOO требуется всего один семпл, однако и итераций будет столько, сколько семплов в данных, поэтому метод невозможно применять на средних и больших задачах. Кроме того, для оценки

эффективности модели на одном объекте подходит не всякая метрика, например, коэффициент детерминации (см. раздел А.7) вычислить не удастся, так как ему необходимо как минимум два объекта, а для единственного объекта он просто не имеет смысла.

Мы уже отмечали, что кросс-валидация модели существенно зависит от фактора времени (см. выше этот же раздел и ниже, рисунок А.12).

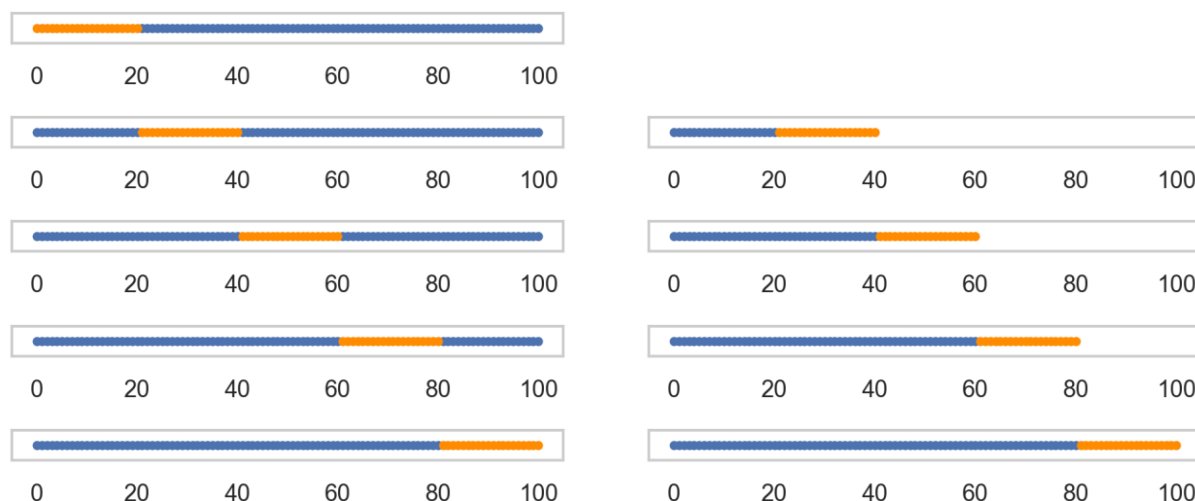


Рисунок А.12 — Классическая кросс-валидация (а) и кросс-валидация временного ряда (б)

На практике ситуация зависимости данных от времени возникает очень часто, это задача прогнозирования временных рядов, которая интуитивно может быть выражена в форме вопроса: «Что будет с наблюдаемыми характеристиками процесса в ближайший день/месяц/год?», и ответ на этот вопрос следует дать, исходя из имеющихся исторических данных. Кросс-валидация моделей для такой задачи осложняется тем, что тренировочные данные должны идти до тестовых, что продемонстрировано на иллюстрации, где в левой части представлена классическая кросс-валидация, а в правой — кросс-валидация временного ряда (см. рисунок А.12). При этом на каждой итерации обучающих данных становится все больше и модель, соответственно, становится все «умнее», но на начальных фазах этого процесса обучающих данных мало, а нулевая итерация вообще оказывается пропущенной из-за невозможности её провести.

ПРИЛОЖЕНИЕ В. БИБЛИОТЕКИ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Анализ данных играет важную роль в современных исследованиях и принятии решений в различных областях, от науки и техники до бизнеса и социальных наук [2–5, 7–9, 12, 17, 35, 173, 204, 218], причём, в связи с экспоненциальным ростом объёмов данных, генерируемых ежедневно, исследователю нужны какие-то эффективные инструменты для их обработки, анализа и интерпретации. Язык **Python**, благодаря своей простоте, гибкости и богатой экосистеме библиотек, стал одним из самых популярных языков программирования для решения этих задач [3, 4, 17, 51, 65, 172, 200, 201]. Он предлагает широкий спектр возможностей: от предварительной обработки и исследования данных до построения моделей машинного обучения [9, 12, 15, 35, 49, 172, 200, 201].

Среди огромного количества **Python**-библиотек, **pandas** и **sklearn** выделяются как ключевые инструменты, формирующие фундамент для большинства задач анализа данных [2, 4, 7, 8, 23, 48, 173, 209, 210]. Так, библиотека **pandas** предоставляет удобные структуры данных, гибкие возможности для их индексации и получения выборок, методы обработки пропусков, группировки, агрегации и т. д. [2, 3, 8, 172]. В свою очередь, библиотека **sklearn** является комплексной библиотекой машинного обучения и предлагает широкий спектр алгоритмов для классификации, регрессии, кластеризации, уменьшения размерности и предварительной обработки данных, что позволяет решать разнообразные задачи машинного обучения с использованием одного инструмента [2, 3, 8, 48, 51, 174].

В этом приложении мы кратко изложим основные понятия этих библиотек, структуры данных, которые в них используются, и дадим беглый обзор их возможностей с указанием наиболее часто используемых модулей и методов.

Безусловно, этот обзор не претендует хоть на какую-нибудь полноту¹¹⁵. Он лишь даёт представление о том, что именно нужно знать исследователю для того, чтобы свободно ориентироваться в алгоритмах методики МССР (Modeling, Challenge, Calibration, Production).

В.1. Библиотека **pandas**

Библиотека **pandas** — это незаменимый и важный инструмент для любого, кто работает с данными в **Python**. Она значительно упрощает весь процесс анализа и какого-либо манипулирования данными, позволяя сосредоточиться на извлечении действительно полезной информации и решении проблем. **Pandas** используется для работы с данными, структурированными в виде таблиц или временных рядов и предоставляет мощные инструменты для:

1. анализа данных — библиотека позволяет исследовать данные, находить закономерности, вычислять статистические показатели и делать выводы;
2. очистки данных — упрощает процесс обработки пропущенных значений, дубликатов, некорректных данных и приведения данных к нужному формату;
3. преобразования данных — предоставляет инструменты для изменения структуры данных, объединения таблиц, изменения типов данных и т.д.;
4. визуализации данных — интегрируется с библиотеками визуализации, такими как **matplotlib** и **seaborn**, чтобы создавать графики и диаграммы для наглядного представления данных;
5. импорта и экспорта данных — поддерживает чтение и запись данных из различных форматов, таких как **csv**, **json** и другие.

¹¹⁵ Для получения полных сведений об этих библиотеках, следует обратиться к официальной документации.

Выражаясь в более конкретных терминах, библиотека **pandas** позволяет исследователю выполнять очень многие важные действия:

1. загружать данные из файлов (**csv**, **excel**, **json** и т.д.) или баз данных;
2. осматривать данные — выводить первые или последние строки, получать статистическую сводку, проверять типы данных и т.д.;
3. фильтровать данные — выбирать строки или столбцы, соответствующие определенным условиям;
4. сортировать данные — располагать строки в определённом порядке;
5. группировать данные — агрегировать данные по определенным категориям;
6. объединять данные — соединять несколько таблиц вместе;
7. выполнять математические и статистические операции — вычислять средние значения, стандартные отклонения, корреляции и т.д.;
8. работать с временными рядами — анализировать данные, изменяющиеся во времени.

К основным структурам данных в библиотеке **pandas** относятся **series**, то есть, одномерный массив данных с метками (индексами), и **dataframe** — двумерная таблица данных, состоящая из **series**.

В.1.1. Структуры данных в библиотеке pandas

В основе **pandas** лежат две ключевые структуры данных: **series** и **dataframe**, и понимание этих структур является необходимым условием для эффективного использования **pandas** в решении задач анализа данных.

Series можно представить как одномерный массив данных, аналогичный списку или массиву **numpy**, но с одним важным отличием: каждому элементу массива сопоставлена метка, называемая индексом, который обеспечивает доступ к элементам **series** по имени, что делает работу с данными более

интуитивной и гибкой. Например, мы можем задать объект `series` при помощи списка `values`, указав в качестве дополнительной опции ещё один список `letters`, указывающий, что значения теперь индексируются буквами:

```
values = [10, 20, 30, 40, 50]
letters = ['A', 'B', 'C', 'D', 'E']
s = pd.Series(data=values, index=letters)
```

и тогда получившийся объект `s` будет иметь вид проиндексированной таблицы с одним столбцом:

A	10
B	20
C	30
D	40
E	50

Если при создании объекта индексы не указывать, то по умолчанию индексация осуществляется как нумерация, начиная с 0. Доступ к элементам серии осуществляется по значению индекса, например, `s['A']` вернёт значение 10, а `s['D']` — значение 40, а если индекс не задавался специально, то доступ осуществляется по номеру элемента.

С сериями в библиотеке `pandas` можно выполнять арифметические и статистические операции, например, если заданы две серии `s1` и `s2`, то выражение `s1 + s2` вернёт их сумму¹¹⁶, `s.mean()` вернёт среднее значение серии `s`, а `s.std()` — её среднее квадратичное отклонение. Кроме того, метод `apply()` позволяет применять любую функцию к каждому элементу серии. Например, выражение

```
def square(x):
    return x * x
s.apply(square)
```

¹¹⁶ Понятно, что длины серий должны быть равны, иначе это выражение вернёт ошибку.

вернёт новую серию, элементами которой будут точные квадраты серии `s`.

Ещё одно важное и очень удобное средство работы с сериями `pandas` состоит в возможности проверять наличие в серии пропусков. Метод `isnull()` возвращает серию с булевыми компонентами: `True` в случае, если соответствующий элемент пропущен в серии, и `False` во всех остальных ситуациях. Метод `notnull()`, наоборот, вернёт `False` для всех пропусков и `True` для всех присутствующих элементов.

Датафрейм — это двумерная табличная структура данных, состоящая из строк и столбцов. Каждый столбец датафрейма представляет собой серию, а строки имеют общий индекс. Датафрейм является основной структурой данных в библиотеке `pandas` и очень широко используется для хранения и обработки табличных данных.

Проще всего получить датафрейм из двумерного массива или списка¹¹⁷, например, следующий код

```
data = [['Alice', 25, 'New York'],
        ['Bob', 30, 'London'],
        ['Charlie', 28, 'Paris'],
        ['David', 32, 'Tokyo']]
df = pd.DataFrame(data, columns=['Name', 'Age', 'City'])
```

вернёт таблицу:

	Name	Age	City
0	Alice	25	New York
1	Bob	30	London
2	Charlie	28	Paris
3	David	32	Tokyo

Здесь индексация строк ведётся по умолчанию (то есть, как и в сериях — пу-

¹¹⁷ Хотя, это не единственный способ получить датафрейм, и чаще всего их получают чтением из внешних файлов.

тём нумерации от 0, хотя мы могли бы задать и специальные имена для индексов), и кроме того, мы указали ещё один дополнительный параметр: имена столбцов `columns` (если бы мы этого не сделали, столбцы тоже получили бы номера, начиная с 0). Таким образом, мы получили двумерную структуру, для обращения к отдельным элементам которой необходимо знать: номер (или индекс) строки и имя (или номер) столбца.

Обычно датафреймы содержат огромное количество строк, и, разумеется, их никто никогда не задаёт вручную. Самый распространённый метод получения датафрейма — это чтение из внешнего файла. Библиотека `pandas` позволяет считывать файлы различных форматов, и хотя, как правило, это файлы `csv`, следует понимать, что читать можно также файлы `json`, `excel` и т. д. Чтение датафрейма из файла `csv` осуществляется при помощи метода `read_csv()` с указанием пути к файлу, из которого берутся данные:

```
df = pd.read_csv('путь к файлу')
```

Итак, объект `dataframe` является основным объектом библиотеки `pandas` и состоит из следующих основных компонентов:

1. индекс (`index`) — метки строк;
2. столбцы (`columns`) — названия столбцов, каждый из которых по отдельности представляет собой объект `series`;
3. данные (`data`) — фактические данные, хранящиеся в таблице.

Понимание этих структур и их операций является необходимым условием для эффективного использования `pandas` в решении задач анализа и манипулирования данными.

В.1.2. Предобработка данных с использованием Pandas

Предобработка данных — это фундаментальный этап в любом процессе анализа данных или машинного обучения. Неважно, насколько сложны алгоритмы или мощны вычислительные ресурсы, если данные содержат ошибки,

неточности, пропущенные значения или имеют неподходящий формат, результаты анализа будут неверными или, в лучшем случае, неполными (здесь мы ещё раз вспоминаем бесспорную истину машинного обучения: «Trash in — trash out», что в переводе на русский язык означает: «Мусор на входе — мусор на выходе»). **Pandas**, мощная и гибкая библиотека **Python** для анализа данных, предоставляет широкий спектр инструментов для эффективной предобработки данных. В этом разделе мы рассмотрим основные этапы предобработки, используя **pandas**.

Прежде всего, следует упомянуть об очистке данных, то есть, процессе выявления и исправления неточностей, ошибок и несоответствий в исходном массиве. Основными задачами очистки являются обработка пропущенных значений (**NaN**, **None**) и удаление дубликатов. Пропущенные значения, часто обозначаемые как **NaN** (англ. — Not a Number) в **pandas**, или, синонимично, **None**, представляют собой отсутствующие данные. Они могут возникать по разным причинам, например, из-за ошибок при сборе данных, технических проблем или просто из-за того, что некоторые данные не были доступны. Наличие пропущенных значений может привести к искажению результатов анализа и некорректной работе алгоритмов машинного обучения. **Pandas** предоставляет несколько способов обработки пропущенных значений:

1. **isnull()** и **notnull()** — эти методы позволяют определить, какие ячейки в датафрейме содержат пропущенные значения (**NaN** или **None**), **isnull()** возвращает **True** для ячеек с пропущенными значениями и **False** в противном случае, а **notnull()** делает наоборот;
2. **dropna()** — этот метод позволяет удалить строки или столбцы, содержащие пропущенные значения. Он принимает аргумент **axis**, который указывает, удалять строки (**axis=0**, по умолчанию) или столбцы (**axis=1**), аргумент **how** позволяет указать, какие строки/столбцы удалять: **how='any'** (удалить, если есть хотя бы одно пропущенное значение) или **how='all'** (удалить, если все значения пропущены).

3. `fillna()` — этот метод позволяет заполнить пропущенные значения определенным значением, он принимает аргумент `value`, который указывает, каким значением заполнить пропущенные значения, например, `fillna(0)` заполняет все пропуски нулями.

Выбор метода обработки пропущенных значений зависит от конкретной задачи и характеристик данных. Удаление строк или столбцов с пропущенными значениями может привести к потере значимой информации, поэтому его следует использовать осторожно. Заполнение пропущенных значений также требует внимательного анализа, чтобы не внести искажения в данные. Важно учитывать, почему данные пропущены и какое значение будет наиболее подходящим для заполнения.

Дубликаты — это идентичные строки, присутствующие в датафрейме. Они могут возникать из-за ошибок при сборе данных, объединения данных из разных источников или каких-то других причин. Наличие дубликатов может привести к искажению результатов анализа и некорректной работе алгоритмов машинного обучения. Библиотека `pandas` предоставляет метод `duplicated()` для выявления дубликатов и метод `drop_duplicates()` для их удаления:

1. `duplicated()` — этот метод возвращает `True` для строк, которые являются дубликатами, и `False` в противном случае. Он принимает аргумент `subset`, который позволяет указать, по каким столбцам искать дубликаты;
2. `drop_duplicates()` — этот метод удаляет дубликаты из датафрейма, он принимает те же аргументы, что и `duplicated()`.

Покончив с очисткой, следует рассмотреть преобразование данных — процесс изменения формата, структуры или значений данных, с тем чтобы сделать их более пригодными для анализа или машинного обучения. Основными задачами преобразования данных являются изменение типов данных, переименование столбцов и добавление новых столбцов.

Тип данных столбца определяет, какие операции можно выполнять с данными в этом столбце и как они будут интерпретированы. Например, столбец, содержащий числовые данные, должен иметь тип `int` или `float`, а столбец, содержащий текстовые данные, должен иметь тип `string` или `object`. Иногда необходимо изменить тип данных столбца, например, чтобы выполнить арифметические операции с данными, которые были изначально представлены как строки. **Pandas** предоставляет метод `astype()` для изменения типов данных столбцов. Важно помнить, что преобразование типов данных может привести к ошибкам, если данные не соответствуют нужному целевому типу. Например, преобразование строки, содержащей нечисловые символы, в числовой тип приведёт к ошибке.

Переименование столбцов часто может быть полезно для улучшения читаемости кода, упрощения доступа к столбцам или приведения имён столбцов к единому стилю. **Pandas** предоставляет метод `rename()` для переименования столбцов. Он принимает аргумент `columns`, который является словарём, где ключи — это старые имена столбцов, а значения — это новые имена.

Добавление новых столбцов позволяет создавать новые признаки на основе существующих данных. Это может быть полезно для упрощения анализа, создания новых метрик или подготовки данных для машинного обучения.

Новые столбцы можно добавлять несколькими способами:

1. присваивание значения столбцу — это самый простой способ: присвоить новое значение столбцу, используя квадратные скобки.
2. `assign()` — этот метод позволяет добавлять несколько новых столбцов одновременно, он принимает аргументы, которые являются парами: `имя столбца = значение`.
3. `apply()` — этот метод позволяет применять пользовательскую функцию

к каждой строке или столбцу датафрейма, что бывает полезно для выполнения более сложных преобразований (см. выше).

И, наконец, фильтрация и выборка данных позволяют извлекать подмножества данных, удовлетворяющие определенным критериям. Это необходимо для фокусировки анализа на конкретных группах данных, отбора данных для обучения моделей или отбрасывания нерелевантных данных. **Pandas** предоставляет несколько способов фильтрации и выборки данных:

1. самый простой способ — использовать логические условия для выбора строк, удовлетворяющих определенным критериям, например:

```
df[df['A'] > 2] # Строки, где A > 2
df[(df['A'] > 2) & (df['B'] < 9)] # Строки, где A > 2 и B < 9
df[df['A'].isin([1, 3, 5])] # Строки, где A равно 1, 3 или 5
```

2. Использование методов **loc** и **iloc** — методы **loc** и **iloc** позволяют выбирать строки и столбцы по меткам или по индексам.
3. Группировка данных — позволяет агрегировать и анализировать данные по группам, определенным одним или несколькими столбцами, это бывает полезно для выявления закономерностей, сравнения групп и вычисления сводных показателей, **pandas** предоставляет метод **groupby()** для группировки данных.

После группировки можно применять различные агрегирующие функции, такие как

sum(), mean(), median(), count(), min(), max(), std(), и т.д.

Также можно использовать метод **agg()** для применения нескольких агрегирующих функций одновременно или для применения разных агрегирующих функций к разным столбцам.

В.2. Библиотека `sklearn`

Библиотека `scikit-learn` (часто сокращаемая до `sklearn`) в `Python` используется для машинного обучения. Она предоставляет инструменты для решения задач классификации, регрессии, кластеризации, уменьшения размерности, выбора моделей и предварительной обработки данных. По сути, эта библиотека предоставляет настолько полный инструментарий машинного обучения, что многие специалисты по анализу данных не используют методы других библиотек. Ключевые области применения `sklearn` таковы:

1. классификация — задача определения, к какой категории принадлежит объект (мы подробно говорили об этом в приложении А, уделяя при этом особое внимание бинарной классификации, как наиболее важному типу таких задач в контексте настоящей диссертационной работы);
2. регрессия — задача предсказания числового значения, являющаяся ключевой в настоящей диссертации, так как именно при помощи решения множества регрессионных задач нам удаётся выполнить подходящий прогноз для выявления аномальных состояний;
3. кластеризация — задача группировки объектов на основе их схожести;
4. уменьшение размерности — задача сокращения количества признаков в данных, с сохранением наиболее важной информации;
5. выбор моделей — инструменты для оценки производительности различных моделей машинного обучения и выбора наилучшей для конкретной задачи (этот приём мы активно используем на этапе моделирования процесса);
6. предварительная обработка данных — преобразование данных в формат, подходящий для обучения моделей, включая масштабирование данных (например, `MinMaxScaler`, `StandardScaler`), кодирование категориальных признаков (например, `OneHotEncoder`), обработку пропущенных значений.
7. оценка моделей — измерение производительности обученных моделей с

использованием различных метрик (например, `accuracy`, `precision`, `recall`, `F1-score` и др.).

В.2.1. Модули и архитектура библиотеки `sklearn`

`Sklearn` — это мощная и широко используемая библиотека `Python` для машинного обучения. Она предоставляет широкий спектр инструментов для решения задач классификации, регрессии, кластеризации, уменьшения размерности, предварительной обработки данных и многого другого. Благодаря своей простоте использования, хорошо документированному API и обширной коллекции алгоритмов, `sklearn` является отличным выбором как для начинающих, так и для опытных специалистов в области машинного обучения. Библиотека организована в виде нескольких основных модулей, каждый из которых отвечает за определённую задачу машинного обучения.

1. Классификация (Classification). Этот модуль содержит алгоритмы для решения задач классификации, то есть присвоения входным данным одной из предопределённых категорий или классов. Примеры алгоритмов:
 - a) логистическая регрессия (Logistic Regression) — используется для бинарной классификации, оценивает вероятность принадлежности к одному из двух классов;
 - b) метод опорных векторов (Support Vector Machines) — находит оптимальную гиперплоскость, разделяющую данные на классы с максимальным зазором;
 - c) деревья решений (Decision Trees) — строят древовидную структуру, принимающую решения на основе последовательности вопросов об атрибутах данных;
 - d) случайный лес (Random Forest) — ансамбль деревьев решений, уменьшающий переобучение и повышающий точность;
 - e) метод k ближайших соседей (k-Nearest Neighbors) — классифицирует

объект на основе классов его k ближайших соседей в пространстве признаков.

- f) наивный байесовский классификатор (Naive Bayes) — основан на теореме Байеса и предполагает независимость признаков.

2. Регрессия (Regression). Этот модуль содержит алгоритмы для решения задач регрессии, то есть предсказания числового значения на основе входных данных. Примеры алгоритмов:

- a) линейная регрессия (Linear Regression) — ищет лучшую линейную зависимость между входными признаками и целевой переменной;
- b) полиномиальная регрессия (Polynomial Regression) — расширяет линейную регрессию, используя полиномиальные признаки;
- c) гребневая регрессия (Ridge Regression) и Lasso регрессия — регуляризованные версии линейной регрессии, предотвращающие переобучение;
- d) деревья решений (Decision Trees) — могут использоваться и для регрессии, предсказывая среднее значение целевой переменной в листе дерева;
- e) случайный лес (Random Forest) — это ансамбль деревьев решений, используемый для регрессии;
- f) градиентный бустинг (Gradient Boosting) — строит ансамбль слабых моделей (обычно деревьев решений) последовательно, каждая из которых исправляет ошибки предыдущей;
- g) метод опорных векторов (Support Vector Regression) — адаптация SVM для задач регрессии.

3. Кластеризация (Clustering). Этот модуль содержит некоторые алгоритмы для группировки данных в кластеры на основе их схожести. Примеры алгоритмов:

- a) k-средних (k-Means) — алгоритм разделяет данные на k кластеров, минимизируя расстояние от точек до центроидов кластеров на основании какой-либо метрики (ка правило, евклидовой);
 - b) иерархическая кластеризация (Hierarchical Clustering) — строит иерархию кластеров, позволяя анализировать данные на разных уровнях granularity;
 - c) DBSCAN (Density-Based Spatial Clustering of Applications with Noise) — находит кластеры на основе плотности точек, выделяя выбросы;
 - d) агломеративная кластеризация (Agglomerative Clustering) — начинает с каждого объекта в отдельном кластере и постепенно объединяет наиболее близкие кластеры до достижения желаемого количества.
4. Понижение размерности (Dimensionality Reduction). Это ещё один модуль, он содержит алгоритмы для уменьшения количества признаков в данных, сохраняя при этом важную информацию. Примеры алгоритмов:
- a) метод главных компонент (Principal Component Analysis) — находит главные компоненты, которые объясняют наибольшую дисперсию в данных;
 - b) линейный дискриминантный анализ (Linear Discriminant Analysis) — находит линейные комбинации признаков, которые наилучшим образом разделяют классы;
 - c) метод независимых компонент (Independent Component Analysis) — разделяет мультивариативный сигнал на аддитивные подкомпоненты, предполагая, что они статистически независимы.
5. Предварительная обработка (Preprocessing). Этот модуль содержит инструменты для подготовки данных к машинному обучению, включая масштабирование, кодирование категориальных признаков, обработку пропущенных значений и разделение данных на обучающую и тестовую выборки. Примеры инструментов:

- a) **StandardScaler** — масштабирует данные, приводя их к нулевому среднему и единичному стандартному отклонению;
- b) **MinMaxScaler** — масштабирует данные в заданный диапазон (обычно от 0 до 1);
- c) **RobustScaler** — масштабирует данные, используя медиану и межквартильный размах, что делает его более устойчивым к выбросам;
- d) **OneHotEncoder** — кодирует категориальные признаки в числовые, создавая отдельные столбцы для каждой категории;
- e) **LabelEncoder** — кодирует категориальные признаки в числовые значения, назначая уникальный номер каждой категории;
- f) **SimpleImputer** — заполняет пропущенные значения средним, медианой или наиболее часто встречающимся значением;
- g) **train_test_split** — разделяет данные на обучающую и тестовую выборки.

В основе архитектуры Sklearn лежат три ключевых понятия: Estimator, Transformer и Predictor.

1. Estimator (оценщик). Это базовый класс для всех моделей машинного обучения в **sklearn**. Он инкапсулирует процесс обучения модели на данных. Все оценщики имеют два основных метода:
 - a) **fit(X, y)** — этот метод обучает модель на данных X (признаки) и y (целевая переменная). Параметры модели настраиваются на основе предоставленных данных;
 - b) **get_params()** — возвращает словарь с параметрами модели;
 - c) **set_params()** — позволяет установить параметры модели;
2. Transformer (преобразователь). Это тип Estimator, который преобразует входные данные X в новое представление. Трансформеры используются

для предварительной обработки данных, уменьшения размерности и извлечения признаков. Они имеют следующие методы:

- a) `fit(X, y)` — обучает трансформер на данных `X` и `y` (в некоторых случаях `y` может не использоваться) и вычисляет необходимые параметры для преобразования;
- b) `transform(X)` — преобразует данные `X` на основе параметров, вычисленных в `fit()`, возвращает преобразованные данные;
- c) `fit_transform(X, y)` — объединяет `fit()` и `transform()` в один шаг, этот метод эффективен, когда обучение и преобразование выполняются на одном и том же наборе данных;
- d) `inverse_transform(X_transformed)` — этот метод преобразует уже преобразованные данные обратно в исходное представление (не все трансформеры поддерживают этот метод).

3. **Predictor (Предсказатель).** Это тип `Estimator`, который может делать предсказания на новых данных после обучения. `Predictor` имеют следующие методы:

- a) `fit(X, y)` — обучает модель на данных `X` и `y`;
- b) `predict(X_new)` — делает предсказания для новых данных `X_new` и возвращает предсказанные значения или классы;
- c) `predict_proba(X)` — возвращает вероятности принадлежности к каждому классу для задач классификации (не все предикторы поддерживают этот метод);
- d) `score(X, y)` — оценивает производительность модели на данных `X` и `y` и возвращает оценку, зависящую от типа задачи (например, точность для классификации, детерминацию для регрессии).

В.2.2. Оценка и выбор моделей в `sklearn`

В машинном обучении задача выбора оптимальной модели для конкретной задачи является одной из ключевых. Невозможно заранее знать, какая модель покажет наилучшие результаты на новых, ранее не виденных данных. Поэтому необходимо тщательно оценивать разные модели и выбирать ту, которая обладает наилучшей обобщающей способностью. Библиотека `sklearn` предоставляет широкий набор инструментов и техник для оценки и выбора моделей машинного обучения. В этом разделе мы рассмотрим основные этапы и методы оценки, предотвращения переобучения и сравнения моделей в `sklearn`, уделив внимание разделению данных, кросс-валидации, регуляризации, выбору признаков и выбору наилучшей модели на основе метрик качества.

Прежде чем приступить к обучению и оценке модели, необходимо разделить имеющийся набор данных на две независимые части: обучающую (`training set`) и тестовую (`test set`). Обучающая выборка используется для тренировки модели, то есть для настройки её параметров на основе известных исторических данных. Тестовая выборка используется для оценки обобщающей способности модели, то есть того, насколько хорошо модель сможет предсказывать результаты на новых, ранее не виденных данных.

Наиболее простой и часто используемый способ разделения данных — функция `train_test_split` из модуля `model_selection`. Эта функция случайным образом разделяет данные на обучающую и тестовую выборки с заданным соотношением.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)
```

Важно отметить, что параметр `random_state` позволяет фиксировать случайное разделение данных, что обеспечивает воспроизводимость результатов при повторном запуске кода. Выбор оптимального соотношения между размерами обучающей и тестовой выборок зависит от размера исходного набора данных. В целом, рекомендуется выделять на тестовую выборку от 20% до 30% данных.

Однократное разделение данных на обучающую и тестовую выборки может привести к смещённой оценке обобщающей способности модели, особенно если размер исходного набора данных невелик. Чтобы получить более надёжную оценку, используется кросс-валидация. Кросс-валидация — это метод оценки, при котором данные разделяются на несколько (K) подмножеств (фолдов). Затем модель обучается K раз, каждый раз используя один из фолдов в качестве тестовой выборки, а остальные $K-1$ фолды в качестве обучающей выборки. В конце рассчитывается средняя оценка качества модели по всем K итерациям. Наиболее распространённым типом кросс-валидации является K -fold cross-validation. В этом методе данные случайным образом разделяются на K непересекающихся фолдов:

```
model = LogisticRegression(solver='liblinear')
kf = KFold(n_splits=5, shuffle=True, random_state=42)
scores = cross_val_score(model, X, y, cv=kf, scoring='accuracy')
```

После того как наилучшая модель выбрана, её рекомендуется обучить на всем наборе обучающих данных (X_{train} , y_{train}) и затем оценить её обобщающую способность на тестовой выборке (X_{test} , y_{test}). Это позволит получить окончательную оценку производительности модели на новых, ранее не виденных данных.

ПРИЛОЖЕНИЕ С. КЛЮЧЕВЫЕ АЛГОРИТМЫ МЕТОДИКИ МССР

Ключевыми алгоритмами в методике МССР (Modeling, Challenge, Calibration, Production) являются алгоритмы прогнозирования значений измеряемой характеристики какого-либо процесса, причём, поскольку категории нормы/аномальности рассматриваются с двух точек зрения (норма/аномалия по отношению ко времени и норма/аномалия по отношению к локациям), есть два семейства таких алгоритмов. К первому семейству относятся алгоритмы прогнозирования по времени, а ко второму — алгоритмы прогнозирования по локациям. В этом приложении мы подробно разберём и прокомментируем алгоритмы обоих семейств.

С.1. Моделирование по времени

Выполняя моделирование по времени, мы используем данные в формате ML (момент/ локация), когда наблюдениями служат моменты времени, а признаковое описание задаётся при помощи набора локаций. В этих алгоритмах проводится прогнозирование временных рядов, и для его осуществления мы применяем приём авторегрессии.

Очень важным является выбор числа локаций, по которым одновременно осуществляется прогноз. А именно:

1. Мы можем рассматривать только одну отдельно взятую локацию и прогнозировать значение измеряемой характеристики в этой локации, ориентируясь исключительно на знания о её предыдущих значениях в этой локации. Такой прогноз мы называем прогнозом на локальных данных.
2. Мы можем взять сразу все локации и прогнозировать значение измеряемой

характеристики в каждой отдельно взятой локации, ориентируясь на знание её предыдущих значений во всех локациях. Тогда мы говорим, что осуществляем прогноз по времени на глобальных данных.

3. И, наконец, возможен промежуточных вариант (точнее, целая серия промежуточных вариантов), когда мы используем знания о нескольких локациях, наиболее близких той, которую мы прогнозируем. Такой подход мы называем прогнозированием по времени на ближайших локациях, число ближайших локаций варьируется от одной (тогда получается алгоритм на локальных данных) до общего числа локаций (тогда получается алгоритм на глобальных данных).

Основу для всех трёх подходов составляет алгоритм на локальных данных, именно его листинг мы приведём полностью, а алгоритмы прогнозирования по времени на глобальных данных и на ближайших локациях мы будем рассматривать как его модификации, которые нетрудно осуществить самостоятельно (см. прил. С.1.2).

С.1.1. Листинг кода

Алгоритм прогнозирования по времени на локальных данных реализуется как функция `modeling_by_moment`, принимающая на вход несколько параметров:

- `path_in` — путь к входному файлу данных в формате ML;
- `history_end` — номер момента времени, в который заканчивается период исторических наблюдений;
- `challenge_end` — номер момента времени, в который заканчивается период челленджа;
- `autoregression_dive_max` — максимальная глубина авторегрессии;
- `model_for_fit` — модель, которая используется для обучения;
- `path_out` — путь к выходному файлу.

В теле функции выполняется следующий код.

```
# Чтение входного файла
df = pd.read_csv(path_in)

# Формирование исторического фрагмента данных
df_history = df[:history_end]
# Формирование фрагмента данных для челленджа
df_challenge = df[history_end:challenge_end]
# Формирование пустого датафрейма для записи прогнозов
df_challenge_pred = pd.DataFrame(index = df_challenge.index, columns = df_challenge.columns)

# Цикл по глубине авторегрессии
for dive in range(1, autoregression_dive_max):

    # Объявление исторических данных внутри цикла
    df_history = df[:history_end]

    # Индикатор выполнения программы
    print('autoregression_dive =', dive)

    # Цикл по моментам наблюдений
    for i in df_challenge.index:

        # Формирование пустого списка для записи строки
        # предсказаний
        y_pred_list = []

        # Цикл по локациям
        for loc in df.columns:

            # Формирование пустого датафрейма для записи
            # авторегрессионных столбцов
            d = pd.DataFrame()
            # Назначение целевой функции
            d['y'] = df_history[loc]

            # Цикл по глубине авторегрессии
```

```

for k in range (1, dive + 1):

    # Добавление очередного авторегрессионного
    # столбца
    d['x_'+ str(k)] =
    df_history[loc].shift(periods=k)

    # Удаление строк с пропусками
    d = d.dropna()

    # Данные для регрессионной задачи
    X = d.drop(columns = ['y'])
    y = d['y']

    # Построение модели
    model = model_for_fit()
    model.fit(X, y)

    # Данные, по которым осуществляется прогноз
    X_pred = [d.iloc[-1][: -1].values]

    # Вычисление прогнозируемого значения
    y_pred = model.predict(X_pred)[0]
    # Добавление прогноза в список
    y_pred_list.append(y_pred)

    # Занесение строки в датафрейм прогнозов
    df_challenge_pred.loc[i] = y_pred_list
    # Увеличение исторических данных на один момент времени
    df_history = df_history.append(df_challenge.loc[i])

    # Вывод датафрейма прогнозов во внешний файл
    df_challenge_pred.to_csv(path_out)

```

С.1.2. Комментарии к коду

Алгоритм на локальных данных. В формате ML моменты времени хранятся построчно. Суть алгоритма состоит в том, что в текущий момент времени мы

строим авторегрессионную задачу для каждой локации, используя для смещения данные по этой локации, и предсказываем значение измеряемой характеристики в этой локации на следующий момент времени. Когда мы проделываем эту процедуру со всеми локациями, у нас появляется строка с предсказаниями для всех локаций, и мы записываем предсказанные значения в специально подготовленный для этих целей датафрейм `df_challenge_pred`, то есть, пополняем предсказания одной строкой. После этого мы считаем, что текущий момент времени прошёл, и наступил следующий момент времени. Теперь мы знаем истинные значения предыдущего момента, следовательно, они теперь относятся к историческим сведениям, и мы отправляем строчку данных из `df_challenge` в `df_history` в качестве последней строки исторических наблюдений. По итогу работы алгоритма мы получаем полностью заполненный датафрейм предсказаний, который отправляем во внешний файл.

Алгоритм на глобальных данных. Отличие алгоритма состоит в том, как строятся столбцы датафрейма `d`. Там, где в алгоритме на локальных данных использовался цикл по глубине авторегрессии, на глобальных данных этот цикл пополняется ещё одним вложенным циклом:

```
# Цикл по глубине авторегрессии
for k in range (1, dive + 1):
    # Добавление очередного авторегрессионного блока столбцов
    for col in df_history.columns:
        # Добавление столбца в блок
        d['x_' + col + str(k)] = df_history[col].shift(periods=k)
```

В остальном этот алгоритм ничем не отличается от алгоритма на локальных данных. Однако следует иметь в виду, что это только формальное сходство. На самом деле, в отличие от локального варианта, алгоритм на глобальных данных приводит в регрессионным задачам, размерность которых очень быстро растёт вместе с увеличением глубины авторегрессии. Это приводит к

интересным и, подчас, неожиданным эффектам¹¹⁸, которые следует учитывать.

Алгоритм на ближайших локациях. Наконец, этот алгоритм занимает промежуточное положение между локальным и глобальным вариантами прогнозирования. Фактически, для каждой локации нам нужно подобрать некоторое количество локаций, ближайших к ней в смысле евклидова расстояния между их портретами, а это означает, что прогнозирование по времени должно учитывать ещё один входной параметр:

`k_nearest` — число ближайших локаций

Код алгоритма снабжается фрагментом поиска ближайших локаций:

```
# Пустой список для записи расстояний
distance_list = []

# Цикл по всем столбцам
for col in df_history.columns:
    # Расстояние между текущей локацией и другой локацией
    dist = np.norm(df_history[loc] - df_history[col])
    # Пополнение списка расстояний
    distance_list.append(dist)

# Буферный датафрейм с парами (имя локации/ расстояние до нее)
bufer = pd.DataFrame()
bufer['location'] = df_history.columns
bufer['distance'] = distance_list

# Буферный датафрейм, упорядоченный по убыванию расстояния
bufer_sorted = bufer.sort_values('distance', ascending = False)
# Упорядоченный список ближайших локаций, начиная с самой близкой
nearest_location = list(bufer_sorted['location'])[:k_nearest]
```

¹¹⁸ Например, в случаях, когда размерность задачи приближается к числу наблюдений, прогнозирующая способность модели `LinearRegression` катастрофически падает (практически, до минус бесконечности), в тот момент как модели других семейств продолжают давать адекватные прогнозы.

После того, как список ближайших локаций `nearest_location` получен, мы можем просто применить алгоритм прогнозирования по времени на полных данных, считая, что полные данные ограничены этим списком:

```
# Цикл по глубине авторегрессии
for k in range(1, dive + 1):
    # Добавление очередного авторегрессионного блока столбцов
    for col in nearest_location:
        # Добавление столбца в блок
        d['x_' + col + str(k)] = df_history[col].shift(periods=k)
```

С.2. Моделирование по локациям

Выполняя моделирование по локациям, мы используем данные в формате LM (локация/ момент), когда наблюдениями служат локации, а признаковое описание задаётся при помощи набора моментов времени. В этих алгоритмах проводится прогнозирование на неупорядоченном множестве локаций (в чем состоит его принципиальное отличие от алгоритма прогнозирования по времени), но моменты времени, являясь предикторами модели, играют большую и важную роль как сами по себе, так и в смысле учитываемого числа прошлых наблюдений. В отличие от словосочетания «глубина авторегрессии», которое мы использовали выше, при описании алгоритма моделирования по времени, теперь мы будем использовать выражение «глубина истории наблюдений», чтобы подчеркнуть, что теперь модели ни в коем случае не являются авторегрессионными.

Важным является выбор объёма исторических данных, ориентируясь на которую мы будем строить прогноз. А именно:

1. Мы можем рассматривать один отдельно взятый момент времени и прогнозировать значение измеряемой характеристики процесса для каждой локации, ориентируясь исключительно на знания о её значениях в других локациях в этот и только этот момент времени. Такой прогноз мы называем прогнозом на локальных данных.

2. Мы можем взять сразу все историю наблюдений и прогнозировать значение измеряемой характеристики в каждой локации, ориентируясь на значение всех её предыдущих значений во всех локациях. Тогда мы говорим, что осуществляем прогноз по локациям на глобальных данных.
3. И, наконец, возможен промежуточных вариант (точнее, целая серия промежуточных вариантов), когда мы используем знания о нескольких моментах, наиболее близких тому моменту, который является текущим на данной итерации алгоритма. Такой подход к прогнозированию мы называем прогнозированием по времени на ближайших моментах, число ближайших моментов варьируется от одного (тогда получается алгоритм на локальных данных) до общего числа наблюдений (тогда получается алгоритм на глобальных данных).

Основу для всех трёх подходов составляет алгоритм на глобальных данных, именно его листинг мы приведем полностью, а алгоритмы прогнозирования по времени на локальных данных и на ближайших моментах времени мы будем рассматривать как его модификации, которые нетрудно осуществить самостоятельно (см. прил. С.2.2).

С.2.1. Листинг кода

Алгоритм прогнозирования по локациям на глобальных данных реализуется как функция `modeling_by_location`, принимающая на вход несколько параметров:

`path_in` — путь к входному файлу данных в формате LM;
`history_end` — номер момента времени, в который заканчивается период исторических наблюдений;
`challenge_end` — номер момента времени, в который заканчивается период челленджа;
`history_dive_max` — максимальная глубина истории наблюдений;
`model_for_fit` — модель, которая используется для обучения;

`path_out` — путь к выходному файлу.

В теле функции выполняется следующий код.

```
# Чтение входного файла
df = pd.read_csv(path_in)

# Формирование исторического фрагмента данных
df_history = df[df.columns[:history_end]]
# Формирование фрагмента данных для челленджа
df_challenge = df[df.columns[history_end:challenge_end]]
# Формирование пустого датафрейма для записи прогнозов
df_challenge_pred = pd.DataFrame(index = df_challenge.index, columns = df_challenge.columns)

# Цикл по глубине истории наблюдений
for k in range(1, history_dive_max):

    # Объявление исторических данных внутри цикла
    df_history = df[df.columns[:history_end]]

    # Индикатор выполнения программы
    print('history_dive =', k)

    # Цикл по моментам наблюдений
    for col in df_challenge.columns:

        # Формирование пустого списка для записи столбца
        # предсказаний
        y_pred_list = []

        # Цикл по локациям
        for i in range(len(df_challenge)):
            # Формирование буфера с последними k столбцами
            X_bufer = df_history[df_history.columns[-k:]]

            # Данные для регрессионной задачи
            X = X_bufer.drop(index = [i])
            y = df_challenge[col].drop(index = [i])
```

```

# Построение модели
model = model_for_fit()
model.fit(X, y)

# Данные, по которым осуществляется прогноз
X_pred = X_bufer.loc[i].values.reshape(1, -1)

# Вычисление прогнозируемого значения
y_pred = model.predict(X_pred)[0]
# Добавление прогноза в список
y_pred_list.append(y_pred)

# Занесение столбца в датафрейм прогнозов
df_challenge_pred[col] = y_pred_list
# Увеличение исторических данных на один момент времени
df_history[col] = df_challenge[col]

# Вывод датафрейма прогнозов во внешний файл
df_challenge_pred.to_csv(path_out)

```

С.2.2. Комментарии к коду

Алгоритм на глобальных данных. В формате LM моменты времени хранятся по столбцам. Суть алгоритма состоит в том, что в текущий момент времени мы строим регрессионную задачу для каждой локации, используя в качестве целевой функции значения измеряемой характеристики в данный момент времени во всех локациях, кроме прогнозируемой, а в качестве предикторов — значения измеряемой характеристики в этих локациях в предыдущие моменты времени вплоть до нужной нам глубины, в результате чего получаем прогнозируемое значение на следующий момент времени. Когда мы проделываем эту процедуру со всеми локациями, у нас появляется столбец с прогнозами для всех локаций, и мы записываем предсказанные значения в специально заготовленный для этих целей датафрейм `df_challenge_pred`, то есть, пополняем предсказания одним столбцом. После этого мы считаем, что текущий момент времени прошёл, и наступил следующий момент времени. Теперь

мы знаем истинные значения предыдущего момента, следовательно, они теперь относятся к историческим сведениям, и мы отправляем столбец данных из `df_challenge` в `df_history` в качестве последнего столбца исторических наблюдений. По итогу работы алгоритма мы получаем полностью заполненный датафрейм предсказаний, который отправляем во внешний файл.

Алгоритм на локальных данных. Локальные данные означают наблюдения во всех локациях в один единственный момент времени: в текущий момент. Это значит, что, при использовании локального формата данных, у нас вообще не оказывается истории наблюдений, а значит — и предикторов регрессионной задачи. Единственное, что можно седлать в таких условиях — это предсказывать значение в локации как среднее значение по всем остальным локациям, что, безусловно, нельзя назвать хорошей моделью. В коде это означает, что вместо построения модели, ее обучения, формирования данных для прогноза и собственно прогноза мы используем одну строчку кода:

```
# Вычисление прогнозируемого значения  
y_pred = y.mean()
```

Отметим ещё раз, что прогнозирование по среднему значению не обеспечивает должной точности прогнозов, и, следовательно, использование прогнозирования по локациям на локальных данных нецелесообразно.

Алгоритм на ближайших моментах. Наконец, этот алгоритм занимает промежуточное положение между локальным и глобальным вариантами прогнозирования по локациям. Фактически, для каждого момента нам нужно подобрать некоторое количество моментов, ближайших к нему в смысле евклидова расстояния между портретами моментов¹¹⁹, и это значит, что функция прогнозирования по времени должна учитывать ещё один входной параметр:

¹¹⁹ Определения того, что мы называем портретами моментов и локаций, мы привели в главе 2, и там же рассмотрели ряд примеров.

`k_nearest` — число ближайших локаций

Код алгоритма снабжается фрагментом поиска ближайших локаций:

```
# Пустой список для записи расстояний
distance_list = []

# Цикл по всем столбцам
for col in df_history.columns:
    # Расстояние между текущей локацией и другой локацией
    dist = np.norm(df_history[loc] - df_history[col])
    # Пополнение списка расстояний
    distance_list.append(dist)

# Буферный датафрейм с парами (имя локации/ расстояние до нее)
bufer = pd.DataFrame()
bufer['moment'] = df_history.columns
bufer['distance'] = distance_list

# Буферный датафрейм, упорядоченный по убыванию расстояния
bufer_sorted = bufer.sort_values('distance', ascending = False)
# Упорядоченный список ближайших локаций, начиная с самой близкой
nearest_moment = list(bufer_sorted['location'])[:k_nearest]
```

После того, как список ближайших локаций `nearest_location` получен, мы можем просто применить алгоритм прогнозирования по локациям на полных данных, считая, что полные данные ограничены этим списком:

```
# Объявление исторических данных внутри цикла
df_history = df[nearest_moment]
```

С.3 Калибровка по экспертному мнению

Этот алгоритм используется, если у нас есть экспертная оценка (если экспертное мнение отсутствует, применяется калибровка по локтевой точке). Для его применения необходимо, чтобы процедура моделирования была уже проведена и у нас в распоряжении уже была выбранная на этапе моделирования

наилучшая схема прогнозирования.

С.3.1. Листинг кода

Алгоритм калибровки по экспертному мнению реализуется как функция `calibration`, принимающая на вход несколько параметров:

`df_expert_true` — датафрейм в формате ML/LM, индексирующий наличие или отсутствие аномалии для каждой пары момент-локация на этапе челленджа с точки зрения экспертного сообщества;

`df_challenge` — датафрейм в формате ML/LM с истинными значениями наблюдаемой характеристики процесса на этапе челленджа;

`df_challenge_pred` — датафрейм в формате ML/LM со значениями, которые были спрогнозированы на этапе челленджа.

В теле функции выполняется следующий код.

```
# Вычисление дельт
delta = df_challenge - df_challenge_pred
# Преобразование дельт в список
delta_list = []
for moment in delta.columns:
    for location in delta.index:
        delta_list.append(delta[moment][location])

# Вычисление сигмы дельт
sigma = np.array(delta_list).std()

# Преобразование экспертных оценок в список
expert_true = []
for moment in df_expert_true.columns:
    for location in df_expert_true.index:
        expert_true.append(df_expert_true[moment][location])

# Списки для метрик
ROC_AUC_list = []
accuracy_list = []
```



```

f1_list = []

# Стартовые значения наилучших характеристик
best_delta_cr = None
best_ROC_AUC = 0
best_accuracy = 0
best_F1 = 0

# Сетка критических порогов от 1 сигмы до 3 сигм
delta_cr_list = np.arange(sigma, 3*sigma, 0.05)
# Цикл по сетке
for delta_cr in delta_cr_list:

    # Список оценок детектора
    dtector = []

    # Цикл по списку отклонений
    for d in delta_list:

        # Условие аномальности
        if abs(d) > delta_cr:
            dtector.append(1)
        else:
            dtector.append(0)

    # Вычисление метрик в сравнении с экспертным мнением
    ROC_AUC = roc_auc_score(expert_true, dtector)
    accuracy = accuracy_score(expert_true, dtector)
    F1 = f1_score(expert_true, dtector)

    # Пополнение списков метрик. Это для картинки
    ROC_AUC_list.append(ROC_AUC)
    accuracy_list.append(accuracy)
    f1_list.append(F1)

# Лучшее значение ROC AUC и лучший порог
if ROC_AUC > best_ROC_AUC:
    best_ROC_AUC = ROC_AUC
    best_delta_cr = delta_cr

```

```

# Лучшее значение accuracy
if accuracy > best_accuracy:
    best_accuracy = accuracy

# Лучшее значение F1
if F1 > best_F1:
    best_F1 = F1

# Вывод лучших метрик и лучшего критического порога
print('best_ROC_AUC =', best_ROC_AUC)
print('best_accuracy =', best_accuracy)
print('best_F1 =', best_F1)
print('')
print('best_delta_cr =', best_delta_cr)

```

С.3.2. Комментарии к коду

Формат данных. Все три входящих датафрейма должны иметь один и тот же формат. Если речь идёт о калибровке по времени, то это формат ML, если о калибровке по локациям — то формат LM.

С.4. Пилотная эксплуатация детектора аномалий по времени

Алгоритм пилотной эксплуатации намного проще алгоритма моделирования, потому что нам уже не нужно выбирать параметры моделирования: модель и глубину авторегрессии — они нам уже известны по итогам этапа челленджа, а по итогам этапа калибровки нам известен ещё один в высшей степени важный показатель: критический порог отклонения реальности от прогноза, превышение которого алгоритм должен признать аномальным.

Поэтому в нашем алгоритме, во-первых, не будет цикла по глубине авторегрессии, а во-вторых, он сможет выводить заключение о норме или аномалии сразу же, как только увидит очередное реально зафиксированное значение наблюдаемой характеристики процесса.

С.4.1. Листинг кода

Алгоритм эксплуатации детектора аномалий по времени реализуется как функция `production_by_moments`, принимающая на вход несколько параметров:

`path_in` — путь к входному файлу;
`end_of_history_end` — момент времени, в который начинается этап эксплуатации детектора;
`delta_cr` — критический порог отклонения;
`best_model` — наилучшая модель;
`best_autoregression_dive` — оптимальная глубина авторегрессии.

В теле функции выполняется следующий код.

```
# Чтение данных
df = pd.read_csv('путь к входному файлу')
# Формирование датафрейма исторических данных
df_history = df[:history_end]
# Формирование датафрейма текущих данных
df_production = df[history_end:]
# Указание критической дельты
delta_cr = 1.228955

# Цикл по моментам времени
for i in df_production.index:

    # Список для записи прогнозов
    y_pred_list = []

    # Цикл по локациям
    for loc in df.columns:

        # Датафрейм регрессионной задачи
        d = pd.DataFrame()
        d['y'] = df_history[loc]

        # Добавление авторегрессионных предикторов
        for k in range(1, best_autoregression_dive + 1):
```

```

        d['x_' + str(k)] = df_history[loc].shift(periods = k)

d = d.dropna()

# Обучение модели
X = d.drop(columns = ['y'])
y = d['y']

model = best_model()
model.fit(X, y)

# Данные для вычисления прогноза
X_pred = [d.iloc[-1][: -1].values]

# Прогноз
y_pred = model.predict(X_pred)[0]

# Проверка условия аномальности и вывод
if abs(y_pred - df_production[loc][i]) > delta_cr:
    print('moment    =', i)
    print('location =', loc)
    print('anomaly')
    print('')
# Пополнение списка прогнозов
y_pred_list.append(y_pred)

# Перевод прошедшего момента времени в историю
df_history = df_history.append(df_production.loc[i])

```

С.4.2. Комментарии к коду

Модель. Поскольку лучшая модель нам известна после этапа моделирования, мы её напрямую указываем в теле функции.

Глубина авторегрессии. Оптимальную глубину мы тоже подобрали на этапе моделирования. Поэтому при добавлении авторегрессионных предикторов мы используем `range(1, best_autoregression_dive + 1)`.

С.5. Пилотная эксплуатация детектора аномалий по локациям

Этот алгоритм является упрощённой версией алгоритма моделирования по локациям, в нем не производится поиск лучшей модели (она известна после этапа моделирования) и не подбирается оптимальная глубина истории наблюдений (она тоже известна после этапа моделирования).

С.5.1. Листинг кода

Алгоритм эксплуатации детектора аномалий по времени реализуется как функция `production_by_locations`, принимающая на вход несколько параметров:

`path_in` — путь к входному файлу;
`end_of_history_end` — момент времени, в который начинается этап эксплуатации детектора;
`delta_cr` — критический порог отклонения;
`best_model` — наилучшая модель;
`best_history_dive` — оптимальная глубина истории наблюдений.

В теле функции выполняется следующий код.

```
# Чтение данных
df = pd.read_csv('Путь к входному файлу')

# Формирование исторических данных
df_history = df[df.columns[:75]]
# Формирование текущих данных
df_production = df[df.columns[75:]]
# Указание критического порога
delta_cr = 1.310583

# Цикл по моментам времени
for col in df_production.columns:

    # Список для прогнозов
```

```

y_pred_list = []

# Цикл по локациям
for i in range(len(df_production)):

    # Регрессионная задача
    d = df_history
    X = d[d.columns[-best_history_dive:]].drop(index = [i])
    y = df_production[col].drop(index = [i])

    model = best_model()
    model.fit(X, y)

    # Данные для прогноза
    d = df_history[df_history.columns[-best_history_dive:]]
    X_pred = d.loc[i].values.reshape(1, -1)

    # Прогноз
    y_pred = model.predict(X_pred)[0]

    # Проверка условия аномальности и вывод
    if abs(y_pred - df_production[col][i] > delta_cr):
        print('moment   =', col)
        print('location =', i)
        print('')

    # Пополнение списка прогнозов
    y_pred_list.append(y_pred)

    # Перевод прошедшего момента в историю
    df_history[col] = df_production[col]

```

С.5.2. Комментарии к коду

Модель. Поскольку лучшая модель нам известна после этапа моделирования, мы её напрямую указываем в теле функции.

Глубина истории наблюдений. Оптимальную глубину мы тоже подо-

брали на этапе моделирования, и при формировании регрессионной задачи используем срез `df_history.columns[-best_history_divide:]`.

ПРИЛОЖЕНИЕ D. КАЛЕНДАРИ ВЫЯВЛЕННЫХ НЕСООТВЕТСТВИЙ В НЕСТАБИЛЬНЫХ ПРОЦЕССАХ ОДНОГО ИЗ ВЕДУЩИХ ОТЕЧЕСТВЕННЫХ АВТОПРОИЗВОДИТЕЛЕЙ

В этом приложении представлены итоговые отчёты о работе детектора, в которых каждый момент наблюдения содержит описание и визуализацию зафиксированных аномалий в протекании соответствующего процесса. Такой формат отчёта мы называем календарём выявленных несоответствий. Календарь обновляется в каждый момент времени, демонстрирует локации с аномальными значениями наблюдаемой характеристики и предоставляет возможность лицу, ответственному за качество процесса, реагировать на его аномальные состояния в порядке рейтинга по выраженности аномального состояния.

Выраженность аномалии — величина, обладающая той же размерностью, что и наблюдаемая характеристика (например, число обработанных заявок сервисного центра измеряется номинально, а суммарная стоимость работ по гарантийному обслуживанию в течение рабочего дня — сумма, измеряемая в рублях). Вместе с тем, для того, чтобы можно было ориентироваться в календарях разных процессов и сравнивать их между собой, мы используем в качестве универсальной единицы масштаба психологически значимую величину 3σ , где в качестве σ используется среднее квадратичное отклонение прогнозируемых значений процесса от его реальных значений, вычисленное по результатам всех наблюдений. Это позволяет сразу же судить о высокой (или, наоборот, низкой) критичности аномалии, хотя и не влияет на положение аномальной локации в рейтинге аномалий.

D.1. Производство

Анализировались данные о производительности труда когорты работников,

принятых в декабре 2022 года, и выполняющих операцию «Установка уплотнителей». В качестве наблюдаемой характеристики использовалось количество установок, выполненных тем или иным работником в течение смены. Данные аккумулировались ежедневно по итогам рабочего дня, для анализа использовалось 515 наблюдений. Первые 100 наблюдений были отнесены к периоду обучения персонала, последние 415 наблюдений использовались для пилотной эксплуатации детектора аномалий. Ниже приведены последние 10 моментов аномальных значений и их рейтинги (в случае, если аномальные значения демонстрируют сразу несколько локаций).

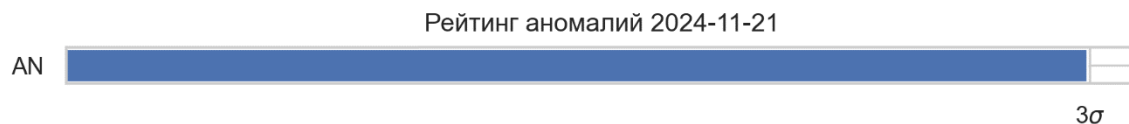
D.1.1. Календарь аномалий по времени

moment = 2024-11-21

location = AN

power = 55.088454792030745

sigma_power = 2.9926491622089495



moment = 2024-11-25

location = DL

power = 63.64406107351158

sigma_power = 3.4591240823601126

location = AM

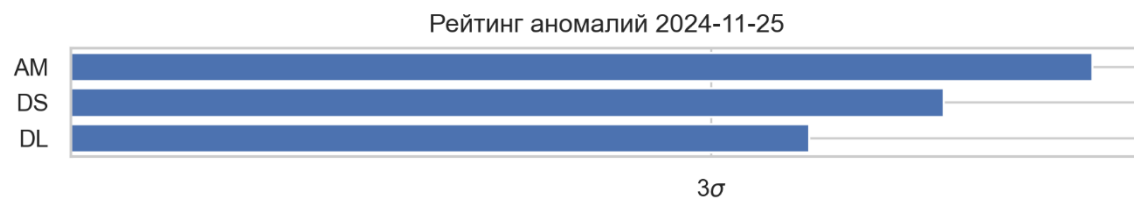
power = 88.06747265375901

sigma_power = 4.7865631198069

location = DS

power = 75.27824057264715

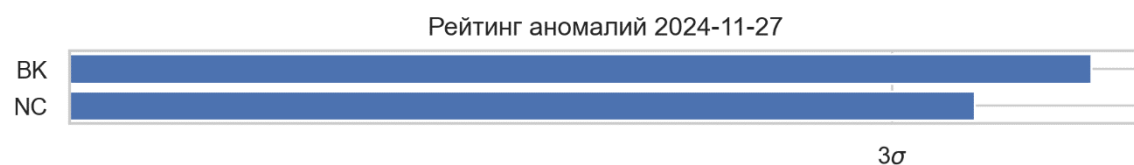
sigma_power = 4.091454417746419



moment = 2024-11-27

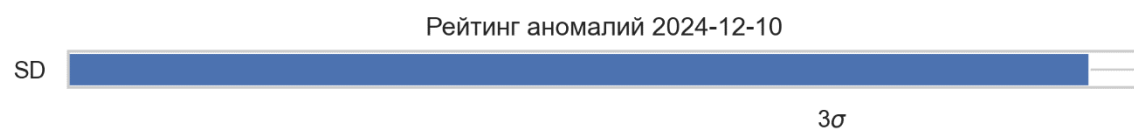
location = NC
 power = 60.86957119108206
 sigma_power = 3.3010258671932164

location = BK
 power = 68.73202582463563
 sigma_power = 3.727415697401122



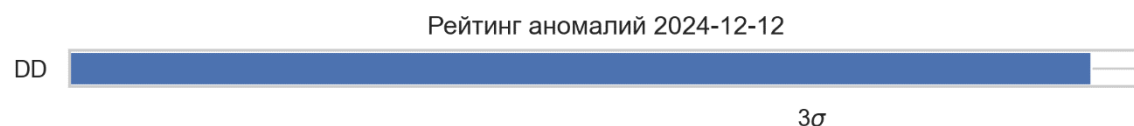
moment = 2024-12-10

location = SD
 power = 73.84487093791188
 sigma_power = 4.009007617796917



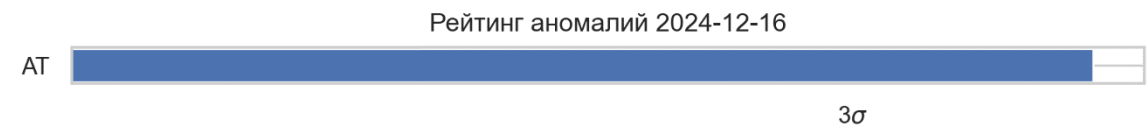
moment = 2024-12-12

location = DD
 power = 76.08925938677945
 sigma_power = 4.128707654701285



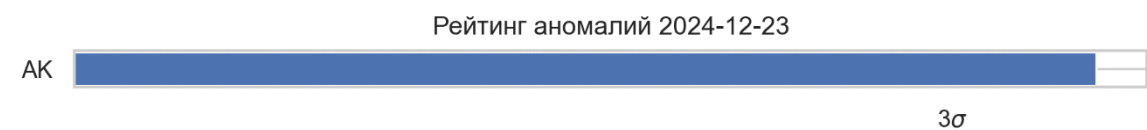
moment = 2024-12-16

location = AT
power = 72.27864309286116
sigma_power = 3.923867349575717



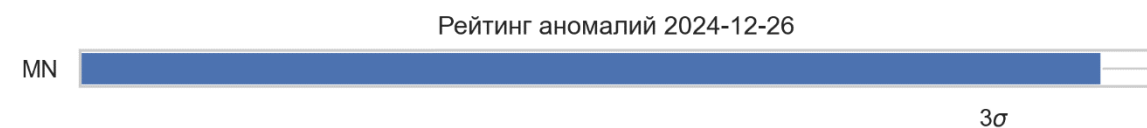
moment = 2024-12-23

location = AK
power = 64.22536455779507
sigma_power = 3.4914088603614575



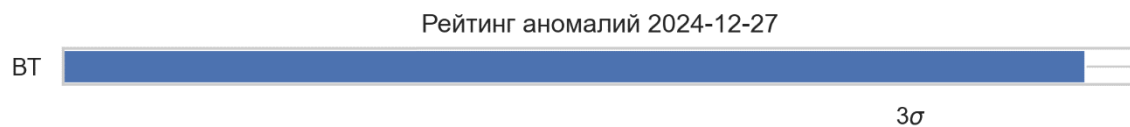
moment = 2024-12-26

location = MN
power = 61.736952777439654
sigma_power = 3.353379590090334



moment = 2024-12-27

location = BT
power = 66.60502508939624
sigma_power = 3.617249755066227



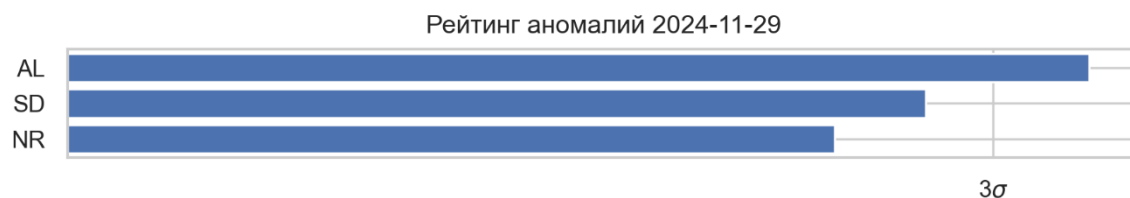
D.1.2. Календарь аномалий по локациям

moment = 2024-11-29

location = NR
power = 48.800000000000001
sigma_power = 2.485247948640372

location = AL
power = 65.0
sigma_power = 3.3102687840496743

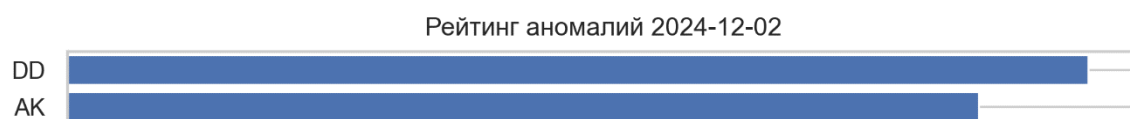
location = SD
power = 54.599999999999994
sigma_power = 2.7806257786017263



moment = 2024-12-02

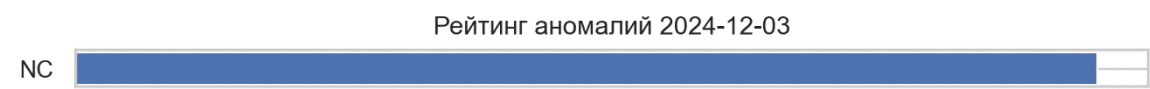
location = DD
power = 54.0
sigma_power = 2.749063740940323

location = AK
power = 48.199999999999999
sigma_power = 2.453793931728214



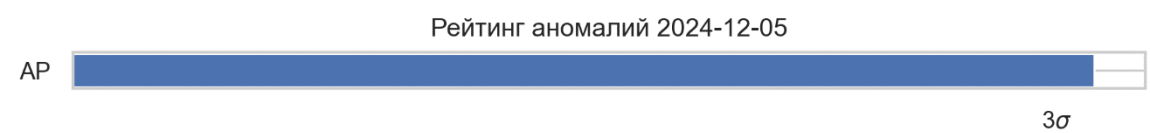
moment = 2024-12-03

location = NC
power = 44.599999999999994
sigma_power = 2.2698101579019534



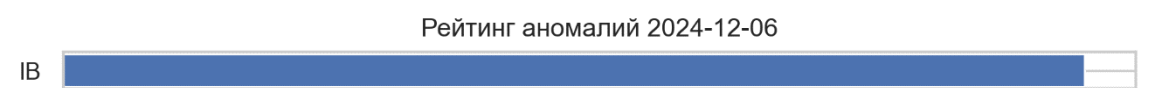
moment = 2024-12-05

location = AP
power = 61.199999999999999
sigma_power = 3.1142766430488034



moment = 2024-12-06

location = IB
power = 54.0
sigma_power = 2.748147652478167



moment = 2024-12-12

location = NR
power = 50.800000000000001
sigma_power = 2.592677394676079

location = NC
power = 50.0
sigma_power = 2.551847829405589

Рейтинг аномалий 2024-12-12



moment = 2024-12-13

location = AT
power = 54.0
sigma_power = 2.7561004110956544

location = VV
power = 45.800000000000001
sigma_power = 2.3375814597811297

Рейтинг аномалий 2024-12-13



moment = 2024-12-18

location = AT
power = 57.400000000000006
sigma_power = 2.9332582223647625

Рейтинг аномалий 2024-12-18



3σ

moment = 2024-12-20

location = SD
power = 56.599999999999994
sigma_power = 2.895377972124201

Рейтинг аномалий 2024-12-20



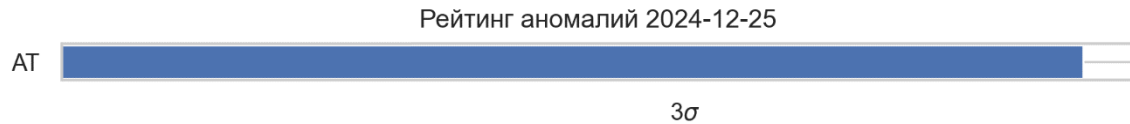
3σ

moment = 2024-12-25

location = AT

power = 96.0

sigma_power = 4.9181936906418295



D.2. Сервис

Анализировались данные о работе сервисных центров, производящих гарантийное обслуживание автомобилей известного отечественного бренда. В качестве наблюдаемых характеристик использовались:

1. число заявок, обработанных сервисом в течение рабочего дня,
2. общая сумма затрат на проведение мероприятий по гарантийному обслуживанию, накопленная сервисом в течение рабочего дня.

Данные аккумулировались ежедневно, для анализа использовалось 716 наблюдений с 2023-01-14 по 2024-12-30. Последние 415 наблюдений использовались для пилотной эксплуатации детектора аномалий. Ниже приведены последние 10 моментов аномальных значений и их рейтинги (в случае, если аномальные значения демонстрируют сразу несколько локаций).

D.2.1. Календарь аномалий по времени для числа обработанных заявок

moment = 2024-12-21

location = 14260

power = 36.094681391105574

sigma_power = 3.603991532036814

Рейтинг аномалий 2024-12-21



moment = 2024-12-22

location = 35060

power = 27.030940427832427

sigma_power = 2.7003749846537604

Рейтинг аномалий 2024-12-22



moment = 2024-12-23

location = 39263

power = 22.894871000051932

sigma_power = 2.2881658108701295

location = 12160

power = 28.632462163805798

sigma_power = 2.8615938043112092

Рейтинг аномалий 2024-12-23



moment = 2024-12-24

location = 45460

power = 29.21704482139978

sigma_power = 2.9213917787340096

location = 12965

power = 28.659993174127557

sigma_power = 2.8656925760042604

location = 62060

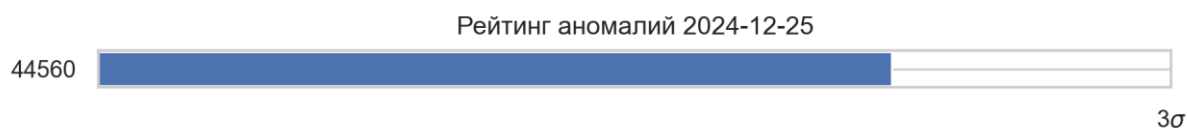
power = 30.804229013985246

sigma_power = 3.0800932107200256



moment = 2024-12-25

location = 44560
power = 22.177496905085455
sigma_power = 2.2178536337353862



moment = 2024-12-26

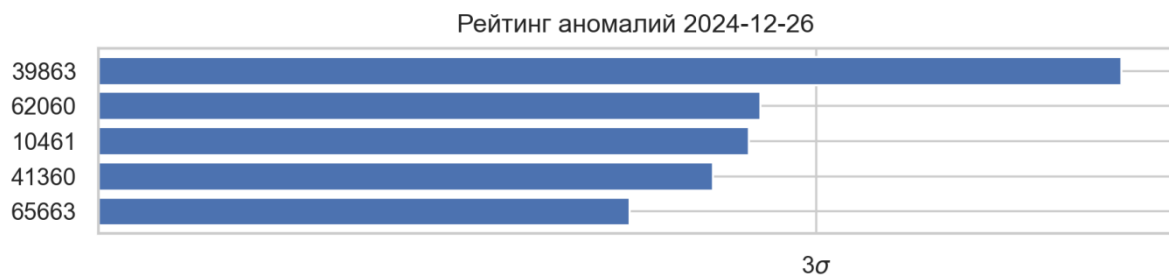
location = 41360
power = 25.668411424551184
sigma_power = 2.5683646475129414

location = 62060
power = 27.643210192007853
sigma_power = 2.7659617350926036

location = 10461
power = 27.171355284523777
sigma_power = 2.7187482381958517

location = 39863
power = 42.73881918532152
sigma_power = 4.276418608712049

location = 65663
power = 22.191131749750056
sigma_power = 2.220430291990951



moment = 2024-12-27

location = 44560

power = 22.72116012393191

sigma_power = 2.2726876170547845

location = 45070

power = 29.879179850974168

sigma_power = 2.988669666719053



moment = 2024-12-28

location = 11785

power = 23.48080742477505

sigma_power = 2.349520866883537

location = 44560

power = 31.5195563986465

sigma_power = 3.1538887966601483

location = 45460

power = 32.83667126497953

sigma_power = 3.2856810645557917

location = 60360

power = 26.894418623984308

sigma_power = 2.691091350337486

location = 12965

power = 25.72814805326624

sigma_power = 2.5743927635826256

location = 62060

power = 23.142295139430665

sigma_power = 2.3156488767204606

location = 35060

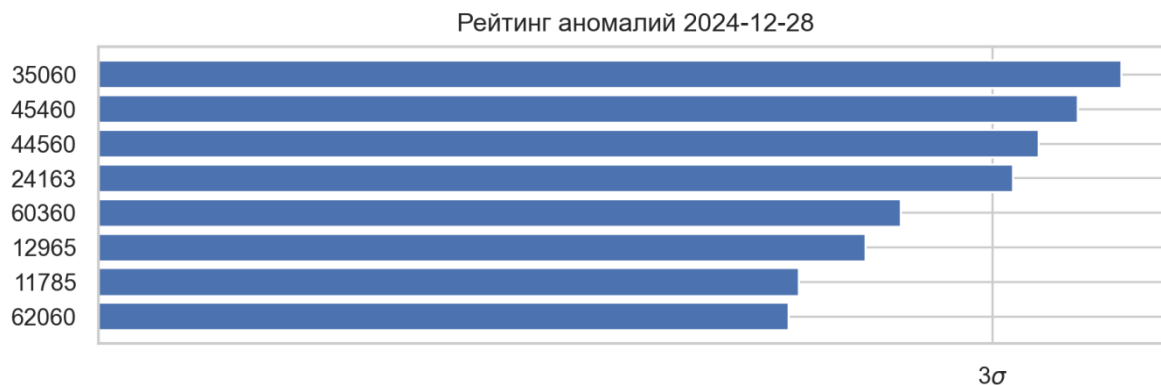
power = 34.294867403462895

sigma_power = 3.431590112460199

location = 24163

power = 30.66999635162402

sigma_power = 3.0688806867582805



moment = 2024-12-29

location = 44360

power = 30.75409204659141

sigma_power = 3.0768199373488887

location = 44560

power = 30.43682072875957

sigma_power = 3.0450782518920376

location = 60360

power = 28.171207217515136

sigma_power = 2.8184129739458297

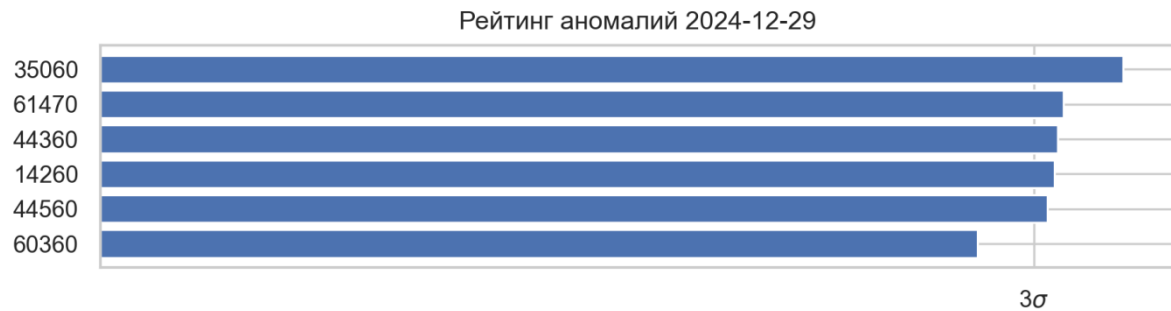
location = 14260

power = 30.64840626532636

sigma_power = 3.06624651126959

location = 35060
power = 32.854061321932434
sigma_power = 3.286913193374813

location = 61470
power = 30.935724872129065
sigma_power = 3.094991551651292



moment = 2024-12-30

location = 44360
power = 27.26089468759976
sigma_power = 2.7269706492890395

location = 44560
power = 41.816257254937554
sigma_power = 4.1829773932256336

location = 45460
power = 31.57429376523013
sigma_power = 3.1584499832640494

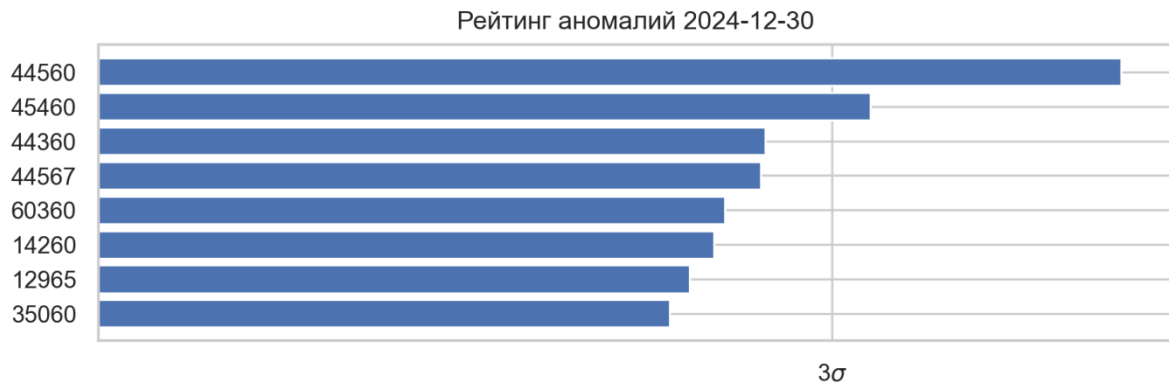
location = 60360
power = 25.609395883459168
sigma_power = 2.561767385865851

location = 12965
power = 24.16193635745779
sigma_power = 2.416974645617489

location = 14260
power = 25.175973103046132
sigma_power = 2.5184110978766228

location = 44567
power = 27.076115093157096
sigma_power = 2.708486717033438

location = 35060
power = 23.360121195049246
sigma_power = 2.3367672115218787



D.2.2. Календарь аномалий по локациям для числа обработанных заявок

moment = 2024-12-21

location = 44360
power = 32.84706041222232
sigma_power = 2.484424911609123

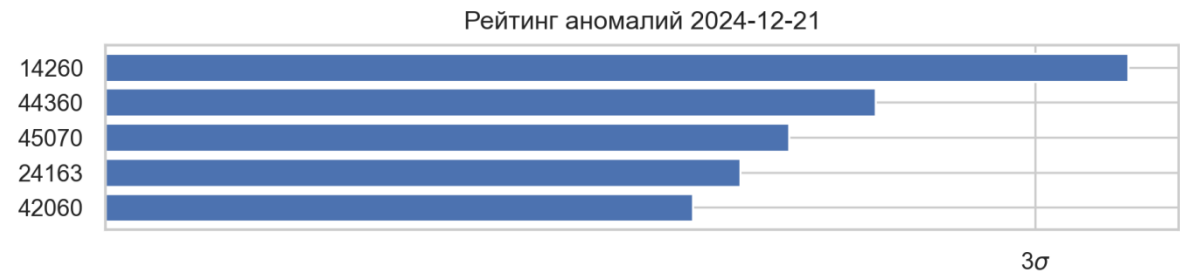
location = 42060
power = 25.042734050748862
sigma_power = 1.8941357780476387

location = 45070
power = 29.14706465799791
sigma_power = 2.204571508921598

location = 14260
power = 43.622044443609745
sigma_power = 3.299403129258349

location = 24163
power = 27.071161561254044

sigma_power = 2.0475582084036037



moment = 2024-12-22

location = 44360

power = 28.267799324387795

sigma_power = 2.13868727310472

location = 30563

power = 34.81668650459402

sigma_power = 2.634163468636579

location = 39263

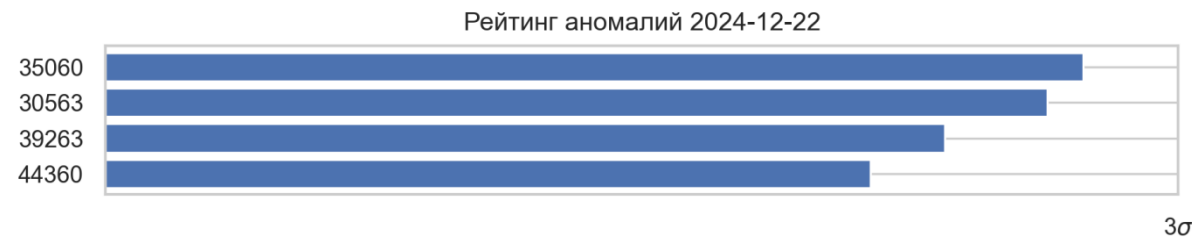
power = 31.017615828772197

sigma_power = 2.3467330956256993

location = 35060

power = 36.14207785972667

sigma_power = 2.7344400268000473



moment = 2024-12-23

location = 44560

power = 28.431026029310747

sigma_power = 2.1515633807984345

location = 39263

power = 25.315914700471858

sigma_power = 1.915822347206106

location = 10461
power = 26.020856650298235
sigma_power = 1.9691699570767953

location = 35060
power = 27.372226900324648
sigma_power = 2.071437062768293



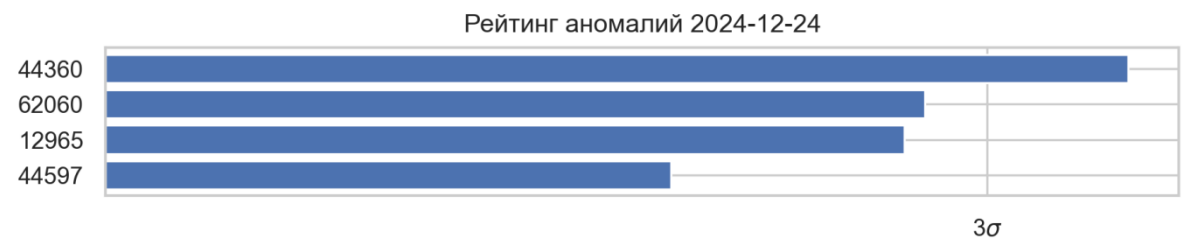
moment = 2024-12-24

location = 44360
power = 45.95616976636862
sigma_power = 3.4796498059320884

location = 12965
power = 35.90828631095399
sigma_power = 2.718857166044848

location = 62060
power = 36.830252365930605
sigma_power = 2.7886653989890804

location = 44597
power = 25.41136564959854
sigma_power = 1.9240649079463505



moment = 2024-12-25

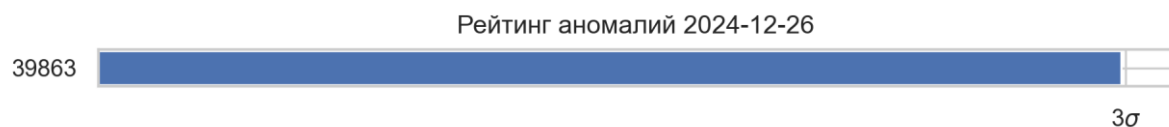
location = 44360
power = 49.02390712806524
sigma_power = 3.7119537197329264

moment = 2024-12-25
location = 14160
power = 27.145043517195276
sigma_power = 2.055347098156551



moment = 2024-12-26

location = 39863
power = 39.419895527151716
sigma_power = 2.985716267038054



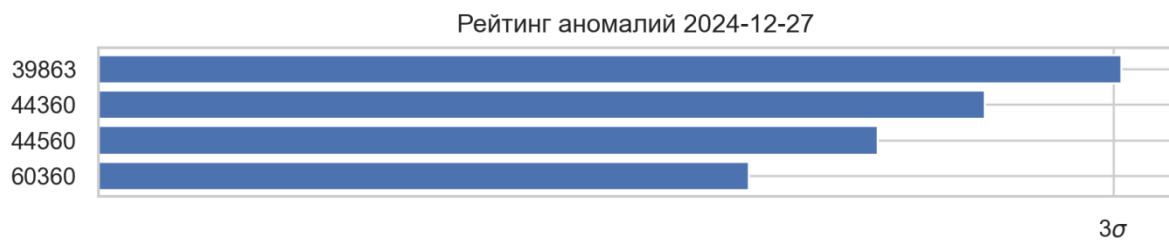
moment = 2024-12-27

location = 44360
power = 34.57927354260087
sigma_power = 2.619771239623553

location = 44560
power = 30.40008705384902
sigma_power = 2.3031505750868293

location = 60360
power = 25.363239409509752
sigma_power = 1.9215523734719444

location = 39863
power = 39.903374682786136
sigma_power = 3.0231321438577266



moment = 2024-12-28

location = 44360

power = 59.30571959313401

sigma_power = 4.494573292181129

location = 44560

power = 37.89347846006835

sigma_power = 2.8718143444326736

location = 12965

power = 37.78511195723098

sigma_power = 2.8636016257815657

location = 35060

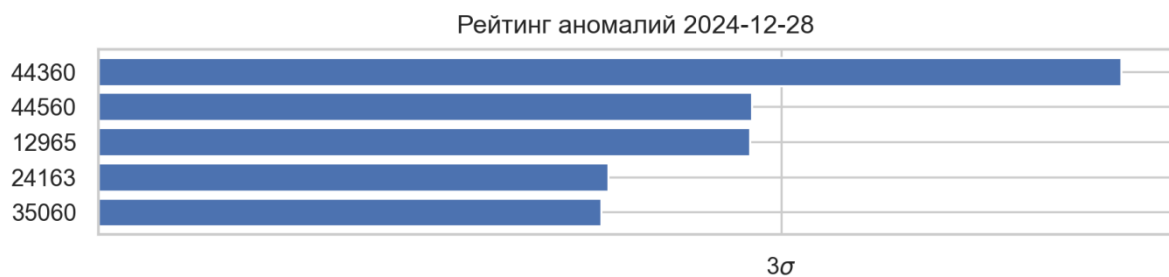
power = 29.136828393922734

sigma_power = 2.2081784289377584

location = 24163

power = 29.552057768685117

sigma_power = 2.2396472125684044



moment = 2024-12-29

location = 44360

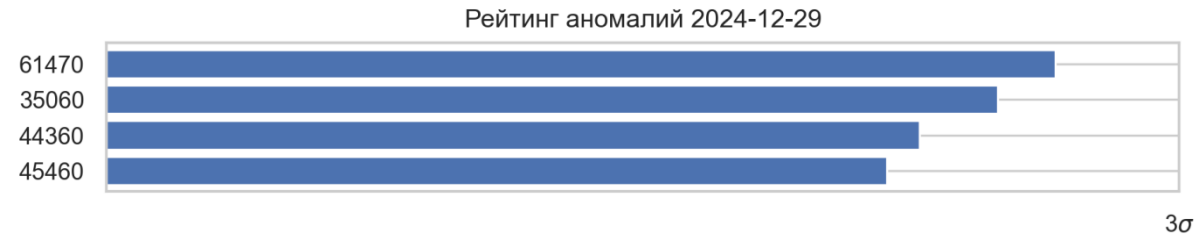
power = 29.996980067942154

sigma_power = 2.2730231863851698

location = 45460
power = 28.785922224579892
sigma_power = 2.181255196681502

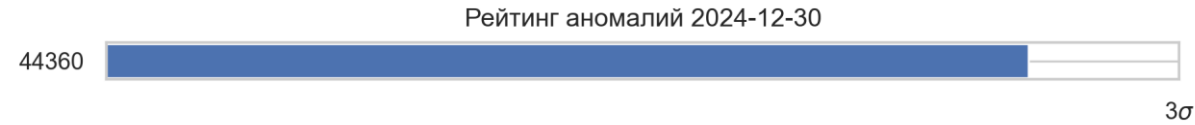
location = 35060
power = 32.89291217257318
sigma_power = 2.4924626367936558

location = 61470
power = 35.01771825139447
sigma_power = 2.6534699606240464



moment = 2024-12-30

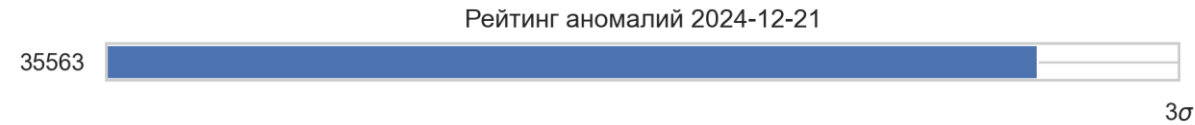
location = 44360
power = 34.00300209525544
sigma_power = 2.5779843095145507



D.2.3. Календарь аномалий по времени для суммы затрат

moment = 2024-12-21

location = 35563
power = 36222.70584114461
sigma_power = 2.6038240041364085



moment = 2024-12-22

location = 62060

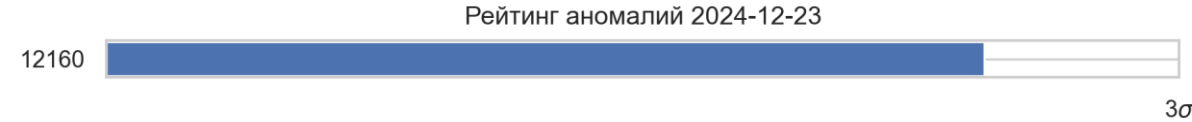
power = 32529.800005674188
sigma_power = 2.3393536426515222

location = 35060
power = 52197.183408608995
sigma_power = 3.753717241476423



moment = 2024-12-23

location = 12160
power = 34126.01358176796
sigma_power = 2.4551570554084634



moment = 2024-12-24

location = 42060
power = 47362.04063244518
sigma_power = 3.4090286028040633

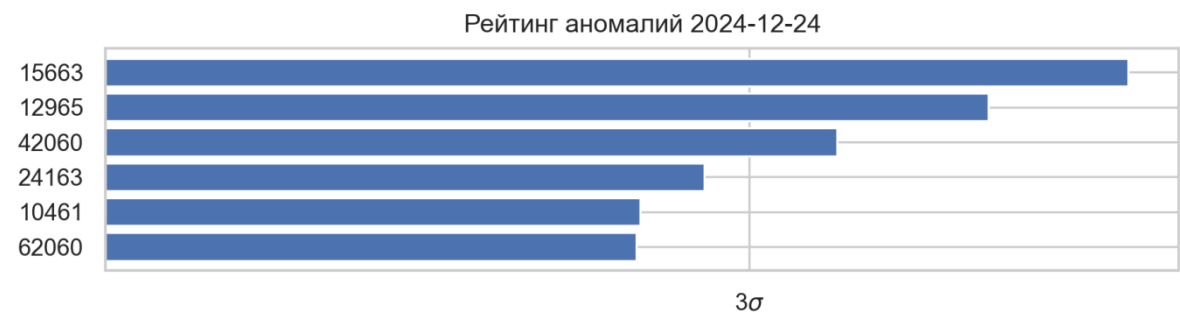
location = 12965
power = 57178.73090140583
sigma_power = 4.115615089891147

location = 62060
power = 34361.03332145997
sigma_power = 2.473241098790743

location = 10461
power = 34641.064640518074
sigma_power = 2.4933972146084398

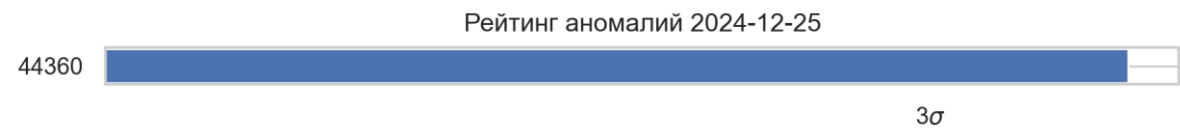
location = 15663
power = 66219.632631972
sigma_power = 4.766361809903191

location = 24163
power = 38766.024974675354
sigma_power = 2.790303926751691



moment = 2024-12-25

location = 44360
power = 51716.32327198825
sigma_power = 3.719708426144556



moment = 2024-12-26

location = 44360
power = 38436.00277319923
sigma_power = 2.7658532942837435

location = 44560
power = 35900.34231050821
sigma_power = 2.583387264054158

location = 35060
power = 50683.12109487613
sigma_power = 3.647155461821113

location = 44597
power = 33189.53768230418
sigma_power = 2.388317866351186

location = 43263
power = 62911.9007924419

$\sigma_{\text{power}} = 4.52713798266656$

location = 39863

power = 58750.773392402996

$\sigma_{\text{power}} = 4.227703413592248$



moment = 2024-12-27

location = 12965

power = 53132.64731098815

$\sigma_{\text{power}} = 3.8204974878340607$

location = 45070

power = 35754.209338942426

$\sigma_{\text{power}} = 2.570902709955376$



moment = 2024-12-28

location = 44560

power = 52058.819253558264

$\sigma_{\text{power}} = 3.7444793290085516$

location = 45460

power = 40225.439933776914

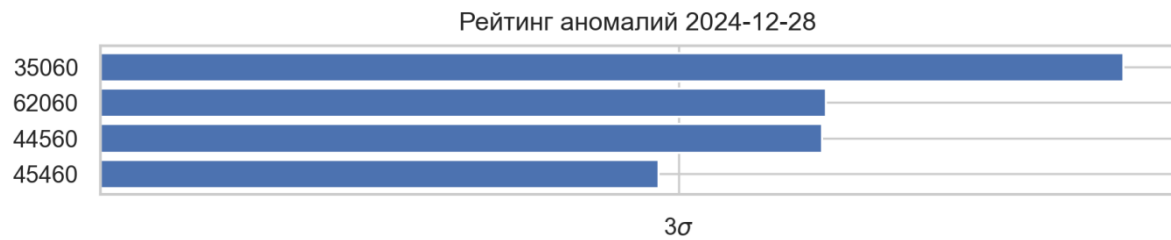
$\sigma_{\text{power}} = 2.893329708433745$

location = 62060

power = 52344.28348574021

$\sigma_{\text{power}} = 3.7650121596010075$

location = 35060
power = 73790.58256313324
sigma_power = 5.307598501943846



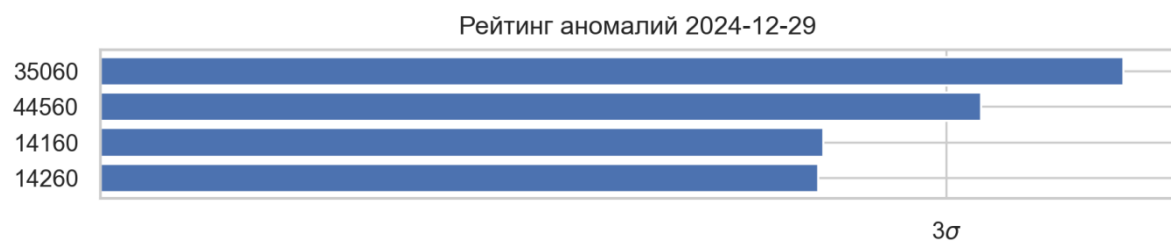
moment = 2024-12-29

location = 44560
power = 43398.412359209
sigma_power = 3.1212061160212934

location = 14260
power = 35373.65954611682
sigma_power = 2.5440673176599855

location = 35060
power = 50417.458479875226
sigma_power = 3.6260146675214533

location = 14160
power = 35637.1965885759
sigma_power = 2.5630208549901803



moment = 2024-12-30

location = 44360
power = 48272.049592687865
sigma_power = 3.4715752415689836

location = 44560
power = 66898.28002533803

`sigma_power = 4.811115637291935`

`location = 45460`

`power = 39040.58811031939`

`sigma_power = 2.807677325567865`

`location = 12965`

`power = 33465.5238503116`

`sigma_power = 2.4067360931464665`

`location = 44567`

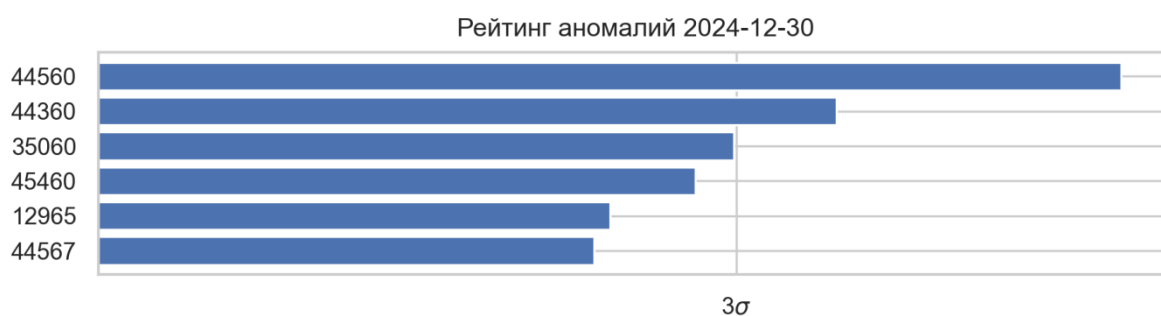
`power = 32418.17600427442`

`sigma_power = 2.331414102837523`

`location = 35060`

`power = 41559.61517913145`

`sigma_power = 2.9888378952706005`



D.2.4. Календарь аномалий по локациям для суммы затрат

`moment = 2024-12-21`

`location = 44597`

`power = 43399.98409882434`

`sigma_power = 2.6823947901360534`

`location = 35563`

`power = 40435.58737058159`

`sigma_power = 2.4991762359165803`

Рейтинг аномалий 2024-12-21



moment = 2024-12-22

location = 44360

power = 34781.19740959344

sigma_power = 2.150867486754783

location = 35060

power = 58641.54492292529

sigma_power = 3.626390168873268

location = 40060

power = 40803.67633355933

sigma_power = 2.523297278480449

Рейтинг аномалий 2024-12-22



moment = 2024-12-23

location = 44360

power = 56470.44606601406

sigma_power = 3.4927494173737306

location = 39263

power = 41068.55912971735

sigma_power = 2.540128438245639

location = 10461

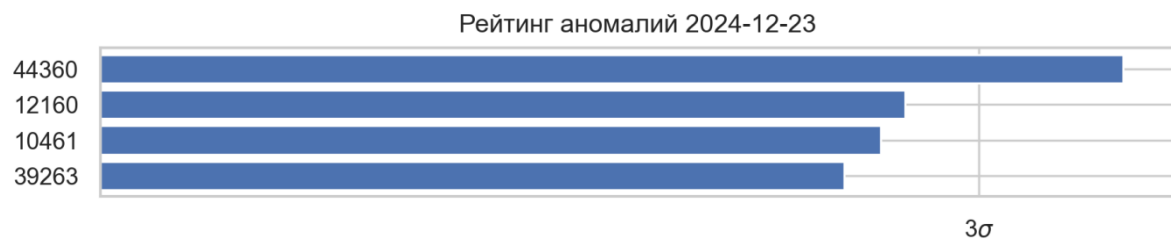
power = 43091.697298414154

sigma_power = 2.6652614087152147

location = 12160

power = 44428.5315450393

sigma_power = 2.747945845642929



moment = 2024-12-24

location = 44360

power = 41481.21992905307

sigma_power = 2.566267772142892

location = 42060

power = 64100.28532181721

sigma_power = 3.9656137569698218

location = 12965

power = 45973.83032604466

sigma_power = 2.8442065910665524

location = 62060

power = 40386.77472679432

sigma_power = 2.4985590727426152

location = 10461

power = 37144.25600890272

sigma_power = 2.297958142959756

location = 40060

power = 42066.91368281387

sigma_power = 2.6025021694723938

location = 44563

power = 35209.72228666813

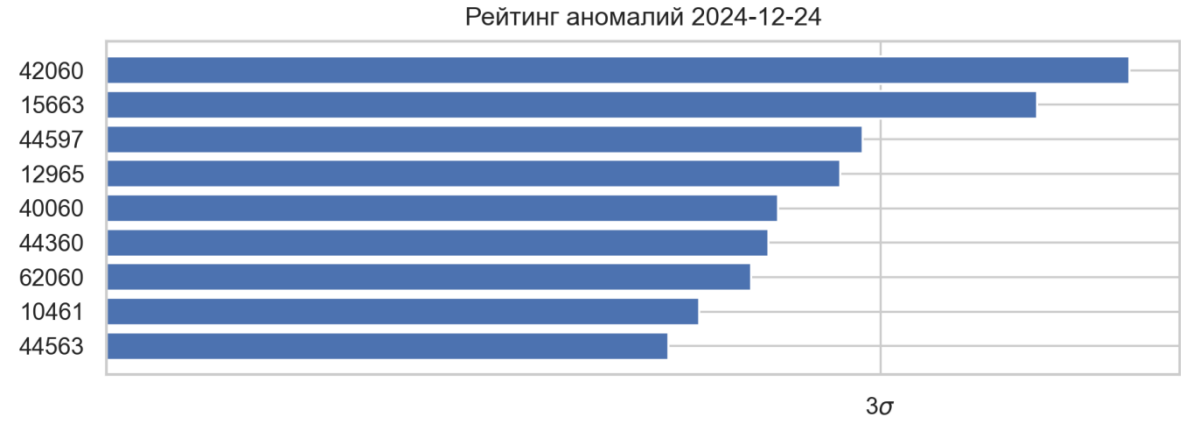
sigma_power = 2.1782767171486235

location = 44597

power = 47366.31902181859

sigma_power = 2.9303539818412165

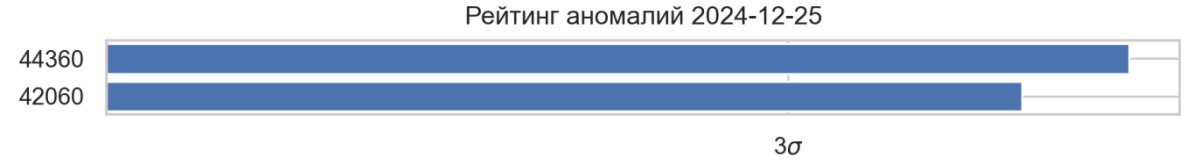
location = 15663
power = 58308.94804729489
sigma_power = 3.607328194716934



moment = 2024-12-25

location = 44360
power = 72786.89409435604
sigma_power = 4.4996172143913515

location = 42060
power = 65160.6660485856
sigma_power = 4.028170981898741



moment = 2024-12-26

location = 44360
power = 36519.39839040005
sigma_power = 2.257627926528348

location = 35060
power = 42399.395966956836
sigma_power = 2.621129170301389

location = 44597
power = 42883.00558416491
sigma_power = 2.6510258998607026

location = 43263
power = 62337.436630564436
sigma_power = 3.8536981442265157

location = 10901
power = 37162.91941456624
sigma_power = 2.2974103736523297

location = 39863
power = 63650.91138048771
sigma_power = 3.9348971071589425



moment = 2024-12-27

location = 12965
power = 52590.43329539413
sigma_power = 3.2494439852287846

location = 10461
power = 40619.12102465982
sigma_power = 2.509763662860343



moment = 2024-12-28

location = 44560
power = 50959.58500335185
sigma_power = 3.1500491676394455

location = 45460
power = 40117.2627618529

sigma_power = 2.479834758517669

location = 30563

power = 40500.5656172917

sigma_power = 2.503528492300026

location = 62060

power = 45767.37087547455

sigma_power = 2.829094242463939

location = 35060

power = 71768.31978133142

sigma_power = 4.436333929626677



moment = 2024-12-29

location = 44360

power = 34872.21137615862

sigma_power = 2.1551569244685163

location = 45460

power = 49639.50717781925

sigma_power = 3.067798783033941

location = 35060

power = 54210.941521946894

sigma_power = 3.350320538687085



moment = 2024-12-30

location = 12965
power = 55112.82021083339
sigma_power = 3.4065611627254966



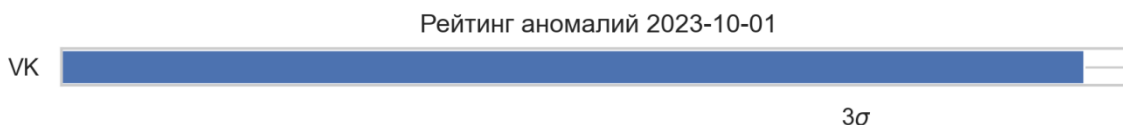
D.3. Маркетинг

Исследовались сведения о просмотрах нарративного контента на канале одного из ведущих отечественных производителей. В качестве наблюдаемой характеристики использовалось число просмотров контента, которое было зарегистрировано сервером в течение недели, а в качестве локаций — источники трафика (такие, как поиск на YouTube, Google, Yandex и т. д.). Всего было использовано 556 наблюдений с 2014-08-10 по 2025-03-23 в пяти локациях, на последних 415 наблюдениях проводилась процедура пилотной эксплуатации.

D.3.1. Календарь аномалий по времени

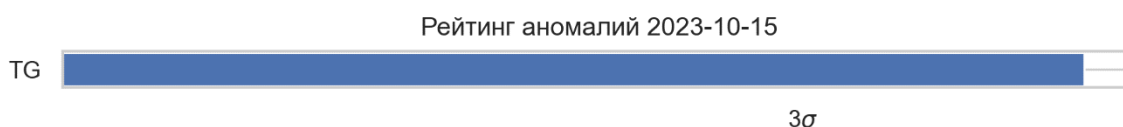
moment = 2023-10-01

location = VK
power = 3517.249620644633
sigma_power = 3.8601494355257784



moment = 2023-10-15

location = TG
power = 3790.2302817378695
sigma_power = 4.142995397450171



moment = 2023-10-22

location = VK

power = 5465.092575031776

sigma_power = 5.956551956764635

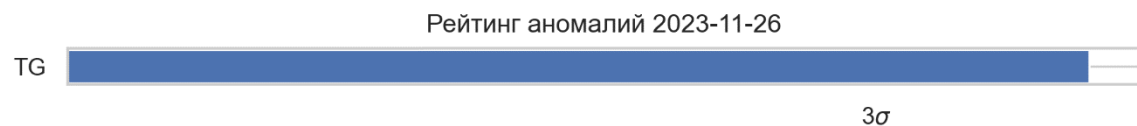


moment = 2023-11-26

location = TG

power = 3506.7305474595646

sigma_power = 3.790155381520001

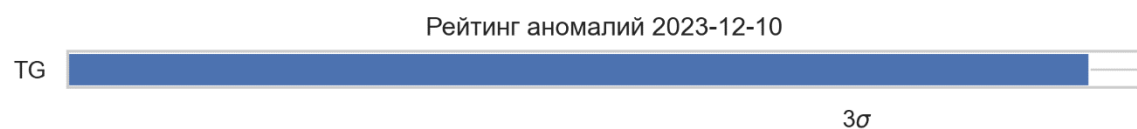


moment = 2023-12-10

location = TG

power = 3596.8188512683882

sigma_power = 3.8838391881565664

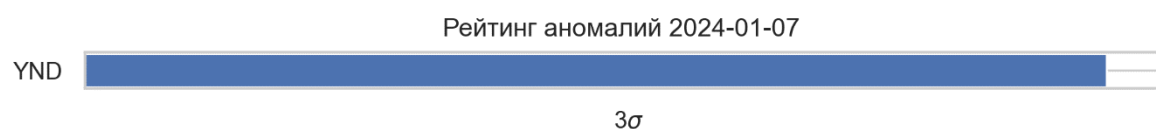


moment = 2024-01-07

location = YND

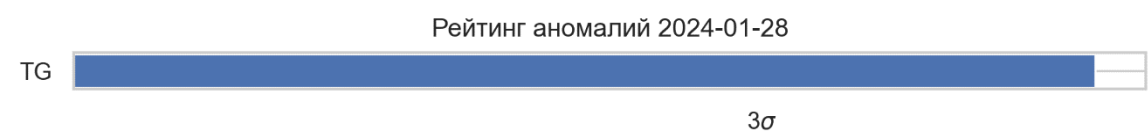
power = 5246.344740508404

sigma_power = 5.640895783339141



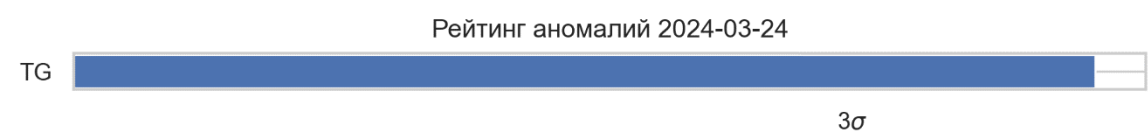
moment = 2024-01-28

location = TG
power = 4211.140538943542
sigma_power = 4.455587013257521



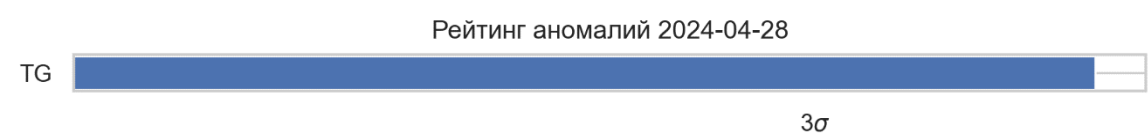
moment = 2024-03-24

location = TG
power = 3729.970370755076
sigma_power = 3.937306828739478



moment = 2024-04-28

location = TG
power = 3917.1281934027643
sigma_power = 4.134518560907973



moment = 2024-05-19

location = YND
power = 4074.542804903826
sigma_power = 4.289721870040514

Рейтинг аномалий 2024-05-19



moment = 2024-10-13

location = TG

power = 4000.320624089182

sigma_power = 4.237785589257192

Рейтинг аномалий 2024-10-13



D.3.2. Календарь аномалий по локациям

moment = 2024-09-22

location = TG

power = 2309.9000266748044

sigma_power = 2.1423204162184017

Рейтинг аномалий 2024-09-22



moment = 2024-09-29

location = TG

power = 3667.8766439833535

sigma_power = 3.4012045451609834

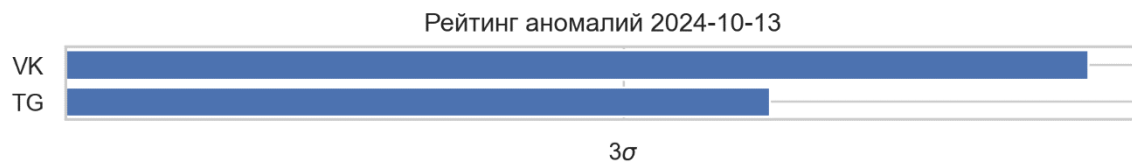
Рейтинг аномалий 2024-09-29



moment = 2024-10-13

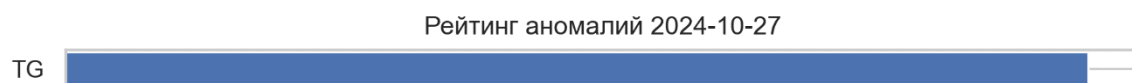
location = VK
power = 5939.215359878938
sigma_power = 5.503050480338555

location = TG
power = 4088.4402024708306
sigma_power = 3.788192792608411



moment = 2024-10-27

location = TG
power = 2220.97890050384
sigma_power = 2.0506538536097256

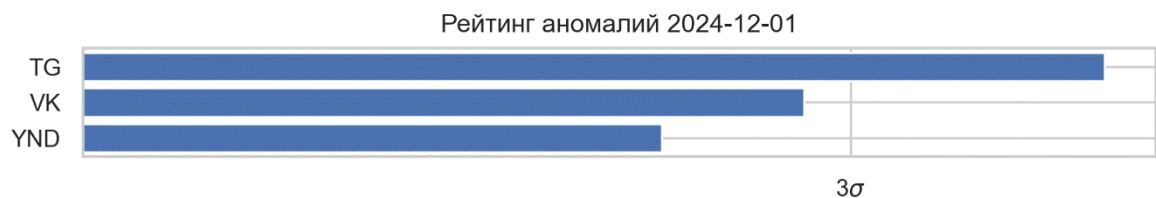


moment = 2024-12-01

location = YND
power = 2439.424262625946
sigma_power = 2.262035110405273

location = VK
power = 3038.263778581397
sigma_power = 2.8173284356963637

location = TG
power = 4307.570242419422
sigma_power = 3.994333941075416



moment = 2024-12-15

location = YND

power = 2043.4110271206466

sigma_power = 1.8887199632470704

location = VK

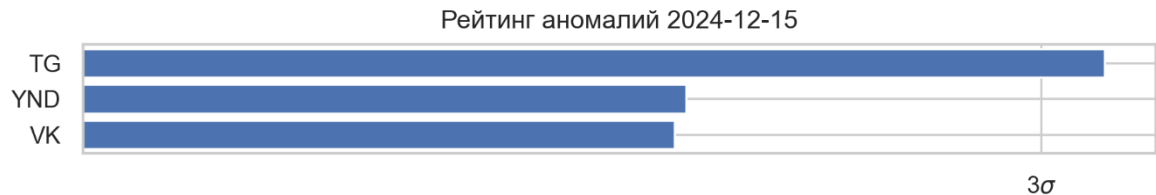
power = 2004.4097951851036

sigma_power = 1.8526712171209978

location = TG

power = 3462.162418177547

sigma_power = 3.2000685072301938



moment = 2024-12-22

location = VK

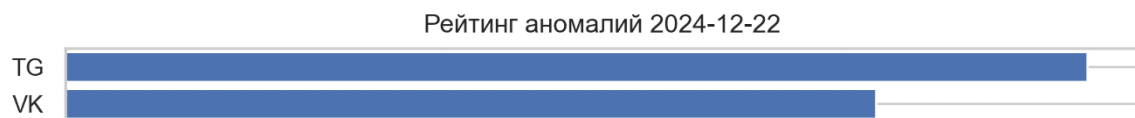
power = 2148.1640944730543

sigma_power = 1.9805208838835466

location = TG

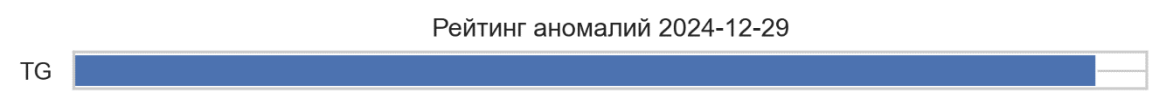
power = 2708.9885553036693

sigma_power = 2.497578477260846



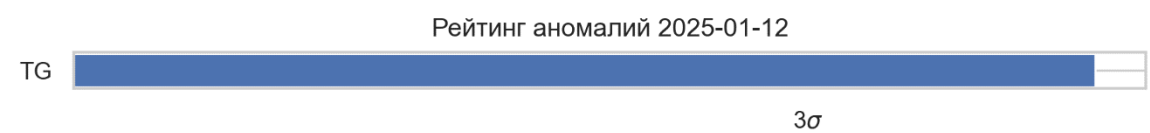
moment = 2024-12-29

location = TG
power = 2912.7477577154536
sigma_power = 2.68089347029469



moment = 2025-01-12

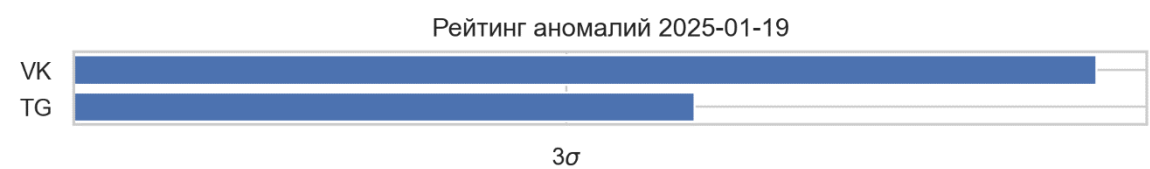
location = TG
power = 4539.97801366664
sigma_power = 4.175223337553971



moment = 2025-01-19

location = VK
power = 6794.289699156094
sigma_power = 6.2328180410232905

location = TG
power = 4124.506294716313
sigma_power = 3.7836622196452736



ПРИЛОЖЕНИЕ Е. АКТЫ ВНЕДРЕНИЯ

Результаты проведенного диссертационного исследования апробированы и внедрены в деятельность подразделений АО «АВТОВАЗ» и ПАО «КАМАЗ». Подтверждение факта внедрения представлено в прилагаемых документах.



АКЦИОНЕРНОЕ ОБЩЕСТВО «АВТОВАЗ»

ОГРН 1026301983113 Южное шоссе, 36, Тольятти
ИНН 5320002223 Самарская область, 445024
Факс (8482) 73-82-21 www.lada.ru | info.avtovaz.ru
Тел. (8482) 73-92-95 E-mail avtovaz@vaz.ru

Справка о внедрении
результатов работ
Мосина Владимира
Геннадьевича

19.09.25 № 3.М440-30

На № _____ от _____

443100, Россия,
г. Самара, СамГТУ
ул. Молодогвардейская, 244

Справка о внедрении

СПРАВКА О ВНЕДРЕНИИ ОСНОВНЫХ РЕЗУЛЬТАТОВ НАУЧНО-ТЕХНИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ В ПРАКТИКУ АО «АВТОВАЗ»

Настоящей справкой подтверждается что ряд разработанных, кандидатом физико-математических наук, доцентом ФГБОУ ВО «Самарский государственный технический университет» Мосиным Владимиром Геннадьевичем, технических решений вошли в устойчивую практику АО «АВТОВАЗ».

В арсенал инструментов по исследованию качества автомобилей проводимых бюро исследования электромагнитной совместимости, службы первого исполнительного Вице-президента по стратегии и техническому развитию вошли следующие научно-прикладные разработки Мосина В.Г.:

1. комплекс программных инструментов по сбору и обработке статистических данных о качестве продукции автосборочного производства. С использованием предложенного комплекса в лаборатории электромагнитной совместимости формируются цифровые двойники характеристик автомобильной техники по ключевым параметрам электромагнитной совместимости продукции;
2. методы моделирования временных рядов и моделирования состояния семейства процессов, протекающих в совокупности локаций. Позволяют проводить анализ стабильности ключевых параметров качества автомобилей с позиции электромагнитной совместимости как для перспективной, так и для серийно выпускаемой продукции. В качестве совокупности локаций рассматриваются отдельные автомобили одной и той же модели и модификации в партии выпущенных в заданный период производства, а также автомобили различных марок и моделей для обеспечения сравнительного анализа качества

продукции на конкурентном рынке;

3. научно-программный инструмент позволяющий реализовать детекцию аномалий в управлении качеством автомобильного производства, используется при проведении анализа данных по результатам испытаний автомобилей по ключевым параметрам качества с позиции электромагнитной совместимости и помехоустойчивости;

4. метод принятия решений о качестве автомобилей по параметрам электромагнитной совместимости, обеспечивающий рационализацию выводов по результатам контроля автомобилей и позволяющий исключить человеческий фактор при проведении оценки.

Все предложенные Мосиным Владимиром Геннадьевичем научно-прикладные инструменты прошли апробацию, доказали свою производственную эффективность и в настоящее время используются в качестве основного инструментария проведения анализа данных о качестве автомобилей с позиции электромагнитной совместимости. Применение предложенной группы инструментов в практике подразделения, позволило снизить уровень трудоемкости процессов исследования и анализа данных о качестве автомобилей на 40% при одновременном улучшении достоверности и точности результатов.

Д.т.н., начальник бюро исследования электромагнитной совместимости,
Службы первого исполнительного
Вице-президента по стратегии и
техническому развитию АО «АВТОВАЗ»



П.А. Николаев

АО «АВТОВАЗ», г. Тольятти, Южное шоссе 36, 445024.

Николаев Павел Александрович.

начальник бюро исследования электромагнитной совместимости.

Доктор технических наук, специальность 05.09.03 – «Электротехнические комплексы и системы»

Тел.: 8-8482-643314

e-mail: Pavel.Nikolaev@vaz.ru

СПРАВКА
О ВНЕДРЕНИИ ОСНОВНЫХ РЕЗУЛЬТАТОВ НАУЧНО-ТЕХНИЧЕСКОЙ
ДЕЯТЕЛЬНОСТИ В ПРАКТИКУ ПАО «КАМАЗ»

В ПАО «КАМАЗ» последовательно реализуется комплексная программа по повышению удовлетворенности потребителей качеством продукции и услуг, проводимая Департаментом качества (ДК) ПАО «КАМАЗ» совместно с ФГБОУ ВО «СамГТУ».

Под общим научным руководством д.т.н., профессора ФГБОУ ВО «СамГТУ» Козловского В.Н. решаются задачи по развитию организационной структуры подразделений службы управления качеством ПАО «КАМАЗ», проводится разработка и реализация системы мониторинга качества автомобилей в эксплуатации, а также проводится разработка технических решений направленных на повышение качества продукции.

Настоящей справкой подтверждается, что в рамках реализации комплексной программы реализуемой на предприятии, к.ф.-м.н., старшим научным сотрудником ФГБОУ ВО «СамГТУ» Мосиным Владимиром Геннадьевичем в период с 2022 по 2024гг. были разработаны и внедрены в устойчивую производственную практику инженерно-технические инструменты составляющие методологию детекции аномалий.

Мосиным В.Г. предложены и внедрены в практику ПАО «КАМАЗ» программно-технические решения обеспечивающие обоснование и автоматизацию поиска позиций дефектов автомобилей в эксплуатации имеющих наибольший прирост по параметрам частоты проявления и затрат на устранение в анализируемый период на основе использования корпоративных электронных баз гарантийного обслуживания автомобилей. Также предложены инженерные решения, направленные на автоматизированный поиск и устранение нарушений связанных с необоснованными ремонтами автомобильной техники. Также предложен инженерно-технический и программный комплекс определения аномалий в электронных базах отражающих информацию о результативности процессов системы менеджмента предприятия, применение которого снижает риски принятия не достаточно обоснованных управленческих решений в системе управления качеством.

Применение разработанных технических решений обеспечивает эффективное расширение базы инженерно-технических инструментов управления качеством в рамках действующей системы менеджмента предприятия.

Директор по качеству
Директор Департамента
качества ПАО «КАМАЗ»



И.В. Тимонин