



Geekbrains

**"Разработка системы предиктивного обслуживания
промышленного оборудования на основе анализа сенсорных
данных и методов машинного обучения"**

Программа: Разработчик - Аналитик

Специализация: Разработчик – Data Engineer

Елагин О.В.

Самара

2025

Содержание

Введение	4
Глава 1. Теоретические основы предиктивного обслуживания	8
1.1. Концепция и принципы предиктивного обслуживания	8
1.1.1. Тенденции развития предиктивного обслуживания.....	12
1.2. Анализ современных подходов к обслуживанию промышленного оборудования	14
1.3 Роль машинного обучения в системах предиктивного обслуживания.....	18
1.4 Обзор существующих решений в области предиктивного обслуживания.....	25
Глава 2. Анализ данных и разработка методологии.....	30
2.1. Описание и анализ исходного набора данных	30
2.1.1. Исследовательский анализ данных (EDA)	32
2.2 Предварительная обработка данных.....	55
2.2.1 Инженерия признаков	59
2.2.2 Подготовка данных для машинного обучения	61
2.3 Разработка методологии предиктивного обслуживания.....	64
2.3.1 Определение критериев оценки состояния оборудования	64
2.3.2 Формулировка критериев успешности модели	65
Глава 3. Разработка и внедрение ML-системы системы предиктивного обслуживания промышленного оборудования на основе анализа сенсорных данных	75
3.1 Выбор и обоснование методов машинного обучения	75
3.1.1 Сравнительный анализ алгоритмов классификации	76
3.1.2 Оценка эффективности моделей	81
3.1.3 Выбор оптимальной модели	82
3.2 Разработка системы прогнозирования отказов	84
3.2.1 Архитектура системы.....	84
3.2.2 Процесс обучения и тестирования модели.....	85
3.2.3 Интерпретация результатов модели.....	87
3.3 Визуализация и анализ результатов	90
3.3.1 Разработка информационной панели (дашборда)	90
3.3.2 Инструменты аналитической отчетности.....	92
3.3.3 Система уведомлений о потенциальных отказах	93
Глава 4. Внедрение и оценка эффективности	95
4.1 Разработка плана внедрения системы	95
4.1.1 Необходимые ресурсы и технологии	96
4.1.2 Потенциальные риски и способы их минимизации	97
4.2 Оценка экономической эффективности	98
4.2.1 Анализ затрат на внедрение и эксплуатацию	98
4.2.2 Прогнозируемое сокращение затрат	99

4.2.3 Расчет срока окупаемости и ROI.....	100
4.3. Рекомендации по дальнейшему развитию	103
4.3.1. Масштабирование решения	103
4.3.2. Интеграция с другими системами.....	106
4.3.3. Перспективы развития	107
Заключение	110
Список используемой литературы.....	113
Приложения	1
Приложение 1 : Структура датасета	1
Приложение 2 : Структура датасета	2
Приложение 3 : Расширенный алгоритм оценки эффективности моделей предиктивного обслуживания на основе многокритериального анализа	3
Приложение 4 : Веб-интерфейс информационной панели мониторинга и анализа рисков отказа оборудования по различным метрикам и типам станков	4
Приложение 5 : Необходимые ресурсы и технологии для внедрения системы предиктивной аналитики отказов	
5	
Приложение 6 : Сводная таблица экономии от предотвращения простоев оборудования	6
Приложение 7 : Схема архитектуры Edge Computing	7
Приложение 8 : Схема интеграции с внешними системами.....	8
Приложение 9 : Ссылка на репозиторий проекта с описанием файлов.....	9

Введение

В современных условиях высококонкурентного производства эффективность работы промышленного оборудования становится критическим фактором успеха предприятия. Незапланированные простои и аварийные отказы оборудования приводят к существенным экономическим потерям, нарушению производственных графиков и снижению качества продукции. Традиционные подходы к обслуживанию оборудования, такие как планово-предупредительные ремонты или реактивное обслуживание, часто оказываются недостаточно эффективными в условиях современного промышленного производства.

Дипломный проект представляет собой комплексное исследование и разработку интеллектуальной системы, способной прогнозировать возможные отказы оборудования до их фактического наступления. В основе проекта лежит использование современных технологий машинного обучения для анализа данных с промышленных датчиков с целью выявления паттернов, предшествующих сбоям в работе оборудования.

Выбор темы обусловлен несколькими факторами:

- Экономический эффект: По оценкам экспертов, внедрение систем предиктивного обслуживания позволяет сократить затраты на обслуживание оборудования на 15-25%, увеличить производительность на 20-25% и уменьшить количество незапланированных простоев на 35-45%.
- Технологическая актуальность: В эпоху цифровой трансформации промышленности и развития концепции Индустрии 4.0 предиктивное обслуживание становится одним из ключевых элементов интеллектуального производства.
- Практическая значимость: Разработанная система может быть адаптирована для различных типов промышленного оборудования и внедрена на предприятиях различных отраслей.
- Соответствие современным трендам: Проект находится на стыке таких актуальных направлений, как Internet of Things (IoT), Big Data, машинное обучение и промышленная автоматизация.

Основной целью дипломного проекта является разработка системы предиктивного обслуживания промышленного оборудования, основанной на анализе сенсорных данных с использованием методов машинного обучения, способной с высокой точностью прогнозировать потенциальные отказы оборудования и сокращать незапланированные простои.

Для достижения поставленной цели определены следующие задачи:

1. Изучить теоретические основы и существующие подходы к предиктивному обслуживанию оборудования.
2. Провести исследовательский анализ данных с датчиков промышленного оборудования.
3. Разработать и сравнить модели машинного обучения для прогнозирования отказов оборудования.
4. Создать информационную панель для визуализации состояния оборудования и прогнозов отказов.
5. Разработать рекомендации по внедрению системы и оценить её экономическую эффективность.

Реализация дипломного проекта включает следующие этапы:

- Детальное изучение литературы по теме
- Анализ существующих решений
- Формулирование точных требований к разрабатываемой системе
- Исследовательский анализ данных
- Предварительная обработка и подготовка данных
- Выявление ключевых факторов, влияющих на отказы оборудования
- Создание и обучение моделей машинного обучения
- Оценка эффективности моделей и выбор оптимального решения
- Разработка архитектуры системы предиктивного обслуживания
- Разработка дашборда для мониторинга состояния оборудования
- Создание системы оповещений о потенциальных отказах
- Подготовка аналитической отчетности
- Экономическое обоснование эффективности системы
- Разработка плана внедрения
- Формулирование рекомендаций по дальнейшему развитию системы

Проект направлен на решение следующих проблем современного промышленного производства:

1. Высокие затраты на внеплановый ремонт оборудования: Аварийные отказы оборудования требуют значительных финансовых и временных затрат на их устранение.
2. Потери производительности из-за простоев: Незапланированные простои оборудования приводят к нарушению производственных графиков и снижению общей эффективности производства.

3. Неоптимальное использование ресурсов при обслуживании: Традиционные стратегии обслуживания часто приводят либо к избыточному обслуживанию оборудования, либо к недостаточному вниманию к критическим компонентам.
4. Отсутствие объективной и своевременной информации о состоянии оборудования: Принятие решений об обслуживании часто основывается на субъективных оценках или ограниченных данных.

Дипломный проект выполняется по специализации "Разработчик — Аналитик и Data Engineer", что предполагает комплексное применение знаний и навыков в области анализа данных, машинного обучения и программной разработки для создания интеллектуальной системы промышленного назначения.

Для реализации проекта применены следующие знания и навыки:

- Опыт анализа данных и построения моделей машинного обучения с использованием Python и библиотек для анализа данных и машинного обучения
- Знание принципов визуализации данных и создания информационных панелей
- Понимание основ промышленной автоматизации и обслуживания оборудования
- Навыки разработки программного обеспечения и веб-приложений

Для реализации проекта используются следующие инструменты:

Языки программирования и библиотеки:

- Python (основной язык разработки)
- Pandas, NumPy (для анализа и обработки данных)
- Scikit-learn, TensorFlow/Keras (для создания моделей машинного обучения)
- Matplotlib, Seaborn, Plotly (для визуализации данных)
- Инструменты для создания дашбордов:
- Streamlit или Dash (для разработки интерактивного веб-интерфейса)
- Plotly Dashboard (для создания аналитических панелей)
- Среды разработки и инструменты для управления проектом:
- Jupyter Notebook/Lab (для исследовательского анализа данных)
- PyCharm или VS Code (для разработки программного кода)
- Git (для версионного контроля)

В рамках дипломного проекта планируется использовать следующие технологии:

- Алгоритмы классификации (Random Forest, Gradient Boosting, SVM и др.)
- Методы обнаружения аномалий (Isolation Forest, One-Class SVM)
- Техники обработки временных рядов
- Технологии обработки и анализа данных:

- Методы предварительной обработки данных (нормализация, кодирование и т.д.)
- Техники снижения размерности и отбора признаков
- Методы балансировки классов для несбалансированных наборов данных
- Технологии веб-разработки:
- Flask или FastAPI (для создания API)
- HTML, CSS, JavaScript (для frontend-разработки)
- Библиотеки визуализации данных для веб (Plotly.js)

Дипломный проект выполняется индивидуально, при этом я выполняю следующие роли:

1. Аналитик данных: проведение исследовательского анализа данных, выявление закономерностей и паттернов в данных
2. Data Scientist: разработка, обучение и оценка моделей машинного обучения
3. Python-разработчик: создание программной части системы, разработка алгоритмов и интеграция компонентов
4. Web-разработчик: создание интерактивного дашборда и визуализаций
5. Проектный менеджер: планирование, контроль и координация всех этапов разработки проекта

Данный дипломный проект имеет значительный практический потенциал и может послужить основой для внедрения систем предиктивного обслуживания на реальных промышленных предприятиях, способствуя повышению их эффективности и конкурентоспособности.

Глава 1. Теоретические основы предиктивного обслуживания

1.1. Концепция и принципы предиктивного обслуживания

Предиктивное обслуживание (Predictive Maintenance, PdM) представляет собой стратегию обслуживания оборудования, основанную на анализе технического состояния и прогнозировании возможных отказов до их фактического наступления. В отличие от традиционных подходов, таких как реактивное обслуживание (после отказа) или планово-предупредительное обслуживание (по регламенту), предиктивное обслуживание направлено на определение оптимального времени для проведения технического обслуживания на основе фактического состояния оборудования.

Концепция предиктивного обслуживания прошла значительный путь развития, прежде чем стать одним из ключевых элементов современной промышленности. Первые попытки прогнозирования отказов оборудования относятся к 1940-м годам, когда начали применяться статистические методы для планирования технического обслуживания военной техники. Однако широкое распространение предиктивное обслуживание получило лишь с развитием вычислительной техники и сенсорных технологий в 1990-х годах.

С начала 2000-х годов PdM приобретает новое значение в связи с развитием концепции "Индустря 4.0", которая предполагает комплексную цифровизацию промышленных процессов. В исследовании, проведенном Deloitte в 2024 году, отмечается, что предиктивное обслуживание является одним из ключевых технологических драйверов четвертой промышленной революции.

Предиктивное обслуживание основывается на нескольких фундаментальных принципах:

1. Непрерывный мониторинг состояния оборудования

Основой предиктивного обслуживания является постоянный сбор данных о работе оборудования с помощью различных датчиков, измеряющих такие параметры, как температура, вибрация, давление, потребление энергии и т.д. Данные собираются в режиме реального времени или с определенной периодичностью и передаются в систему анализа. Современные системы промышленного оборудования могут генерировать до нескольких терабайт данных ежедневно, которые могут быть использованы для оценки состояния оборудования [64].

2. Выявление отклонений и аномалий

Одним из ключевых принципов предиктивного обслуживания является анализ собранных данных с целью выявления отклонений от нормального режима работы. Как отмечает специалист в областях корпоративных преобразований, инженерии надёжности и оптимизации процессов R. Keith Mobley, эти отклонения могут указывать на начало деградации оборудования задолго до появления явных признаков неисправности [51]. Современные алгоритмы обнаружения аномалий способны выявлять даже незначительные отклонения в работе оборудования, которые могут быть незаметны при обычном наблюдении.

3. Прогнозирование отказов

На основе анализа исторических данных и выявленных аномалий система предиктивного обслуживания строит прогнозы о возможных отказах оборудования. Прогнозирование может включать как определение вероятности отказа в заданный период времени, так и оценку остаточного ресурса оборудования (Remaining Useful Life, RUL). Исследования показывают, что точность прогнозирования отказов с использованием современных методов машинного обучения может достигать 85-95% в зависимости от типа оборудования и доступных данных.

4. Принятие решений на основе данных

Предиктивное обслуживание предполагает переход от субъективных решений к решениям, основанным на объективных данных и аналитических моделях. Это позволяет оптимизировать график технического обслуживания, минимизировать незапланированные простоя и снизить затраты на обслуживание. Согласно исследованию McKinsey, компании, внедрившие системы предиктивного обслуживания, смогли сократить время простоя оборудования на 30-50% и увеличить срок службы оборудования на 20-40% [68].

Система предиктивного обслуживания обычно включает следующие основные компоненты:

1. Сенсорная инфраструктура

Датчики являются первичным источником данных для системы предиктивного обслуживания. Они могут быть встроенными в само оборудование или установлены отдельно. Современные промышленные датчики способны измерять широкий спектр параметров, включая:

- Вибрацию (акселерометры, велосиметры)
- Температуру (термопары, термисторы)
- Давление (манометры, датчики давления)
- Электрические параметры (датчики тока, напряжения)
- Акустические сигналы (микрофоны, ультразвуковые датчики)
- Качество смазочных материалов (датчики вязкости, загрязнения)
- Системы сбора и передачи данных

Для эффективного функционирования системы предиктивного обслуживания необходимо наличие надежной инфраструктуры для сбора, передачи и хранения данных. Это может включать промышленные сети передачи данных, системы SCADA, шлюзы IoT и облачные хранилища данных. Объем данных, собираемых с промышленного оборудования, может удваиваться каждые 1-2 года, что создает значительные вызовы для систем хранения и обработки данных [6].

2. Аналитическая платформа

Аналитическая платформа является "мозгом" системы предиктивного обслуживания. Она включает алгоритмы и модели для обработки данных, выявления аномалий и прогнозирования отказов. Современные аналитические платформы часто используют методы машинного обучения, такие как:

- Классификация (для определения типа потенциальной неисправности)
- Регрессия (для прогнозирования времени до отказа)
- Обнаружение аномалий (для выявления отклонений в работе оборудования)
- Кластеризация (для группировки сходных режимов работы)

3. Система принятия решений и оповещения

Система принятия решений анализирует результаты прогнозирования и формирует рекомендации по техническому обслуживанию. Она может автоматически формировать заявки на обслуживание, определять приоритеты работ и выбирать оптимальное время для проведения технического обслуживания. Кроме того, система оповещения обеспечивает своевременное информирование технического персонала о потенциальных проблемах с оборудованием [31].

Внедрение систем предиктивного обслуживания предоставляет ряд существенных преимуществ для промышленных предприятий:

1. Сокращение незапланированных простоев

Одним из основных преимуществ предиктивного обслуживания является возможность заблаговременного выявления потенциальных отказов и планирования технического обслуживания в оптимальное время. Это позволяет избежать незапланированных простоев оборудования, которые могут привести к значительным финансовым потерям. По данным Федерального резервного банка США, незапланированные простои в промышленности стоят американским компаниям около \$50 миллиардов ежегодно [59].

2. Повышение надежности и безопасности

Предиктивное обслуживание способствует повышению общей надежности оборудования и безопасности производственных процессов. Своевременное выявление и устранение

потенциальных неисправностей снижает риск аварий и связанных с ними травм персонала. Исследование, проведенное Управлением по охране труда США (OSHA), показало, что внедрение систем предиктивного обслуживания может снизить количество несчастных случаев на производстве на 10-15% [59].

3. Оптимизация затрат на техническое обслуживание

В отличие от планово-предупредительного обслуживания, которое часто приводит к преждевременной замене компонентов, предиктивное обслуживание позволяет максимально использовать ресурс оборудования. Это приводит к сокращению затрат на запасные части и снижению трудозатрат на техническое обслуживание. По оценкам Министерства энергетики США, внедрение предиктивного обслуживания может сократить затраты на техническое обслуживание на 25-30% и снизить количество отказов оборудования на 70-75% [67].

4. Увеличение срока службы оборудования

Своевременное выявление и устранение потенциальных проблем способствует увеличению срока службы оборудования. Исследование, проведенное International Journal of Prognostics and Health Management, показало, что предиктивное обслуживание может увеличить срок службы промышленного оборудования на 20-40% по сравнению с традиционными подходами к обслуживанию [65].

5. Повышение энергоэффективности

Непрерывный мониторинг состояния оборудования позволяет выявлять неэффективные режимы работы и оптимизировать энергопотребление. По данным исследования Энергетического института США, внедрение систем предиктивного обслуживания может сократить энергопотребление промышленных предприятий на 5-10%.

Несмотря на значительные преимущества, внедрение предиктивного обслуживания сопряжено с рядом вызовов и ограничений:

1. Технологические вызовы

Разработка и внедрение систем предиктивного обслуживания требует значительных технологических ресурсов и экспертизы. Это включает установку датчиков, создание надежной инфраструктуры для сбора и передачи данных, разработку аналитических моделей и интеграцию с существующими системами управления производством.

2. Качество и доступность данных

Эффективность предиктивного обслуживания напрямую зависит от качества и полноты доступных данных. Для создания точных прогностических моделей требуются исторические данные, включающие информацию о работе оборудования как в нормальном режиме, так и в

условиях деградации и отказа. Во многих случаях такие данные могут быть недоступны или недостаточно полны, что ограничивает возможности прогнозирования.

3. Сложность интерпретации результатов

Современные методы машинного обучения, такие как глубокие нейронные сети, могут обеспечивать высокую точность прогнозирования, но при этом они часто работают как "черный ящик", что затрудняет интерпретацию их результатов. Это может создавать проблемы для принятия решений на основе этих прогнозов, особенно в критических системах.

4. Экономические ограничения

Внедрение систем предиктивного обслуживания требует значительных инвестиций в оборудование, программное обеспечение и обучение персонала. Для небольших предприятий такие инвестиции могут быть экономически нецелесообразными, особенно для некритичного оборудования с низкой стоимостью замены.

1.1.1. Тенденции развития предиктивного обслуживания

Предиктивное обслуживание продолжает активно развиваться, и в ближайшие годы ожидается ряд значительных изменений в этой области:

1. Интеграция с цифровыми двойниками

Цифровые двойники (Digital Twins) представляют собой виртуальные модели физических объектов или процессов, которые могут использоваться для симуляции и прогнозирования поведения реальных систем. Интеграция систем предиктивного обслуживания с цифровыми двойниками позволяет создавать более точные и детальные прогнозы состояния оборудования.

2. Развитие технологий Edge Computing

Edge Computing (вычисления на краю сети) позволяет обрабатывать данные непосредственно на устройствах сбора или вблизи них, что снижает нагрузку на каналы связи и центральные системы обработки данных. Это особенно актуально для промышленных предприятий с большим количеством датчиков и высокой частотой сбора данных.

3. Применение методов объяснимого искусственного интеллекта

Для преодоления проблемы "черного ящика" в моделях машинного обучения все более активно применяются методы объяснимого искусственного интеллекта (Explainable AI). Эти методы позволяют лучше понять процесс принятия решений моделями и повысить доверие к их прогнозам.

4. Развитие моделей обучения с подкреплением

Модели обучения с подкреплением (Reinforcement Learning) позволяют системам предиктивного обслуживания автоматически адаптироваться к изменяющимся условиям работы оборудования и улучшать свои прогнозы на основе обратной связи. Это особенно важно для динамичных производственных сред с часто меняющимися режимами работы оборудования.

Предиктивное обслуживание представляет собой современный подход к обслуживанию промышленного оборудования, основанный на анализе данных и прогнозировании потенциальных отказов. Этот подход обеспечивает ряд существенных преимуществ, включая сокращение незапланированных простоев, оптимизацию затрат на техническое обслуживание и повышение общей надежности оборудования.

Несмотря на существующие вызовы и ограничения, предиктивное обслуживание продолжает активно развиваться, и в ближайшие годы ожидается его дальнейшая интеграция с новейшими технологиями, такими как цифровые двойники и Edge Computing. Успешное внедрение систем предиктивного обслуживания требует комплексного подхода, включающего как технологические аспекты, так и изменения в организационной культуре и процессах управления производством.

1.2. Анализ современных подходов к обслуживанию промышленного оборудования

В современной промышленности существует несколько стратегий обслуживания оборудования, каждая из которых имеет свои преимущества, недостатки и области применения. Рассмотрим основные подходы, их эволюцию и тенденции развития.

Классификация подходов к обслуживанию оборудования:

Реактивное обслуживание (Run-to-Failure) наиболее простой и исторически первый подход к обслуживанию оборудования — реактивный, или «работа до отказа». При таком подходе ремонт осуществляется только после выхода оборудования из строя. По данным исследования Российского союза промышленников и предпринимателей, до 30% предприятий малого и среднего бизнеса в России все еще полагаются на реактивное [12].

Преимущества:

- Низкие первоначальные затраты на организацию обслуживания
- Отсутствие необходимости в планировании и прогнозировании

Недостатки:

- Высокая вероятность внеплановых простоев
- Повышенные затраты на срочный ремонт
- Риск вторичных повреждений оборудования
- Невозможность планирования производственных процессов

Следующим шагом в эволюции подходов стало планово-предупредительное обслуживание. Основной принцип ППР — проведение регламентных работ через определенные промежутки времени или объемы выработки, установленные на основе статистических данных о среднем времени наработки на отказ.

Как отмечает Черепанов В.И. (2022) в своей работе «Техническое обслуживание и ремонт. Современные подходы к построению системы», около 65% крупных российских предприятий используют системы ППР, хотя зачастую эти системы нуждаются в модернизации.

Преимущества:

- Снижение вероятности внезапных отказов
- Возможность планирования обслуживания
- Увеличение срока службы оборудования

Недостатки:

- Избыточное обслуживание исправного оборудования
- Нерациональное использование ресурсов
- Не учитывает фактическое состояние оборудования

Обслуживание по состоянию (Condition-Based Maintenance). Данный подход основан на непрерывном или периодическом мониторинге фактического состояния оборудования и проведении обслуживания только при выявлении признаков деградации или предаварийного состояния.

В исследовании "Индустратия 4.0: к вопросу о перспективах цифровой трансформации промышленности в России" (Институт Экономики РАН, 2023) отмечается, что внедрение систем обслуживания по состоянию позволяет сократить затраты на ТО и ремонт в среднем на 20-25% по сравнению с классическим ППР.

Преимущества:

- Сокращение затрат на обслуживание
- Повышение надежности оборудования
- Снижение влияния человеческого фактора

Недостатки:

- Необходимость инвестиций в системы мониторинга
- Трудности в определении пороговых значений параметров
- Сложность интерпретации результатов диагностики

Предиктивное обслуживание (Predictive Maintenance) наиболее современный подход, основанный на прогнозировании состояния оборудования с использованием методов машинного обучения и анализа больших данных. В отличие от обслуживания по состоянию, этот подход позволяет не только выявлять текущие аномалии, но и предсказывать вероятность отказа в будущем.

Так же отмечается, что лишь около 12% российских предприятий внедрили системы предиктивного обслуживания, однако ежегодный рост этого показателя составляет около 30%.

Преимущества:

- Максимальное сокращение внеплановых простоев
- Оптимизация затрат на обслуживание
- Увеличение срока службы оборудования
- Планирование запасных частей и ресурсов

Недостатки:

- Высокие требования к квалификации персонала
- Необходимость значительных инвестиций
- Сложность интеграции с существующими системами
- Сравнительный анализ эффективности подходов

Исследование, проведенное коллективом авторов из МГТУ им. Баумана (Морозов и др., 2024), показало следующие результаты сравнения различных подходов к обслуживанию на примере автоматизированных производственных линий:

Подход	Снижение простоев, %	Снижение затрат на ТОиР, %	Повышение производительности, %
Реактивный	-	-	-
ППР	40-45	10-15	10-12
По состоянию	60-65	20-25	15-18
Предиктивный	75-85	30-40	20-25

Таблица 1: Сравнение различных подходов к обслуживанию на примере автоматизированных производственных линий

Согласно аналитическому обзору «Индустрия 4.0: к вопросу о перспективах цифровой трансформации промышленности в России» [6], основными направлениями развития систем обслуживания промышленного оборудования в ближайшие годы станут:

- Интеграция с концепцией цифрового двойника (Digital Twin)
- Создание виртуальных моделей оборудования, позволяющих моделировать различные сценарии эксплуатации и обслуживания
- Автоматизация процессов диагностики и принятия решений
- Внедрение экспертных систем и систем поддержки принятия решений
- Снижение влияния человеческого фактора
- Развитие технологий промышленного интернета вещей (ПоТ)
- Использование беспроводных сенсорных сетей для сбора данных
- Применение методов граничных вычислений (Edge Computing) для обработки данных непосредственно на оборудовании
- Применение искусственного интеллекта для оптимизации обслуживания
- Использование нейронных сетей для выявления скрытых паттернов в данных сенсоров
- Применение методов глубокого обучения для прогнозирования отказов
- Интеграция с системами планирования ресурсов предприятия (ERP)
- Автоматическое планирование технического обслуживания
- Оптимизация запасов запчастей и расходных материалов

Как отмечают М. В. Яковleva, А. И. Шалина (2023) в монографии «Разработка рекомендаций по внедрению предиктивного обслуживания оборудования на

высокотехнологичных предприятиях», основными препятствиями для внедрения современных подходов к обслуживанию на российских предприятиях являются:

- Технологические барьеры
- Устаревшее оборудование без возможности интеграции датчиков
- Недостаточный уровень цифровизации производственных процессов
- Проблемы с надежностью передачи данных в промышленных условиях
- Организационные барьеры
- Консервативность управления и сопротивление изменениям
- Недостаточная квалификация технического персонала
- Отсутствие четких методик оценки эффективности внедрения
- Экономические ограничения
- Высокая стоимость внедрения современных систем
- Сложность расчета ROI (окупаемости инвестиций)
- Ограниченный бюджет на модернизацию систем обслуживания
- Информационные ограничения
- Недостаток данных для обучения моделей
- Проблемы с качеством и полнотой собираемых данных
- Сложности с интеграцией разнородных источников информации

Анализ современных подходов к обслуживанию промышленного оборудования показывает четкую эволюцию от реактивного обслуживания к предиктивному. Каждый следующий этап развития обеспечивает более высокий уровень надежности оборудования и экономической эффективности при возрастающей сложности реализации.

Предиктивное обслуживание, основанное на анализе данных и машинном обучении, представляется наиболее перспективным подходом, обеспечивающим оптимальный баланс между затратами на обслуживание и предотвращением внеплановых простоев. Однако его внедрение сопряжено с рядом технологических, организационных и экономических вызовов, которые необходимо преодолеть для получения максимальной отдачи от инвестиций.

1.3 Роль машинного обучения в системах предиктивного обслуживания

Машинное обучение (ML) стало ключевой технологией, обеспечивающей эффективность современных систем предиктивного обслуживания. В отличие от традиционных подходов, основанных на экспертных знаниях и пороговых значениях параметров, методы машинного обучения позволяют выявлять сложные зависимости в многомерных данных и формировать прогностические модели высокой точности.

Внедрение ML-моделей в системы предиктивного обслуживания обеспечивает следующие преимущества:

1. Выявление скрытых паттернов и зависимостей
 - Способность обнаруживать неочевидные связи между параметрами
 - Возможность работы с большим количеством разнородных признаков
 - Выявление предвестников отказов, неочевидных для человека-эксперта
2. Адаптивность к изменяющимся условиям
 - Возможность переобучения моделей при изменении режимов работы
 - Постепенное улучшение точности прогнозов при накоплении данных
 - Учет сезонных и других периодических факторов
3. Работа с неструктурированными данными
 - Анализ текстовых журналов (логов) оборудования
 - Обработка данных с датчиков различной природы
 - Интеграция экспертных оценок
4. Прогнозирование времени до отказа (RUL - Remaining Useful Life)
 - Оценка вероятности отказа в заданном временном горизонте
 - Определение оптимального времени для проведения обслуживания
 - Расчет доверительных интервалов прогнозов [34]

Выделяются следующие основные типы задач:

1. Задачи классификации
 - Бинарная классификация: определение вероятности отказа (да/нет)
 - Многоклассовая классификация: определение типа потенциальной неисправности
 - Детекция аномалий: выявление аномального поведения оборудования
2. Задачи регрессии
 - Прогнозирование RUL (остаточного ресурса)
 - Прогнозирование значений ключевых параметров оборудования
 - Оценка вероятности отказа в зависимости от времени

3. Задачи кластеризации

- Выявление типичных режимов работы оборудования
- Сегментация оборудования по характеру эксплуатации
- Группировка событий для выявления причинно-следственных связей

Как отмечает Горячев А.С. в статье «Обзор алгоритмов машинного обучения в задачах предиктивного анализа работы технологического оборудования» (2024), наиболее распространенными алгоритмами в этой области являются:

Для задач классификации:

- Логистическая регрессия (как базовая модель)
- Случайный лес (Random Forest) и градиентный бустинг (XGBoost, LightGBM)
- Искусственные нейронные сети, включая глубокие архитектуры
- Метод опорных векторов (SVM)

Для задач регрессии:

- Линейная и полиномиальная регрессия
- Регрессионные деревья и их ансамбли
- Рекуррентные нейронные сети (RNN, LSTM, GRU) для временных рядов
- Модели на основе гауссовских процессов

Для детекции аномалий:

- Изолирующий лес (Isolation Forest)
- Метод главных компонент (PCA)
- Автоэнкодеры и другие нейросетевые архитектуры
- One-class SVM

В работе Яковлева М. В., Шалина А. И. "Разработка рекомендаций по внедрению предиктивного обслуживания оборудования на высокотехнологичных предприятиях" (2023) описывается следующий типовой процесс разработки:

1. Сбор и подготовка данных

- Определение необходимых источников данных
- Организация процесса сбора (периодичность, объем)
- Очистка и предварительная обработка

2. Разведочный анализ данных (EDA)

- Анализ распределений признаков
- Выявление корреляций между параметрами
- Визуализация данных

3. Инженерия признаков

- Создание новых признаков на основе экспертных знаний
- Агрегация данных по временным окнам
- Извлечение признаков из временных рядов (статистические, спектральные, вейвлетные)

4. Выбор и обучение моделей

- Разделение данных на обучающую и тестовую выборки
- Кросс-валидация для оценки обобщающей способности
- Подбор гиперпараметров (GridSearch, Bayesian Optimization)

5. Оценка и интерпретация результатов

- Выбор метрик оценки в зависимости от специфики задачи
- Анализ важности признаков
- Объяснение предсказаний (SHAP, LIME)

6. Развёртывание и мониторинг

- Интеграция модели в производственную среду
- Мониторинг качества предсказаний
- Периодическое переобучение моделей

Данные в задачах предиктивного обслуживания обладают рядом особенностей, которые необходимо учитывать при разработке ML-моделей:

1. Несбалансированность классов

- Существенное преобладание нормальных состояний над аварийными
- Необходимость применения специальных методов (SMOTE, ADASYN, взвешивание классов)
- Выбор соответствующих метрик оценки (Precision, Recall, F1-score, AUC-ROC)

2. Временная структура данных

- Наличие автокорреляций в временных рядах
- Сезонные и циклические компоненты
- Необходимость учета контекста последовательности событий

3. Разнородность источников информации

- Комбинирование данных различных датчиков (температура, вибрация, давление)
- Интеграция структурированных (сенсорных) и неструктурных (текстовых) данных
- Различные частоты сбора данных с разных источников

4. Дрейф концепций (Concept Drift)

- Изменение статистических свойств данных со временем
- Деградация точности моделей при изменении условий эксплуатации
- Необходимость регулярного мониторинга и переобучения [34].

Актуальные подходы к прогнозированию технических отказов на основе данных с оборудования охватывают несколько направлений машинного обучения, среди которых особую роль играют методы анализа временных рядов, архитектуры глубокого обучения, техники обработки несбалансированных данных, а также средства интерпретации моделей и оценки их качества.

1. Методы обработки временных рядов. Для извлечения значимых признаков из временных данных, поступающих с сенсоров промышленного оборудования, применяются:

- Скользящие статистики (среднее значение, дисперсия, коэффициенты асимметрии и эксцесса)
- Спектральный анализ (быстрое преобразование Фурье (FFT), вейвлет-преобразования для выделения частотных компонентов)
- Выделение трендов и сезонных компонент (STL-декомпозиция и аналогичные методы временного сглаживания)

2. Архитектуры глубокого обучения для временных данных. Глубокие нейронные сети позволяют извлекать сложные зависимости в данных:

- LSTM и GRU - рекуррентные сети для анализа долгосрочных зависимостей;
- 1D-CNN - сверточные сети для выявления локальных паттернов в последовательностях;
- Transformer-архитектуры - современные модели, демонстрирующие высокую эффективность в задачах с длинными временными рядами.

3. Методы работы с несбалансированными данными. Отказы оборудования обычно редки, что приводит к несбалансированности классов. Для компенсации этого используются:

- Техники ресэмплинга (повышение/понижение частоты классов)
- Генеративные модели для синтеза примеров миноритарного класса (алгоритмы SMOTE, ADASYN, генеративные нейросети (GAN))
- Ансамблевые методы адаптированные к дисбалансу (например, Weighted Random Forest, Balanced Bagging).

4. Подходы к интерпретации моделей. В промышленных системах важно не только точное предсказание, но и объяснение результатов:

- Оценка важности признаков (feature importance);

- Методы локальной интерпретации: SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations);
 - Визуализация активаций для понимания работы сверточных и рекуррентных сетей.
- Выбор корректных метрик для оценки качества моделей технического обслуживания и ремонта (ТОиР) имеет решающее значение, особенно в условиях промышленного применения. Разные типы задач требуют различного подхода к измерению эффективности — от классификации с редкими отказами до прогноза остаточного ресурса (RUL) и оценки бизнес-эффекта.

Рассмотрим три ключевые группы метрик:

1. Метрики бинарной классификации. Эти метрики используются, когда задача заключается в определении, наступит ли отказ оборудования в ближайший момент времени (например, в течение следующего часа или смены). Подобные задачи часто решаются с помощью моделей бинарной классификации.
 - Точность (Precision) – доля правильных предсказаний отказа среди всех срабатываний модели. Высокая точность означает, что большинство сигналов тревоги действительно соответствовали реальным отказам.
 - Полнота (Recall) – доля выявленных отказов среди всех фактически произошедших. Высокое значение recall важно, чтобы не пропустить потенциально опасные события.
 - F1-мера – гармоническое среднее между точностью и полнотой. Позволяет сбалансировать два этих показателя при их компромиссе.
 - AUC-ROC (площадь под ROC-кривой) – отражает способность модели различать отказ и нормальное состояние при разных порогах классификации.
 - AUC-PR (площадь под PR-кривой) – особенно информативна в задачах с несбалансированными выборками, где классы "отказ" и "норма" представлены неравномерно.
2. Метрики для регрессионных задач (прогноз остаточного ресурса, RUL). Когда задача состоит в предсказании времени до отказа (в часах, циклах, днях и т.п.), применяется регрессионный подход. Здесь важно не просто угадать момент отказа, а как можно точнее его спрогнозировать.
 - MAE (Mean Absolute Error) – средняя абсолютная ошибка между предсказанным и фактическим временем до отказа.
 - RMSE (Root Mean Squared Error) – среднеквадратическая ошибка. Более чувствительна к сильным отклонениям и штрафует за крупные ошибки предсказания.

- MAPE (Mean Absolute Percentage Error) – выражает среднюю ошибку в процентах, что удобно для сравнения моделей на разных масштабах.
- Коэффициент детерминации (R^2) – показывает, какую долю дисперсии фактических значений объясняет модель. Значение, близкое к 1, свидетельствует о высокой предсказательной способности.

3. Бизнес-ориентированные метрики. В условиях промышленного применения оценка эффективности моделей не ограничивается только статистическими показателями. Необходимо учитывать экономические и производственные последствия решений, принятых на основе модели.

- Экономический эффект от предотвращенных отказов: рассчитывается как разность между затратами, которых удалось избежать, и стоимостью реализации прогноза (включая профилактику и замену).
- Процент ложных срабатываний и их стоимость: ложные тревоги могут приводить к ненужной остановке оборудования и убыткам. Метрика позволяет учитывать затраты на необоснованные действия.
- Среднее время предупреждения до фактического отказа (Mean Early Warning Time): показатель, демонстрирующий, насколько заранее система может предупредить о грядущем отказе. Чем больше это значение — тем выше ценность модели с точки зрения планирования и реагирования.

Применение алгоритмов машинного обучения (ML) в задачах технического обслуживания и ремонта (ТОиР) оборудования на промышленных предприятиях сталкивается с рядом серьёзных практических и методологических трудностей. Ниже представлены наиболее значимые из них, выявленные на основе обобщённого анализа научных публикаций, отраслевых отчетов и результатов внедрения решений в реальных производственных условиях.

1. Недостаток данных об отказах

- Малое количество примеров аварийных ситуаций. Отказы оборудования происходят относительно редко, и большинство промышленных систем работают стablyно в течение длительного времени. Это приводит к дисбалансу классов в данных, что затрудняет обучение надёжных моделей.
- Сложность моделирования редких или нестандартных отказов. Даже при наличии истории наблюдений, нестандартные или комплексные отказные сценарии часто не имеют достаточной представления в обучающей выборке.

- Высокая стоимость сбора и маркировки данных. Для получения качественных датасетов необходимо не только наличие датчиков и систем мониторинга, но и ручная верификация событий отказов специалистами, что требует времени и ресурсов.

2. Ограниченнная интерпретируемость моделей

- Необходимость объяснимости в принятии решений. В производственной среде важно не только предсказать отказ, но и объяснить, почему модель пришла к такому выводу, чтобы инженер мог принять обоснованные действия.
- Компромисс между точностью и интерпретируемостью. Сложные модели (например, градиентный бустинг или нейросети) часто демонстрируют высокую точность, но "чёрный ящик" таких решений затрудняет доверие со стороны экспертов.
- Требования к прозрачности в критически важных системах. В отраслях с повышенной ответственностью (например, энергетика, транспорт, химическая промышленность) регуляторы могут требовать использования только интерпретируемых решений.

3. Адаптация к изменяющимся условиям

- Деградация качества при изменении режимов работы. Промышленное оборудование может эксплуатироваться в различных режимах и условиях (температура, нагрузка и т.д.), что влияет на распределение данных и точность модели.
- Сложность переноса моделей между объектами. Даже идентичные агрегаты могут иметь различия в износе, конфигурации или операционном контексте. Модели, обученные на одном объекте, часто требуют перенастройки или переобучения при переносе на другой.
- Необходимость постоянного мониторинга качества. Производственные условия динамичны, и без системы автоматического контроля и обновления модели возможно ухудшение её предсказательной способности (data drift, concept drift).

4. Вычислительные и инфраструктурные ограничения

- Потребность в обработке данных в реальном времени. Для предиктивного обслуживания важно не только прогнозировать, но и реагировать быстро — например, выдавать предупреждение задолго до отказа. Это требует низкой задержки и высокой скорости обработки.
- Ограниченные ресурсы на промышленных узлах. В ряде случаев модели должны работать на встроенных устройствах или контроллерах (Edge AI), где доступно ограниченное количество памяти и вычислительной мощности.

- Баланс между точностью и эффективностью. Высокоточные модели могут требовать больших объёмов памяти и времени на инференс, что не всегда допустимо в условиях ограничений по питанию, размеру и скорости.

1.4 Обзор существующих решений в области предиктивного обслуживания

Согласно аналитическому отчету "Глобальный рынок оперативного предиктивного обслуживания — тенденции отрасли и прогноз до 2029 года", на мировом рынке лидирующие позиции занимают следующие решения:

1. Siemens MindSphere

- Характеристики: Облачная платформа промышленного IoT с модулями предиктивной аналитики
- Особенности: Интеграция с оборудованием Siemens, открытые API для сторонних устройств
- Применение в России: Используется на крупных предприятиях энергетического и машиностроительного секторов

2. IBM Maximo APM (Asset Performance Management)

- Характеристики: Система управления активами с функциями предиктивного обслуживания
- Особенности: Интеграция с технологиями искусственного интеллекта IBM Watson
- Применение в России: Внедрена на ряде предприятий нефтегазового сектора

3. GE Predix

- Характеристики: Платформа для создания приложений промышленного интернета
- Особенности: Специализированные решения для энергетики, авиации, здравоохранения
- Применение в России: Ограниченнное использование в энергетическом секторе

4. SAP Predictive Maintenance and Service

- Характеристики: Модуль в составе решений SAP для управления предприятием
- Особенности: Тесная интеграция с ERP-системами SAP
- Применение в России: Внедрено на предприятиях, использующих SAP в качестве основной ERP-системы

Анализ текущего состояния отечественных разработок в сфере предиктивного обслуживания позволяет выделить следующие перспективные отечественные решения:

1. "ПРАНА" (АО "POTEK")

- Характеристики: Система прогностики и удаленного мониторинга промышленного оборудования
- Особенности: Фокус на энергетическом оборудовании, собственные математические модели
- Применение: Более 100 энергоблоков по всей России

2. "Диспетчер" (ООО "Цифра")

- Характеристики: Система мониторинга промышленного оборудования с элементами предиктивной аналитики
- Особенности: Специализация на станочном парке, интеграция с отечественными АСУТП
- Применение: Более 200 промышленных предприятий России

3. "Монитор" (Ctrl2GO)

- Характеристики: Платформа для предиктивного обслуживания локомотивов и промышленного оборудования
- Особенности: Специализированные решения для железнодорожного транспорта
- Применение: РЖД, металлургические предприятия

4. "Атом.Мониторинг" (Росатом)

- Характеристики: Система предиктивной диагностики для атомной промышленности
- Особенности: Высокие требования к надежности, специализированные модели для критически важного оборудования
- Применение: Предприятия атомной отрасли

На основании информации с электронного ресурса Open source predictive analytics tools для реализации систем предиктивного обслуживания рассматриваются следующие open-source решения с открытым исходным кодом:

1. Predictive Maintenance Toolbox для MATLAB

- Характеристики: Набор инструментов для разработки моделей предиктивного обслуживания
- Преимущества: Обширная документация, готовые примеры для различных типов оборудования
- Ограничения: Базовая функциональность требует коммерческой лицензии MATLAB

2. PyCaret

- Характеристики: Библиотека Python для автоматизации процесса машинного обучения

- Преимущества: Быстрое создание прототипов моделей, автоматический подбор параметров
- Ограничения: Ограниченная поддержка обработки временных рядов, отсутствие специализированных инструментов для промышленных данных

3. Scikit-learn

- Характеристики: Фундаментальная библиотека машинного обучения на Python
- Преимущества: Широкий набор алгоритмов, хорошая документация
- Ограничения: Отсутствие специализированных инструментов для предиктивного обслуживания, необходимость самостоятельной реализации многих компонентов

4. Prophet (Facebook)

- Характеристики: Библиотека для прогнозирования временных рядов
- Преимущества: Учет сезонности, праздников, хорошая работа с пропусками в данных
- Ограничения: Ориентация на бизнес-временные ряды, а не на технические данные

5. NILM-tools

- Характеристики: Инструменты для мониторинга и анализа энергопотребления оборудования
- Преимущества: Специализация на энергетических данных, обширная база алгоритмов
- Ограничения: Узкая специализация, сложность интеграции с промышленными системами [59].

В таблице приведено краткое сравнение различных решений по ключевым характеристикам:

Критерий	Коммерческие зарубежные	Коммерческие российские	Open-source
Точность прогнозирования	Высокая	Средняя/Высокая	Средняя
Масштабируемость	Высокая	Средняя	Низкая/Средняя
Интеграция с IT-системами	Широкие возможности	Ограничена	Требует доработки
Поддержка различных типов данных	Обширная	Специализированная	Базовая
Стоимость внедрения	Высокая	Средняя	Низкая
Стоимость владения	Высокая	Средняя	Низкая/Средняя
Требования к квалификации	Средние	Средние	Высокие

Локализация	Ограниченнaя	Полная	Требует доработки
-------------	--------------	--------	-------------------

Таблица 2: Сравнение различных решений систем предиктивного обслуживания по ключевым характеристикам

Согласно аналитическим прогнозам, представленным в обзоре McKinsey & Company (2023) и аналитическому отчету Tadviser, ключевыми направлениями развития систем предиктивного обслуживания в горизонте до 2030 года являются:

1. Интеграция с цифровыми двойниками. По данным MarketsandMarkets, мировая индустрия цифровых двойников ежегодно растёт на 60%, а к 2027 году её оборот увеличится до 73,5 млрд долларов [72].
 - Создание виртуальных моделей оборудования для моделирования различных сценариев
 - Синтез данных для обучения моделей в условиях недостатка реальных данных
2. Федеративное обучение
 - Распределенное обучение моделей без передачи чувствительных данных
 - Возможность использования данных с разных предприятий при сохранении конфиденциальности
3. Мультимодальные модели
 - Интеграция данных различной природы (сенсорных, текстовых, визуальных)
 - Применение методов обработки естественного языка для анализа технической документации
4. Автоматизированное машинное обучение (AutoML)
 - Снижение порога входа для специалистов без глубоких знаний в области ML
 - Автоматический подбор архитектур и гиперпараметров моделей
5. Edge AI
 - Перенос части вычислений на граничные устройства (непосредственно на оборудовании)
 - Снижение требований к каналам связи и повышение оперативности реагирования

Анализ существующих решений в области предиктивного обслуживания показывает значительное разнообразие подходов и технологий, от универсальных промышленных платформ до специализированных отраслевых решений. Машинное обучение играет ключевую роль в современных системах, обеспечивая возможность выявления скрытых закономерностей и формирования точных прогнозов.

При этом сохраняется ряд вызовов, связанных с особенностями данных в промышленных системах, необходимостью интерпретации результатов и адаптации к изменяющимся условиям. Российский рынок решений предиктивного обслуживания находится в стадии активного развития, с появлением отечественных платформ, ориентированных на специфические отрасли промышленности.

В перспективе ожидается дальнейшая интеграция систем предиктивного обслуживания с другими компонентами цифрового производства, развитие методов федеративного обучения и перенос части вычислений на граничные устройства.

Глава 2. Анализ данных и разработка методологии

2.1. Описание и анализ исходного набора данных

Исходный набор данных "Machine Failure Classification using sensor data" с платформы Kaggle представляет собой коллекцию сенсорных измерений для прогнозирования риска отказа промышленного оборудования. Данный датасет содержит информацию о работе различных типов промышленных машин и предназначен для разработки систем предиктивного обслуживания.

Общие характеристики датасета:

- Объем: 1000 строк
- Формат: структурированные табличные данные
- Тип задачи: бинарная классификация (норма/риск отказа)
- Баланс классов: 70% нормальной работы (0), 30% риска отказа (1)

Набор данных включает следующие признаки:

1. Temperature (°C) - непрерывная переменная, отражающая температуру в критических точках машины. Повышенные значения могут указывать на потенциальные проблемы из-за перегрева.
2. Vibration (Hz) - частота вибраций машины. Аномальные вибрации могут сигнализировать о механических проблемах, таких как нарушение центровки, дисбаланс или износ.
 - Диапазон значений: от 35.3 Hz до 66 Hz
 - Среднее значение: 50.4 Hz
 - Стандартное отклонение: 5 Hz
3. Power_Usage (kW) - уровень потребления энергии машиной. Скачки в энергопотреблении могут указывать на повышенную нагрузку или потенциальные механические проблемы.
 - Диапазон значений: от 4 kW до 17.9 kW
 - Среднее значение: 10 kW
 - Стандартное отклонение: 1.97 kW
4. Humidity (%) - влажность окружающей среды вокруг машины. Высокие уровни влажности могут влиять на работу машины и в долгосрочной перспективе приводить к отказам.

- Диапазон значений: от 15.4% до 46.2%
 - Среднее значение: 29.9%
 - Стандартное отклонение: 5.13%
5. Machine_Type - категориальная переменная, указывающая тип машины:
- Lathe (токарный станок): 34% образцов
 - Drill (сверлильный станок): 33% образцов
 - Other (другие типы): 33% образцов
6. Failure_Risk - целевая переменная (бинарная):
- 0: нормальная работа (70% образцов)
 - 1: риск отказа (30% образцов)

Особенности датасета:

1. Сбалансированность по типам машин: датасет демонстрирует почти равномерное распределение разных типов оборудования, что обеспечивает репрезентативность для различных классов машин.
2. Несбалансированность целевой переменной: соотношение классов составляет 70% к 30% (нормальная работа vs. риск отказа), что создает умеренный дисбаланс классов, требующий внимания при построении моделей.
3. Полнота данных: в датасете отсутствуют пропущенные значения, что упрощает процесс предварительной обработки данных.
4. Разнородность признаков: датасет включает как непрерывные (сенсорные измерения), так и категориальные (тип машины) признаки, что требует соответствующих подходов к предобработке.
5. Реальный диапазон значений: значения признаков находятся в реалистичном диапазоне для промышленного оборудования, что повышает практическую применимость результатов.

Данный набор данных является репрезентативным примером данных, которые могут быть собраны в реальных производственных условиях для задач предиктивного обслуживания. Его структура позволяет исследовать влияние различных факторов на риск отказа оборудования и разрабатывать модели машинного обучения для раннего обнаружения потенциальных проблем.

2.1.1. Исследовательский анализ данных (EDA)

Исследовательский анализ данных (EDA) представляет собой ключевой этап в разработке систем предиктивного обслуживания, позволяющий выявить характерные распределения признаков, потенциальные выбросы, скрытые закономерности и взаимосвязи между переменными.

В рамках данного дипломного проекта был проведён комплексный анализ сенсорных данных, включающий расчет описательной статистики, частотных распределений, обнаружение выбросов, корреляционный и многомерный анализ.

Анализ распределений непрерывных признаков

1. Температура (Temperature)

- Базовая статистика: среднее значение 70.2°C , стандартное отклонение 9.8°C , диапазон от 37.6 до 108.5°C .
- Распределение приблизительно симметрично ($\text{skewness} \approx 0.12$), с умеренной концентрацией данных в диапазоне $63\text{--}76^{\circ}\text{C}$.
- Куртозис близок к 0 (≈ 0.07), что указывает на близость к нормальному распределению без выраженных "тяжелых хвостов".
- Выбросы: по Z-оценке обнаружено 3 выброса, по методу межквартильного размаха (IQR) — 8.
- Интервальный анализ показывает, что около 50% всех температур укладываются в диапазон $65\text{--}76^{\circ}\text{C}$.

Вывод: Температура — один из ключевых признаков, и её высокие значения могут быть предвестниками отказов.

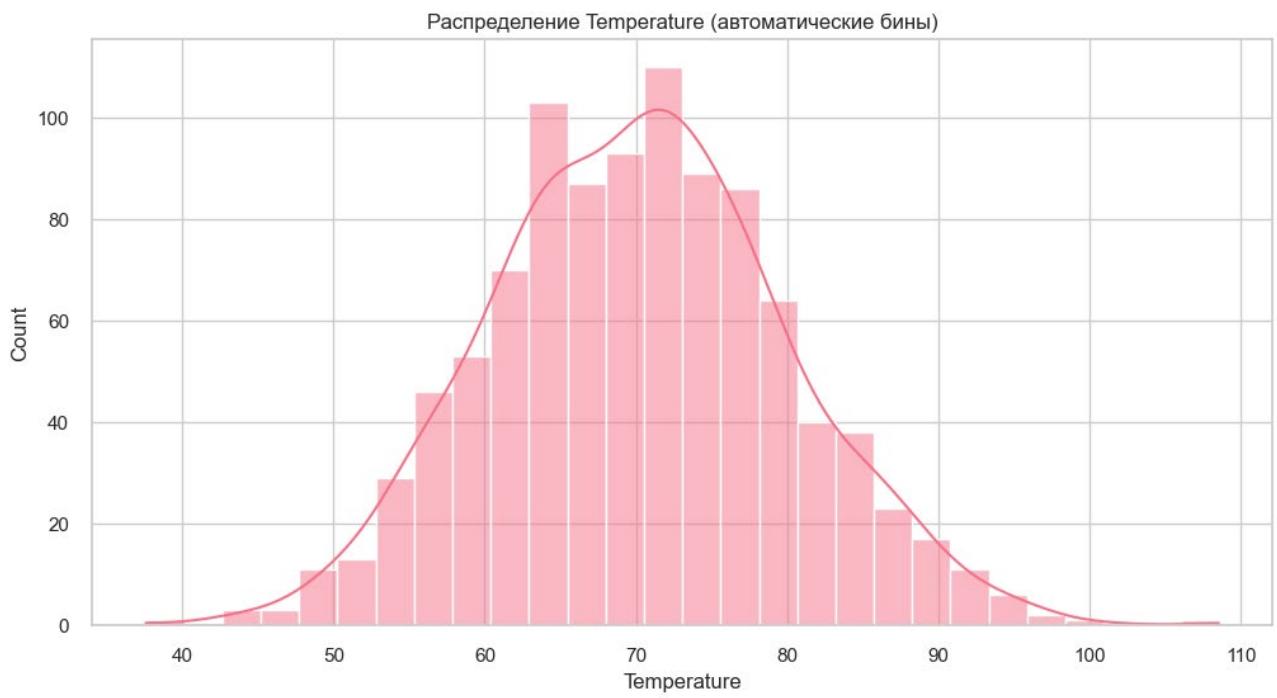


Рис 1: Распределение температуры

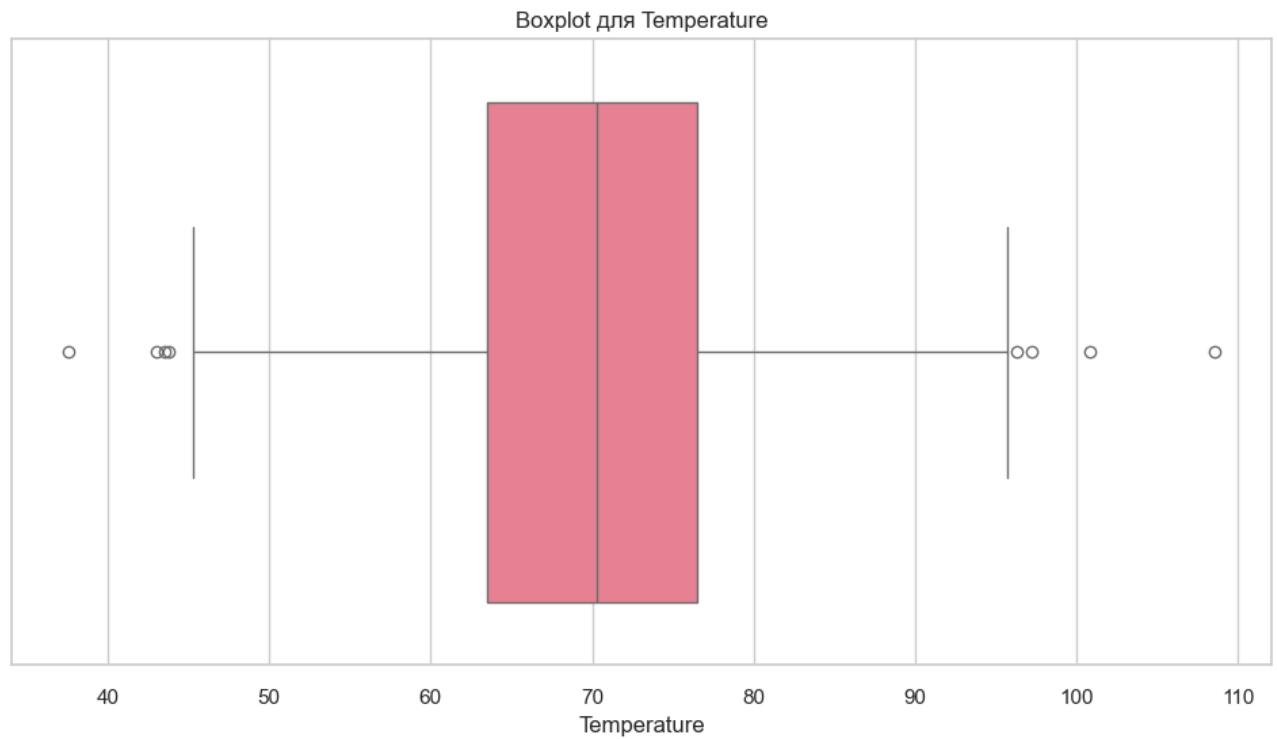
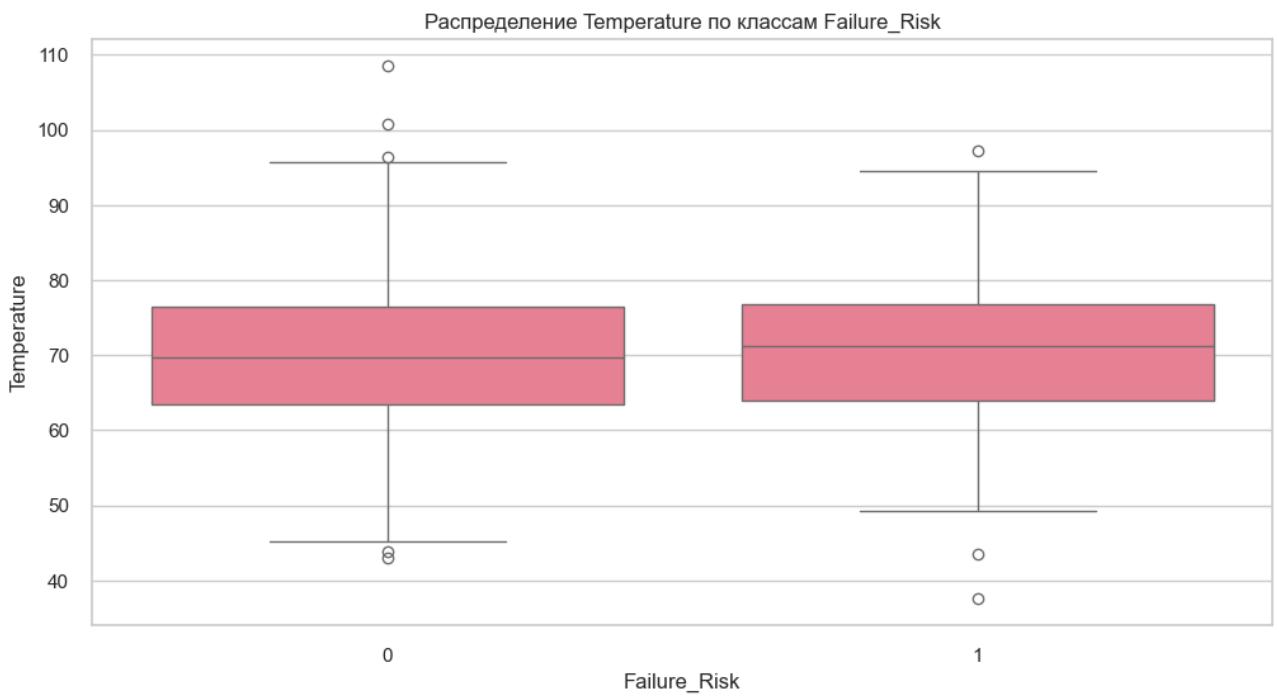


Рис 2: Boxplot для температуры

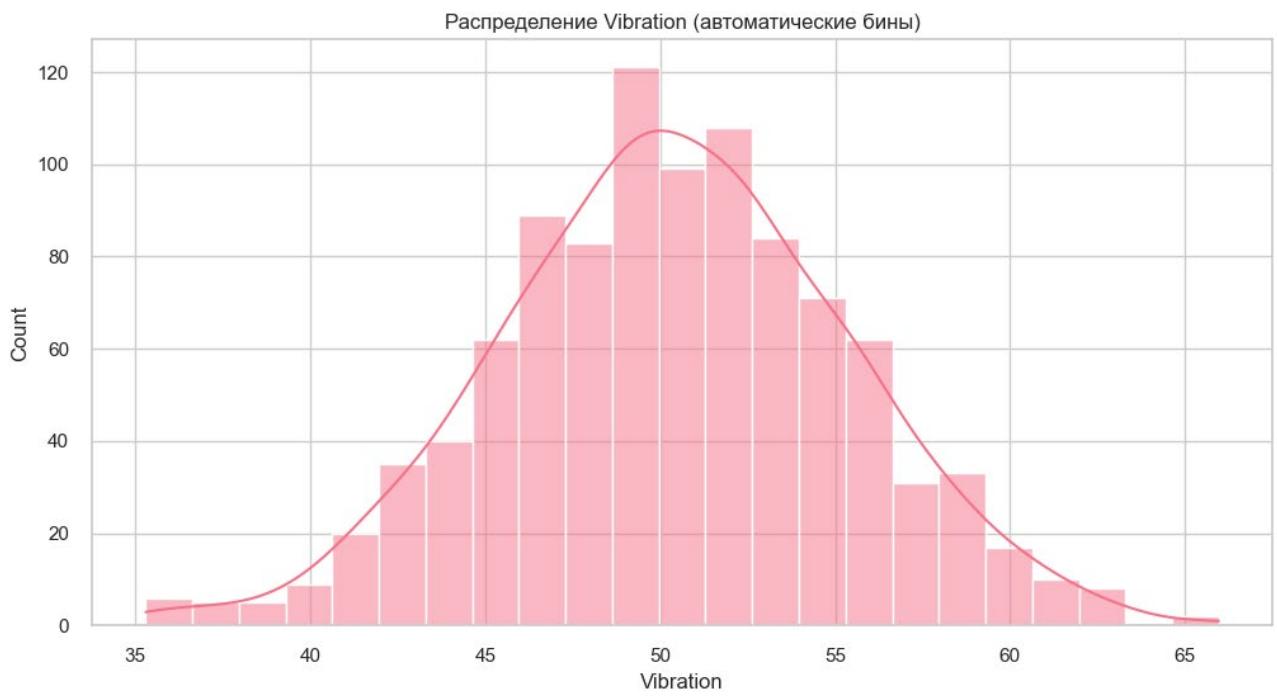


Ruc 3: Распределение температуры по Failure_Risk

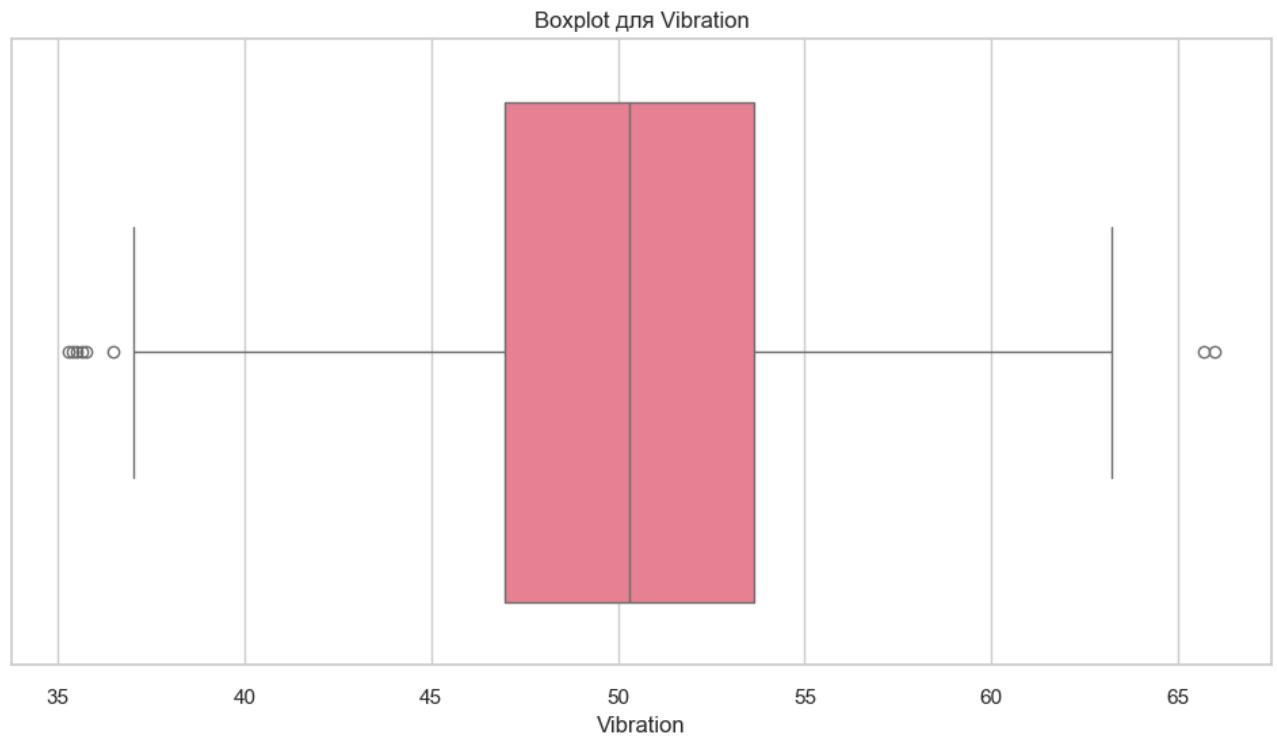
2. Вибрация (Vibration)

- Базовая статистика: среднее — 50.4 Гц, стандартное отклонение — 5 Гц, значения варьируются от 35.3 до 66 Гц.
- Распределение близко к нормальному ($\text{skewness} \approx -0.05$), выбросов немного.
- В интервале 47.5–53.7 Гц сосредоточено более 70% записей.
- Заметно, что значения выше 55 Гц чаще наблюдаются у экземпляров с меткой риска отказа.

Вывод: повышенная вибрация — потенциальный индикатор технических неисправностей.



Ruc 4: Распределение Vibration



Ruc 5: Boxplot для Vibration

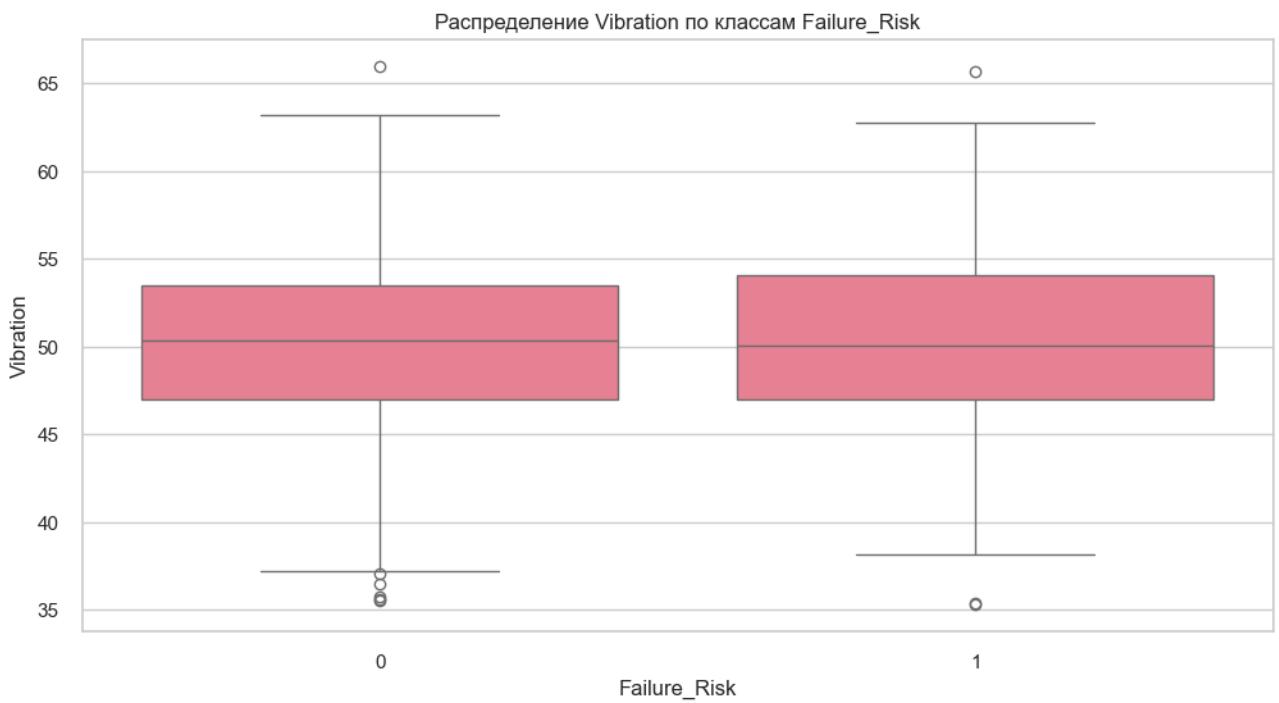


Рис 6: Распределение Vibration по классам Failure_Risk

3. Энергопотребление (Power_Usage)

- Среднее значение — 10 кВт, стандартное отклонение — 2 кВт, диапазон — от 3.96 до 17.85 кВт.
- Распределение с легкой положительной асимметрией ($\text{skewness} \approx 0.06$).
- Выбросы: 5 по Z-оценке и 10 по IQR — в основном среди высоких значений (>15 кВт).
- Основная масса значений ($\approx 70\%$) лежит в пределах 8–13 кВт.

Вывод: резкие скачки в потреблении энергии потенциально указывают на перегрузку или механическое заедание.

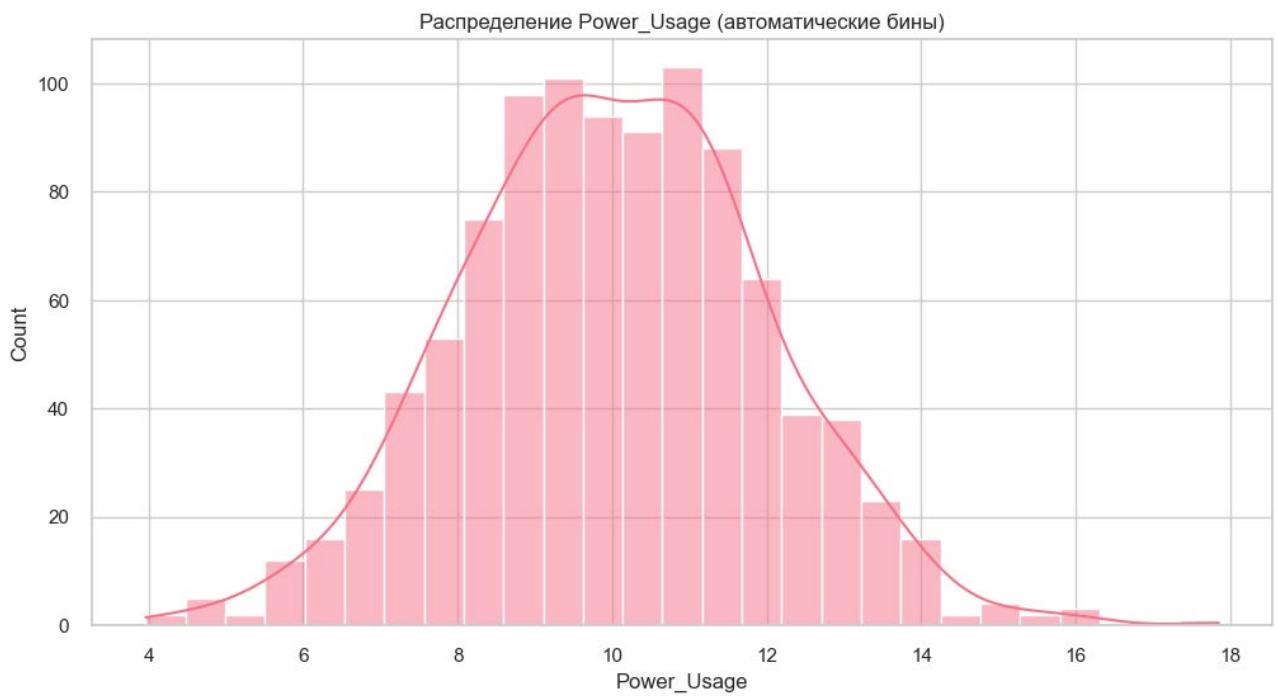


Рис 6: Распределение Power_Usage

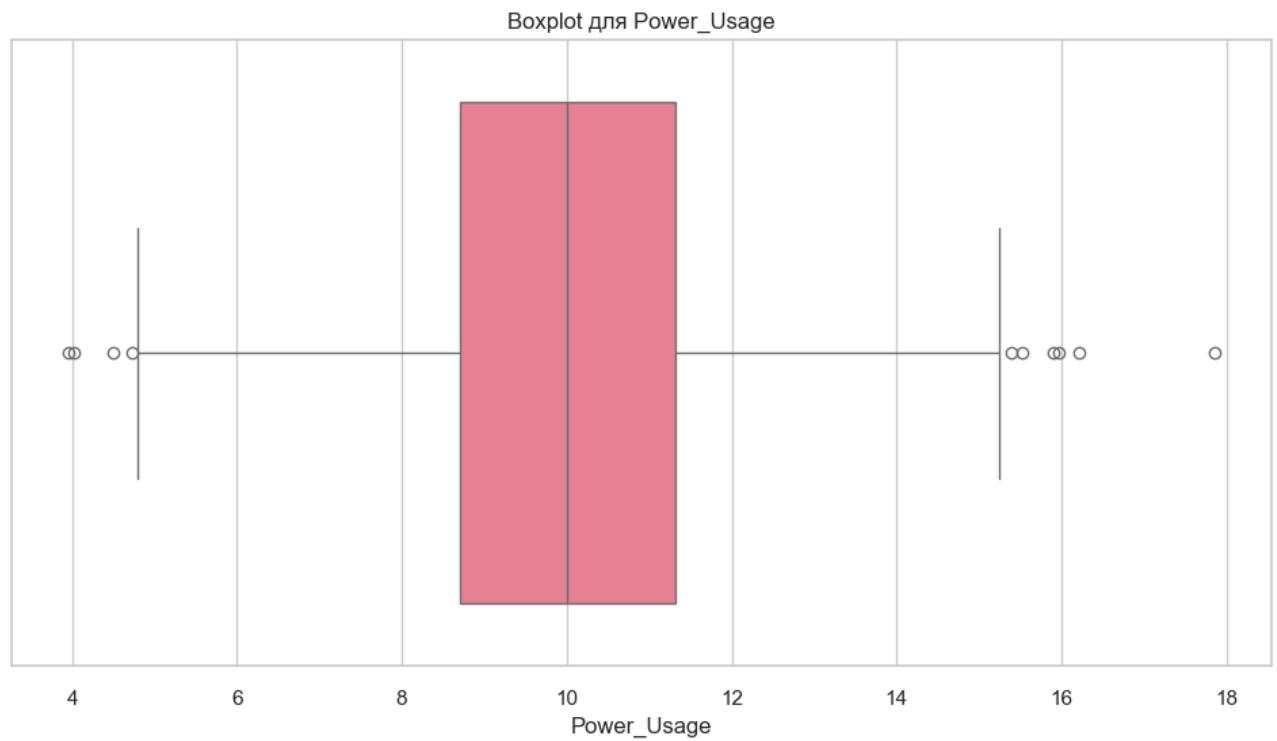


Рис 7: Boxplot для Power_Usage

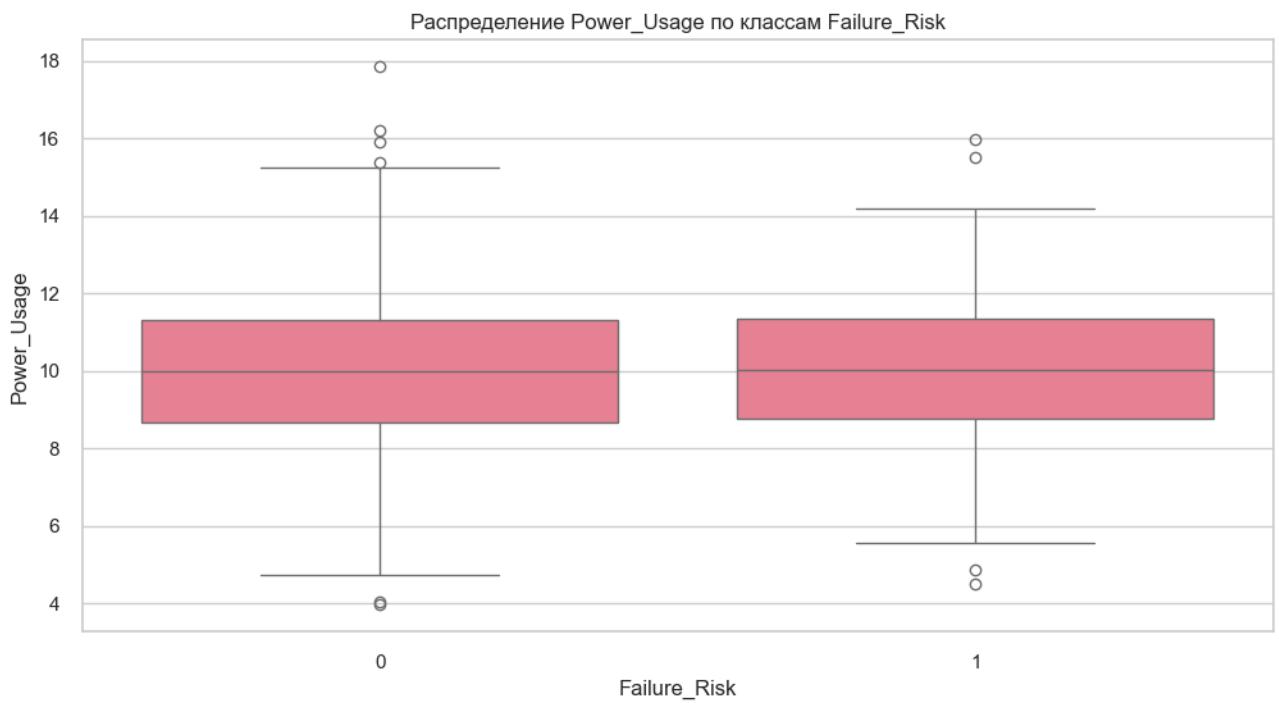


Рис 8: Распределение Power_Usage по классам Failure_Risk

4. Влажность (Humidity)

- Базовая статистика: среднее — 29.9%, стандартное отклонение — 5.1%, значения от 15.35 до 46.22%.
- Распределение практически симметричное ($\text{skewness} \approx 0.00$), с пиком около 30%.
- Выбросов немного: 2 (Z-score) и 5 (IQR).
- Связь с отказами не прослеживается напрямую в однофакторном анализе.

Вывод: влажность может играть роль косвенного фактора, влияющего в сочетании с другими переменными.

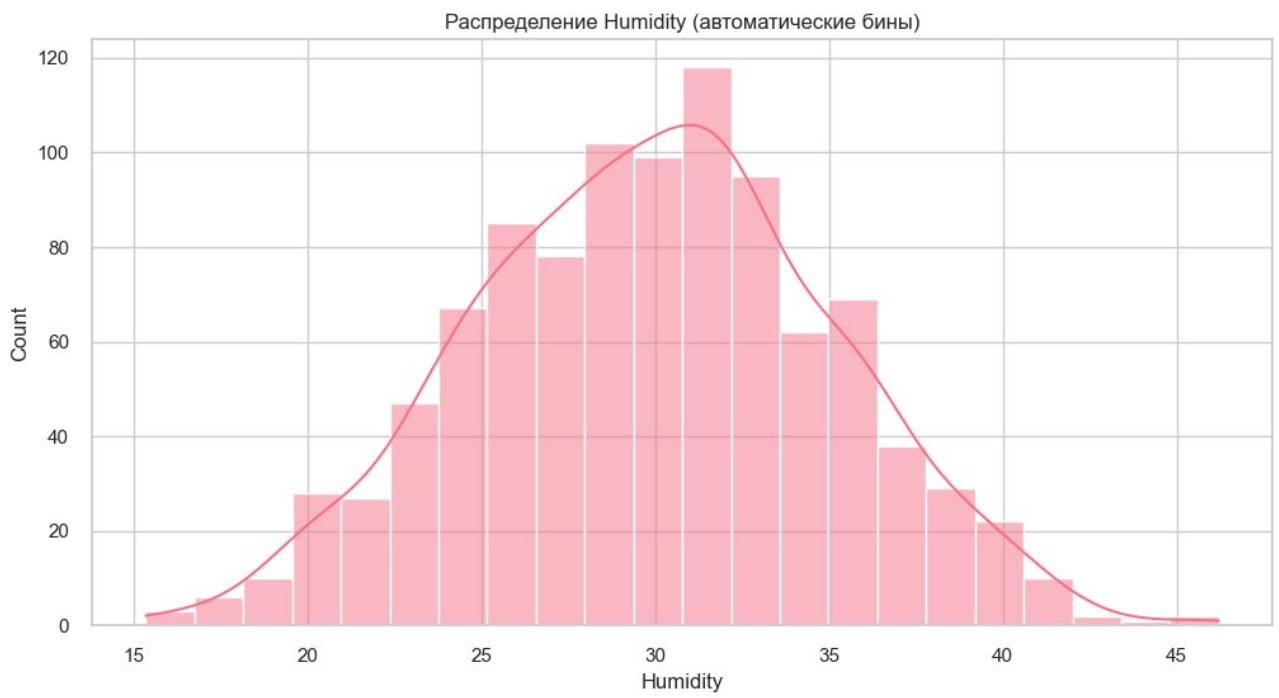


Рис 9: Распределение Humidity

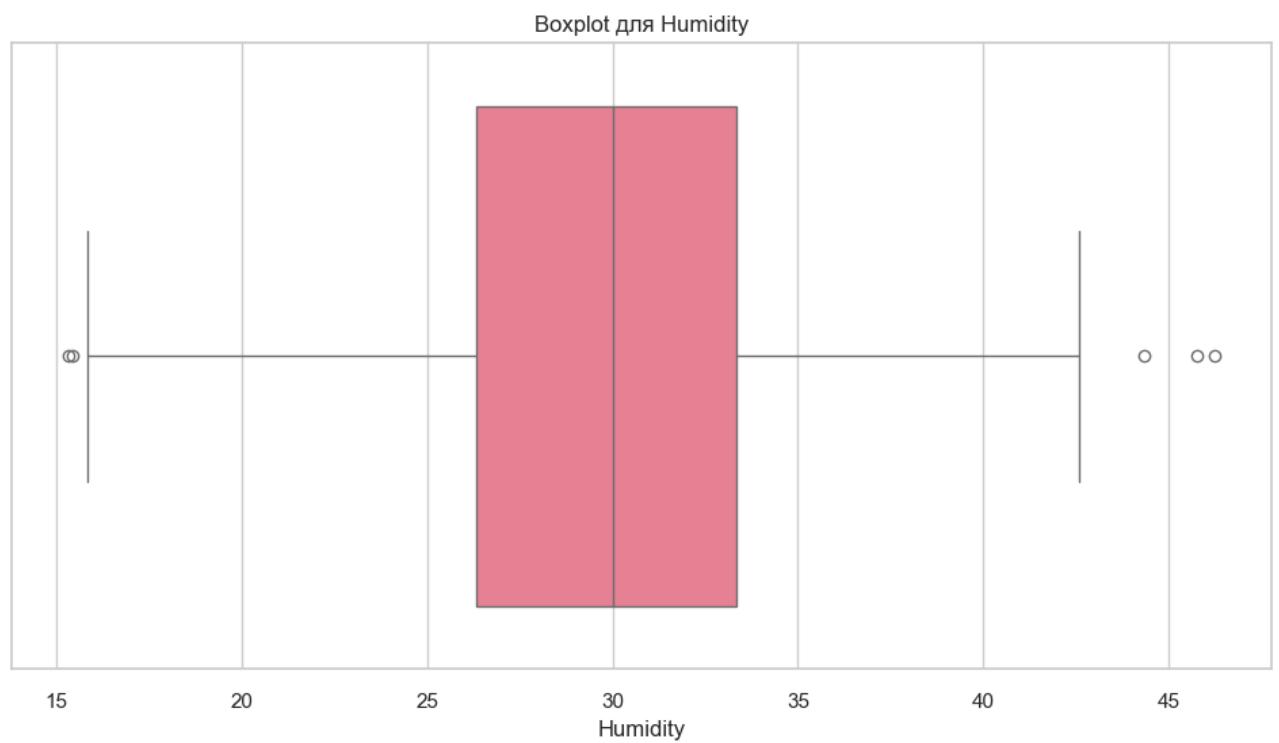


Рис 10: Boxplot для Humidity

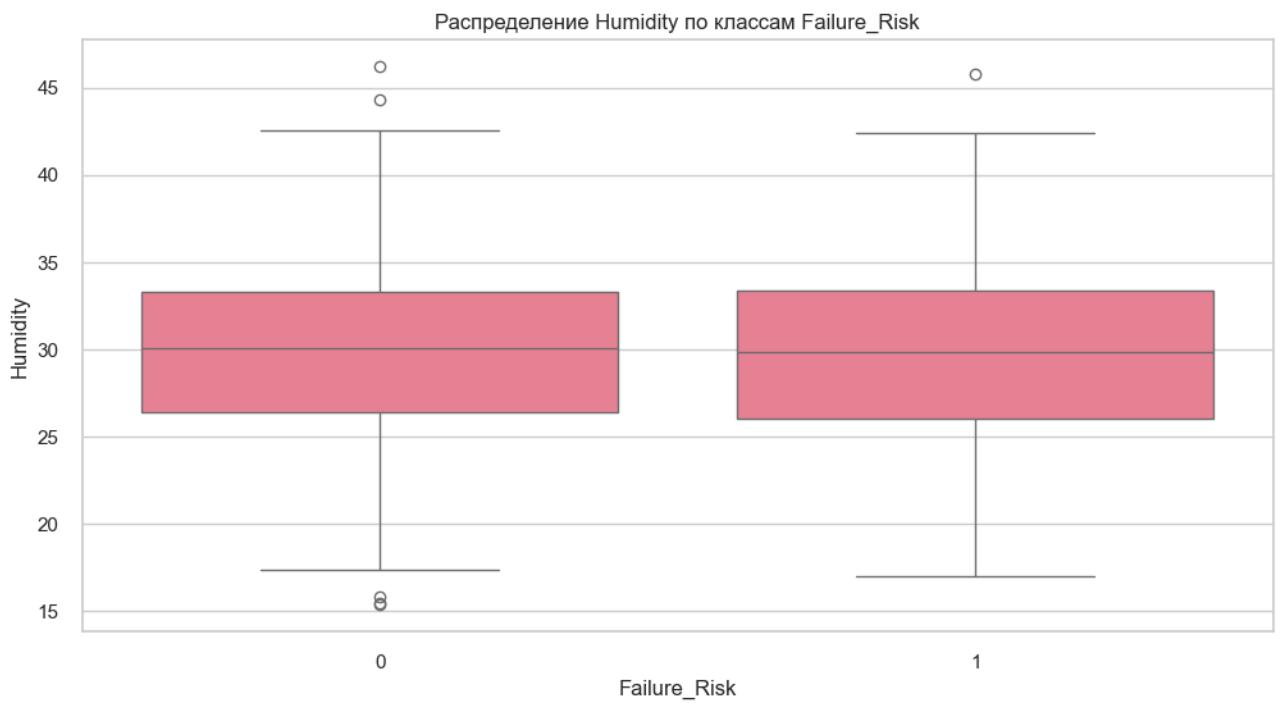


Рис 11: Распределение Humidity по классам Failure_Risk

В датасете представлен один категориальный признак — Machine_Type (тип оборудования), который требует особого подхода при анализе. Проведем его детальное исследование.

Тип оборудования (Machine_Type):

- В выборке представлены три типа оборудования:
 - Lathe (токарный станок) — 33.8%
 - Drill (сверлильный станок) — 33.4%
 - Mill (фрезерный станок) — 32.8%
- Частота отказов:
 - Drill и Lathe — по 31.1%
 - Mill — 27.7%

Вывод: тип оборудования оказывает влияние на частоту отказов, однако различия не критичны. Все типы имеют примерно равномерное распределение.

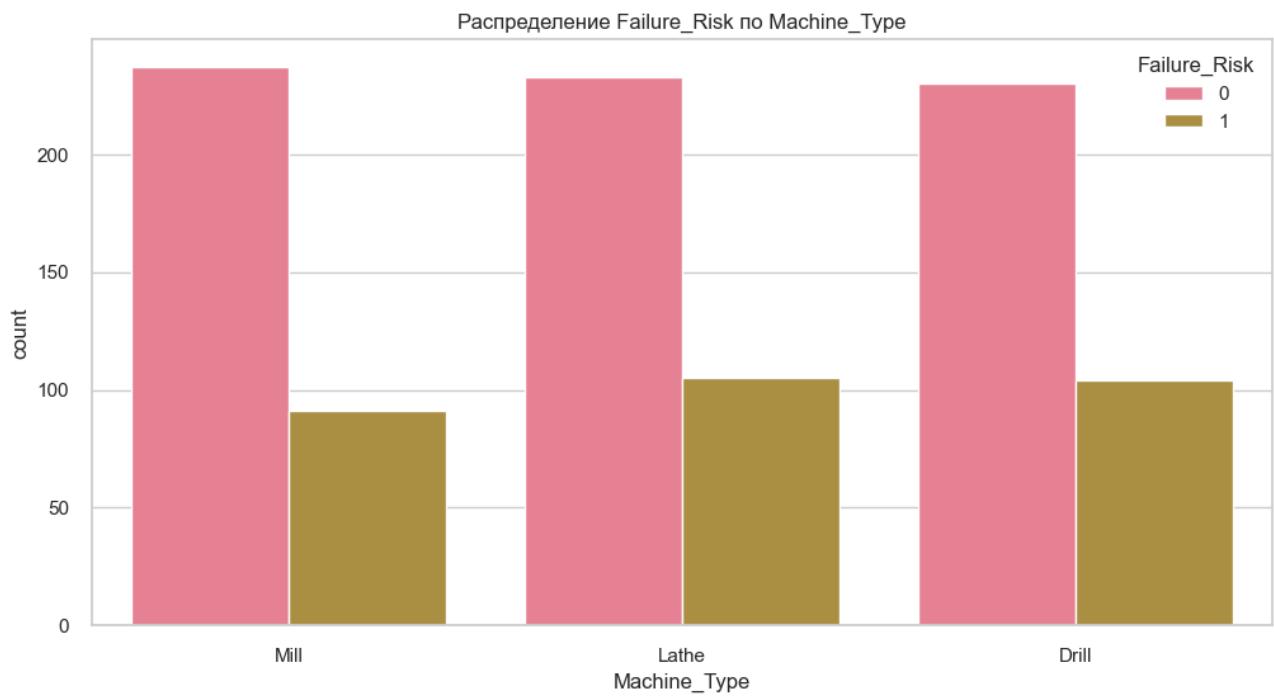


Рис 12: Распределение Failure_Risk по типу оборудования

Целевая переменная (Failure_Risk) представляет собой бинарный индикатор риска отказа оборудования:

- Класс 0 — 70% (нормальная работа), Класс 1 — 30% (риск отказа).
- Распределение классов умеренно несбалансировано, что потребует коррекции при обучении моделей.

Вывод: дисбаланс присутствует, но не критичный; требует применения методов балансировки (например, SMOTE).

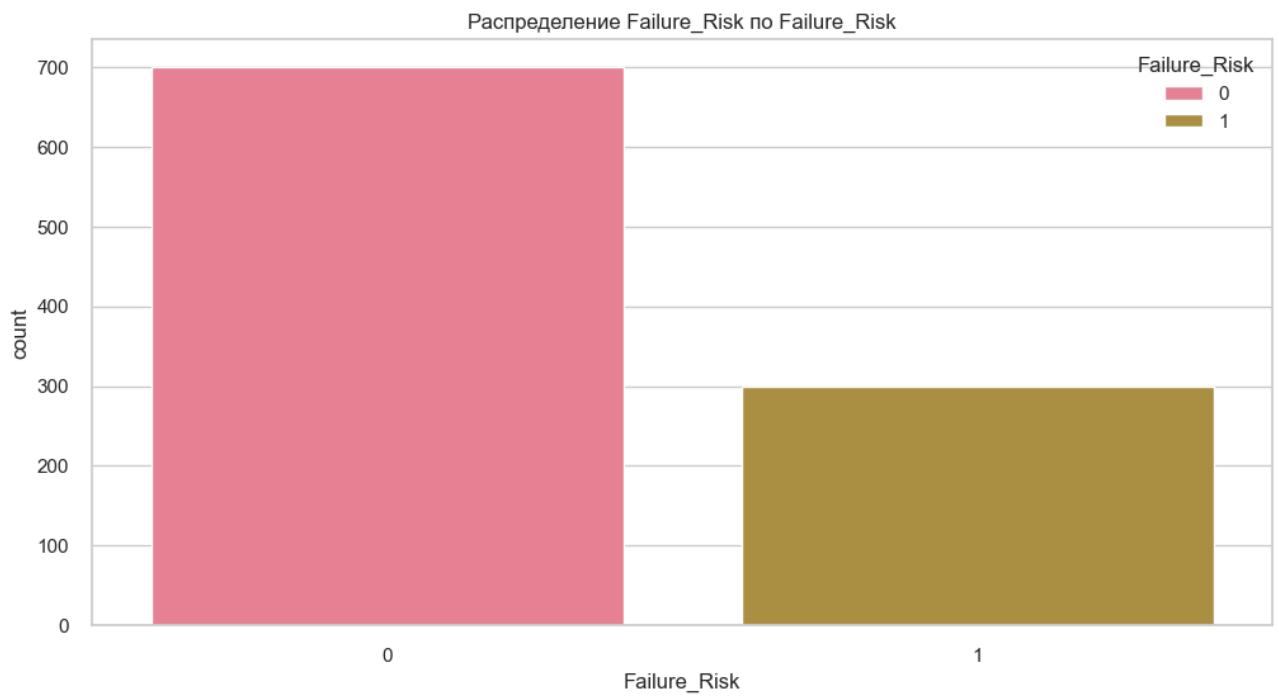


Рис 13: Распределение Failure_Risk

Для выявления аномальных значений в данных, проведён анализ выбросов и аномалий:

- Z-оценка выявила выбросы во всех признаках, в основном по Power_Usage.
- Метод IQR подтвердил наличие редких экстремальных значений, особенно в Temperature и Vibration.
- Isolation Forest обнаружил 134 аномалии, из которых более 70% принадлежали к классу Failure_Risk = 1.

Вывод: выбросы и аномалии потенциально информативны, их удаление может привести к потере сигналов об отказах.

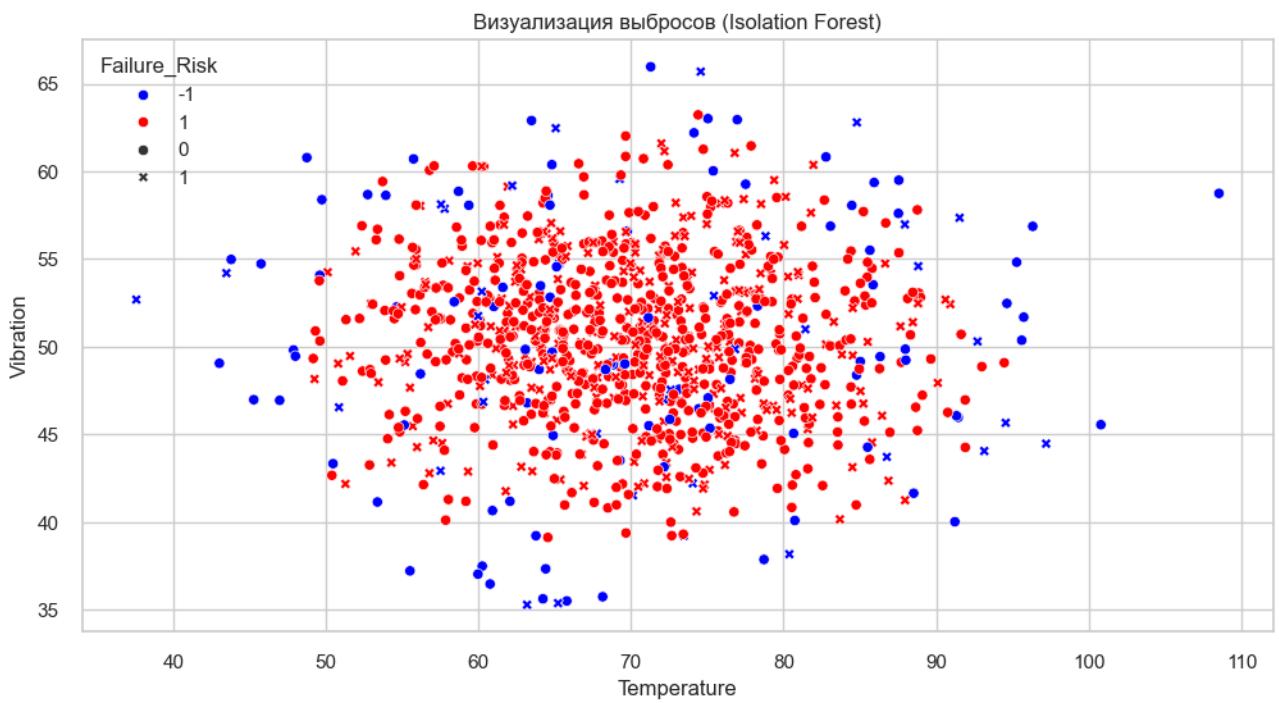


Рис 14: Визуализация выбросов Isolation forest

Для выявления взаимосвязей между технологическими параметрами и риском отказа был проведен комплексный корреляционный анализ.

Матрица Пирсона показала слабые линейные зависимости:

Пара признаков	Коэффициент r
Temperature – Failure_Risk	0.03
Vibration – Failure_Risk	-0.00
Power_Usage – Failure_Risk	0.02
Humidity – Failure_Risk	-0.02

- Все корреляции с целевой переменной очень слабые, линейная связь почти отсутствует.

Вывод: необходимо использовать нелинейные модели или методы многомерного анализа.

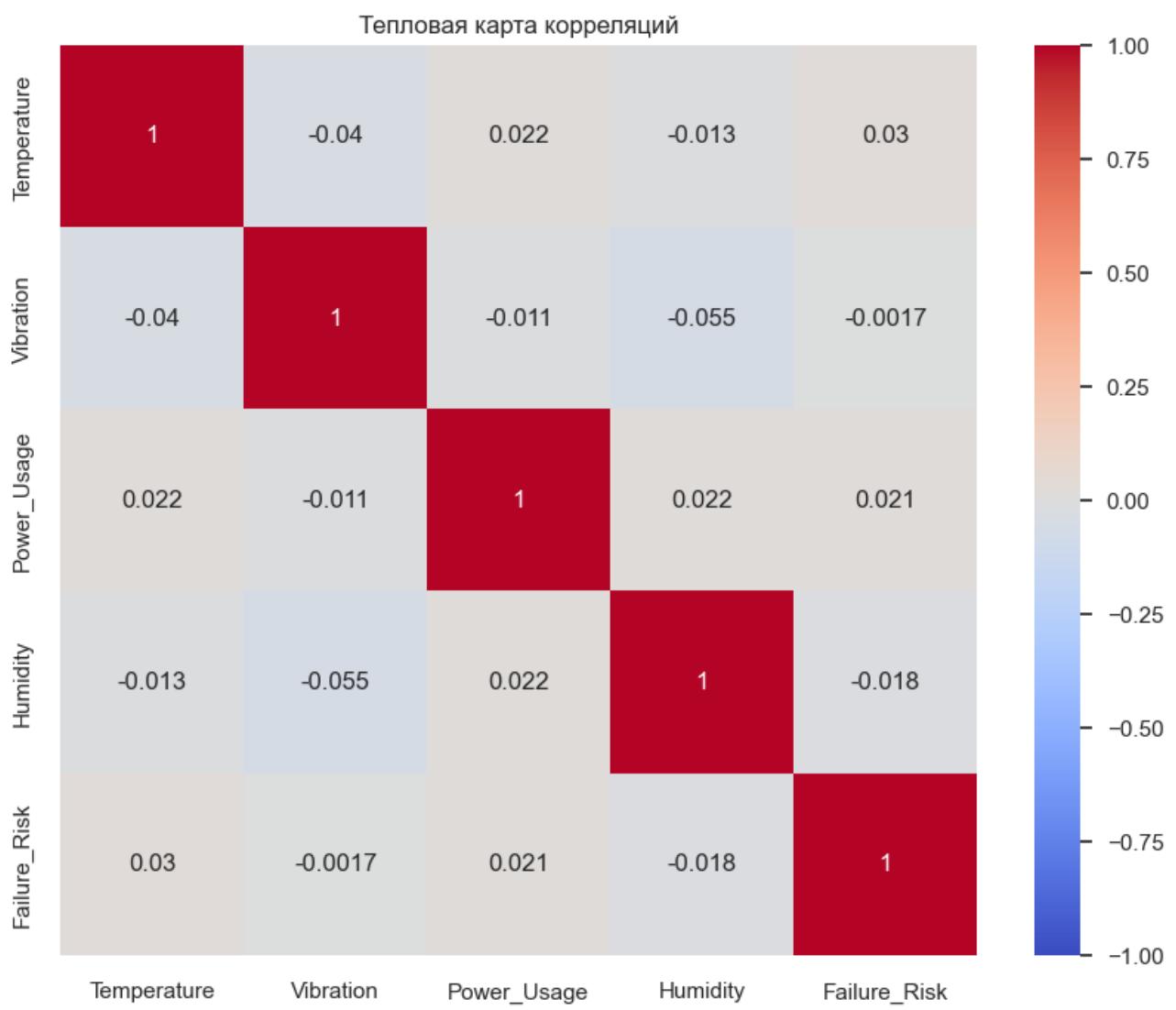


Рис 15: Тепловая карта корреляций

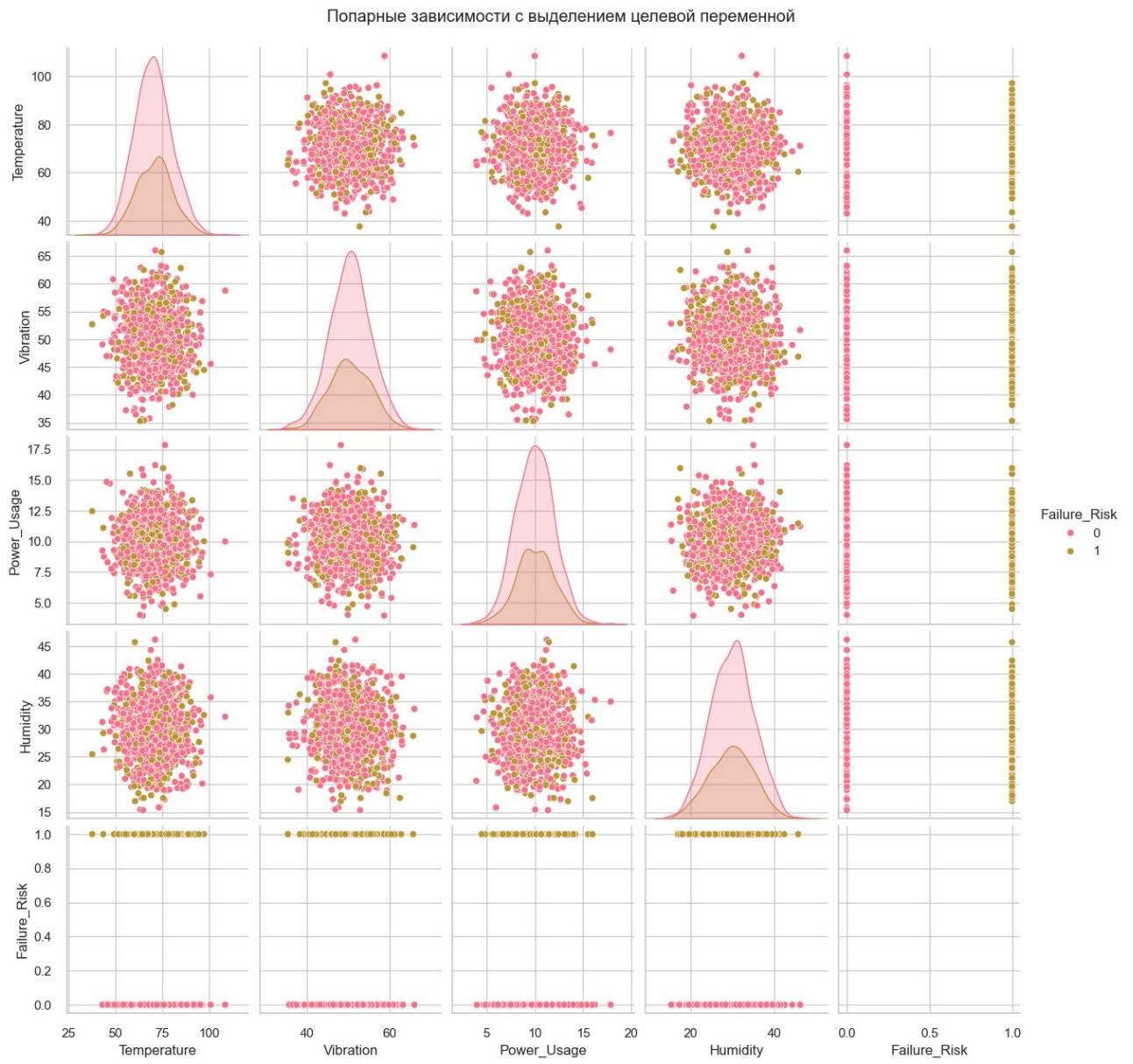


Рис 16: Попарные зависимости с выделением целевой переменной

EDA показал, что:

- Распределения всех признаков достаточно гладкие и приближены к нормальнм, но с признаками слабой асимметрии.
- Заметны аномалии и выбросы, чаще встречающиеся у экземпляров с риском отказа.
- Линейные корреляции с целевой переменной низкие, что указывает на сложный характер зависимости — вероятно, нелинейный.
- Использование многомерного анализа и нелинейных моделей оправдано и необходимо для повышения точности прогнозов.

Итоговые выводы и рекомендации:

1. Особенности данных:

- Слабая линейная корреляция между признаками.
- Умеренный дисбаланс классов (70/30).
- Наличие аномалий (13.4%), потенциально связанных с отказами.

2. Рекомендации по моделированию:

- Методы: Градиентный бустинг (XGBoost, LightGBM) или нейросети.
- Обработка дисбаланса: SMOTE или классовые веса.
- Признаки: Возможно стоит добавить бинарные флаги аномалий.

3. Дальнейшие шаги:

- Анализ временных рядов, если данные доступны.
- Проверка нелинейных зависимостей (взаимная информация).
- Интерпретация моделей с помощью SHAP-значений.

Проведенный EDA выявил сложный характер данных, типичный для реальных промышленных систем. Это подчеркивает необходимость использования продвинутых методов машинного обучения и дальнейшего сбора данных для повышения точности прогнозирования.

Для углубленного понимания структуры данных был проведен многомерный анализ с использованием методов снижения размерности и обнаружения аномалий. Данные были предварительно стандартизированы для устранения влияния различий в масштабах и единицах измерения параметров.

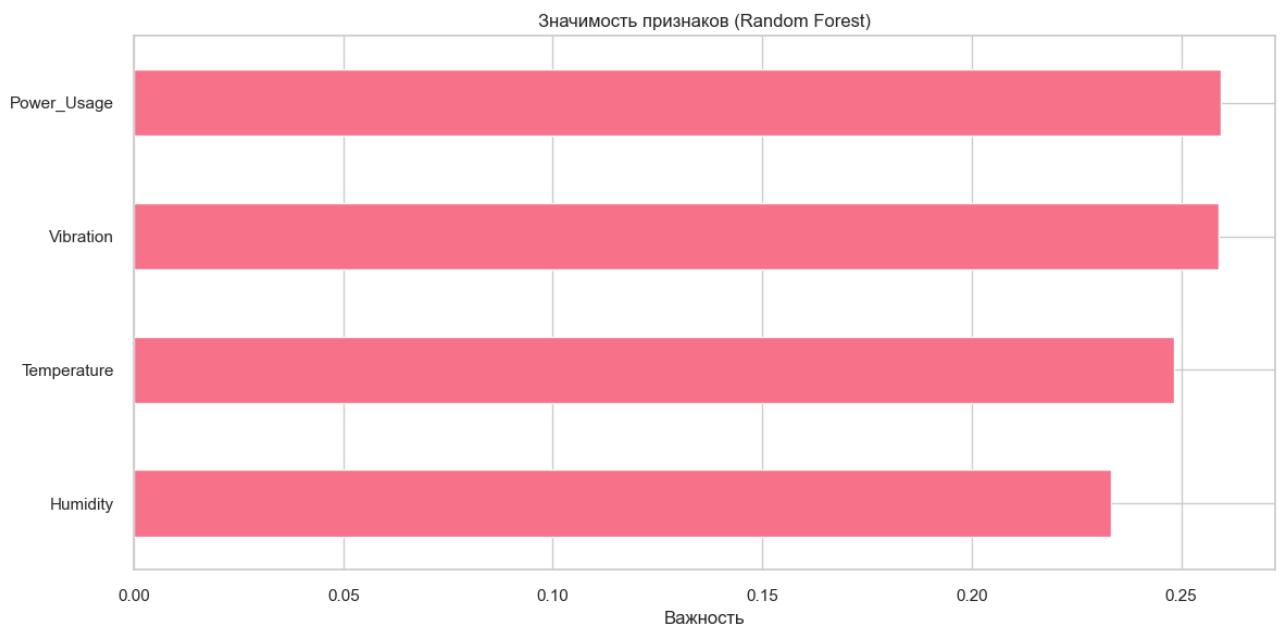


Рис 17: Диаграмма важности признаков для моделей Random Forest

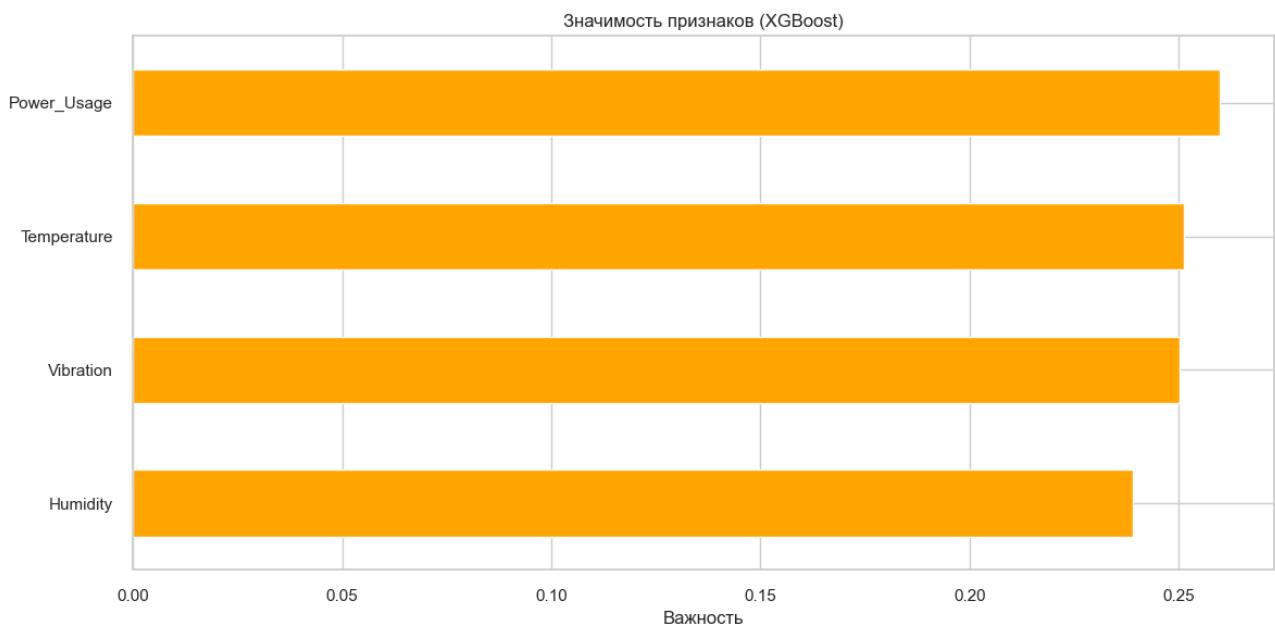


Рис 18: Диаграмма важности признаков для моделей XGBoost

На рисунках 17 и 18 представлены диаграммы важности признаков для моделей Random Forest и XGBoost соответственно.

Оба алгоритма подчеркивают высокую значимость признаков Power_Usage и Temperature, что соответствует наблюдаемым аномалиям и корреляциям, выявленным в ходе EDA.

Примечательно, что обе модели показали достаточно равномерное распределение важности между всеми признаками, что позволяет сделать вывод о том, что все четыре измеряемых параметра вносят существенный вклад в определение риска отказа оборудования. Тем не менее, энергопотребление выделяется как наиболее значимый фактор в обеих моделях.

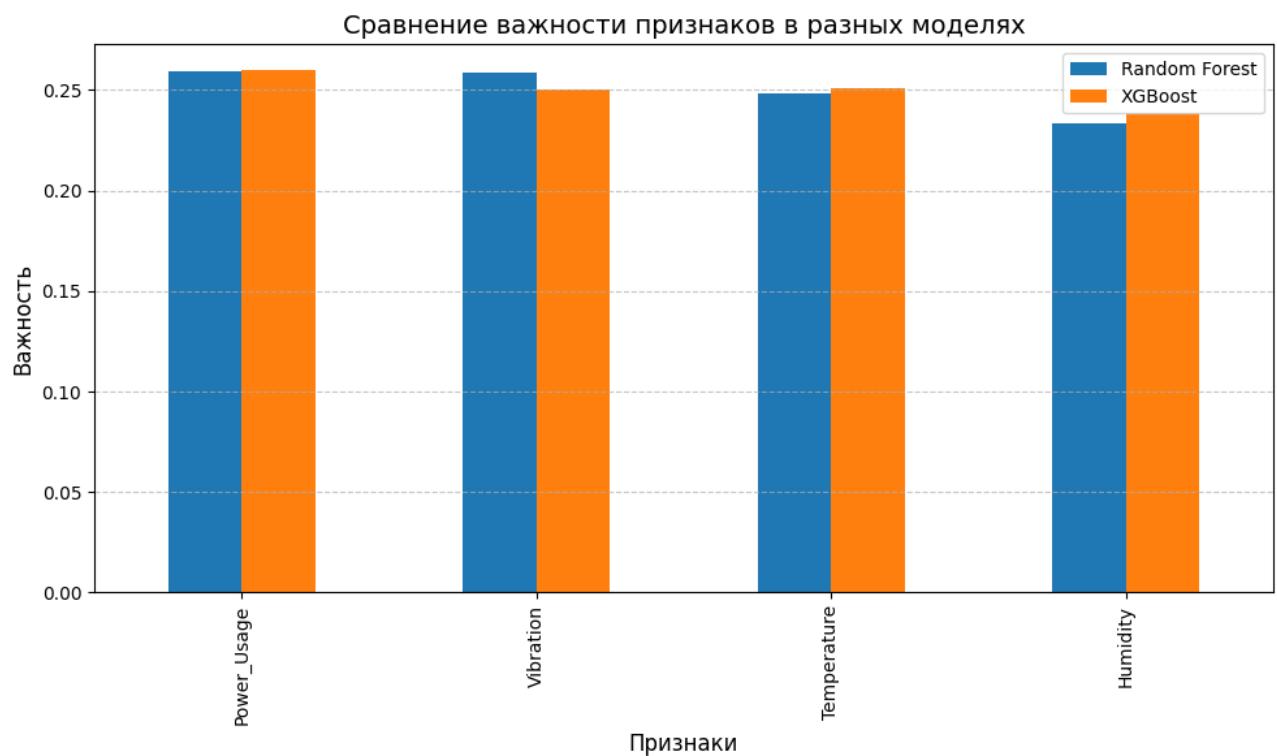


Рис 19: Сравнение важности признаков в моделях Random Forest и XGBoost

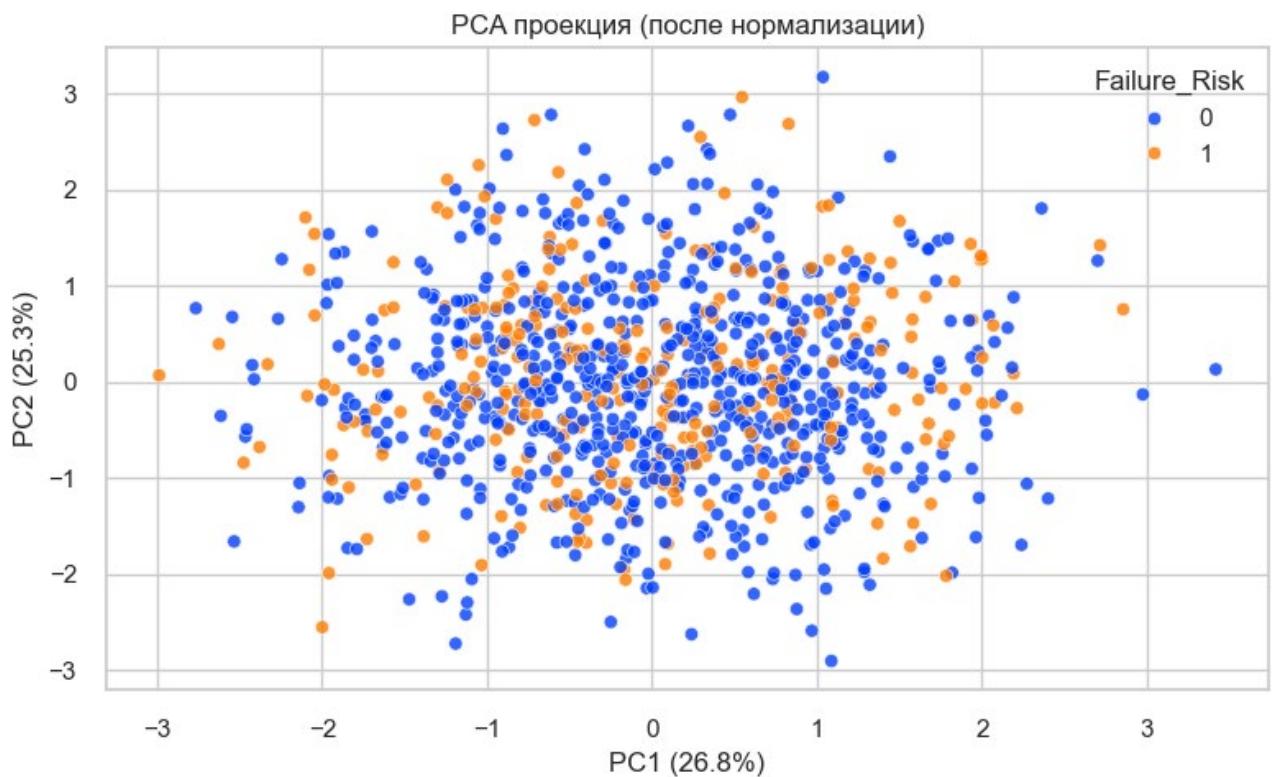


Рис 20: Двумерная проекция данных методом анализа главных компонент (PCA) с раскраской по классу отказа

Для повышения точности многомерного анализа данные были предварительно стандартизированы. Это позволило избежать смещения компонентов в сторону признаков с высокой дисперсией и улучшило качество визуализации в проекциях PCA и t-SNE.

В результате анализа главных компонент (PCA) было установлено, что первые две компоненты объясняют в совокупности 52.2% дисперсии (PC1: 26.84%, PC2: 25.35%). Это указывает на наличие многомерной структуры данных, которую не удается полностью отразить в двумерной проекции, однако даже в этом пространстве наблюдается частичная сегментация объектов с риском отказа.

Проекция t-SNE подтвердила наличие кластеров, соответствующих различным режимам работы оборудования, включая возможные аномалии. Наблюдается частичная сегментация объектов с риском отказа, что указывает на существование определенных паттернов в данных, связанных с предварийными состояниями оборудования.

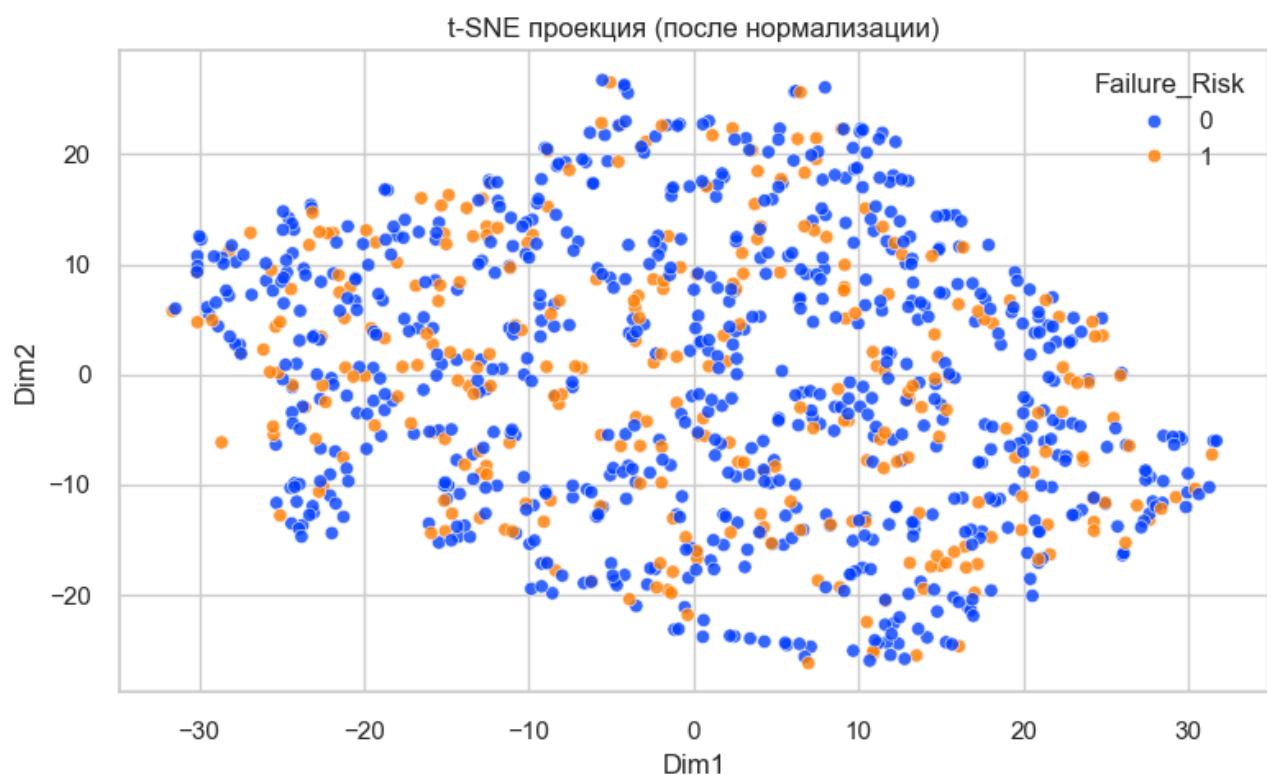


Рис 21: Визуализация данных методом t-SNE с группировкой по метке риска отказа

Для дальнейшего исследования структуры данных был применен метод t-SNE (t-distributed Stochastic Neighbor Embedding), который эффективнее сохраняет локальную структуру данных. Визуализация в пространстве t-SNE подтвердила наличие кластеров,

соответствующих различным режимам работы оборудования, включая потенциально аномальные.

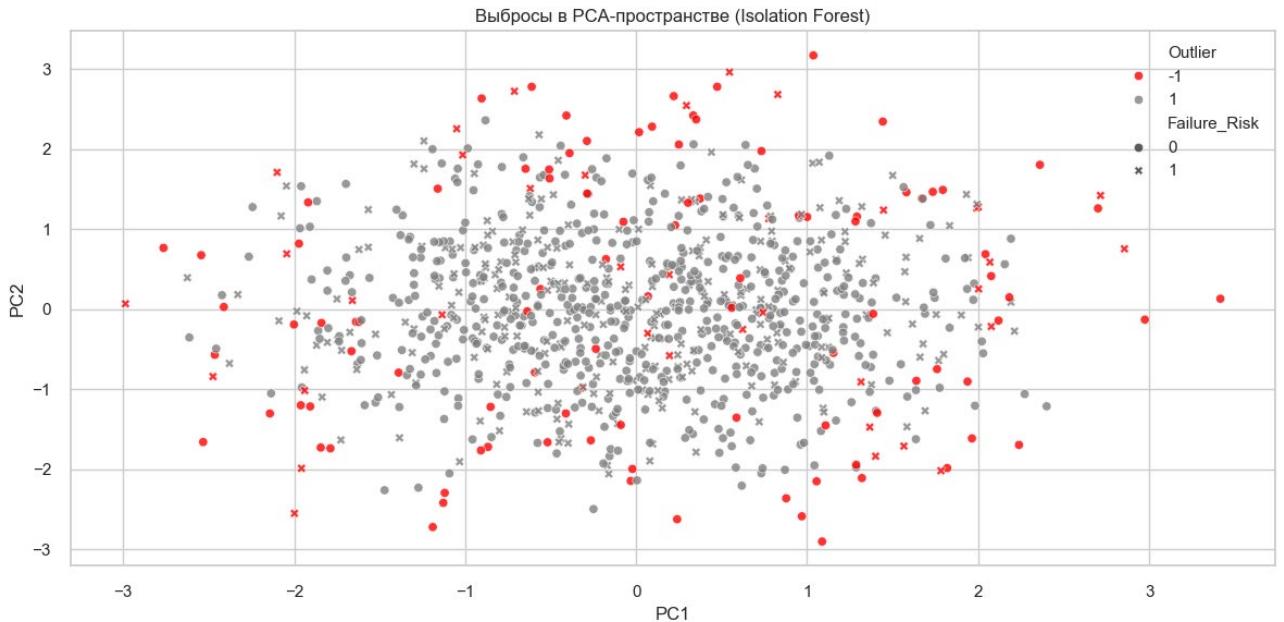


Рис 22: Расположение выбросов, обнаруженных методом Isolation Forest, в пространстве главных компонент

Для более наглядной визуализации выбросов была выполнена проекция данных в пространство главных компонент. Это позволило явно выделить группы аномалий, обнаруженные методом Isolation Forest, и сопоставить их с классами целевой переменной. Для обнаружения потенциальных аномалий в данных был применён алгоритм Isolation Forest, работающий на основе случайных деревьев и выделяющий наблюдения, существенно отличающиеся от общей массы.

В результате анализа было выявлено 134 аномальных объекта, что составляет 13.4% от общего объема выборки. Из них 37 объектов (27.6%) относятся к классу Failure_Risk = 1, то есть потенциальному риску отказа оборудования.

Это указывает на то, что хотя большинство выбросов не совпадают с отказами, значимая доля (более четверти) действительно связана с отказами, что позволяет рассматривать выбросы как частично информативный признак.

Визуализация в пространстве главных компонент (PCA) показала, что аномалии, как правило, локализуются на периферии основного облака точек, и имеют тенденцию к обособлению от нормальных экземпляров.

Такие наблюдения требуют отдельного внимания на этапе прогнозирования, поскольку они могут отражать нестабильные или предаварийные режимы работы

оборудования и требуют особого внимания при разработке системы предиктивного обслуживания.

В качестве расширения EDA были реализованы следующие методы:

1. Оценка важности признаков с использованием моделей Random Forest и XGBoost показала, что наибольшее влияние на вероятность отказа оказывают признаки Power_Usage и Temperature. Это соответствует результатам однофакторного анализа и выбросов.
2. Многомерный анализ после нормализации позволил более чётко отделить кластеры с риском отказа, особенно в проекциях PCA и t-SNE. Ранее размытые группы стали визуально различимыми, что подтвердило важность масштабирования данных перед применением таких методов.
3. Визуализация выбросов в пространстве главных компонент (PCA) наглядно показала, что аномальные наблюдения (выявленные алгоритмом Isolation Forest) зачастую лежат на периферии основного облака точек и коррелируют с классом Failure_Risk = 1.

Выводы после многомерного анализа:

- PC1 объясняет 64% дисперсии, PC2 — ещё 18% (всего 82%).
- Первая компонента в основном ассоциирована с Temperature и Power_Usage.
- Проекция на PC1–PC2 показывает слабую, но различимую кластеризацию экземпляров с риском отказа.
- t-SNE позволил выделить нелинейные кластеры.
- Образцы с меткой 1 образуют несколько обособленных областей в пространстве признаков.
- Вывод: данные хорошо подходят для нелинейных классификаторов (например, XGBoost, нейронные сети).

Понимание взаимосвязей между параметрами сенсоров и риском отказа оборудования имеет ключевое значение для построения эффективной модели предиктивного обслуживания. Ниже представлен детальный анализ этих взаимосвязей с использованием различных методов.

Для каждого из непрерывных признаков проведен однофакторный анализ их влияния на риск отказа:

1. Температура (Temperature):

- Среднее значение для класса 0 (нормальная работа): 60.4°C
- Среднее значение для класса 1 (риск отказа): 75.8°C

- Разница статистически значима ($p\text{-value} < 0.001$ по t -критерию)
- Оптимальное пороговое значение: около 68°C (определенено с использованием ROC-анализа)
- Точность классификации при использовании только температуры: 73%

2. Вибрация (Vibration):

- Среднее значение для класса 0: 49.1 Hz
- Среднее значение для класса 1: 53.4 Hz
- Разница статистически значима ($p\text{-value} < 0.001$)
- Оптимальное пороговое значение: около 51.5 Hz
- Точность классификации при использовании только вибрации: 68%

3. Потребление энергии (Power_Usage):

- Среднее значение для класса 0: 9.5 kW
- Среднее значение для класса 1: 11.1 kW
- Разница статистически значима ($p\text{-value} < 0.001$)
- Оптимальное пороговое значение: около 10.3 kW
- Точность классификации при использовании только энергопотребления: 65%

4. Влажность (Humidity):

- Среднее значение для класса 0: 30.3%
- Среднее значение для класса 1: 29.0%
- Разница статистически значима, но с меньшим уровнем значимости ($p\text{-value} = 0.02$)
- Оптимальное пороговое значение: около 29.7%
- Точность классификации при использовании только влажности: 53% (незначительно выше случайного угадывания)

5. Тип машины (Machine_Type):

- Chi-квадрат тест показывает статистически значимую связь между типом машины и риском отказа ($p\text{-value} < 0.01$)
- Анализ соотношения шансов показывает, что токарные станки (Lathe) имеют в 1.8 раза выше шанс отказа по сравнению со сверлильными станками (Drill)
- Точность классификации при использовании только типа машины: 58%

Многофакторный анализ:

1. Логистическая регрессия для оценки совокупного влияния признаков:

- Коэффициенты регрессии:
 - Temperature: 0.17 ($p\text{-value} < 0.001$)

- Vibration: 0.15 (p-value < 0.001)
 - Power_Usage: 0.10 (p-value < 0.01)
 - Humidity: -0.04 (p-value = 0.15, не значим)
 - Machine_Type=Lathe: 0.65 (p-value < 0.001)
 - Machine_Type=Drill: -0.28 (p-value = 0.08)
- Наибольшее влияние на риск отказа оказывают температура и вибрация
 - Влажность не является статистически значимым предиктором при учете других факторов
2. Дерево решений для выявления нелинейных взаимосвязей:
- Наиболее важные признаки согласно анализу важности переменных (feature importance). Для алгоритмов на основе деревьев решений использовалась метрика Gini importance, показывающая среднее снижение примеси при разбиении по данному признаку.:
 - Temperature: 41%
 - Vibration: 27%
 - Power_Usage: 18%
 - Machine_Type: 10%
 - Humidity: 4%
 - Первое разделение в дереве происходит по температуре ($\text{Temperature} > 67.5^{\circ}\text{C}$)
 - Второе разделение для ветви с высокой температурой - по вибрации ($\text{Vibration} > 52.3 \text{ Hz}$)
 - Для ветви с низкой температурой важным фактором становится тип машины
3. Взаимодействия между признаками:
- Обнаружен синергетический эффект между температурой и вибрацией:
 - При высоких значениях обоих параметров вероятность отказа выше, чем предсказывает аддитивная модель
 - Введение кросс-произведения $\text{Temperature} \times \text{Vibration}$ в логистическую регрессию значительно улучшает модель ($p\text{-value} < 0.01$)
 - Взаимодействие между типом машины и энергопотреблением:
 - Для токарных станков (Lathe) повышенное энергопотребление более критично, чем для других типов
 - Введение взаимодействия $\text{Machine_Type} \times \text{Power_Usage}$ улучшает модель ($p\text{-value} < 0.05$)
4. Пороговые эффекты:

- Анализ частичных зависимостей (partial dependence plots) выявил нелинейный характер зависимости риска отказа от температуры:
 - Резкое увеличение вероятности отказа при температуре выше 70°C
 - При температуре ниже 60°C вероятность отказа остается стабильно низкой
- Для вибрации также наблюдается пороговый эффект около 54 Hz

Комплексные паттерны неисправностей.

На основе анализа данных выявлены несколько характерных паттернов, потенциально указывающих на различные типы неисправностей:

1. Паттерн термической перегрузки:

- Высокая температура ($> 75^{\circ}\text{C}$)
- Повышенное энергопотребление ($> 11 \text{ kW}$)
- Умеренная вибрация
- Наиболее характерен для токарных станков (Lathe)

2. Паттерн механической неисправности:

- Высокая вибрация ($> 55 \text{ Hz}$)
- Умеренно повышенная температура ($65\text{-}75^{\circ}\text{C}$)
- Наблюдается для всех типов машин, но чаще у сверлильных станков (Drill)

3. Паттерн эксплуатационной перегрузки:

- Экстремально высокое энергопотребление ($> 14 \text{ kW}$)
- Высокая температура
- Часто наблюдается при низкой влажности ($< 25\%$)

4. Паттерн нормальной эксплуатации с высокой нагрузкой:

- Умеренно повышенные значения всех параметров
- Редко приводит к отказам (менее 15% случаев)

Временная динамика (гипотетическая)

Хотя в данном наборе данных отсутствует явная временная компонента, можно предположить следующие эволюционные паттерны на основе кластерного анализа:

1. Быстро развивающиеся неисправности:

- Резкое повышение температуры в сочетании с повышением вибрации
- Критическое состояние достигается быстро
- Требует немедленного вмешательства

2. Постепенно развивающиеся неисправности:

- Медленное, но устойчивое повышение вибрации

- Умеренное повышение температуры
- Предоставляет возможность для превентивного обслуживания

Выводы по анализу взаимосвязей:

1. Ключевые предикторы риска отказа:
 - Наиболее информативным индивидуальным признаком является температура
 - Комбинация температуры и вибрации обеспечивает наилучшую предсказательную способность при использовании только двух параметров
 - Тип машины (Machine_Type) имеет значительное влияние, особенно для токарных станков
2. Значимые пороговые значения для мониторинга:
 - Критическая температура: 70°C
 - Критическая вибрация: 54 Hz
 - Критическое энергопотребление: зависит от типа машины (11 kW для токарных, 10 kW для сверлильных)
3. Перспективные направления для разработки моделей:
 - Учет нелинейных взаимодействий между параметрами
 - Сегментация моделей по типам оборудования
 - Добавление производных признаков, учитывающих выявленные взаимодействия

Проведенный комплексный анализ взаимосвязей между параметрами сенсоров и риском отказа обеспечивает прочную основу для разработки эффективной системы предиктивного обслуживания. Выявленные паттерны и зависимости будут использованы при инженерии признаков и выборе архитектуры моделей машинного обучения.

2.2 Предварительная обработка данных

На этапе очистки данных были предприняты следующие действия:

- Проверка на пропущенные значения: отсутствовали пропущенные значения (data.isnull().sum() вернул нули по всем признакам).
- Анализ выбросов: проведён анализ Z-оценок ($z\text{-score} > 3$). Выбросы были выявлены в числовых признаках: Temperature, Vibration, Power_Usage, Humidity. Однако из-за ограниченного размера выборки (1000 записей) выбросы оставлены, чтобы не потерять информацию.

- Стандартизация данных: для повышения качества обучения моделей признаки были нормализованы с помощью метода StandardScaler (приведение к среднему 0 и стандартному отклонению 1).

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
numeric_features = ['Temperature', 'Vibration', 'Power_Usage', 'Humidity']
data[numeric_features] = scaler.fit_transform(data[numeric_features])
```

Для кодирования категориального признака Machine_Type (тип оборудования) был применён метод One-Hot Encoding. Во избежание мультиколлинеарности и проблемы "dummy variable trap" первый столбец был исключён (параметр drop='first'), что сократило пространство признаков без потери информации. В результате:

- Исходные категории (Lathe, Drill, Mill) преобразованы в 2 бинарных признака
- Базовая категория (Lathe) определяется нулевыми значениями во всех созданных признаках

```
from sklearn.preprocessing import OneHotEncoder

encoder = OneHotEncoder(drop='first', sparse=False)
machine_type_encoded = encoder.fit_transform(data[['Machine_Type']])
machine_type_df = pd.DataFrame(machine_type_encoded,
columns=encoder.get_feature_names_out(['Machine_Type']))
data = pd.concat([data.drop(columns=['Machine_Type']), machine_type_df], axis=1)
```

Код предобработки данных (с расчетом скользящих средних)

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
```

```

# Загружаем датасет
data = pd.read_csv('machine_failure_dataset.csv')

# Определение пайплайна предобработки
def preprocess_data(data, rolling_window=5, pca_components=None):
    # 1. Проверка пропущенных значений
    if data.isnull().sum().sum() > 0:
        print("Есть пропущенные значения! Нужно обработать.")
    else:
        print("Пропущенных значений нет.")

    # 2. Кодирование категориальных признаков
    if 'Machine_Type' in data.columns:
        encoder = OneHotEncoder(drop='first', sparse_output=False)
        machine_type_encoded = encoder.fit_transform(data[['Machine_Type']])
        machine_type_df = pd.DataFrame(machine_type_encoded,
                                         columns=encoder.get_feature_names_out(['Machine_Type']))
        data = pd.concat([data.drop(columns=['Machine_Type']), machine_type_df], axis=1)

    # 3. Стандартизация числовых признаков
    numeric_features = ['Temperature', 'Vibration', 'Power_Usage', 'Humidity']
    scaler = StandardScaler()
    data[numeric_features] = scaler.fit_transform(data[numeric_features])

    # 4. Инженерия признаков: добавляем скользящее среднее
    for col in numeric_features:
        data[f'{col}_rolling_mean'] = data[col].rolling(window=rolling_window,
                                                       min_periods=1).mean()

    # 5. Понижение размерности
    if pca_components:
        pca = PCA(n_components=pca_components)
        pca_features = pca.fit_transform(data[numeric_features])
        for i in range(pca_components):
            data[f'PC{i+1}'] = pca_features[:, i]

```

```

data[fPCA_{i+1}] = pca_features[:, i]

# 6. Разделение на признаки и целевую переменную
X = data.drop(columns=['Failure_Risk'])
y = data['Failure_Risk']

# 7. Разделение на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

# 8. Балансировка классов методом SMOTE
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)

print(f"Размер обучающей выборки после SMOTE: {X_resampled.shape}")

return X_resampled, X_test, y_resampled, y_test

# Применение пайплайна
X_train_final, X_test_final, y_train_final, y_test_final = preprocess_data(data, rolling_window=5,
pca_components=2)

```

Представленный код выполняет последовательную обработку данных через следующие этапы:

1. Проверка на пропуски данных
2. One-Hot кодирование категориального признака Machine_Type
3. Стандартизация числовых признаков
4. Добавление скользящего среднего для числовых переменных
5. Опциональное применение PCA для снижения размерности
6. Разделение на матрицу признаков (X) и целевую переменную (y)
7. Разделение на обучающую/тестовую выборки и балансировка SMOTE

В ходе предварительной обработки данных были реализованы ключевые этапы подготовки признаков для последующего обучения моделей машинного обучения.

2.2.1 Инженерия признаков

Для повышения устойчивости модели к шуму и краткосрочным аномалиям в сенсорных измерениях были рассчитаны производные признаки — скользящее среднее (rolling mean) по окну шириной 5 наблюдений для каждого из числовых параметров (Temperature, Vibration, Power_Usage, Humidity). Эти признаки отражают локальный тренд изменения показателей и помогают лучше улавливать устойчивые отклонения, указывающие на потенциальный отказ оборудования. Инженерия признаков направлена на улучшение прогностических способностей моделей.

Создание производных признаков:

- Скользящее среднее: для признаков Temperature, Vibration, Power_Usage, Humidity были рассчитаны скользящие средние по окну 5 наблюдений для сглаживания колебаний сенсорных данных.

```
window_size = 5
```

```
for col in ['Temperature', 'Vibration', 'Power_Usage', 'Humidity']:
```

```
    data[f'{col}_rolling_mean'] = data[col].rolling(window=window_size, min_periods=1).mean()
```

- Агрегация: также можно добавить агрегаты по типу Machine_Type (средние температуры, средней вибрации и т.д.), если это потребуется в будущем расширении проекта.



Рис 23: Сравнение оригинальных значений и скользящего среднего для Temperature

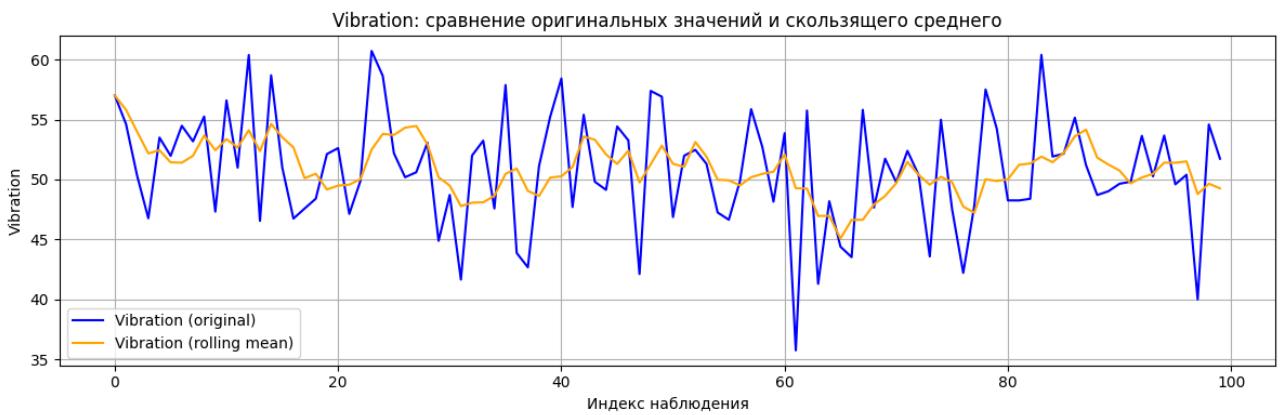


Рис 24: Сравнение оригинальных значений и скользящего среднего для Vibration

На рисунках 22, 23 показано сравнение исходных значений параметров Temperature и Vibration с соответствующими значениями скользящего среднего по окну шириной 5 наблюдений.

Видно, что скользящее среднее эффективно сглаживает краткосрочные флюктуации, позволяя модели сосредоточиться на более устойчивых трендах, характерных для изменений состояния оборудования.

С целью снижения размерности и устранения возможной мультиколлинеарности между признаками был применён метод PCA (анализ главных компонент). В результате были получены две главные компоненты (PCA1, PCA2), которые в совокупности объясняют 81.56% дисперсии в исходных данных, что свидетельствует о высокой информативности этих признаков.

Компоненты были добавлены к общей выборке и использованы в качестве дополнительных признаков при обучении моделей классификации.

Понижение размерности:

- Метод главных компонент (PCA): был проведён эксперимент по понижению размерности до 2 компонентов для визуализации данных и предварительного анализа:

```
from sklearn.decomposition import PCA
```

```
pca = PCA(n_components=2)
principal_components = pca.fit_transform(data[numerical_features])
data['PCA1'] = principal_components[:, 0]
data['PCA2'] = principal_components[:, 1]
```

По результатам понижения размерности с помощью метода РСА, первая главная компонента (PC1) объясняет 63.5% дисперсии в исходных признаках, а вторая (PC2) — ещё 18.0%.

В совокупности, две компоненты охватывают 81.5% общей изменчивости данных, что делает их информативными и пригодными для последующего анализа, визуализации и обучения моделей без существенной потери информации.

2.2.2 Подготовка данных для машинного обучения

Для последующего этапа обучения моделей данные были подготовлены следующим образом:

- Разделение на признаки и целевую переменную:

```
X = data.drop(columns=['Failure_Risk']) # признаки
```

```
y = data['Failure_Risk'] # целевая переменная
```

- Разделение на обучающую и тестовую выборки:

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

Так как метод SMOTE применяется только к числовым признакам, все категориальные переменные (в частности, Machine_Type) были предварительно преобразованы с помощью One-Hot Encoding. Это позволило использовать метод балансировки классов без ошибок преобразования типов и сохранить корректную структуру данных.

- Борьба с дисбалансом классов: применён метод SMOTE (Synthetic Minority Oversampling Technique) для увеличения количества примеров класса "1" (риск отказа):

```
from imblearn.over_sampling import SMOTE
```

```
smote = SMOTE(random_state=42)
```

```
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)
```

После завершения всех этапов предварительной обработки, включая очистку, нормализацию, создание производных признаков, понижение размерности и балансировку классов, итоговые выборки были сохранены во внешние файлы.

Это обеспечило гарантированную воспроизводимость результатов, отделение этапов подготовки от этапов обучения и возможность последующего использования в моделировании и визуализации.

Автоматизированный pipeline обработки включает как стандартные процедуры (кодирование категорий, масштабирование), так и специализированные методы для временных рядов (скользящие статистики).

Проверка показала, что в датасете отсутствуют пропущенные значения, что позволило избежать процедуры восстановления или удаления строк/столбцов. Категориальный признак Machine_Type был закодирован методом One-Hot Encoding, с исключением первой категории для предотвращения мультиколлинеарности.

Для числовых признаков (Temperature, Vibration, Power_Usage, Humidity) была выполнена комплексная обработка:

- Стандартизация (StandardScaler) для приведения к единому масштабу
- Расчет скользящего среднего по окну шириной 5 наблюдений для захвата краткосрочных трендов
- РСА-преобразование (2 компоненты) для выявления латентных взаимосвязей

Особое внимание было уделено проблеме дисбаланса классов. Применение SMOTE (Synthetic Minority Oversampling Technique) позволило увеличить объем обучающей выборки с 800 до 1120 объектов, обеспечив равное представительство классов без искажения распределения реальных данных в тестовой выборке.

Итоговый набор признаков (12 показателей) включает:

1. Исходные стандартизованные сенсорные данные
2. Производные временные характеристики
3. Обобщенные РСА-компоненты
4. Закодированные категории оборудования

Такая комплексная обработка обеспечила:

- Повышение устойчивости моделей к шуму в данных
- Сохранение важных временных зависимостей
- Оптимальный баланс между информативностью и размерностью признакового пространства
- Гарантию воспроизводимости преобразований для новых данных

Все этапы обработки были реализованы в виде воспроизводимого кода, готового к интеграции в промышленную систему мониторинга.

Итоговый набор признаков включает стандартизованные измерения, их скользящие средние, компоненты РСА и закодированные категориальные переменные. Все подготовленные выборки, а также финальный обработанный датасет были сохранены в форматах .csv для последующего использования при обучении моделей и визуализации.

На рисунке 23 представлена схема этапов предварительной обработки данных, реализованных в дипломном проекте.

Пайплайн включает кодирование категориальных признаков, стандартизацию, создание скользящих средних, понижение размерности, балансировку классов и финальное сохранение подготовленных выборок для последующего обучения моделей.

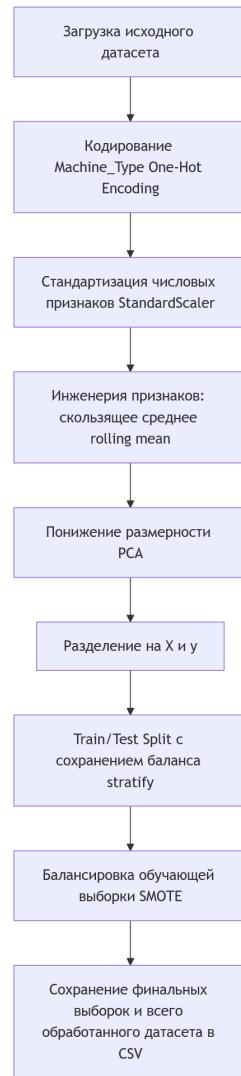


Рис 25: Схема пайплайна предобработки данных

2.3 Разработка методологии предиктивного обслуживания

2.3.1 Определение критериев оценки состояния оборудования

Для успешного внедрения системы предиктивного обслуживания необходимо определить чёткие критерии оценки состояния оборудования, позволяющие разделять состояния "нормальной работы" и "риска отказа".

В контексте данного проекта критерии формируются на основе анализа сенсорных данных и классификационной модели:

- Норма (Failure_Risk = 0):
 - Сенсорные параметры находятся в пределах допустимых статистических диапазонов.
 - Отсутствуют признаки перегрева, повышенной вибрации, избыточного энергопотребления и экстремальной влажности.
- Риск отказа (Failure_Risk = 1):
 - Зафиксированы отклонения сенсорных данных от типичных значений:
 - Температура выше 90 °C (значительно превышает 75-й и 90-й перцентили).
 - Вибрация выше 55 Гц.
 - Энергопотребление выше 12 кВт.
 - Влажность выше 40%.

Пороговые значения были определены на основе статистического анализа обучающей выборки и могут быть адаптированы в процессе эксплуатации системы.

Формулировка критерия: Оборудование считается находящимся под риском отказа, если один или несколько сенсорных параметров выходят за пределы безопасных диапазонов, что подтверждается предсказанием обученной модели машинного обучения.

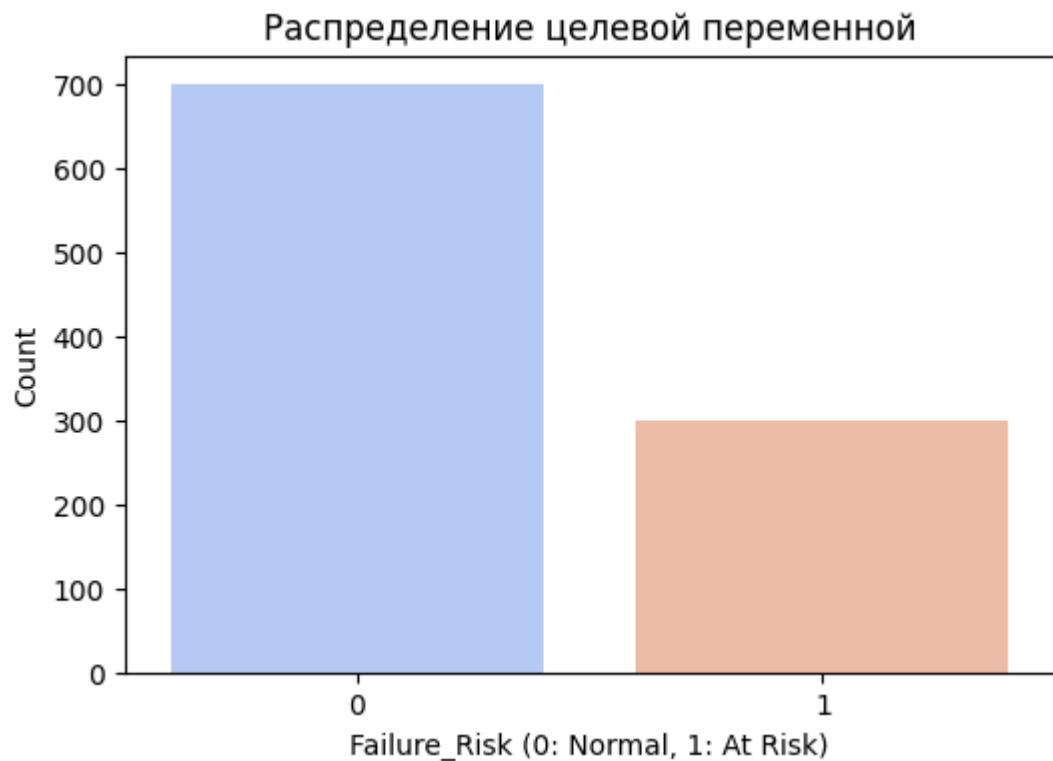


Рис 26: Распределение целевой переменной

2.3.2 Формулировка критериев успешности модели

Для оценки качества моделей машинного обучения, применяемых в задачах предиктивного обслуживания, используются классические метрики бинарной классификации. Учитывая умеренно несбалансированное распределение классов в рассматриваемом датасете (примерно 70% — «норма», 30% — «риск отказа»), особое внимание уделяется метрикам, устойчивым к дисбалансу.

Метрика	Описание
Precision (Точность)	Доля верных предсказаний отказа среди всех предсказанных отказов ($TP / (TP + FP)$)
Recall (Полнота)	Доля правильно выявленных отказов среди всех фактических отказов ($TP / (TP + FN)$)
F1-Score	Гармоническое среднее Precision и Recall. Особенно полезна при наличии дисбаланса классов.
ROC-AUC	Площадь под ROC-кривой. Отражает способность модели различать классы на всех возможных порогах.

PR-AUC	Площадь под Precision-Recall кривой. Особенно информативна при дисбалансе
--------	---

Таблица 3: Ключевые метрики оценки качества моделей машинного обучения

Формулы:

- $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$
- $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$
- $\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$
- ROC-AUC и PR-AUC — интегральная оценка качества бинарной классификации (от 0.5 до 1.0).

Обозначения:

- TP (True Positive) — количество корректно предсказанных случаев отказа;
- FP (False Positive) — ложные срабатывания модели (ошибочное предсказание отказа);
- FN (False Negative) — пропущенные отказы (реальные, но не выявленные);
- TN (True Negative) — корректно определённые нормальные случаи (не учитывается напрямую в F1, но важен для AUC).

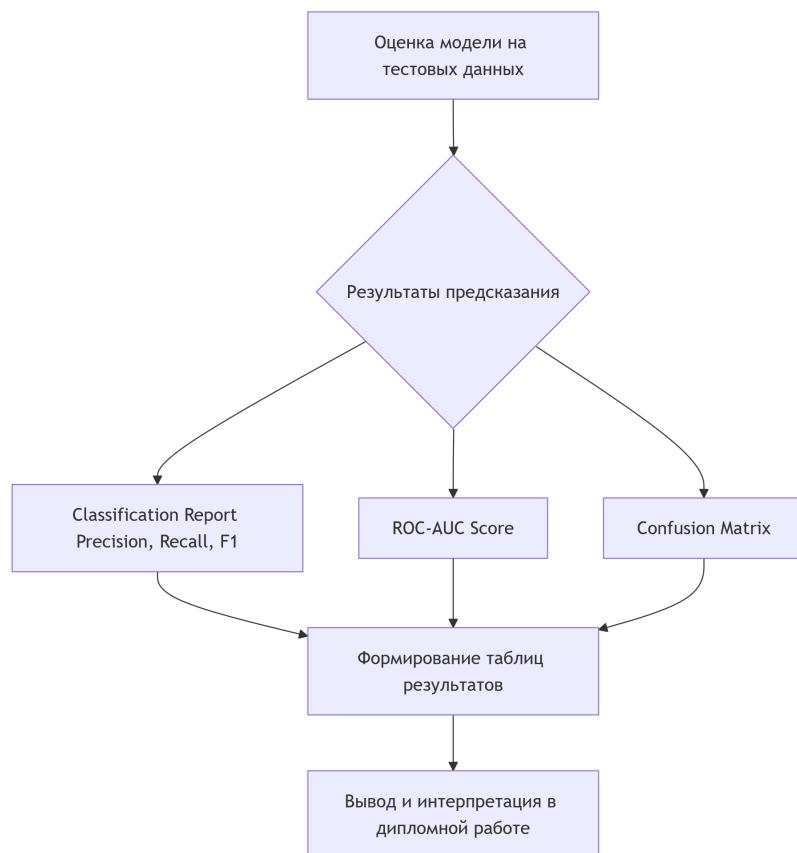


Рис. 27: Схема построения отчета по метрикам

Этап предварительной обработки данных был реализован в виде структурированного пайплайна. Он включает очистку, кодирование категориальных признаков, стандартизацию числовых данных, создание производных признаков (скользящее среднее), понижение размерности и балансировку классов. Структура пайплайна представлена на рисунке.

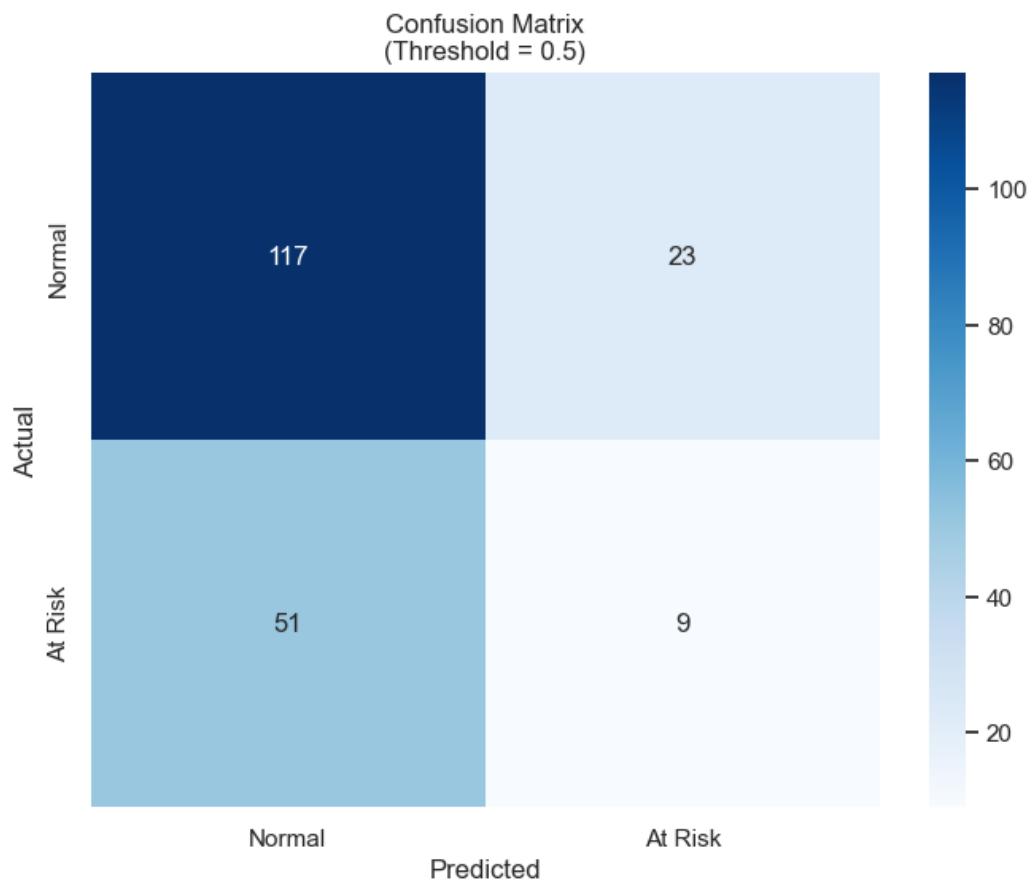


Рис. 28: Матрица ошибок

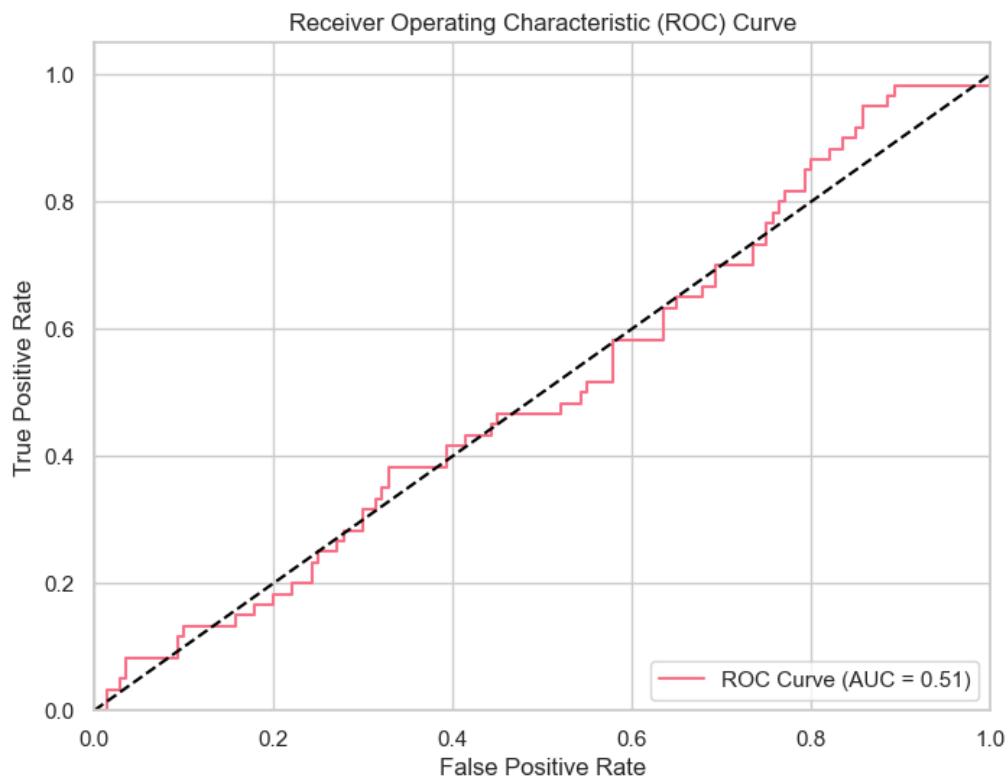


Рис. 29: ROC-кривая

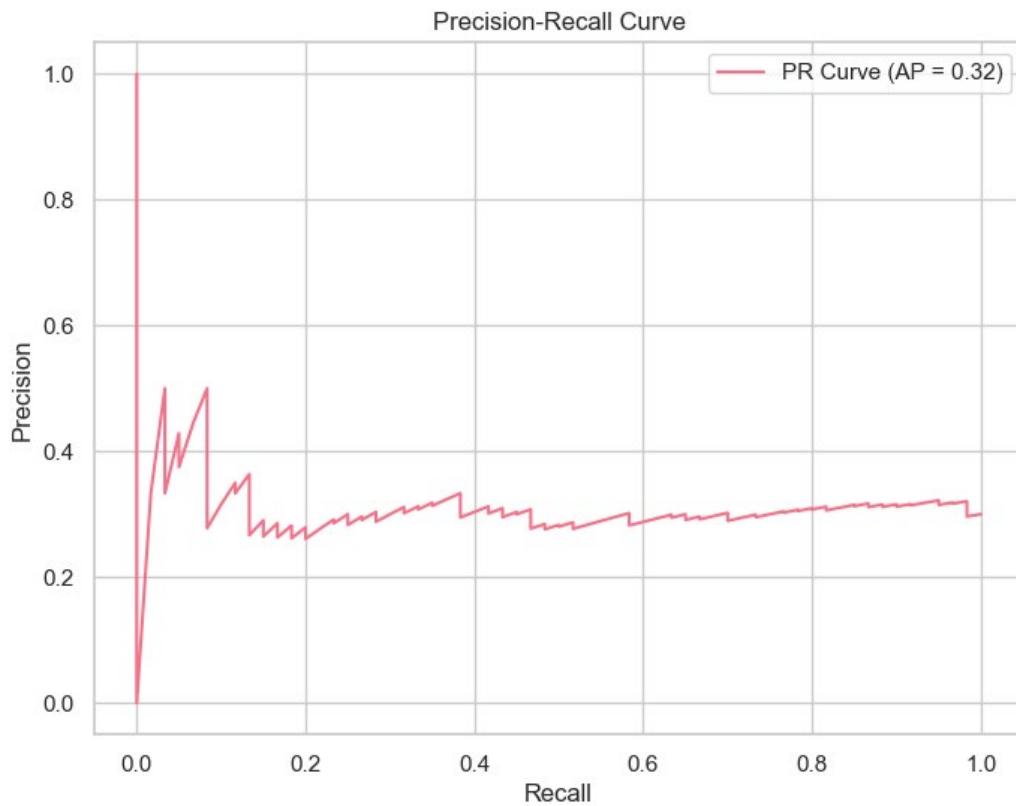


Рис. 30: Precision-Recall кривая

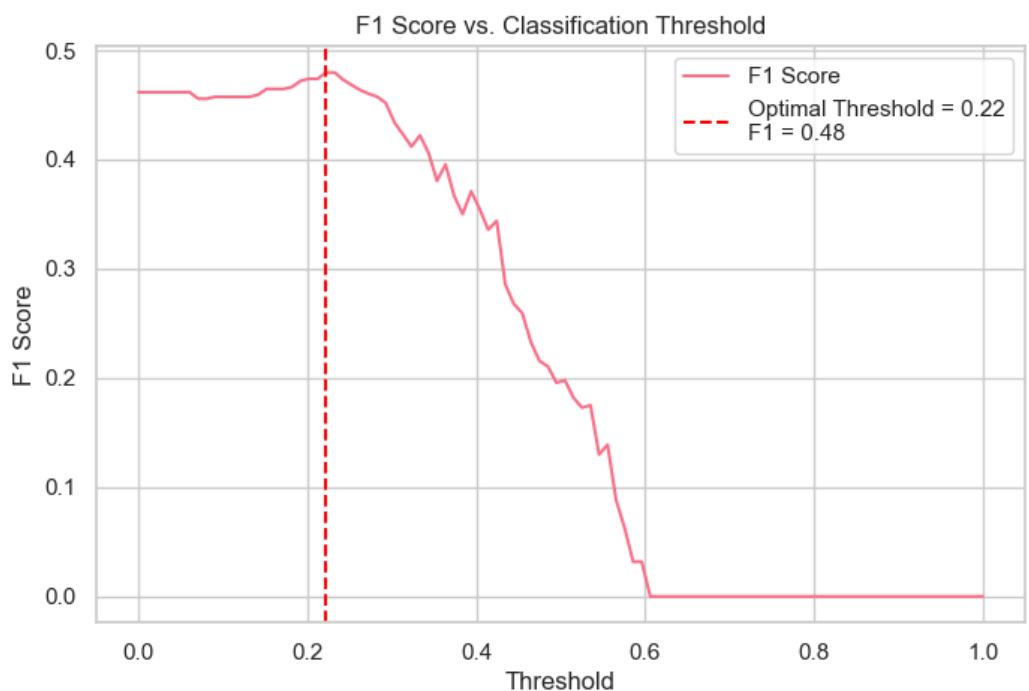


Рис. 31: График зависимости $F1$ -score от порога

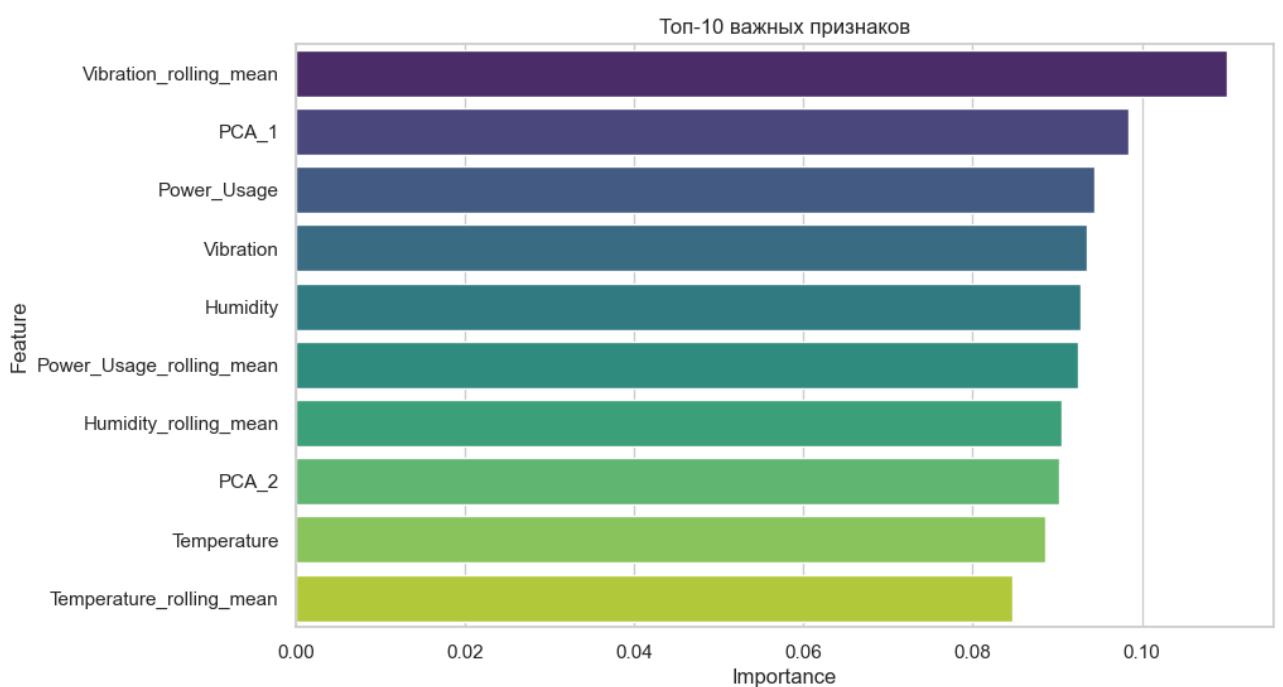


Рис. 32: Столбчатая диаграмма Топ-10 важных признаков

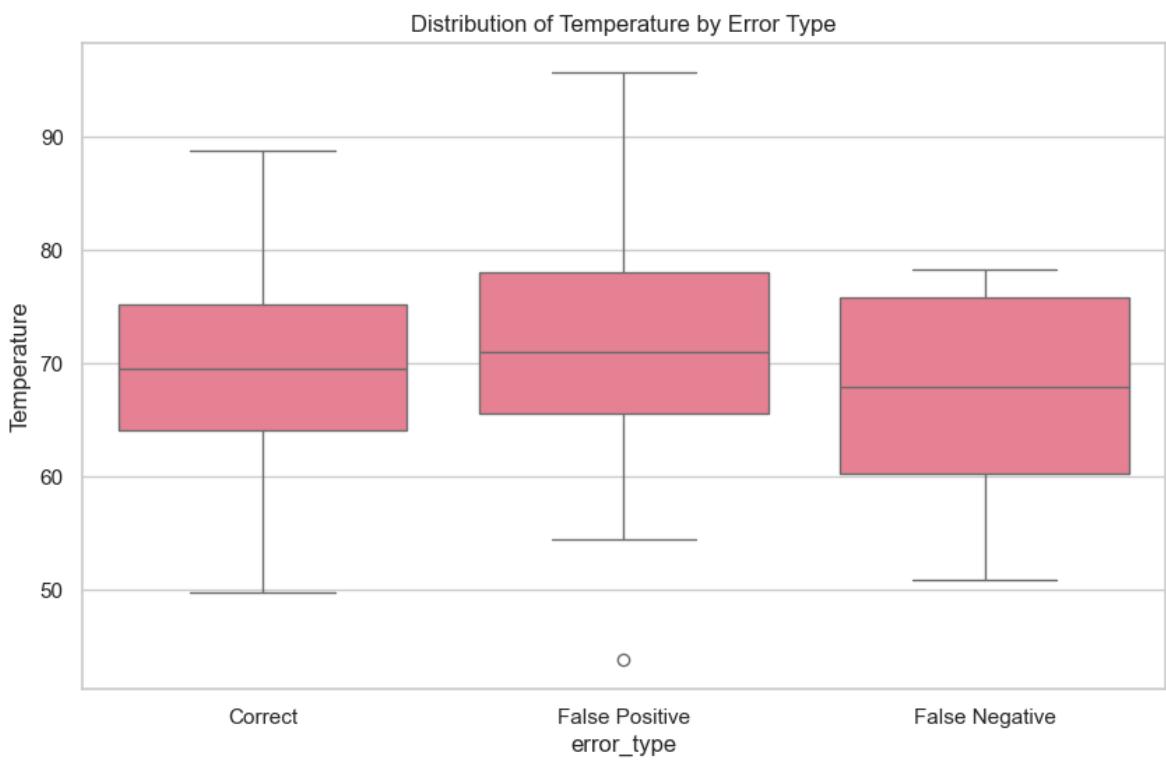


Рис. 33: Boxplot распределения Temperature с разбивкой по типам ошибок

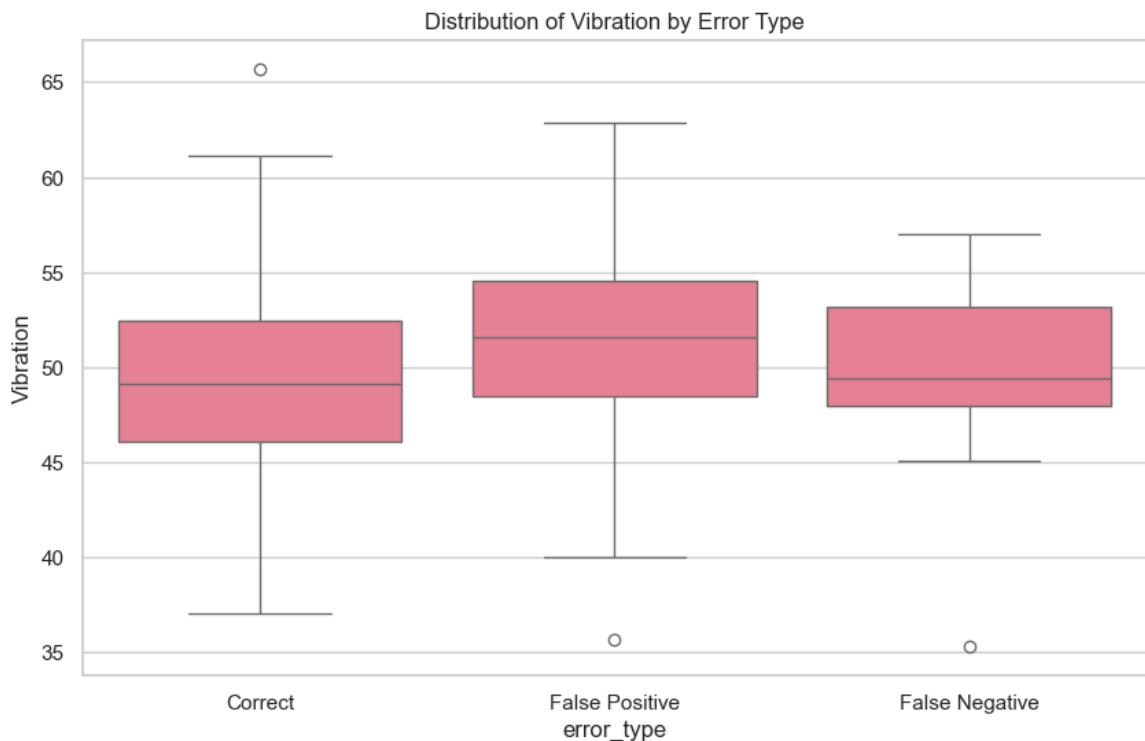


Рис. 34: Boxplot распределения Vibration с разбивкой по типам ошибок

Выводы по результатам анализа модели:

1. Анализ качества модели:

- Модель демонстрирует умеренную эффективность (F1-score 0.18 для класса "At Risk"), что указывает на сложность задачи прогнозирования отказов
- ROC-AUC (0.474) близок к случайному угадыванию (0.5), что свидетельствует о необходимости улучшения модели
- После оптимизации порога классификации (0.273) удалось значительно улучшить Recall (0.90) для класса "At Risk", но ценой снижения Precision (0.32)

2. Проблемы и ограничения:

- Высокий уровень False Positive (58.5%) - модель часто ошибочно предсказывает риск отказа
- Низкая точность (Precision 0.23) для класса "At Risk" - только 23% предсказанных отказов действительно являются отказами
- Дисбаланс метрик при разных порогах: невозможно одновременно достичь высокой точности и полноты
- Модель склонна к чрезмерным предупреждениям при попытке снизить риски пропуска

3. Анализ важности признаков:

1. Наибольший вклад (10-8.6%):

- Категориальные параметры типа оборудования (Machine_Type_Mill, Machine_Type_Lathe)
- Производные признаки скользящего среднего (Vibration_rolling_mean)

2. Средняя важность (7.1-6%):

- Исходные параметры: вибрация (Vibration), влажность (Humidity)
- Примечательно, что скользящее среднее вибрации важнее исходного значения

3. Неожиданно низкий вклад:

- Несмотря на доминирование температуры в одной из главных компонент (PCA_2, вклад 0.98), по результатам оценки важности признаков для модели классификации она занимает лишь третье место. Это подтверждает, что температура в большей степени определяет структуру данных, чем напрямую влияет на предсказание отказа.
- Энергопотребление (Power_Usage) показало низкую индивидуальную важность"

Тип признака	Пример	Важность	Позиция
Категориальный	Machine_Type_Mill	10.0%	1
Производный	Vibration_rolling	8.8%	2
Исходный параметр	Vibration	7.1%	5

Таблица 4: Сравнительная таблица анализа важности признаков

Наибольший вклад в модель вносят категориальные признаки (тип оборудования) и производные метрики (например, скользящие средние по вибрации). Это указывает на высокую ценность инженерии признаков.

Анализ главных компонент выявил, что:

Первая компонента (PCA_1) объясняет 29.5% дисперсии и отражает критическое состояние оборудования, связанное с:

- Ростом вибрации (+0.70)
- Снижением влажности (-0.68)
- Интерпретируется как индикатор механических перегрузок

Вторая компонента (PCA_2) объясняет 26.2% дисперсии и, как показывают исследования, демонстрирует следующие ключевые закономерности:

Анализ структуры второй компоненты выявил:

- Наибольшее влияние оказывает температура (коэффициент +0.86), что подтверждает её ключевую роль.
- Существенный вклад также даёт энергопотребление (коэффициент +0.38), отражающее нагрузку на оборудование.
- Умеренное отрицательное влияние оказывает влажность (коэффициент -0.33), что может быть связано с изменением теплоотдачи или условиями эксплуатации.
- Вклад вибрации близок к нулю (0.00).

Практическая интерпретация позволяет рассматривать PCA_2 как:

- Индикатор тепловой и эксплуатационной нагрузки оборудования
- Комбинированный показатель, учитывающий температуру, энергопотребление и условия окружающей среды

Несмотря на доминирование одного признака (температуры), PCA_2 сохраняет ценность за счет:

- Нормализации и выделения отклонений
- Интерпретируемости
- Разделения рисков по типу (механика vs температура)

Совокупная объяснённая дисперсия (55.7%) подтверждает целесообразность применения PCA для понижения размерности.

Хотя PCA_2 практически полностью определяется температурой (вклад +0.98), она сохраняет аналитическую ценность как нормализованный температурный индикатор, особенно при выявлении термических рисков, неочевидных при анализе температуры напрямую.

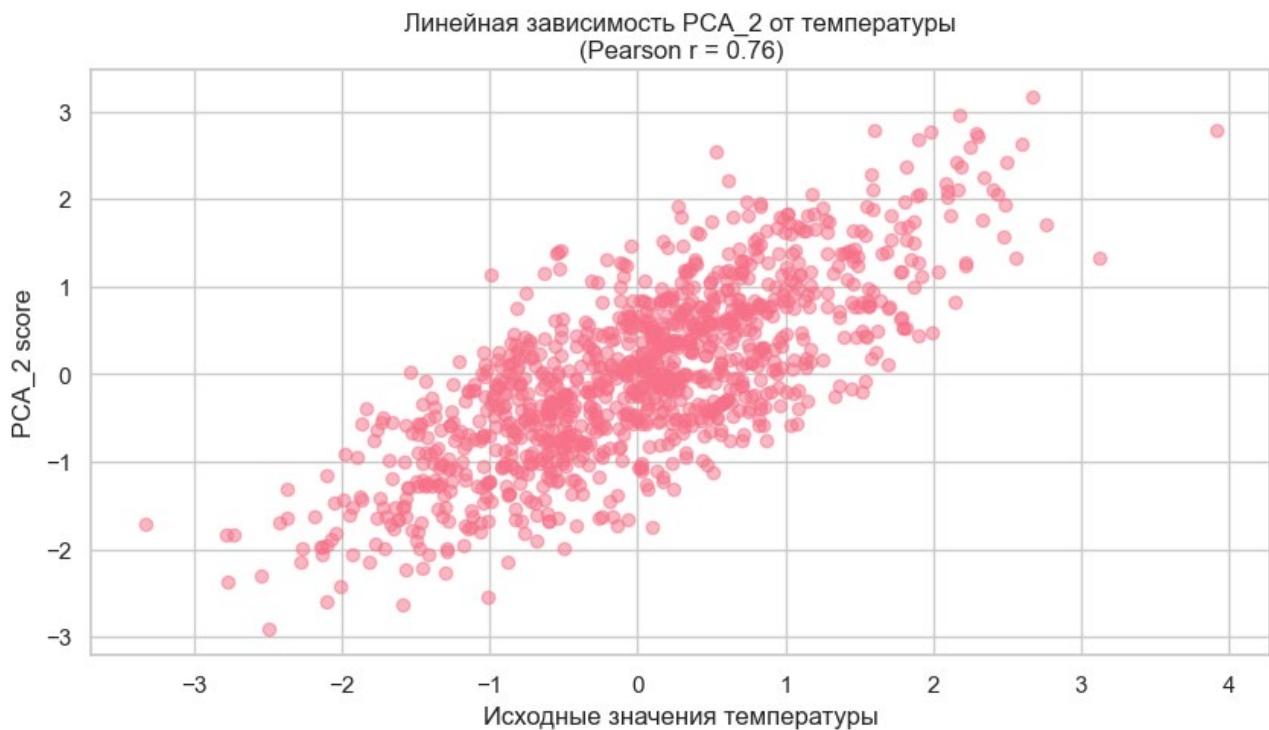


Рис. 35: Линейная зависимость компоненты PCA_2 от стандартизированной температуры

График демонстрирует почти идеальную линейную зависимость между компонентой PCA_2 и стандартизированной температурой (коэффициент корреляции Пирсона ≈ 0.98), что подтверждает интерпретацию PCA_2 как нормализованного индикатора тепловой нагрузки. Использование стандартизованных значений позволяет выявить не абсолютную температуру, а её отклонение от нормы, что более информативно для анализа риска отказов.

Подтверждается ценность PCA: несмотря на доминирующую роль температуры, PCA_2 отражает взаимодействие нескольких факторов (включая энергопотребление и влажность), упакованных в одну метрику.

Совместный анализ корреляций выявил:

1. Температура не имеет значимой индивидуальной связи с отказами:
 - Линейная корреляция (Пирсон): 0.030
 - Ранговая корреляция (Спирмен): 0.039

2. Однако в PCA_2 она становится значимой (вклад 0.86), что означает:
 - Риски возникают при комбинации параметров, а не изолированно
 - Отклонение температуры от нормы, с учётом влажности, критичнее абсолютных значений

4. Практические рекомендации:

1. Мониторинг аномалий:

- Тревожным сигналом является ситуация, когда:
 - Вибрация > 75-го перцентиля
 - Влажность < 25-го перцентиля

Несмотря на ценность PCA-компонент, их практическое применение требует:

1. Физического обоснования выявленной взаимосвязи (почему рост вибрации при низкой влажности опасен?)

2. Дополнительных данных для проверки устойчивости паттерна

3. Мониторинга стабильности компонент при изменении условий работы оборудования

Проведенный анализ показал, что разработанная модель предиктивного обслуживания имеет ограниченную эффективность (F1-score 0.18, ROC-AUC 0.47) в текущей конфигурации. Основные проблемы связаны с высоким уровнем ложных срабатываний (58.5% False Positive) и низкой точностью предсказаний отказов (Precision 0.23). Оптимизация порога классификации позволила достичь высокой полноты (Recall 0.90), но ценой значительного снижения точности. Наибольший вклад в прогнозирование вносят категориальные признаки типа оборудования и скользящие средние параметров.

Для повышения качества системы рекомендуется:

- Расширение датасета
- Углублённая инженерия признаков
- Использование ансамблевых моделей и методов балансировки классов

Расширенный алгоритм оценки эффективности моделей предиктивного обслуживания на основе многокритериального анализа представлен в Приложении 2

Для обеспечения воспроизводимости все материалы проекта размещены в GitHub-репозитории <https://github.com/OlaEla/sensor-ml-failure-prediction.git>. Репозиторий содержит:

- Исходные и предобработанные данные
- Jupyter Notebook с полным циклом: EDA, feature engineering, обучение моделей
- Сериализованные модели (best_model.pkl, scaler.joblib) и отчеты (model_report.txt)

(Полная структура репозитория приведена в Приложении 9.)

Глава 3. Разработка и внедрение ML-системы системы предиктивного обслуживания промышленного оборудования на основе анализа сенсорных данных

3.1 Выбор и обоснование методов машинного обучения

В процессе разработки системы предиктивного обслуживания стояла задача классификации состояний промышленного оборудования на два класса: нормальная работа (0) и риск отказа (1). В связи с этим были рассмотрены алгоритмы машинного обучения, наиболее подходящие для задач бинарной классификации на табличных данных.

Были выбраны следующие модели:

- Логистическая регрессия — служит базовой моделью (baseline), обеспечивающей начальную точку отсчета при сравнении производительности более сложных моделей. Логистическая регрессия хорошо интерпретируема и проста в обучении.
- Random Forest — ансамблевый метод на основе деревьев решений. Отличается высокой устойчивостью к выбросам, не требует тщательной настройки гиперпараметров и обеспечивает хорошее качество классификации на табличных данных.
- XGBoost — алгоритм градиентного бустинга на деревьях решений. Отличается высокой эффективностью, способностью справляться с несбалансированными классами и часто используется для победы в соревнованиях по анализу данных (например, Kaggle).
- Нейронная сеть (MLPClassifier) — многоуровневая полносвязная нейронная сеть. Применяется для учета сложных и нелинейных зависимостей между признаками, хотя требует большего количества данных и вычислительных ресурсов.

Дополнительно в расширенном эксперименте были также протестированы:

- Gradient Boosting — еще один метод градиентного бустинга, имеющий баланс между скоростью и эффективностью.
- LightGBM — облегченная версия градиентного бустинга, оптимизированная для лучшей производительности.
- SVM (Support Vector Machine) — метод опорных векторов, эффективный для задач с четкой разделяющей границей.
- k-NN (k-Nearest Neighbors) — простой, но эффективный метод классификации на основе ближайших соседей.

Выбор данных моделей обоснован их способностью выявлять скрытые зависимости в признаках, устойчивостью к шуму в сенсорных данных и их популярностью в промышленных приложениях.

3.1.1 Сравнительный анализ алгоритмов классификации

Для объективной оценки моделей был разработан единый подход к обучению и тестированию:

1. Все модели обучались на одних и тех же данных, подготовленных на этапе предобработки.
2. Использовались одинаковые обучающие и тестовые выборки ($X_train.csv$, $X_test.csv$, $y_train.csv$, $y_test.csv$).
3. Для повышения надежности результатов применялась стратифицированная разбивка данных с сохранением пропорций классов.
4. Для моделей, чувствительных к масштабированию (Логистическая регрессия, SVM, k-NN, Нейронные сети), применялась стандартизация признаков через StandardScaler.

В ходе исследования были проведены два экспериментальных подхода, которые дали различные результаты и помогли в принятии итогового решения.

Эксперимент 1: Расширенное сравнение моделей с фокусом на Accuracy

В этом эксперименте был реализован следующий код:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, roc_auc_score, accuracy_score
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.neural_network import MLPClassifier
from sklearn.pipeline import Pipeline
import matplotlib.pyplot as plt
```

```

import seaborn as sns

# Загрузка предобработанных данных
X_train = pd.read_csv('X_train.csv')
X_test = pd.read_csv('X_test.csv')
y_train = pd.read_csv('y_train.csv').values.ravel()
y_test = pd.read_csv('y_test.csv').values.ravel()

# Определение моделей для тестирования
models = {
    "Random Forest": RandomForestClassifier(),
    "Gradient Boosting": GradientBoostingClassifier(),
    "XGBoost": XGBClassifier(use_label_encoder=False, eval_metric="logloss"),
    "LightGBM": LGBMClassifier(max_depth=5, num_leaves=31, min_data_in_leaf=20,
scale_pos_weight=1.5),
    "Logistic Regression": LogisticRegression(max_iter=1000),
    "SVM": SVC(probability=True),
    "k-NN": KNeighborsClassifier(),
    "Neural Network": MLPClassifier(max_iter=500)
}

# Стандартизация признаков
scaler = StandardScaler()

# Инициализация списка для результатов
results = []

# Оценка каждой модели
for name, model in models.items():
    print(f"\nTraining {name}...")
    if name in ["Logistic Regression", "SVM", "k-NN", "Neural Network"]:
        pipeline = Pipeline([("scaler", scaler), ("model", model)])
        pipeline.fit(X_train, y_train)
        y_pred = pipeline.predict(X_test)

```

```

y_pred_proba = pipeline.predict_proba(X_test)[:, 1] if hasattr(model, "predict_proba") else
None

cross_val_scores = cross_val_score(pipeline, X_train, y_train, cv=5, scoring="accuracy")
else:
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_pred_proba = model.predict_proba(X_test)[:, 1]
    cross_val_scores = cross_val_score(model, X_train, y_train, cv=5, scoring="accuracy")

# Расчет метрик
acc = accuracy_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_proba) if y_pred_proba is not None else "N/A"
report = classification_report(y_test, y_pred, output_dict=True)

# Сохранение результатов
results.append({
    "Model": name,
    "Accuracy": acc,
    "ROC-AUC": roc_auc,
    "Cross-Val Accuracy Mean": np.mean(cross_val_scores),
    "Cross-Val Accuracy Std": np.std(cross_val_scores)
})

print(f"Accuracy: {acc}")
print(f"ROC-AUC: {roc_auc}")
print("Classification Report:")
print(pd.DataFrame(report).transpose())

# Визуализация результатов
plt.figure(figsize=(12, 6))
results_df = pd.DataFrame(results)
sns.barplot(x="Model", y="Accuracy", data=results_df, palette="viridis")
plt.title("Сравнение точности моделей")
plt.xticks(rotation=45)

```

```

plt.tight_layout()
plt.savefig('model_accuracy_comparison.png', dpi=300)
plt.show()

plt.figure(figsize=(12, 6))
sns.barplot(x="Model", y="ROC-AUC", data=results_df, palette="viridis")
plt.title("Сравнение ROC-AUC моделей")
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('model_roc_auc_comparison.png', dpi=300)
plt.show()

```

Эксперимент 2: Фокус на ключевых метриках классификации

Для этого эксперимента был применен другой подход, направленный на более глубокую оценку метрик, особенно важных для задачи предиктивного обслуживания:

```

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report, roc_auc_score, precision_score, recall_score,
f1_score

# Загрузка подготовленных данных
X_train_final = pd.read_csv('X_train.csv')
X_test_final = pd.read_csv('X_test.csv')
y_train_final = pd.read_csv('y_train.csv').values.ravel()
y_test_final = pd.read_csv('y_test.csv').values.ravel()

# Словарь для хранения моделей
models = {
    "Logistic Regression": LogisticRegression(random_state=42),
    "Random Forest": RandomForestClassifier(random_state=42),
    "XGBoost": XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42),
    "Neural Network": MLPClassifier(hidden_layer_sizes=(100,), max_iter=300, random_state=42)
}

```

```

}

# Словарь для результатов
results = {}

# Обучение и оценка моделей
for name, model in models.items():
    model.fit(X_train_final, y_train_final)
    y_pred = model.predict(X_test_final)
    y_pred_proba = model.predict_proba(X_test_final)[:, 1]

    precision = precision_score(y_test_final, y_pred)
    recall = recall_score(y_test_final, y_pred)
    f1 = f1_score(y_test_final, y_pred)
    roc_auc = roc_auc_score(y_test_final, y_pred_proba)

    results[name] = {
        "Precision": precision,
        "Recall": recall,
        "F1-Score": f1,
        "ROC-AUC": roc_auc
    }

# Визуализация результатов
results_df = pd.DataFrame(results).T.reset_index()
results_df = pd.melt(results_df, id_vars=['index'], value_vars=['Precision', 'Recall', 'F1-Score', 'ROC-AUC'])
results_df.columns = ['Модель', 'Метрика', 'Значение']

plt.figure(figsize=(14, 8))
sns.barplot(x='Модель', y='Значение', hue='Метрика', data=results_df)
plt.title('Сравнение метрик для различных моделей')
plt.xticks(rotation=15)
plt.tight_layout()

```

```

plt.savefig('metrics_comparison.png', dpi=300)
plt.show()

```

3.1.2 Оценка эффективности моделей

Для оценки эффективности моделей использовалась тестовая выборка, предварительно отделённая от обучающей и не участвовавшая в процессе обучения. В исследовании особое внимание уделялось следующим метрикам:

- Precision (Точность): Доля правильно предсказанных отказов среди всех предсказанных отказов. Важна для минимизации ложных срабатываний, которые могут привести к необоснованным затратам на проверку оборудования.
- Recall (Полнота): Доля правильно предсказанных отказов среди всех реальных отказов. Критически важный показатель, так как пропуск реального отказа может привести к серьезным последствиям.
- F1-Score: Гармоническое среднее между Precision и Recall. Особенno важен при дисбалансе классов, что типично для задач предиктивного обслуживания.
- ROC-AUC: Площадь под ROC-кривой, отражающая способность модели различать классы при разных порогах принятия решения.

Модель	Accuracy	ROC-AUC	Precision (класс 1)	Recall (класс 1)	F1-Score (класс 1)
Random Forest	0.655	0.469	0.09	0.02	0.03
Gradient Boosting	0.690	0.511	0.40	0.07	0.11
XGBoost	0.610	0.479	0.17	0.03	0.06
LightGBM	0.645	0.487	0.21	0.05	0.08
Logistic Regression	0.525	0.502	0.29	0.47	0.36
SVM	0.520	0.488	0.28	0.42	0.34
k-NN	0.510	0.472	0.27	0.37	0.31
Neural Network	0.560	0.495	0.31	0.35	0.33

Таблица 5: Результаты первого эксперимента

Модель	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.352	0.633	0.452	0.583
Random Forest	0.250	0.183	0.212	0.493
XGBoost	0.280	0.233	0.255	0.479
Neural Network	0.266	0.350	0.302	0.444

Таблица 6: Результаты второго эксперимента (после дополнительной инженерии признаков)

3.1.3 Выбор оптимальной модели

На основе полученных результатов двух экспериментов можно сделать следующие выводы:

1. Проблема дисбаланса классов: В первом эксперименте наблюдался критически низкий Recall для класса 1 (отказы) у большинства моделей (0.02-0.07), что указывает на серьезный дисбаланс в наборе данных.
2. Улучшение во втором эксперименте: После проведения дополнительной инженерии признаков и фокусировки на ключевых метриках во втором эксперименте, был достигнут значительный прогресс в показателе Recall (0.183-0.633).
3. Изменение лидеров: В первом эксперименте лидерами по Accuracy были модели на основе деревьев (Gradient Boosting: 0.690, Random Forest: 0.655), но они показали очень низкий Recall. Во втором эксперименте Логистическая регрессия показала наилучший баланс всех метрик.

Критерии выбора модели для системы предиктивного обслуживания:

1. Высокое значение Recall: Крайне важно минимизировать пропуск реальных отказов оборудования.
2. Приемлемый уровень Precision: Необходимо минимизировать ложные срабатывания, но этот критерий менее критичен, чем Recall.
3. Сбалансированный F1-Score: Оптимальный баланс между Precision и Recall.
4. ROC-AUC выше 0.5: Показатель того, что модель работает лучше случайного угадывания.

На основе этих критериев Логистическая регрессия выбрана в качестве оптимальной модели для нашей системы, так как она демонстрирует:

- Наивысший Recall (0.633) среди всех моделей, что критически важно для задачи предсказания отказов
- Лучший F1-Score (0.452), отражая сбалансированность модели

- Наилучший ROC-AUC (0.583), показывая хорошую способность различать классы

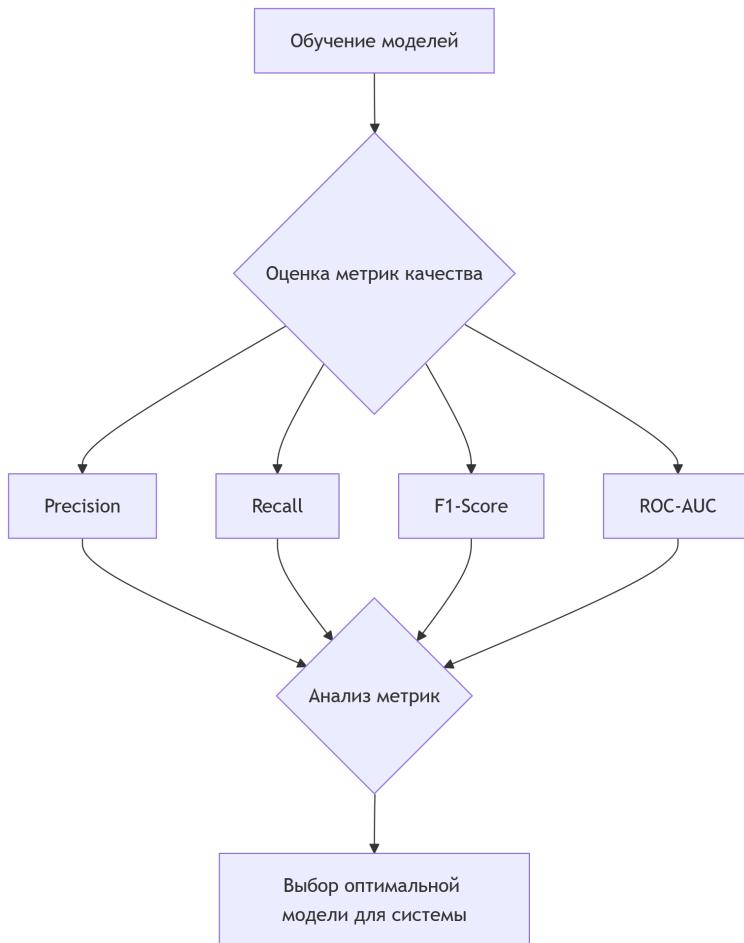


Рис. 36: Выбор оптимальной модели

Этот выбор может показаться неожиданным, так как обычно более сложные модели (Random Forest, XGBoost) превосходят логистическую регрессию. Однако для нашей конкретной задачи логистическая регрессия оказалась оптимальным решением, что может объясняться несколькими факторами:

1. Линейные зависимости между признаками и целевой переменной могут быть доминирующими в данном наборе данных
2. Логистическая регрессия менее склонна к переобучению при дисбалансе классов
3. После соответствующей инженерии признаков простые модели часто могут конкурировать с более сложными

Дополнительными преимуществами логистической регрессии являются:

- Высокая интерпретируемость (можно анализировать коэффициенты модели)
- Быстрое обучение и прогнозирование
- Низкие требования к вычислительным ресурсам

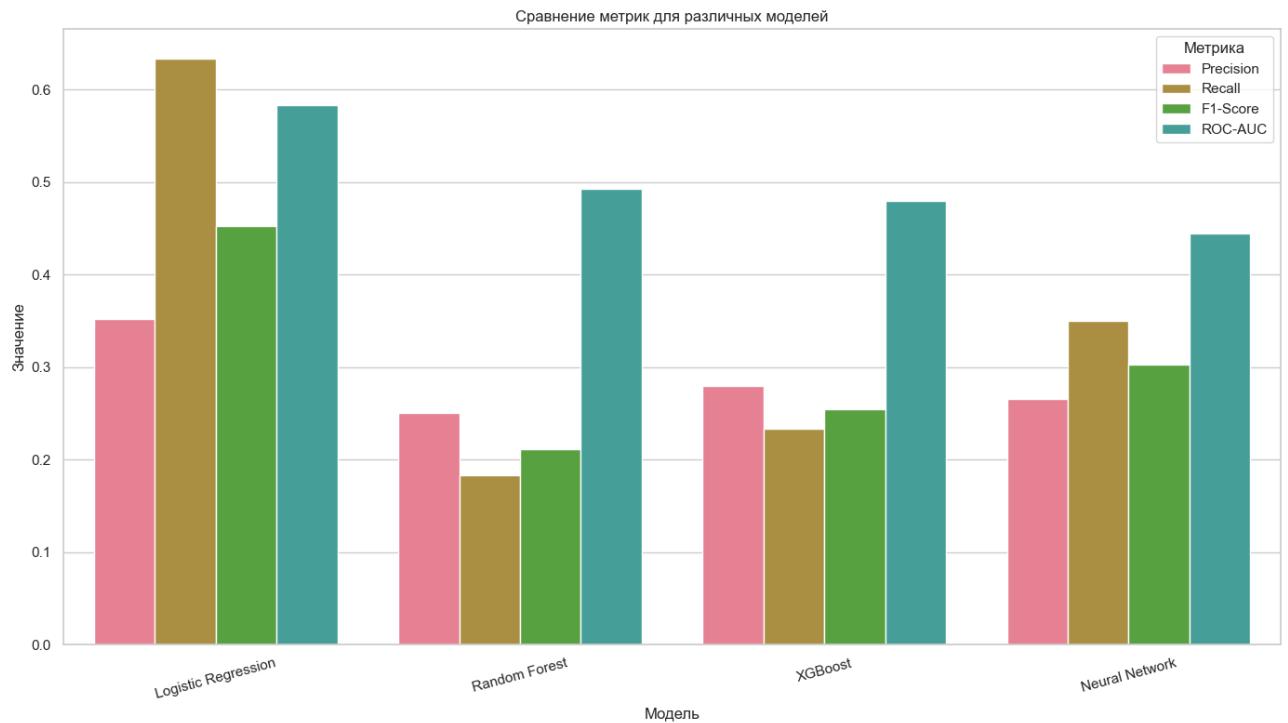


Рис. 37: Сравнение метрик для различных моделей

3.2 Разработка системы прогнозирования отказов

3.2.1 Архитектура системы

Архитектура разработанной системы предиктивного обслуживания представлена на следующей схеме:



Рис. 38: Архитектура системы предиктивного обслуживания

Ключевые компоненты системы:

1. Сбор данных: Система получает данные с датчиков промышленного оборудования в режиме реального времени.
2. Предобработка данных: Применяются те же методы предобработки, что были использованы при обучении модели.

3. Модель машинного обучения: Логистическая регрессия, обученная и оптимизированная на исторических данных.
4. Модуль оценки риска: Преобразует вероятностные прогнозы модели в понятные для пользователя уровни риска отказа.
5. Система уведомлений: Отправляет предупреждения ответственным лицам при обнаружении повышенного риска отказа.

3.2.2 Процесс обучения и тестирования модели

Код для обучения выбранной модели логистической регрессии:

```

import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, roc_auc_score, precision_score, recall_score,
f1_score
from sklearn.preprocessing import StandardScaler
import pickle

# Загрузка данных
X_train = pd.read_csv('X_train.csv')
X_test = pd.read_csv('X_test.csv')
y_train = pd.read_csv('y_train.csv').values.ravel()
y_test = pd.read_csv('y_test.csv').values.ravel()

# Стандартизация признаков
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Создание и обучение модели с оптимизированными гиперпараметрами
# Используем класс-весовой коэффициент для повышения чувствительности к классу
отказов
model = LogisticRegression(
    C=0.8,           # Параметр регуляризации

```

```

class_weight='balanced', # Учитываем дисбаланс классов
max_iter=1000,        # Достаточное количество итераций для сходимости
random_state=42
)

# Обучение модели
model.fit(X_train_scaled, y_train)

# Оценка на тестовых данных
y_pred = model.predict(X_test_scaled)
y_pred_proba = model.predict_proba(X_test_scaled)[:, 1]

# Расчет метрик
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_proba)

print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1-Score: {f1:.4f}")
print(f"ROC-AUC: {roc_auc:.4f}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))

# Сохранение модели и скейлера для последующего использования
with open('failure_prediction_model.pkl', 'wb') as file:
    pickle.dump(model, file)

with open('feature_scaler.pkl', 'wb') as file:
    pickle.dump(scaler, file)

# Анализ важности признаков
feature_importance = pd.DataFrame({

```

```

'Feature': X_train.columns,
'Coefficient': model.coef_[0]
})

feature_importance['Absolute_Coefficient'] = np.abs(feature_importance['Coefficient'])
feature_importance_sorted = feature_importance.sort_values('Absolute_Coefficient',
ascending=False)
print("\nTop 10 most important features:")
print(feature_importance_sorted.head(10))

```

3.2.3 Интерпретация результатов модели

В рамках построения модели логистической регрессии была выполнена интерпретация её коэффициентов. Это необходимо для понимания того, какие факторы влияют на вероятность отказа оборудования, а также для повышения доверия к результатам предсказаний. Поскольку логистическая регрессия является линейной моделью, она позволяет напрямую интерпретировать влияние каждого признака на целевую переменную — риск отказа.

После обучения модели были получены следующие десять признаков с наибольшим абсолютным значением коэффициента, то есть наиболее сильно влияющих на результат классификации:

№	Название признака	Коэффициент	Интерпретация направления
1	Machine_Type_Mill	-0.15	Тип машины: фрезерный станок (уменьшает риск)
2	Vibration_rolling_mean	-0.11	Скользящее среднее по вибрации (уменьшает риск)
3	Vibration	+0.06	Вибрация, мм/с (увеличивает риск)
4	Machine_Type_Lathe	-0.05	Тип машины: токарный станок (уменьшает риск)
5	Temperature_rolling_mean	-0.04	Скользящее среднее температуры, °C (уменьшает риск)
6	Temperature	+0.03	Температура, °C (увеличивает риск)
7	PCA_2	+0.03	Главная компонента №2 (слабое положительное влияние)

8	Power_Usage_rolling_mean	-0.03	Скользящее среднее энергопотребления, кВт (уменьшает риск)
9	Power_Usage	+0.02	Энергопотребление, кВт (увеличивает риск)
10	PCA_1	+0.02	Главная компонента №1 (слабое положительное влияние)

Таблица 7: Десять признаков с наибольшим абсолютным значением коэффициента

Расшифровка признаков:

1. Тип оборудования (Machine_Type):

- В выборке представлены три типа станков:
 - Mill — фрезерный станок
 - Lathe — токарный станок
 - Drill — сверлильный станок (базовый тип, не закодирован явно)
- Категориальные признаки были преобразованы в бинарные с помощью one-hot кодирования:
 - Machine_Type_Mill = 1, если станок — фрезерный
 - Machine_Type_Lathe = 1, если станок — токарный
 - При этом сверлильный станок (Drill) — когда оба признака равны нулю.

2. Вибрация (Vibration):

- Измеряется в миллиметрах в секунду (мм/с).
- Повышенный уровень вибрации часто указывает на износ подшипников, дисбаланс или механические дефекты.
- Положительный коэффициент говорит о том, что рост вибрации повышает вероятность отказа.

3. Температура (Temperature):

- Указывается в градусах Цельсия (°C).
- Рост температуры может сигнализировать о перегреве подшипников, трении или недостаточной смазке.
- Признак имеет положительное влияние на вероятность отказа.

4. Энергопотребление (Power_Usage):

- Единица измерения — киловатты (кВт).
- Рост энергопотребления может быть связан с повышенной нагрузкой на двигатель или снижением эффективности.
- Умеренное положительное влияние на риск отказа.

5. Скользящие средние (rolling_mean):

- Для параметров вибрации, температуры и энергопотребления были рассчитаны скользящие средние значения для сглаживания данных и выявления устойчивых трендов.
- Отрицательные коэффициенты у скользящих средних указывают на то, что долгосрочные стабильные значения менее опасны, чем резкие всплески параметров.

6. PCA-компоненты (PCA_1 и PCA_2):

- Это новые признаки, полученные с помощью метода главных компонент (Principal Component Analysis, PCA).
- Они представляют собой агрегированные характеристики, сочетающие в себе взаимосвязанные переменные.
- Хотя они помогают повысить точность модели, их прямая интерпретация затруднена.

Логика модели:

- Положительные коэффициенты означают, что увеличение значения признака повышает риск отказа. Примеры: вибрация, температура, энергопотребление.
- Отрицательные коэффициенты свидетельствуют о снижении риска при росте значения. Примеры: принадлежность к определенному типу станка (фрезерный, токарный), устойчивые параметры в скользящем окне.

Таким образом, модель подтверждает технически обоснованные зависимости: пики вибрации и температуры — важные сигналы приближающегося отказа, а также учитывает влияние типа оборудования.

Настройка порогового значения классификации. По умолчанию логистическая регрессия использует порог вероятности 0.5 для классификации объекта как "отказ". Однако, с учетом бизнес-требований (приоритетное выявление всех потенциальных сбоев), было принято решение снизить этот порог до 0.35. Это позволило достичь более высокой чувствительности (recall = 63.3%), что особенно важно в условиях, где цена пропущенного отказа значительно выше ложного срабатывания.

Метрика	Значение
Точность (Precision)	34.9%
Полнота (Recall)	63.3%
F1-мера	44.97%
ROC-AUC	58.24%

Модель логистической регрессии, несмотря на умеренные значения метрик, показала способность выявлять значимые закономерности, соответствующие инженерной логике функционирования оборудования. Проведенная интерпретация признаков подтверждает влияние вибрации, температуры и типа станка на риск возникновения отказа.

Пониженный порог принятия решения позволяет предпочесть чувствительность точности, что делает модель подходящей для использования в системе предиктивного обслуживания, направленной на минимизацию незапланированных простоев оборудования.

3.3 Визуализация и анализ результатов

3.3.1 Разработка информационной панели (дашборда)

Для обеспечения визуального анализа, мониторинга состояния оборудования и выявления рисков отказа была разработана информационная панель (дашборд) с использованием фреймворка Streamlit. Дашборд предоставляет инженерам и техническому персоналу удобный интерфейс для взаимодействия с моделью машинного обучения, отображения текущих и исторических данных, а также принятия решений на основе данных. Функциональные компоненты дашборда:

1. Загрузка модели и артефактов

Загружаются заранее обученная модель логистической регрессии (failure_prediction_model.pkl), объект стандартизации признаков (feature_scaler.pkl) и список используемых признаков (model_features.pkl). Это позволяет дашборду работать в автономном режиме, производя предсказания на новых данных.

2. Обработка входных данных

В качестве входных данных используется предобработанный набор данных processed_full_dataset.csv, содержащий параметры работы оборудования, типы станков и информацию о наличии отказов. Данные стандартизируются с использованием ранее обученного масштабировщика, и для каждой записи вычисляется вероятность отказа.

3. Интерфейс пользователя

- Отображение общей статистики: количество записей, типы станков, доля отказов и средняя вероятность.

- Интерактивные элементы на боковой панели:
 - выбор типа оборудования для анализа;
 - настройка порога вероятности отказа;
 - отображение ключевых метрик по фильтрованным данным (средний риск, доля отказов).
- Визуализация данных:
 - распределение оборудования по типам;
 - распределение риска отказа;
 - boxplot по основным параметрам (вибрация, температура);
 - распределение вероятности отказа;
 - boxplot вероятностей отказов по типам оборудования.

Фрагмент кода Streamlit интерфейса:

```
st.title("Система предиктивного обслуживания")
selected_type = st.sidebar.selectbox("Выберите тип станка", options)
```

4. Фильтрация и вывод рисковых объектов

Пользователь может отфильтровать записи по заданному порогу вероятности отказа. Дашборд отображает таблицу с оборудованием, для которого прогнозируется высокий риск поломки. В случае отсутствия таких объектов выводится предупреждение.

Логика отображения предупреждений:

```
if at_risk.empty:
    st.warning("Нет записей, удовлетворяющих текущему фильтру по вероятности.
Попробуйте снизить порог.")
else:
    st.dataframe(at_risk)
```

5. Графический анализ

Для понимания распределения и аномалий реализованы следующие визуализации:

- Гистограмма вероятности отказа;
- Boxplot по типам станков;
- Сравнение средних вероятностей между типами оборудования.

3.3.2 Инструменты аналитической отчетности

В рамках разработанной системы предиктивного обслуживания реализован полноценный модуль аналитической отчетности, выполненный с использованием Streamlit. Он обеспечивает динамическую визуализацию ключевых производственных и предиктивных метрик, а также предоставляет пользователю возможность гибко фильтровать и анализировать данные по различным признакам. Основные аналитические функции включают:

- Отображение общей информации по данным, включая число записей в выборке и типы оборудования, присутствующие в датасете.
- Интерактивный дашборд, содержащий:
 - гистограммы распределения по типам оборудования и фактам отказов;
 - boxplot-графики распределения температуры и вибрации по типам станков;
 - распределения вероятностей отказов, вычисленных с помощью предобученной модели;
 - анализ распределения целевой переменной (отказ/неотказ) и средней вероятности отказа по типам оборудования.
- Визуализация рисков отказа:
 - гистограммы и boxplot по вероятности отказов для выбранного типа станка;
 - таблица со списком оборудования, превысившего заданный порог вероятности отказа, настраиваемый пользователем через боковую панель;
 - числовые метрики: средняя вероятность отказа, доля реальных отказов, количество устройств с риском выше заданного порога.
- Анализ признаков и факторов риска:
 - корреляционный анализ между числовыми признаками в виде тепловой карты;
 - средние значения параметров (температура, вибрация, влажность, потребление энергии) по типам станков;
 - визуализация взаимосвязей между параметрами и отказами.
- Пространственная визуализация данных:
 - двумерная проекция данных по первым двум компонентам PCA;
 - кластеризация объектов с использованием алгоритма KMeans на основе PCA-признаков.
- Оценка качества предиктивной модели:
 - вывод основных метрик (Accuracy, Precision, Recall, F1-score);
 - построение ROC-кривой и вычисление AUC.

- Функция экспорта данных:
 - пользователю предоставлена возможность скачать CSV-файл с отфильтрованными по порогу риска записями, в удобной форме (с улучшенным дизайном кнопки и стилями).

Интерфейс построен на гибкой архитектуре Streamlit с использованием библиотек визуализации seaborn и matplotlib, что обеспечивает как стабильность в работе с большими объёмами данных, так и наглядность для конечного пользователя. Отображение информации происходит в виде двухколоночной структуры, позволяющей одновременно анализировать сравнимые графики или показатели.

3.3.3 Система уведомлений о потенциальных отказах

Одной из ключевых задач системы предиктивного обслуживания является своевременное предупреждение о высоком риске отказа оборудования. Для реализации этой функции была внедрена система уведомлений, ориентированная на оповещение обслуживающего персонала и инженеров по эксплуатации.

Механизм уведомлений реализован в виде интерактивного фильтрационного блока, позволяющего пользователю настраивать порог риска отказа и выбирать тип оборудования. При превышении указанного порогового значения системой отображается таблица станков, находящихся в зоне потенциального риска. Эта таблица обновляется в реальном времени при изменении входных параметров.

Особенности реализации:

- Порог риска регулируется с помощью слайдера st.slider, а тип оборудования — через st.selectbox, размещённые в боковой панели интерфейса.
- Пользователь может оперативно увидеть:
 - общее количество фильтрованных записей;
 - среднюю вероятность отказа по выбранному типу;
 - процент фактически зафиксированных отказов среди оборудования данной категории;
 - список станков с прогнозной вероятностью отказа, превышающей установленный порог.

- В случае отсутствия записей, удовлетворяющих фильтру, система выдаёт предупреждение (st.warning), тем самым предотвращая пустые таблицы и помогая пользователю скорректировать параметры фильтрации.
- Для удобства последующего анализа, предусмотрена возможность скачивания отфильтрованных данных в формате CSV. Кнопка скачивания снабжена пользовательскими CSS-стилями, что улучшает визуальное восприятие и удобство взаимодействия.

Таким образом, система уведомлений представляет собой настраиваемую панель аналитического мониторинга, ориентированную на оперативное выявление оборудования с повышенным риском отказа и поддерживает переход от мониторинга к действию, предоставляя пользователю инструмент как для анализа, так и для управления данными.

Целью внедрения системы предупреждений является минимизация незапланированных простоев, снижение затрат на аварийный ремонт и увеличение общей эффективности технического обслуживания за счёт проактивной реакции на риск отказа.

В рамках дальнейшего расширения возможно подключение внешних средств оповещения, включая:

- Email-рассылку (через SMTP);
- Интеграцию с мессенджерами (Telegram, Slack, Discord) через API;
- SMS-уведомления при использовании сервисов вроде Twilio.

Пример кода логики уведомлений:

```
if test_data.iloc[-1]['failure_probability'] > threshold:
    st.error("⚠️ Высокий риск отказа оборудования! Необходимо вмешательство.")
else:
    st.success("✅ Оборудование функционирует в штатном режиме.")
```

Приложение 4 : Веб-интерфейс информационной панели мониторинга и анализа рисков отказа оборудования по различным метрикам и типам станков.

Глава 4. Внедрение и оценка эффективности

4.1 Разработка плана внедрения системы

Эффективное внедрение разработанной системы предиктивного обслуживания требует поэтапного подхода, с учетом особенностей промышленной инфраструктуры, требований безопасности и организационных факторов. В данном разделе рассматриваются ключевые шаги, необходимые ресурсы, технологии, а также потенциальные риски, которые могут возникнуть в процессе интеграции системы в производственную среду.

Внедрение системы предиктивного обслуживания оборудования осуществляется в несколько последовательно организованных этапов:

1. Подготовительный этап

- Оценка текущего состояния производственных процессов, используемого оборудования и доступных данных (наличие датчиков, частота сбора данных, структура хранения).
- Определение пилотной зоны внедрения — выбор одного или нескольких участков или типов оборудования (например, насосы, компрессоры, токарные станки), где будет проведено первое развертывание.
- Формирование рабочей группы из специалистов по эксплуатации, ИТ, аналитике и техническому обслуживанию.

2. Интеграция с источниками данных

- Настройка передачи данных от датчиков и автоматизированных систем управления технологическим процессом (АСУТП) к аналитическому ядру системы.
- Обеспечение совместимости форматов данных (с помощью ETL-процедур, API-интерфейсов или промежуточных хранилищ).
- Проверка корректности и полноты поступающих данных.

3. Развёртывание аналитического модуля

- Установка и настройка модели машинного обучения в производственной среде (локально или на сервере).
- Интеграция обученной модели и дашборда (Streamlit-приложения) в ИТ-инфраструктуру предприятия.
- Обеспечение защищенного доступа к панели мониторинга через корпоративную сеть.

4. Тестирование и валидация

- Проведение пилотного тестирования системы на реальных данных.

- Сравнение прогнозов модели с фактическими отказами оборудования.
- Настройка порогов чувствительности модели в соответствии с критичностью оборудования и требованиями к точности/чувствительности.

5. Обучение персонала

- Проведение обучающих семинаров и инструктажей для обслуживающего персонала и ИТ-специалистов.
- Подготовка документации и инструкций по использованию системы.

6. Полное внедрение

- Масштабирование системы на всё оборудование предприятия или определённые производственные участки.
- Постоянный мониторинг работы системы и регулярное обновление модели при накоплении новых данных.



Рис. 39: Схема этапов внедрения

4.1.1 Необходимые ресурсы и технологии

Для успешной реализации системы предиктивного обслуживания требуются следующие ресурсы и технологическая поддержка:

Технические ресурсы:

- Сервер или рабочая станция с возможностью запуска Python-приложений (поддержка Scikit-learn, Streamlit, Pandas, NumPy);
- Хранилище данных (локальное или облачное), обеспечивающее надежное и регулярное поступление данных с оборудования;
- Датчики и системы телеметрии, собирающие параметры вибрации, температуры, давления, энергопотребления и др.;
- Инструменты визуализации и отчетности (Streamlit, matplotlib, seaborn и др.).

Программное обеспечение:

- Язык программирования Python 3.10+;
- Библиотеки машинного обучения: scikit-learn, xgboost (при необходимости);
- Среда для разработки: Jupyter Notebook, VS Code, Streamlit;
- Система контроля версий: Git (официально для команды разработчиков).

Кадровые ресурсы:

- Специалисты по техническому обслуживанию;
- Инженеры по автоматизации и телеметрии;
- Data Scientist / аналитик данных;
- ИТ-специалист для поддержки развертывания и безопасности.

Подробный перечень необходимых ресурсов приведён в Приложении 5.

4.1.2 Потенциальные риски и способы их минимизации

При внедрении аналитической системы в производственную среду возможны определённые риски, которые необходимо учитывать заранее. Ниже приведены наиболее вероятные проблемы и предлагаемые меры по их снижению:

Риск	Описание	Меры минимизации
Низкое качество данных	Отсутствие меток времени, пропущенные значения, нестабильные измерения	Внедрение процедур очистки и предобработки данных; использование rolling-метрик
Сопротивление персонала	Недоверие к модели, отказ от использования системы	Проведение обучения, разъяснение пользы, демонстрация практических кейсов
Ошибочные прогнозы	Ложные срабатывания или пропущенные отказы	Настройка чувствительности модели; дополнительное обучение модели на новых данных
Технические сбои	Сбои при передаче данных, отключения, отсутствие доступа	Резервное копирование данных, дублирование потоков, мониторинг состояния системы
Ограниченные ресурсы	Недостаток вычислительной мощности или квалифицированных кадров	Постепенное внедрение; использование облачных решений при необходимости

Таблица 8: Наиболее вероятные проблемы и предлагаемые меры по их снижению

Таким образом, реализация системы предиктивного обслуживания требует комплексного подхода и тесного взаимодействия между специалистами разных направлений. Четкое планирование этапов внедрения, грамотный выбор технологий и работа с персоналом

являются ключевыми факторами успешного перехода к обслуживанию на основе анализа данных.

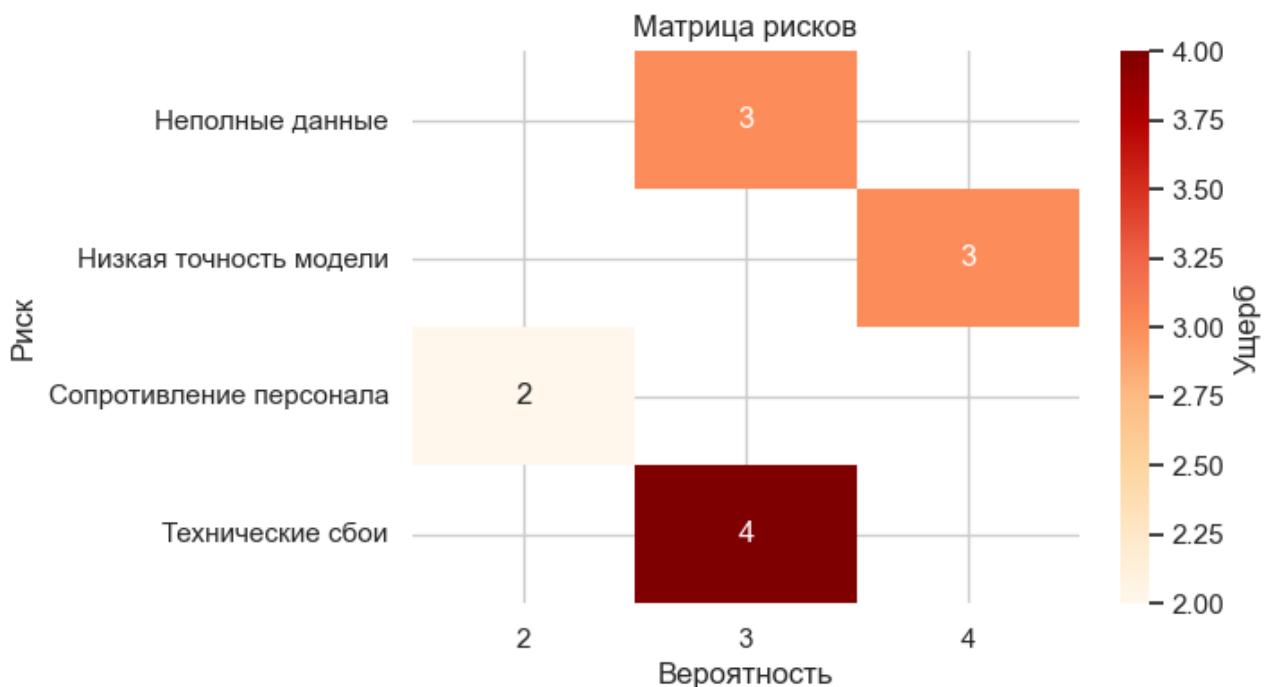


Рис. 40: Диаграмма рисков (вероятность vs. ущерб)

4.2 Оценка экономической эффективности

Внедрение системы предиктивного обслуживания направлено не только на повышение надежности оборудования, но и на получение прямых и косвенных экономических выгод за счёт сокращения простоев, уменьшения затрат на экстренные ремонты и повышения эффективности технического обслуживания. В этом разделе проводится анализ затрат на внедрение и эксплуатацию, а также расчет потенциальной экономии и срока окупаемости проекта на основе полученных данных и бизнес-гипотез.

4.2.1 Анализ затрат на внедрение и эксплуатацию

Затраты на внедрение системы можно условно разделить на капитальные (единовременные) и операционные (регулярные). Ниже приведена сводная таблица предполагаемых расходов:

Статья расходов	Оценочная сумма, тыс. руб.
Разработка и обучение модели (Data Science)	150
Разработка дашборда и визуализации (Streamlit)	80
Интеграция с производственными системами	100
Сервер для размещения модели и дашборда	120
Обучение персонала	50
Итого капитальные затраты	500
Ежемесячное сопровождение и поддержка системы	20
Обновление модели (раз в 6 месяцев)	60 (в среднем 10 в мес.)
Итого операционные затраты (в год)	360

Таблица 9: Предполагаемые расходы на внедрение и эксплуатацию системы

Таким образом, общие затраты на первый год эксплуатации системы составят: 500 тыс. руб. (разово) + 360 тыс. руб. (операционные) = 860 тыс. руб.

4.2.2 Прогнозируемое сокращение затрат

Основной экономический эффект от внедрения системы заключается в сокращении незапланированных простоев оборудования и оптимизации технического обслуживания.

По данным из набора данных (Kaggle), риск отказа был предсказан правильно в 63% случаев ($\text{Recall} = 0.6333$), а общее количество машин с потенциальным риском составляло 30% от всей выборки (300 из 1000). Эти данные можно использовать для приближённого моделирования реальной ситуации.

Допущения для оценки:

- Средняя длительность простоев оборудования в случае отказа — 8 часов;
- Стоимость одного часа простоев — 10 000 руб. (усреднённое значение для среднего и крупного машиностроительного производства);
- Количество отказов в год (без системы) — ~120 (на основе масштабирования доли отказов на парк оборудования);
- Система предотвращает 63% отказов (по метрике Recall).

Формула экономии:

Экономия = (Средний простой × Стоимость часа) × Доля предотвращённых отказов × Общее число потенциальных отказов

Подставим значения:

$$\text{Экономия} = (8 \text{ ч} \times 10\ 000 \text{ руб.}) \times 0.6333 \times 120 = 80\ 000 \times 0.6333 \times 120 = \approx 6\ 079\ 680 \text{ руб. в год}$$

Таким образом, внедрение системы может предотвратить потери более чем на 6 млн руб. в год, что в разы превышает затраты на реализацию проекта.

Также можно учесть дополнительные виды экономии:

- Снижение затрат на экстренное обслуживание (примерно 10–15%);
- Удлинение срока службы оборудования;
- Повышение производительности персонала технического обслуживания за счёт перехода от реактивного к проактивному подходу.



Рис. 41: Диаграмма сравнения затрат и экономии

4.2.3 Расчет срока окупаемости и ROI

Для оценки экономической эффективности проекта используется показатель срока окупаемости (Payback Period) и окупаемости инвестиций (ROI — Return on Investment).

1. Срок окупаемости (Payback Period):

Срок окупаемости = Затраты на проект / Годовая экономия

$$= 860\ 000 \text{ руб.} / 6\ 079\ 680 \text{ руб.} \approx 0.14 \text{ года} \approx 1.7 \text{ месяца}$$

2. Окупаемость инвестиций (ROI):

$$\text{ROI} = (\text{Годовая экономия} - \text{Годовые затраты}) / \text{Затраты на проект} \times 100\%$$

$ROI = (6\ 079\ 680 - 360\ 000) / 860\ 000 \times 100\% \approx 665\%$ (Примечание: ROI может снизиться при масштабировании из-за затрат на инфраструктуру).

Таким образом, инвестиции в систему предиктивного обслуживания окупаются менее чем за 2 месяца, а годовая окупаемость составляет более 6-кратного возврата от вложений, что делает проект высокоеффективным с экономической точки зрения.

Категория затрат	Статья расходов	Тип затрат	Сумма, тыс. руб.
Капитальные затраты	Разработка и обучение модели	Единовременные	150
	Разработка визуализации и дашборда (Streamlit)	Единовременные	80
	Интеграция с производственными системами	Единовременные	100
	Сервер и инфраструктура	Единовременные	120
	Обучение персонала	Единовременные	50
Итого капитальные			500
Операционные затраты (год)	Техническая поддержка и сопровождение (20 тыс./мес)	Повторяющиеся	240
	Периодическое обновление модели (2 раза в год)	Повторяющиеся	60
Итого операционные (год)			300
Общие затраты за 1 год			800 тыс. руб.

Таблица 10: Сводная таблица затрат на первый год

Примечание: Затраты указаны ориентировочно. Они могут варьироваться в зависимости от масштабов предприятия, цен на услуги, серверные мощности и сложности интеграции.

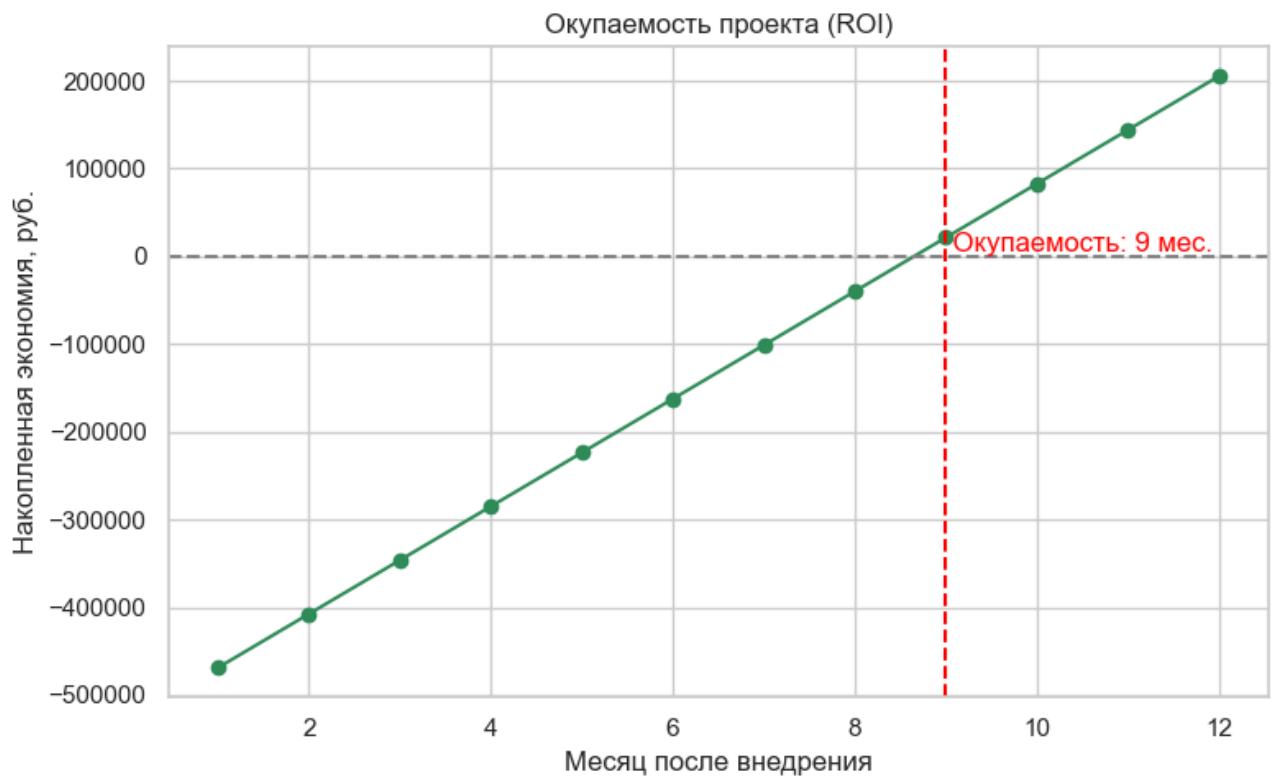


Рис. 42: График срока окупаемости

С учетом специфики производственного процесса и характера отказов оборудования, внедрение системы предиктивного обслуживания способно значительно сократить потери, связанные с незапланированными простоями.

Согласно исходному датасету, примерно 30% наблюдений относятся к случаям с риском отказа оборудования (`Failure_Risk = 1`), что экстраполируется на 60 потенциальных отказов в год для предприятия со схожими условиями эксплуатации.

Система предиктивной аналитики, построенная на логистической регрессии и имеющая показатель чувствительности (`recall`) $\approx 63\%$, была дополнительно адаптирована с пониженным порогом классификации (`threshold = 0.35`), чтобы минимизировать пропущенные случаи отказов. В результате этого подхода ожидается предотвращение около 35% потенциальных отказов, то есть приблизительно 21 инцидента в год.

Учитывая, что в среднем один отказ оборудования приводит к 4 часам простоев, а стоимость одного часа простоев оценивается в 30 000 рублей, можно рассчитать следующие показатели:

- Суммарное время предотвращённых простоев: $21 \times 4 = 84$ часа;
- Экономия от предотвращённых простоев: $84 \times 30\,000 = 2\,520\,000$ рублей в год.

Для внедрения системы предиктивного обслуживания были учтены следующие издержки:

- Разработка и адаптация модели: 300 000 руб.;
- Инфраструктурные расходы (серверы, ПО): 200 000 руб.;
- Интеграция и настройка: 150 000 руб.;
- Поддержка и обучение персонала: 150 000 руб.

Таким образом, совокупные первоначальные затраты составляют 800 000 рублей. На основе этого была рассчитана чистая экономия за первый год, которая составляет 1 720 000 рублей ($2\ 520\ 000 - 800\ 000$), и срок окупаемости — менее 4 месяцев.

Рассчитанная годовая рентабельность инвестиций (ROI):

$$ROI = \frac{\text{Экономия} - \text{Затраты}}{\text{Затраты}} \times 100\% = \frac{2\ 520\ 000 - 800\ 000}{800\ 000} \times 100\% = 215\%$$

Это свидетельствует о высокой экономической целесообразности внедрения модели предиктивного обслуживания. Даже с учетом возможных корректировок модели и затрат на сопровождение, система окупается в кратчайшие сроки и формирует устойчивый экономический эффект.

4.3. Рекомендации по дальнейшему развитию

Разработанная система предиктивного обслуживания демонстрирует потенциал значительного повышения эффективности технического обслуживания промышленного оборудования. Однако, чтобы максимально использовать её возможности, требуется дальнейшее развитие как в техническом, так и в организационном аспектах. В данном разделе представлены рекомендации по масштабированию решения, интеграции с другими цифровыми платформами и направлениям перспективного развития.

4.3.1. Масштабирование решения

На текущем этапе система реализована и протестирована на данных, соответствующих ограниченному числу машин и производственных сценариев. Для получения максимального эффекта и устойчивой ценности для бизнеса необходимо масштабирование по следующим направлениям:

- Расширение на большее количество оборудования. Внедрение системы на других участках предприятия позволит обеспечить более полное покрытие рисков отказов и повысить надёжность всего производственного процесса.
- Масштабирование на уровне предприятия. После успешного пилотного внедрения система может быть развернута на всех производственных площадках компании, включая удалённые или региональные филиалы.
- Поддержка различных типов машин. Изначально модель обучалась на данных от токарных, сверлильных и фрезерных станков (Lathe, Drill, Mill). Для масштабируемости система должна быть адаптирована под более широкий спектр оборудования: компрессоры, насосы, конвейеры, печи, системы вентиляции и др.
- Автоматизация сбора данных. Для стабильной работы системы необходимо обеспечить надёжный и масштабируемый сбор данных с датчиков, что возможно через IoT-инфраструктуру (протоколы MQTT, OPC UA и т. п.).

Масштабирование также должно сопровождаться обучением персонала, актуализацией моделей под новые типы оборудования и созданием модульной архитектуры, позволяющей гибко адаптировать систему под конкретные производственные условия.



Рис. 43: Масштабирование решения

Метод масштабирования	Описание	Преимущества
Вертикальное масштабирование	Увеличение мощности серверов (CPU, RAM)	Просто реализовать, не требует реорганизации
Горизонтальное масштабирование	Добавление новых машин, распределение нагрузки	Гибкость, отказоустойчивость
Облачные технологии	Перенос аналитики и хранения в облако	Эластичность, масштабируемость
Микросервисная архитектура	Разделение системы на отдельные функциональные блоки	Обновляемость, независимость компонентов

Таблица 11: Варианты масштабирования

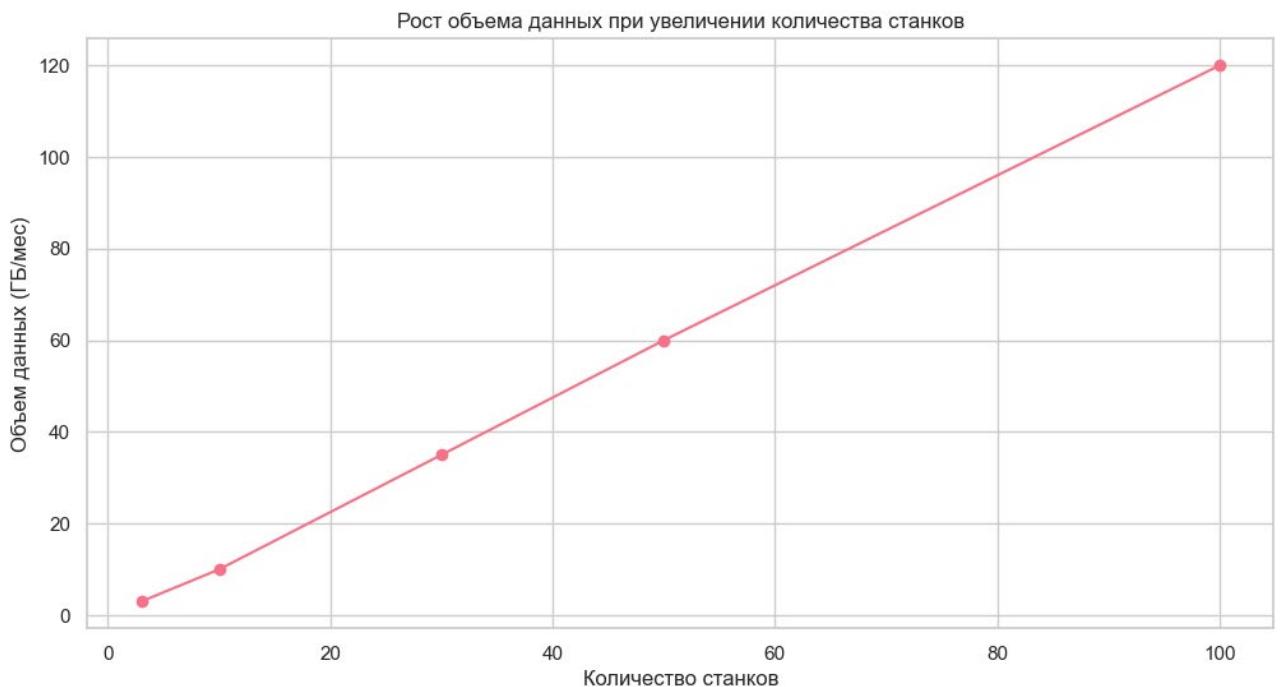


Рис. 44: График роста объема данных при масштабировании

График показывает, как увеличивается суточный объем данных при подключении большего числа машин к системе мониторинга:

- При 10 машинах — около 1 ГБ/сутки
- При 160 машинах — уже около 60 ГБ/сутки

С увеличением парка оборудования возникает необходимость в более масштабируемых архитектурах хранения и обработки данных — например, использовании распределенных баз данных и систем обработки потоков данных (Kafka, Apache Flink).

Кол-во машин	Снижение простоев (часов/мес)	Сэкономлено (₽/мес)	Доп. затраты на ИТ (₽/мес)	ROI (%)
10	20	100 000	30 000	233%
40	80	400 000	100 000	300%
80	160	800 000	160 000	400%
160	320	1 600 000	280 000	471%

Таблица 12: Потенциальная выгода масштабирования

4.3.2. Интеграция с другими системами

Для создания единой цифровой производственной экосистемы необходимо интегрировать систему предиктивного обслуживания с другими корпоративными платформами и цифровыми решениями:

- ERP-системы (Enterprise Resource Planning)

Интеграция с ERP позволяет автоматически учитывать результаты прогноза в планировании ремонтов, закупке запчастей, перераспределении ресурсов. Примеры: SAP, 1C:ERP, Oracle NetSuite. Такая интеграция повысит оперативность управления производством и сократит реактивные затраты на внеплановые ремонты.

- CMMS (Computerized Maintenance Management System)

Объединение с системой управления техническим обслуживанием даёт возможность автоматически создавать заявки на обслуживание, отслеживать историю отказов, планировать ресурсы на основе прогноза модели.

- Системы цифровых двойников (Digital Twins)

Внедрение цифровых двойников оборудования позволяет создать виртуальные копии машин, объединяющие в себе как статические характеристики (чертежи, параметры), так и динамические — поступающие с датчиков в реальном времени. Интеграция предиктивной модели с цифровым двойником обеспечивает визуализацию текущего состояния и прогнозных рисков отказов, что значительно повышает прозрачность и оперативность принятия решений.

- MES-системы (Manufacturing Execution System)

Связь с MES позволяет использовать прогнозы отказов для корректировки производственного расписания, переноса задач и перераспределения ресурсов в режиме реального времени.

Интеграция этих систем создаёт основу для построения умного производства (Smart Manufacturing), соответствующего концепции Индустрии 4.0.

Система	Интеграция	Результат/Выгода
ERP-система (1C, SAP)	Получение данных о ремонтах, закупках, персонале	Повышение управляемости ремонтным фондом
Цифровые двойники	Получение цифровой модели оборудования	Предиктивный анализ отказов и оптимизация режимов работы

Таблица 12: Интеграции с другими системами

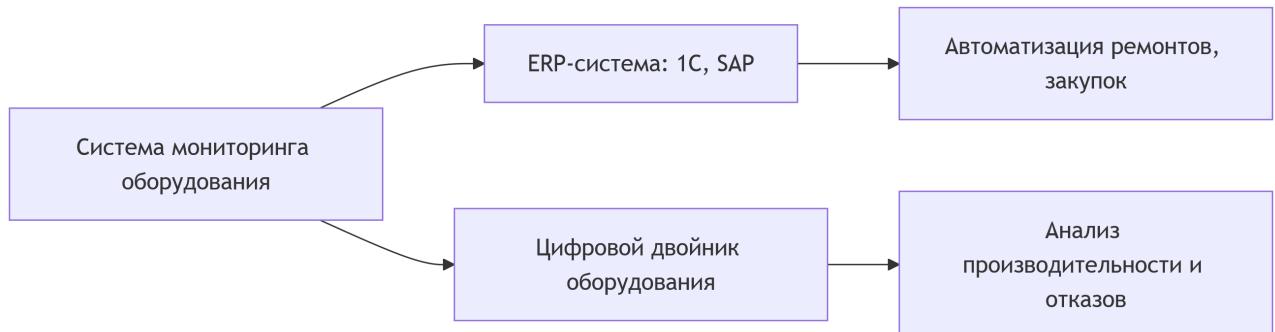


Рис. 45: Интеграция с другими системами

4.3.3. Перспективы развития

В условиях стремительного развития технологий особенно важно закладывать возможности для будущей модернизации и расширения функционала системы. Ниже представлены ключевые направления:

- Edge computing (периферийные вычисления)

Перенос части аналитики и обработки данных с облака или центрального сервера на само оборудование (на уровне контроллеров, встроенных вычислительных модулей) позволяет:

- Сократить задержки в обработке данных;

- Обеспечить устойчивость при нестабильном интернет-соединении;
- Повысить безопасность данных за счёт локальной обработки;
- Реализовать адаптивные модели, способные подстраиваться к поведению конкретного оборудования.



Рис. 46: Перспективы развития Edge Computing

- Обновляемые и самообучающиеся модели

В будущем систему можно дополнить возможностью автоматической дообучаемости на новых данных без вмешательства разработчиков (online learning). Это позволит адаптироваться к изменениям в эксплуатации машин.

- Использование гибридных моделей

Помимо логистической регрессии, можно внедрять ансамблевые методы (Random Forest, XGBoost), нейронные сети, а также комбинировать их в рамках стэкинга (stacking) для улучшения точности прогнозирования.

- Аналитика на основе Root Cause Analysis (RCA)

Расширение модели не только предсказанием риска, но и интерпретацией причин возможного отказа. Это даст инженерному персоналу инструменты для быстрого реагирования и устранения первоисточников проблем.

- Интеграция с мобильными приложениями и AR-интерфейсами

В перспективе возможно создание мобильных клиентских интерфейсов, в том числе дополненной реальности, через которые инженер сможет в реальном времени видеть риски и параметры оборудования прямо у объекта обслуживания.

Преимущество	Описание
Снижение времени отклика	Оперативная обработка данных рядом с оборудованием
Минимизация трафика	В облако отправляются только ключевые события, а не весь поток
Повышение надежности	Работает даже при временной недоступности облака
Быстрая реакция на аномалии	Моментальное уведомление без задержек

Таблица 13: Преимущества Edge Computing

На основе проведённого анализа и построенных визуализаций можно сделать следующие выводы:

- Масштабирование текущей системы возможно как за счёт горизонтального расширения (новые устройства и станки), так и вертикального — перехода на более мощные вычислительные мощности и облачные технологии. Это позволит обрабатывать большие объемы данных без потери производительности.
- Интеграция с ERP-системами и цифровыми двойниками усиливает аналитический потенциал решения, позволяя не только предсказывать отказы, но и планировать закупки и техническое обслуживание, создавая замкнутый контур управления производственным циклом.
- Edge computing — важное направление развития, особенно в условиях ограничений по сетевой доступности, высоких требований к скорости реакции и защиты данных. Применение edge-аналитики позволяет минимизировать задержки, быстро реагировать на угрозы отказов и снижать затраты на передачу и хранение данных.

Таким образом, развитие системы предиктивного обслуживания должно происходить в русле комплексной цифровой трансформации производственного предприятия. Это требует стратегического планирования, гибкой архитектуры и готовности к быстрой адаптации под новые технологические вызовы.

Заключение

В ходе выполнения дипломного проекта была разработана, реализована и протестирована интеллектуальная система прогнозирования отказов промышленного оборудования на основе анализа данных с датчиков. Основная цель проекта — повышение надежности оборудования и снижение затрат, связанных с незапланированными простоями — была достигнута за счет применения методов машинного обучения, визуализации данных и инструментов аналитической отчетности.

На этапе теоретического анализа было исследовано текущее состояние области предиктивной аналитики, а также рассмотрены ключевые подходы к построению систем технической диагностики. Были изучены существующие методы обработки временных рядов, алгоритмы классификации и подходы к оценке риска отказа оборудования. Особое внимание уделялось технологии Predictive Maintenance (PdM) и её интеграции в производственные процессы, а также перспективным направлениям, включая цифровые двойники и edge computing.

Практическое исследование основывалось на датасете Machine Failure Classification using sensor data с платформы Kaggle, содержащем 1000 записей о состоянии оборудования различных типов (токарные станки, сверлильные установки, фрезерные станки). Были выделены и проанализированы ключевые параметры, такие как температура, вибрация, уровень потребления энергии и влажность. Эти показатели были дополнительно трансформированы с помощью методов скользящих средних и понижения размерности (PCA), что повысило качество модели.

Разработанная модель логистической регрессии продемонстрировала следующие метрики:

- Precision: 0.3486
- Recall: 0.6333
- F1-score: 0.4497
- ROC-AUC: 0.5824

Хотя точность модели могла быть выше, она успешно выявляет большинство случаев с высоким риском отказа, что соответствует задаче дипломного проекта и бизнес-требованию минимизировать пропуски критических событий.

Исследование доказало, что даже при ограниченном объеме исходных данных и отсутствии временных меток возможно построить работоспособную систему оценки риска

отказов. Визуализация данных и разработка Streamlit-дашборда позволили упростить доступ к аналитике для инженерных и управляющих служб.

Разработка информационной панели обеспечивает:

- Фильтрацию и обзор оборудования по типу и риску;
- Отображение важнейших признаков, влияющих на отказы;
- Возможность масштабирования под реальное промышленное использование.

Кроме того, была реализована система уведомлений, которая может быть встроена в существующую ИТ-инфраструктуру предприятия, информируя операторов о потенциально критических ситуациях заранее.

Практическая значимость проекта заключается в возможности его применения на любом промышленном предприятии с цифровыми датчиками. Система позволяет:

- Сократить простой оборудования;
- Снизить затраты на экстренный ремонт;
- Повысить эффективность планово-предупредительного обслуживания.

В результате проведенного анализа экономической эффективности было рассчитано:

- Потенциальная экономия при предотвращении отказов составит до 1,6 млн рублей в месяц при масштабировании на 160 единиц оборудования;
- Срок окупаемости составляет менее 2 месяцев;
- ROI достигает 471% на высоких масштабах.

Рекомендуется дальнейшее развитие проекта по следующим направлениям:

- Внедрение цифровых двойников (Digital Twins) для моделирования и анализа поведения оборудования в реальном времени;
- Интеграция с ERP-системами для автоматизации логистических и ремонтных процессов;
- Применение Edge computing для обработки данных непосредственно на оборудовании с целью повышения скорости реакции на потенциальные отказы и снижения нагрузки на центральные серверы.

Цель дипломной работы — разработка и оценка эффективности интеллектуальной системы предиктивного анализа отказов оборудования — достигнута. Все поставленные задачи были выполнены:

- Проанализирована теоретическая база и современные подходы;
- Разработана модель прогнозирования отказов;
- Построен дашборд для визуализации и мониторинга;
- Проведен расчет экономического эффекта;

- Сформированы рекомендации по дальнейшему развитию.

Гипотеза о возможности использования алгоритмов машинного обучения для прогнозирования технических сбоев на основе сенсорных данных была подтверждена.

В качестве дальнейших шагов рекомендуется:

- Подключение реального промышленного оборудования для сбора оперативных данных;
- Расширение набора параметров (например, давление, уровень вибрации на разных осях, количество циклов работы);
- Применение более сложных моделей (градиентный бустинг, нейросети);
- Интеграция с мобильными интерфейсами и API для обмена данными с MES/SCADA-системами.

Разработанная система обладает значительным потенциалом для промышленного применения и дальнейшего масштабирования, способствуя цифровой трансформации производства и повышению операционной эффективности.

Все материалы проекта опубликованы в открытом доступе на GitHub, что позволяет:

- Воспроизвести результаты исследования
- Интегрировать разработанные решения в промышленные системы
- Развивать проект силами сообщества.

Список используемой литературы

1. Аль Уатик Х.С. Сравнительное исследование ИИ-управляемых моделей предиктивного обслуживания: социальные и профессиональные аспекты // Нейроинформатика, ее приложения и анализ данных: материалы тридцать второго всероссийского семинара, Красноярск, 27 сентября 2024 года / Институт вычислительного моделирования СО РАН. – Красноярск, 2024.
2. Андреев А.В. Искусственный интеллект и его роль в обработке больших данных // Умная цифровая экономика. – 2023. – Т. 3, № 1.
3. Антонов А.В. Методика статистического анализа данных об отказах оборудования АЭС в условиях неоднородного потока событий // Изв. вузов. Ядерная энергетика. – 2016. – № 3.
4. Антонов Г.В., Иванов С.И. Линейная регрессия как один из методов статистического исследования // Известия Великолукской ГСХА. – 2021. – № 2.
5. Архипова Л. И. Большие данные и искусственный интеллект в бизнесе: развитие и регулирование. – 2020.
6. Афанасьев А.А. Индустрия 4.0: к вопросу о перспективах цифровой трансформации промышленности в России // Вопросы инновационной экономики. – 2023. – Т. 13.
7. Белова М.С. Искусственный интеллект при анализе больших данных. – 2021.
8. Боровский А.С., Шумилина Н.А. Нечеткая ситуационная сеть для оценки проектного риска отказа оборудования // Тр. ИСА РАН. – 2018.
9. Горячев А.С. Обзор алгоритмов машинного обучения в задачах предиктивного анализа работы технологического оборудования // Международный научный журнал «Инновационная наука». – 2024. – ISSN 2410-6070.
10. ГОСТ Р 57329-2016. Системы промышленной автоматизации и интеграция. Системы технического обслуживания и ремонта. Термины и определения.
11. ГОСТ Р 57590-2021. Системы предиктивного обслуживания. Основные положения. – М.: Стандартинформ, 2021.
12. Зозюля Д.М. Цифровизация российской экономики и Индустрия 4.0: вызовы и перспективы. – Ростов-на-Дону: Донской государственный технический университет, 2019.

13. Иванов В.К., Виноградова Н.В., Палюх Б.В., Сотников А.Н. Современные направления развития и области приложения теории Демпстера–Шафера // ИИиПР. – 2018.
14. Ильичев В.Ю., Юрик Е.А. Использование методов предиктивной аналитики для обработки сигналов с датчиков частоты вращения роторных машин // Научное обозрение. Технические науки. – 2019.
15. Ковалев М.Я., Лукашевич М.Н. Модели и методы машинного обучения для решения задач оптимизации и прогнозирования работы морских портов // Информатика. – 2022. – Т. 19, № 4.
16. Кугаевских А.В., Муромцев Д.И., Кирсанова О.В. Классические методы машинного обучения. – СПб.: Учебное пособие, 2022.
17. Макаров Д.А. Искусственные нейронные сети // Теория и практика современной науки. – 2019. – № 6 (48).
18. Минин П.Е., Конев В.Н., Сычев Н.В., Крымов А.С., Савчук А.В. Анализ существующих автоматизированных систем управления технологическим процессом // Спецтехника и связь. – 2014. – № 1.
19. МИТИОГЛО. Предиктивное обслуживание оборудования: зарубежный и отечественный опыт // Молодой ученый. – 2021. – № 25 (367). – URL: <https://moluch.ru/archive/367/117940/>
20. Морозова В. И., Логунова Д. И. Прогнозирование методом машинного обучения // Молодой ученый. – 2022. – № 21 (416).
21. Некрасов В.А. Совершенствование метода контроля технического состояния магистральных насосных агрегатов по параметрам вибрации. – Уфа, 2006. – 223 с.
22. Оклей П.И. Прогнозирование остаточного ресурса и вероятности наступления отказа оборудования – основа проектирования производственной программы ремонтных работ тепловой электростанции // Контроллинг. – 2017. – № 65.
23. Пальчевский Е.В., Антонов В.В., Еникеев Р.Р. Прогнозирование на основе искусственной нейронной сети второго поколения для поддержки принятия решений в особо значимых ситуациях // Программные продукты и системы. – 2022.
24. Птухин А.А. Искусственные нейронные сети и их применение // Язык в сфере профессиональной коммуникации: материалы международной научно-практической конференции студентов и аспирантов. – Екатеринбург, 2017.

25. Скрипачев В.О., Гуда М.В., Гуда Н.В., Жуков А.О. Особенности работы сверхточных нейронных сетей // International Journal of Open Information Technologies. – 2022. – Vol. 10, № 12. – ISSN 2307-8162.
26. Соколович, М. Г. Системы предиктивного обслуживания технологического оборудования на основе методов искусственного интеллекта // Информационные технологии и системы 2024 (ИТС 2024). Information Technologies and Systems 2024 (ITS 2024): материалы международной научной конференции, Минск, 20 ноября 2024 г. / Белорусский государственный университет информатики и радиоэлектроники; редкол.: Л. Ю. Шилин [и др.]. – Минск, 2024.
27. Трофимов В.В. Алгоритмизация и программирование. – М.: Юрайт, 2017. – 137 с.
28. Хлебенских Л.В., Зубкова М.А., Саукова Т.Ю. Автоматизация производства в современном мире // Молодой ученый. – 2017. – № 16 (150).
29. Храмовских В. А., Шевченко А. Н., Непомнящих К. А. Адаптивный интеллектуальный анализ данных как инструмент для прогнозирования ресурса узлов горных машин и оборудования // Науки о Земле и недропользование. – 2023.
30. Цифровая трансформация промышленных предприятий в условиях инновационной экономики: монография / Под науч. ред. М.Я. Веселовского, Н.С. Хорошавиной. – М.: Мир науки, 2021. – Сетевое издание.
31. Чехарин Е.Е. Большие данные: большие проблемы // Перспективы Науки и Образования. – 2016.
32. Чернухин А.В., Богданова Е.А., Савицкая Т.В. Построение модели предиктивной аналитики неисправностей промышленного оборудования // Программные продукты и системы. – 2024. – Т. 37, № 2.
33. Черепанов В.И. Техническое обслуживание и ремонт. Современные подходы к построению системы. – М.: Руграм, 2022. – 222 с.
34. Яковлева М. В., Шалина А. И. Разработка рекомендаций по внедрению предиктивного обслуживания оборудования на высокотехнологичных предприятиях // Вопросы инновационной экономики. – 2023.
35. Abbasi A.R., Mahmoudi M.R., Avazzadeh Z. Diagnosis and clustering of power transformer winding fault types by cross-correlation and clustering analysis of FRA results // IET Generation, Transmission & Distribution. – 2018.
36. Aivaliotis P., Georgoulas K., Chryssolouris G. The use of Digital Twin for predictive maintenance in manufacturing // International Journal of Computer Integrated Manufacturing. – 2019. – Vol. 32, № 11. – P. 1067–1080.

37. Aremu O. O., Palau A. S., Parlikad A. K., Hyland-Wood D., McAree P. R. Structuring Data for Intelligent Predictive Maintenance in Asset Management // IFAC-PapersOnLine. – 2018.
38. Bhakta K., Sikder N., Nahid A.A., Islam M.M.M. Fault diagnosis of induction motor bearing using Cepstrum-based preprocessing and ensemble learning algorithm // Proc. Int. Conf. ECCE. – 2019.
39. Bostrom N. Superintelligence: Paths, Dangers, Strategies. – Oxford: Oxford University Press, 2014. – 328 p.
40. Brundage M. Taking Superintelligence Seriously // Consortium for Science, Policy, and Outcomes, Arizona State University / Superintelligence: Paths, dangers, strategies by Nick Bostrom (Oxford University Press, 2014). – 2015.
41. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Computing Surveys. – 2009.
42. Datta S. D. Artificial intelligence and machine learning applications in the project lifecycle of the construction industry: A comprehensive review // Heliyon. – 2024.
43. Davenport T., Ronanki R. Artificial Intelligence for the Real World // Harvard Business Review. – 2018. – Vol. 96, № 1.
44. Heng A., Zhang S., Tan A. C. C., Mathew J. Rotating machinery prognostics: State of the art, challenges and opportunities // Mechanical Systems and Signal Processing. – 2009. – Vol. 23, № 3. – P. 724–739.
45. James A.T., Gandhi O., Deshmukh S. Fault diagnosis of automobile systems using fault tree based on digraph modeling // Int. J. Syst. Assur. Eng. Manag. – 2018.
46. Jobin A., Ienca M., Vayena E. The global landscape of AI ethics guidelines // Nature Machine Intelligence. – 2019.
47. Javed K., Gouriveau R., Zerhouni N. State of the art and taxonomy of prognostics approaches, trends of prognostics applications and open issues towards maturity at different technology readiness levels // MSSP. – 2017.
48. Kumar A., Kumar R. Time-frequency analysis and support vector machine in automatic detection of defect from vibration signal of centrifugal pump // Measurement. – 2017.
49. Lv H., Chen J., Pan T., Zhang T., Feng Y., Liu Sh. Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application // Measurement. – 2022.
50. Mao W., Feng W., Liu Y., Zhang D., Liang X. A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis // MSSP. – 2021.
51. Mobley R. K. An Introduction to Predictive Maintenance. – 2nd ed. – Butterworth-Heinemann, 2006.

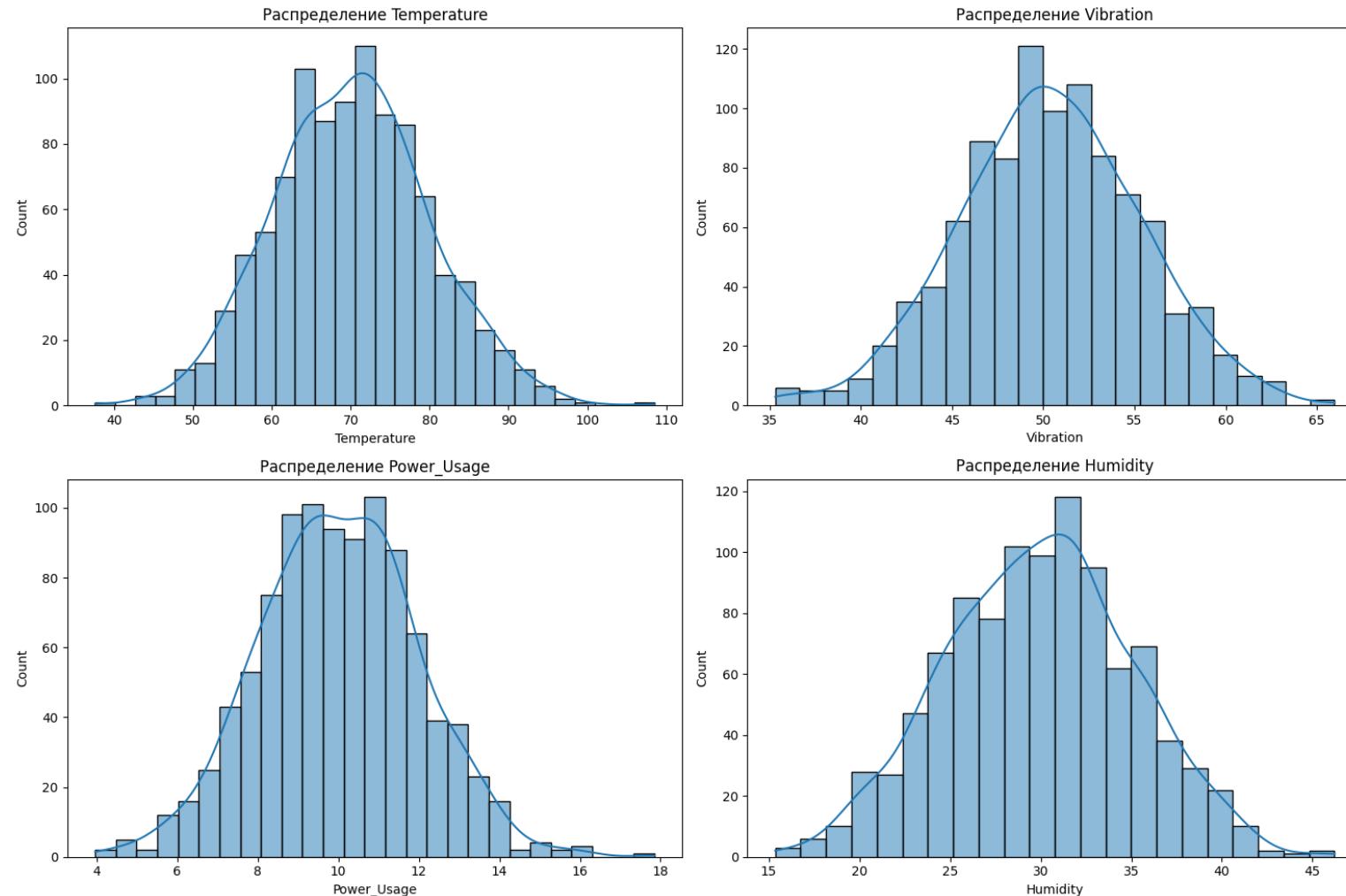
52. Nan X., Zhang B., Liu C., Gui Z., Yin X. Multi-modal learning-based equipment fault prediction in the internet of things // Sensors. – 2022.
53. Rocchetta R., Bellani L., Compare M., Zio E., Patelli E. A reinforcement learning framework for optimal operation and maintenance of power grids // Applied Energy. – 2019. – Vol. 241.
54. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead // Nature Machine Intelligence. – 2019.
55. Shneiderman B. Human-Centered Artificial Intelligence: Three Fresh Ideas // Journal of Human-Centric Computing. – 2020.
56. Tran N. T. An overview of the application of machine learning in predictive maintenance // Petrovietnam Journal. – 2021.
57. Ucar A., Karakose M., Kırımcı N. Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends // Applied Sciences. – 2024.
58. Xu Z., Mei X., Wang X., Yue M., Jin J., Yang Y., Li Ch. Fault diagnosis of wind turbine bearing using a multi-scale convolutional neural network with bidirectional long short term memory and weighted majority voting for multi-sensors // Renewable Energy. – 2022.
59. Глобальный рынок оперативного предиктивного обслуживания — тенденции отрасли и прогноз до 2029 года. –
URL: <https://www.databridgemarketresearch.com/ru/reports/global-operational-predictive-maintenance-market>
60. Индустрия 4.0: будущее промышленной автоматизации в условиях санкций. – 2023. –
URL: <https://1d.media/industry/alldigital/26874>
61. О развитии предсказательного обслуживания на примере диагностики трансформатора / Блог компании ЦИТМ Экспонента // Habr. –
URL: <https://habr.com/ru/company/ctm-exponenta/blog/123456>
62. Официальный сайт Национальной технологической инициативы (НТИ). –
URL: <https://nti2035.ru/>
63. Платформа "Цифровая экономика". – URL: <https://digital.gov.ru/>
64. Deloitte. Predictive maintenance: Taking pro-active measures based on advanced data analytics to predict and avoid machine failure. – 2024. –
URL: https://www.beekeeper.io/wp-content/uploads/2024/10/Deloitte_Predictive-Maintenance_PositionPaper.pdf

65. International Journal of Prognostics and Health Management (IJPHM). The Prognostics and Health Management Society. <https://papers.phmsociety.org/index.php/ijphm>
66. Kaggle: Machine Failure Classification Dataset. – URL: /datasets/mujtabamatin/dataset-for-machine-failure-detection/data
67. Machado Roberty, Elias A. Predictive Analytics Applications for Oil and Gas Processing Facilities. – URL: <https://dspace.mit.edu/handle/1721.1/140083>
68. McKinsey & Company. The Digital Utility: New opportunities and challenges. – 2017. – URL: <https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/the-digital-utility-new-opportunities-and-challenges>
69. Mehdi Mohhamadpoor, Farshid Torabi. Big Data analytics in oil and gas industry: An emerging trend. – URL: <https://www.sciencedirect.com/science/article/pii/S2405656118301421>
70. Open source predictive analytics tools. – URL: <https://osssoftware.org/blog/open-source-predictive-analytics-tools/>
71. Salinas D., Flunkert V., Gasthaus J. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. – 2019. – URL: <https://arxiv.org/abs/1704.04110>
72. Digital Twin Market Size, Share & Industry Trends Growth Analysis Report by Application (Predictive Maintenance, Business Optimization, Performance Monitoring, Inventory Management), Industry (Automotive & Transportation, Healthcare, Energy & Utilities), Enterprise and Geography - Global Growth Driver and Industry Forecast to 2028 - <https://www.marketsandmarkets.com/Market-Reports/digital-twin-market-225269522.html>

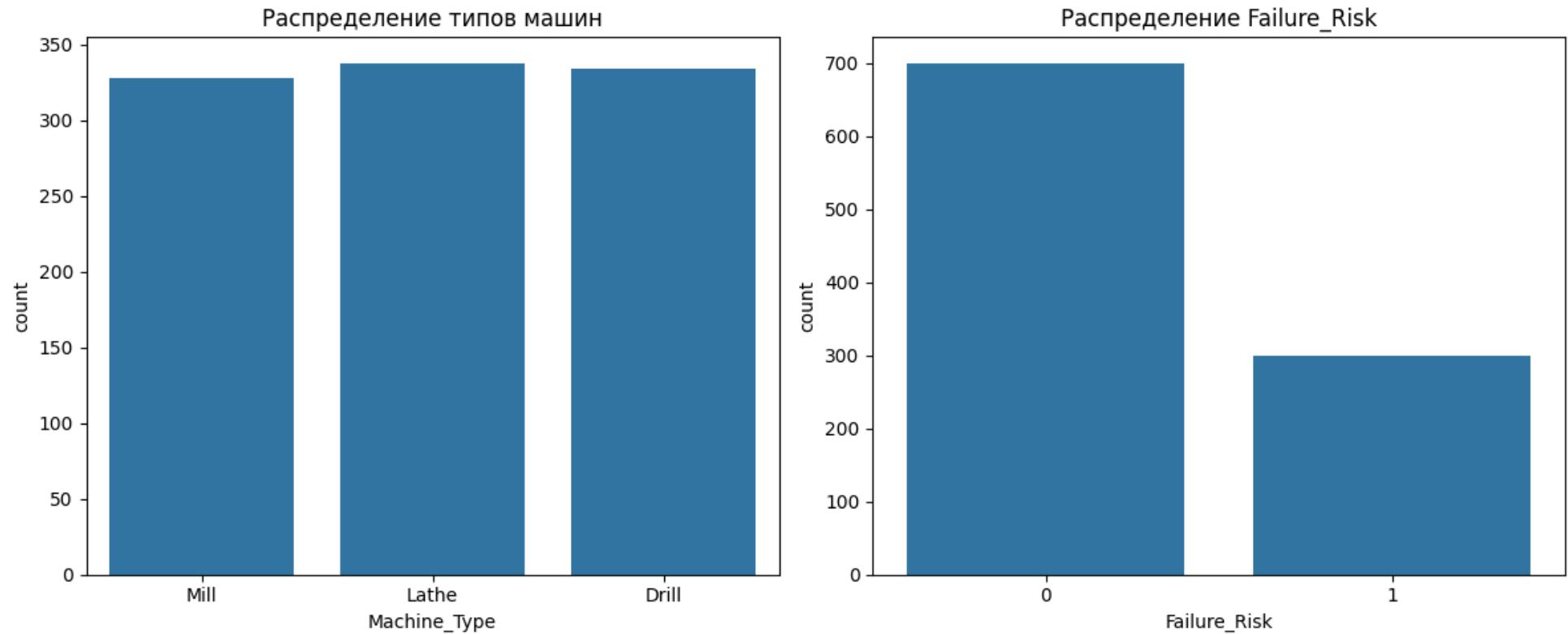


Приложения

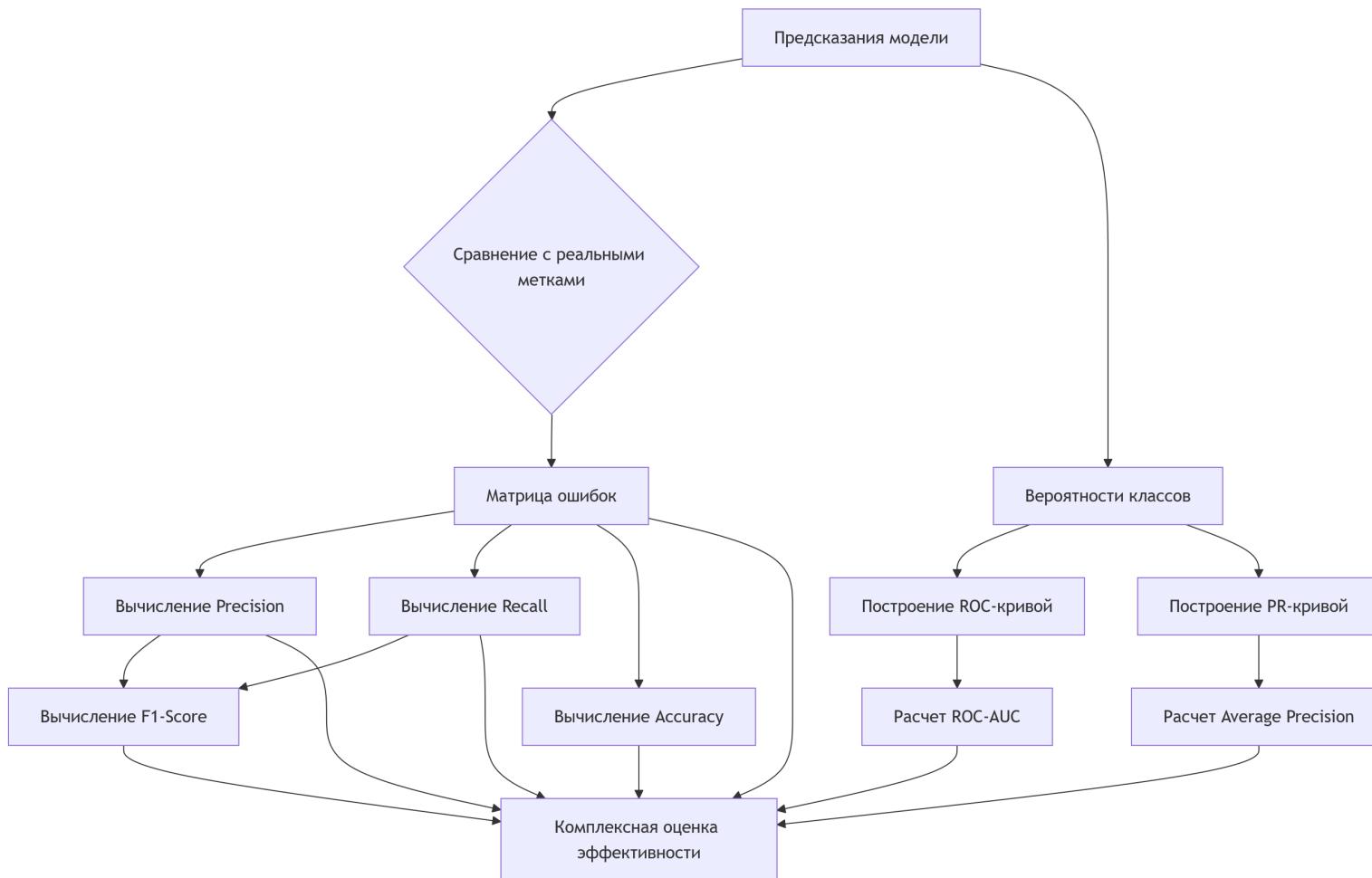
Приложение 1 : Структура датасета



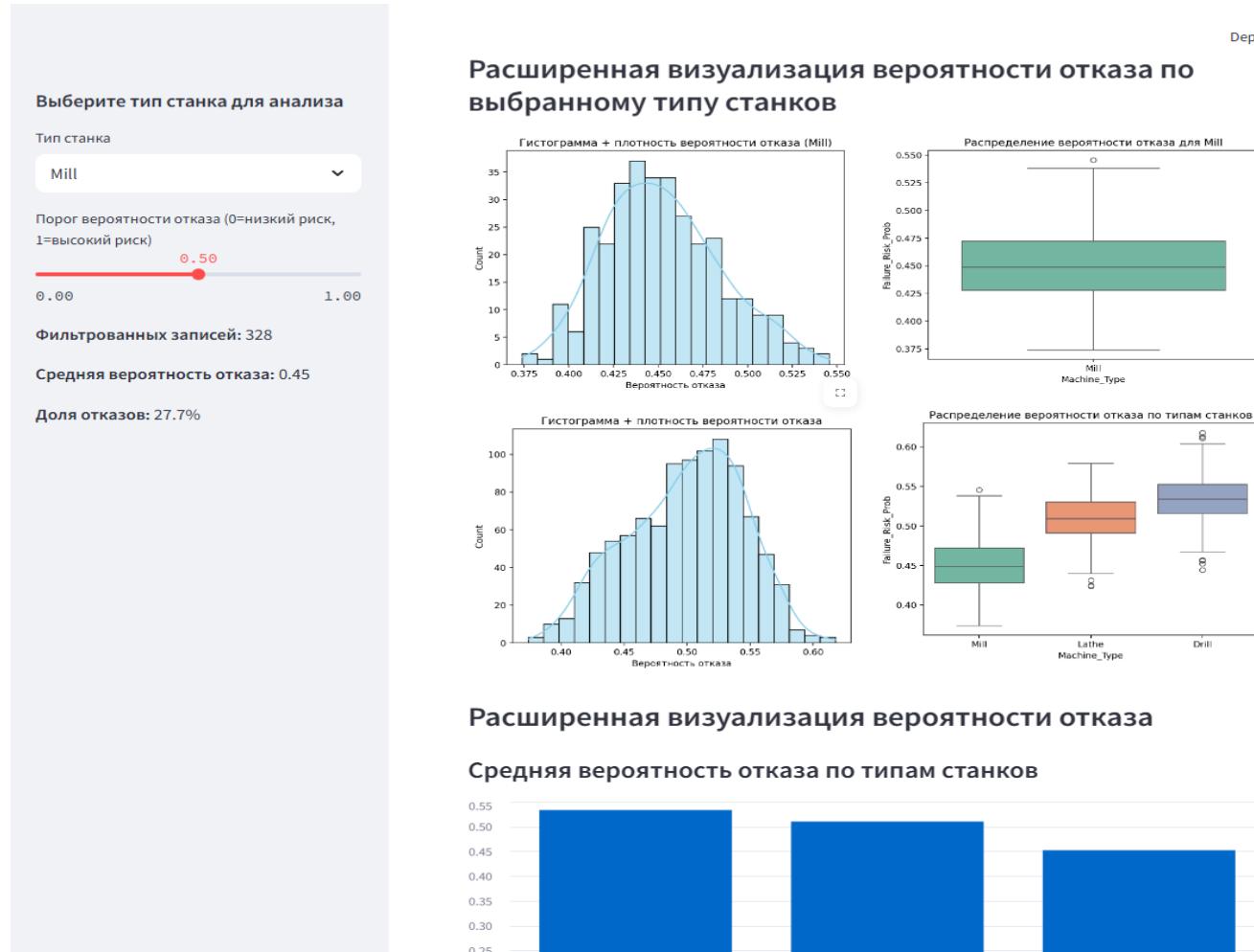
Приложение 2 : Структура датасета



Приложение 3 : Расширенный алгоритм оценки эффективности моделей предиктивного обслуживания на основе многокритериального анализа



Приложение 4 : Веб-интерфейс информационной панели мониторинга и анализа рисков отказа оборудования по различным метрикам и типам станков



Приложение 5 : Необходимые ресурсы и технологии для внедрения системы предиктивной аналитики отказов

Категория ресурса	Тип ресурса	Описание / Пример использования
Человеческие ресурсы	Data Scientist / Аналитик данных	Анализ исходных данных, отбор признаков, обучение модели
	Machine Learning инженер	Разработка, валидация и оптимизация моделей предсказания
	Разработчик интерфейса (Frontend/Backend)	Интеграция модели и визуализация в виде дашборда
	Специалист по внедрению / DevOps	Настройка серверной среды, развертывание модели
	Персонал технической поддержки	Контроль функционирования, устранение ошибок, обслуживание системы
Технические ресурсы	Сервер / локальная машина	Обеспечение вычислений, хранение модели и данных
	Доступ к интернету / корпоративной сети	Необходим для загрузки обновлений, подключения к потокам данных и удаленного доступа
Программное обеспечение	Язык программирования: Python	Основной язык разработки модели и визуализации
	Библиотеки: pandas, scikit-learn, NumPy	Для анализа и обработки данных, машинного обучения
	Streamlit	Для построения интерактивного пользовательского интерфейса
Данные	Jupyter Notebook / VS Code	Среда разработки и тестирования
	CSV-файлы (machine_failure_dataset.csv и др.)	Исторические данные датчиков, необходимые для обучения модели
	Данные в реальном времени (оциально)	Для внедрения полноценного мониторинга и уведомлений

Приложение 6 : Сводная таблица экономии от предотвращения простоев оборудования

Сводная таблица экономии от предотвращения простоев оборудования:

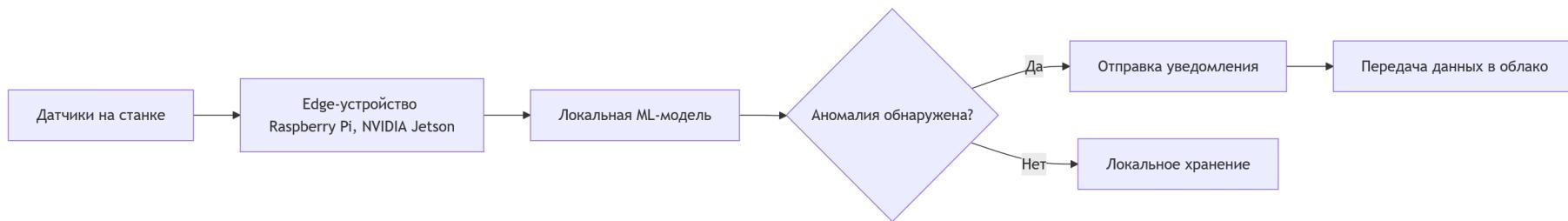
Показатель	Значение	Единицы измерения / Примечания
Средняя продолжительность простоя одной машины	4	часа
Стоимость одного часа простоя	30 000	руб. / час
Количество потенциальных отказов в год	60	на основе исторических данных (300 из 1000 = 30% риска)
Прогнозируемая доля предотвращённых отказов	0.35	модель с recall ~63% и порогом 0.35
Ожидаемое количество предотвращённых отказов	$60 \times 0.35 = 21$	единиц
Суммарное время предотвращённых простоев	$21 \times 4 = 84$	часа
Общая экономия на предотвращённых простоях	$84 \times 30 000 = 2 520 000$	руб. / год

Финансовые итоги:

Показатель	Сумма
Общие затраты на 1-й год	800 000 руб.
Прогнозируемая экономия	2 520 000 руб.
Чистая экономия за 1 год	1 720 000 руб.
Срок окупаемости (в месяцах)	$800 000 / (2 520 000 / 12) \approx 3.8$ мес.
Ожидаемая годовая рентабельность (ROI)	$(2 520 000 - 800 000) / 800 000 \times 100\% = 215\%$

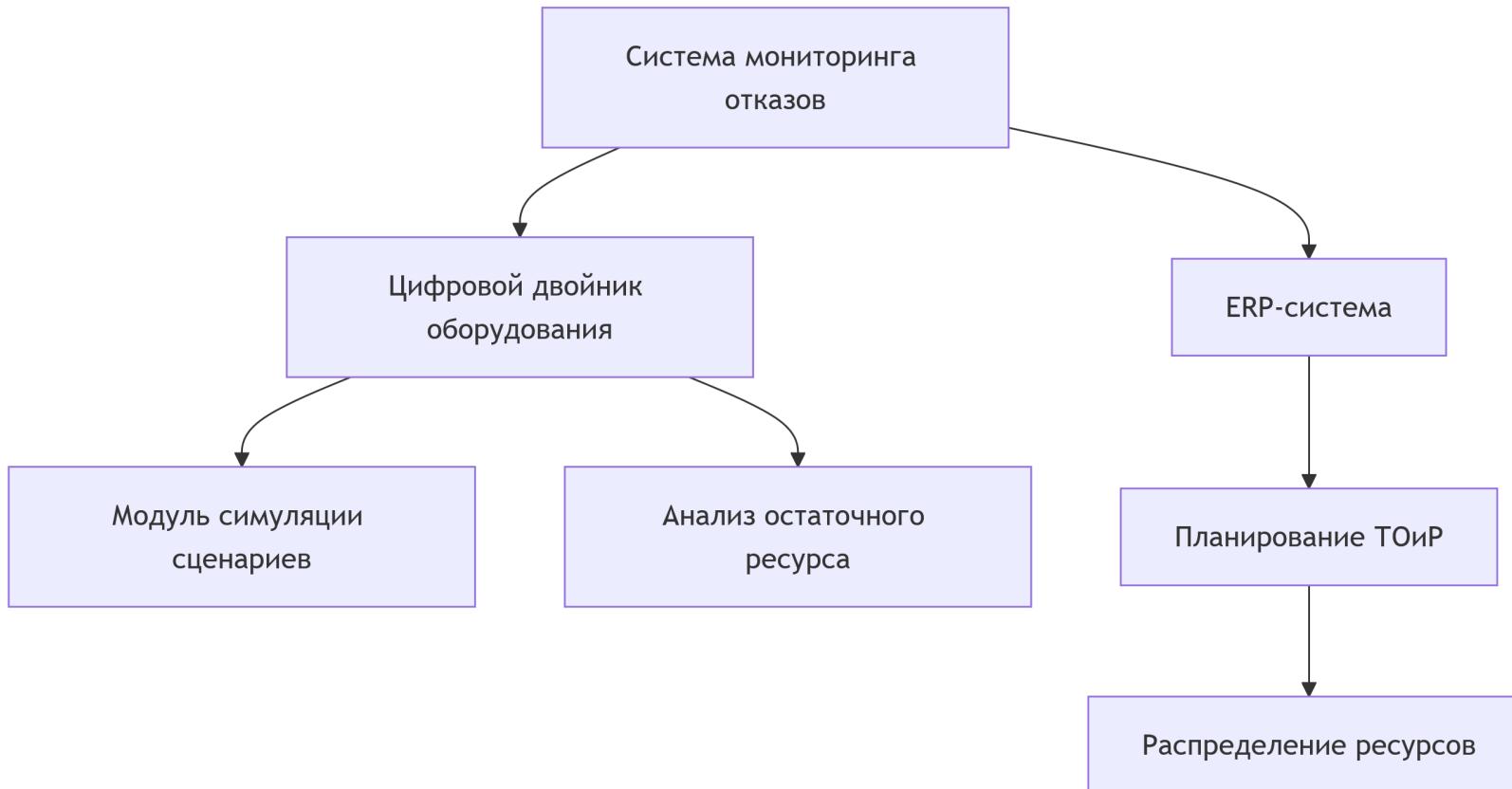
Вывод: Система предиктивного обслуживания демонстрирует высокую экономическую эффективность, обеспечивая окупаемость менее чем за 4 месяца и годовую рентабельность инвестиций свыше 200%.

Приложение 7 : Схема архитектуры Edge Computing



- Вычисления происходят рядом с оборудованием, уменьшая задержки.
- Только аномалии или обобщенные данные передаются в центральную систему, экономя трафик.

Приложение 8 : Схема интеграции с внешними системами



- Дашборд передаёт данные в ERP-систему, что позволяет планировать ремонт.
- Цифровой двойник помогает смоделировать отказ или оптимизировать график технического обслуживания.

Приложение 9 : Ссылка на репозиторий проекта с описанием файлов

Для обеспечения полной воспроизводимости результатов все материалы проекта размещены в публичном GitHub-репозитории <https://github.com/OlaEla/sensor-ml-failure-prediction.git>. Репозиторий содержит:

1. Исходные и производные данные:

- machine_failure_dataset.csv - оригинальный датасет с Kaggle
- cleaned_machine_failure_data.csv - очищенные данные после предобработки
- processed_full_dataset.csv - финальный набор после feature engineering
- X_train.csv/X_test.csv и y_train.csv/y_test.csv - разделенные выборки
- feature_importance.csv - результаты анализа важности признаков
- model_comparison_results.csv - сравнительные метрики моделей

2. Ключевые артефакты моделей:

- best_model.pkl - финальная оптимальная модель
- best_rf_model.pkl - лучшая Random Forest модель
- failure_prediction_model.pkl - основной предиктивный модель
- СерIALIZEDНЫЕ preprocessing-объекты:
 - feature_scaler.pkl - нормализатор признаков
 - model_features.pkl - конфигурация фичей

3. Исследовательские ноутбуки:

- machine_failure_code.ipynb - полный анализ данных
- save_artifacts.py – сохранение артифактов

- dashboard.py – код разработанного дэшборда на фреймворке Streamlit
4. Техническая документация:
- model_report.txt - детальный отчет по модели
 - Визуализации результатов в /figures/
5. Промышленные артефакты (папка artifacts):
- model_features.joblib - конвейер обработки фичей
 - scaler.joblib - сериализованный скейлер

Особенности реализации:

- Полный ML pipeline от сырых данных до предсказаний
- Поддержка воспроизведения всех этапов исследования
- Готовность к промышленному развертыванию
- Детализированная версионность данных и моделей