Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Физико-механичский институт
Высшая школа теоретической механики и математической физики

Работа допущена к защите
Директор ВШТМиМФ
д.ф.-м.н., чл.-корр. РАН
_____ А.М. Кривцов
«_____» _____ 2024 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

ПРОГНОЗИРОВАНИЕ РИСКА МЕХАНИЧЕСКОГО ПОВРЕЖДЕНИЯ СТАНКОВ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

по направлению подготовки
01.03.03 Механика и математическое моделирование
направленность
01.03.03_03 Математическое моделирование процессов нефтегазодобычи

Выполнил
студент гр. 5030103/00301 М.В. Трубачев

Руководитель
доцент ВШТМиМФ,
к.ф.-м.н. С.В. Каштанова

Консультант
ассистент ВШТМиМФ А.Д. Ершов

Санкт-Петербург
2024

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт

Высшая школа теоретической механики и математической физики

УТВЕРЖДАЮ

Директор ВШТМиМФ

А. М. Кривцов

«__»_____20__г.

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студенту Трубачеву Максиму Вячеславовичу, гр. 5030103/00301

1. Тема работы: Прогнозирование риска механического повреждения станков с применением методов машинного обучения.

2. Срок сдачи студентом законченной работы: 30.05.2024

3. Исходные данные по работе: Справочные данные, синтетический набор данных, который отражает реальное профилактическое обслуживание, встречающееся в отрасли.

4. Содержание работы (перечень подлежащих разработке вопросов): Анализ исходных данных, реализация модели для прогнозирования различными методами. Обработка и анализ полученных результатов.

5. Перечень графического материала (с указанием обязательных чертежей): не предусмотрено

6. Консультанты по работе: Ершов А.Д., ассистент ВШТМиМФ

7. Дата выдачи задания 26.02.2024

8. Руководитель ВКР _____ Каштанова С.В., доцент ВШТМиМФ, к.ф.-м.н.

Задание принял к исполнению 26.02.2024

Студент _____ Трубачев М.В.

РЕФЕРАТ

На 38 с., 5 рисунков, 11 таблиц

КЛЮЧЕВЫЕ СЛОВА: МАШИННОЕ ОБУЧЕНИЕ, МЕХАНИЧЕСКИЕ ПОВРЕЖДЕНИЯ, ПРОГНОЗИРОВАНИЕ РИСКА, ПРЕДОБРАБОТКА ДАННЫХ, ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ, K-БЛИЖАЙШИХ СОСЕДЕЙ, АЛГОРИТМ ОПОРНЫХ ВЕКТОРОВ, ДЕРЕВО РЕШЕНИЙ, СЛУЧАЙНЫЙ ЛЕС, ПРОМЫШЛЕННОЕ ОБОРУДОВАНИЕ.

В данной дипломной работе рассматривается применение методов машинного обучения для прогнозирования риска механических повреждений станков. Описаны типы, причины и последствия механических повреждений, а также методы их профилактики. Проведен анализ данных, выявлены ключевые признаки и проведена их предобработка, включая удаление выбросов и устранение пропущенных значений. Обучены и оценены различные модели машинного обучения, такие как логистическая регрессия, K-ближайших соседей, алгоритм опорных векторов, дерево решений и случайный лес.

ABSTRACT

38 pages, 5 figures, 11 tables

KEYWORDS: MACHINE LEARNING, MECHANICAL FAILURES, RISK PREDICTION, DATA PREPROCESSING, LOGISTIC REGRESSION, K-NEAREST NEIGHBORS, SUPPORT VECTOR MACHINE, DECISION TREE, RANDOM FOREST, INDUSTRIAL EQUIPMENT.

The subject of the graduate qualification work is «Predicting the risk of mechanical damage to machine tools using machine learning methods».

This thesis explores the application of machine learning methods for predicting the risk of mechanical failures in machinery. It describes the types, causes, and consequences of mechanical failures, as well as methods for their prevention. Data analysis was conducted, key features were identified, and preprocessing was performed, including outlier removal and handling of missing values. Various machine learning models were trained and evaluated, including logistic regression, k-nearest neighbors, support vector machines, decision trees, and random forests.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В эпоху быстрого технологического развития и роста промышленной автоматизации, эффективное функционирование производственных систем становится ключевым фактором успеха для предприятий во всех отраслях. В этом контексте, станки, являющиеся сердцем многих производственных процессов, играют особенно важную роль. Однако, вместе с повышением автоматизации, возрастает и риск механических повреждений оборудования, что может привести к серьезным простоям и убыткам для предприятия.

Одним из подходов к обслуживанию оборудования является стратегия "Работа до отказа". Она предполагает использование оборудования до тех пор, пока оно не выйдет из строя, после чего производится его ремонт или замена. Этот подход имеет свои минусы, включая неожиданные простои и высокие затраты на внеплановые ремонтные работы, что может привести к существенным потерям как в финансовом, так и в производственном плане.

В отличие от работы до отказа, профилактическое обслуживание предполагает регулярное и систематическое техническое обслуживание и проверку оборудования с целью выявления потенциальных проблем и их предотвращения до того, как они приведут к выходу оборудования из строя. Оценка состояния оборудования и своевременные профилактические мероприятия позволяют предотвратить возможные поломки, минимизировать простои и снизить общие эксплуатационные расходы.

Таким образом, профилактическое обслуживание представляет собой эффективный и экономически выгодный подход к управлению оборудованием, который способствует повышению его надежности, продолжительности службы и общей эффективности производственных процессов.

Хотя профилактическое обслуживание считается эффективным методом поддержания надежности оборудования, у него также есть свои минусы. Одним из главных недостатков является трудность определения оптимального расписания технического обслуживания. Традиционные методы определения интервалов обслуживания основаны на общих статистических данных или опыте специалистов, что может привести к недооценке или переоценке реального состояния оборудования.

Вторым минусом является возможность пропуска потенциально критических проблем при обслуживании, основанном на расписании, если не учитывать специ-

фические особенности и индивидуальные характеристики каждого оборудования.

Применение методов машинного обучения позволяет эффективно решить эти проблемы. Алгоритмы машинного обучения могут анализировать большие объемы данных, включая данные о состоянии оборудования, параметры его работы и историю технического обслуживания. На основе этого анализа можно разработать модели, которые способны предсказывать оптимальные временные интервалы для проведения профилактических мероприятий.

Более того, с помощью машинного обучения можно создать системы мониторинга состояния оборудования в режиме реального времени. Эти системы могут автоматически обнаруживать аномалии в работе оборудования и выдавать предупреждения о возможных проблемах, что позволит оперативно реагировать и предотвращать возможные поломки.

Таким образом, использование методов машинного обучения в профилактическом обслуживании позволяет улучшить точность прогнозирования состояния оборудования, сократить затраты на его обслуживание и уменьшить риск непредвиденных простоев в производственных процессах.

Именно здесь встает необходимость разработки системы прогнозирования и предотвращения риска механических повреждений станков. Это важное направление исследований становится все более актуальным в сфере промышленности, где предприятия стремятся минимизировать потери и повысить безопасность производственных процессов.

Актуальность исследования — Исследование имеет практическую значимость для предприятий, позволяя им не только повысить надежность и эффективность производственных процессов, но и сократить затраты на обслуживание и ремонт оборудования, что способствует повышению конкурентоспособности предприятия в условиях современного рынка.

Цель исследования — Основной целью работы является исследование возможностей применения современных методов машинного обучения для создания эффективной системы прогнозирования риска нарушения работы оборудования.

Задачи работы:

A. Анализ исходных данных: Первоначальный этап работы предполагает тщательный анализ полученных данных с целью выявления важных признаков, влияющих на риск механических повреждений станков.
B. Подготовка данных для обучения модели: Второй этап работы включает в себя предварительную обработку и подготовку данных для обучения

модели машинного обучения. Это может включать в себя удаление выбросов, заполнение пропущенных значений, масштабирование признаков.

C. Выбор и обучение модели: Далее проводится выбор подходящей модели машинного обучения, которая наилучшим образом справится с поставленной задачей прогнозирования риска механических повреждений станков. После выбора модели производится её обучение на подготовленных данных.

D. Оценка модели и анализ результатов: После обучения модели производится оценка её качества с использованием различных метрик, таких как точность, полнота, F1-мера и т. д. Затем проводится анализ полученных результатов с целью выявления основных факторов, влияющих на риск механических повреждений станков, а также определения эффективности разработанной модели.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ ПО МЕХАНИЧЕСКИМ ПОВРЕЖДЕНИЯМ СТАНКОВ

Понимание механических повреждений станков требует анализа их типов, причин и последствий. В данной главе мы рассмотрим механические повреждения, с которыми сталкиваются производственные предприятия, и изучим основные причины их возникновения. Погружение в эту тему поможет лучше понять, какие методы машинного обучения можно применить для прогнозирования риска механических повреждений станков.

1.1. Анализ механических повреждений станков: типы, причины и последствия.

Механические повреждения станков представляют собой серьезную проблему в промышленности и могут иметь различные типы, причины и последствия.[4]

Типы механических повреждений:

Износ: Постепенное разрушение поверхности станка вследствие трения, избыточной нагрузки или химического воздействия.

Трещины: Образование разрывов или трещин в материале станка, что может привести к его поломке или неправильному функционированию.

Искажение формы: Изменение формы или геометрии станка, что может привести к снижению точности обработки или ухудшению качества продукции.

Причины механических повреждений:

Износ материала: Повышенный износ материала станка вследствие интенсивной эксплуатации, неправильного технического обслуживания или низкого качества материала.

Недостаточная смазка и охлаждение: Отсутствие или недостаточное количество смазки и охлаждения может привести к увеличению трения и повышенному износу деталей станка.

Перегрузки: Превышение допустимых нагрузок на станок может привести к его деформации, трещинам или поломке.

Последствия механических повреждений: Простои в производстве: Необходимость остановки работы станка для ремонта или замены поврежденных деталей приводит к временным простоям в производственном процессе.

Увеличение затрат на обслуживание: Ремонт или замена поврежденных деталей

станка требует дополнительных затрат на запасные части, трудовые ресурсы и время. Снижение качества продукции: Механические повреждения станков могут привести к ухудшению качества обработки и, как следствие, к браку или отказу продукции.

Важно иметь в виду разнообразие типов, причин и последствий механических повреждений станков при разработке методов и моделей для их прогнозирования и предотвращения.[1]

1.2. Основные методы профилактики и предотвращения механических повреждений оборудования

Механические повреждения оборудования могут привести к серьезным простоям в производстве и значительным финансовым потерям.[6] Эффективная профилактика и предотвращение таких повреждений играют критическую роль в обеспечении непрерывной работы производственных процессов. Рассмотрим более подробно основные методы, которые применяются для уменьшения риска механических повреждений оборудования:

Регулярное техническое обслуживание:Регулярные плановые обслуживание и технические проверки являются основой для предотвращения механических повреждений. Это включает в себя проверку и замену изношенных деталей, смазку подшипников и других движущихся частей, а также контроль за параметрами работы оборудования. Для визуализации этого процесса можно использовать схемы или фотографии, демонстрирующие процесс обслуживания и замены деталей.

Использование современных материалов и технологий:Применение современных материалов с высокой износостойкостью и технологий позволяет увеличить срок службы оборудования и снизить вероятность механических повреждений. Для иллюстрации этого метода можно вставить сравнительные таблицы или диаграммы, демонстрирующие преимущества современных материалов перед более старыми.

Оптимизация режимов работы:Рациональное использование оборудования с учетом его технических характеристик и требований производства помогает снизить нагрузку на механизмы и предотвратить излишнее изнашивание. Для визуализации этого метода можно использовать графики, показывающие зависимость между режимами работы и износом оборудования.

Обучение персонала:Подготовленные сотрудники способны правильно обращаться с оборудованием и реагировать на предупреждающие признаки его

неисправности. Для иллюстрации этого можно использовать фотографии или видео, демонстрирующие процессы обучения персонала и правильного использования оборудования.

Мониторинг состояния оборудования:Применение систем мониторинга и диагностики позволяет оперативно выявлять отклонения от нормы и предупреждать возможные поломки. Для визуализации этого метода можно вставить схемы работы систем мониторинга и диагностики или графики, демонстрирующие изменения параметров работы оборудования во времени.

Разработка резервных планов:Создание резервных планов действий в случае аварийных ситуаций позволяет быстро реагировать на проблемы и минимизировать их последствия. Для иллюстрации этого метода можно вставить примеры резервных планов и схемы действий в аварийных ситуациях.

Использование вышеописанных методов в комбинации позволяет создать эффективную систему профилактики и предотвращения механических повреждений оборудования, что в итоге способствует повышению эффективности производственных процессов и снижению финансовых потерь предприятия.

1.3. Использование методов машинного обучения для прогнозирования риска механических повреждений станков

В современном мире методы машинного обучения становятся все более важным инструментом для прогнозирования и предотвращения различных видов повреждений оборудования. Применение этих методов открывает новые возможности для раннего обнаружения потенциальных проблем и оптимизации процессов обслуживания. Рассмотрим основные методы машинного обучения, которые могут быть использованы для прогнозирования риска механических повреждений станков[8, 2]:

Классификация и регрессия:Методы классификации и регрессии позволяют определить вероятность возникновения механических повреждений на основе исторических данных и текущих параметров работы оборудования. Для иллюстрации этого метода можно вставить графики или диаграммы, демонстрирующие зависимость между параметрами станка и вероятностью его поломки.

Кластеризация и ассоциативные правила:Методы кластеризации позволяют выявлять группы оборудования с похожими характеристиками и поведением, что может помочь в выявлении общих закономерностей и факторов, влияющих на

вероятность повреждений. Для визуализации этого метода можно использовать диаграммы рассеяния или деревья принятия решений.

Нейронные сети и глубокое обучение:Применение нейронных сетей и глубокого обучения позволяет обрабатывать большие объемы данных и выявлять сложные закономерности, что может быть полезно при анализе многомерных параметров работы оборудования. Для иллюстрации этого метода можно вставить примеры архитектур нейронных сетей и результаты их работы.

Методы временных рядов:Анализ временных рядов позволяет предсказывать динамику изменения параметров работы станка во времени и выявлять аномалии, которые могут свидетельствовать о возможных повреждениях. Для иллюстрации этого метода можно использовать графики временных рядов и результаты прогнозирования.

Использование указанных методов машинного обучения позволяет эффективно прогнозировать риск механических повреждений станков и принимать меры по их предотвращению до того, как они приведут к серьезным последствиям для производства. Кроме того, развитие современных методов машинного обучения открывает новые перспективы для автоматизации процессов обслуживания и оптимизации работы промышленного оборудования.

Для решения задачи прогнозирования риска механических повреждений оборудования мы планируем использовать метод классификации. Подобный выбор обоснован характером задачи и особенностями предоставленных данных.

В данной задаче наша цель - предсказать вероятность возникновения поломки оборудования на основе различных параметров, таких как температура, частота вращения, износ инструмента и другие. Это является задачей бинарной классификации, где мы пытаемся определить, произойдет ли отказ оборудования (положительный класс) или нет (отрицательный класс) на основе имеющихся данных.

Методы классификации позволяют нам обучить модель на данных с известным исходом (отказ или нормальная работа оборудования) и использовать эту модель для прогнозирования вероятности отказа на новых данных. Модели классификации способны выявлять сложные зависимости между входными признаками и целевой переменной, что делает их подходящими для нашей задачи.

Таким образом, выбор метода классификации обусловлен целями задачи и характером доступных данных, и мы уверены, что этот подход позволит нам

достичь высокой точности в прогнозировании риска механических повреждений оборудования.

1.4. Классификация и регрессия

Методы классификации и регрессии являются одними из наиболее распространенных и широко используемых в машинном обучении. Эти методы позволяют моделировать зависимость между входными данными и выходными переменными, что делает их эффективными инструментами для прогнозирования риска механических повреждений станков.[10]

Классификация:

В задачах классификации модель стремится предсказать принадлежность объекта к одному из заданных классов на основе набора характеристик или признаков. В контексте прогнозирования риска механических повреждений станков, классификация может использоваться для определения вероятности возникновения повреждения в заданный временной интервал.[15]

Примеры классификационных методов, которые могут быть применены к данной задаче, включают в себя:

- Логистическая регрессия
- Метод ближайших соседей (k-Nearest Neighbors)
- Метод опорных векторов (Support Vector Machines)
- Деревья решений и их ансамбли (Random Forest, Gradient Boosting)

Применение этих методов позволяет моделировать зависимость между различными характеристиками работы станка (например, скорость, нагрузка, температура) и вероятностью возникновения механических повреждений.

Регрессия:

В отличие от классификации, задача регрессии заключается в предсказании непрерывной числовой переменной на основе входных данных. В контексте прогнозирования риска механических повреждений станков, регрессия может использоваться для определения вероятности или оценки степени тяжести повреждений.

Примеры регрессионных методов, которые могут быть применены к данной задаче, включают в себя:

- Линейная регрессия
- Регрессия с использованием полиномиальных признаков

- Регрессия на основе деревьев (Decision Tree Regression)
- Нейронные сети

Применение регрессионных методов позволяет предсказывать непрерывные переменные, такие как время до возникновения поломки или степень износа деталей станка, что позволяет оперативно принимать меры по их предотвращению или замене.

Из-за характера данных, в дальнейшем будет решаться задача многоклассовой классификации, то рассмотрим более подробно методы решения.

Логистическая регрессия

Хотя в названии присутствует слово регрессия, на самом деле данный алгоритм не имеет никакого отношение к данному методу.[12] Алгоритм решает задачи бинарной классификации, так как алгоритм применяет сигмоидальную функцию, который расположен между y ∈ 0,1. Отсюда можно сделать вывод, что цель данного алгоритма не восстановление значений или предсказание, а классификация. В данном методе выполняется условие, где $0 \leqslant Y \leqslant 1$, что достигается применением сигмоидальной (логистической) функции:

$$Y = \frac{1}{1 + e^{-F(x)}}$$

где F(x) – стандартное уравнение регрессии. При этом нужно учесть что если значение не равно 0 или 1, значение аппроксимируется.

Алгоритм k ближайших соседей

Этот жадный алгоритм требует значительных вычислений по объектам обучающей выборки каждого класса в пространстве. После подсчета он выбирает пространство из k объектов с наименьшим расстоянием, в центре которого находится распознаваемый объект. Затем классифицируемый объект относится к классу, у которого в данном пространстве больше всего объектов.[9] Однако перед началом подсчета необходимо определить, какой алгоритм будет использоваться для этих вычислений. Расстояние между классифицируемыми объектами может быть рассчитано как расстояние в декартовом пространстве (евклидова метрика), но также можно использовать и другие метрики, такие как манхэттенская, метрика Чебышева, Минковского и др. В качестве классического алгоритма вычисления расстояния чаще всего выбирается евклидова метрика, которая выражается следу-

ющим образом:

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

где a и b это точки (объекты), состоящий из координат (признаков). В целом это один из самых простых, но часто неточных алгоритмов классификации. Алгоритм также отличается высокой вычислительной сложностью. Объем вычислений при использовании евклидовой метрики пропорционален квадрату от числа обучающих примеров. Важно отметить что алгоритм применяется не только для задач классификации но и для регрессии.

Алгоритм опорных векторов

Алгоритм опорных векторов (Support Vector Machines) относится к группе методов, которые определяют классы путем построения границ областей. Основой этого метода является концепция плоскостей решений, которые разделяют объекты разных классов. В случае пространств с большим числом измерений, вместо линий используются гиперплоскости - пространства, одномерность которых на единицу меньше, чем размерность исходного пространства. Например, в трехмерном пространстве R3 гиперплоскость является двумерной плоскостью.

Метод опорных векторов ищет образцы, которые находятся на границах классов (не менее двух), так называемые опорные векторы. Он решает задачу разделения множества объектов на классы с использованием линейной решающей функции. Алгоритм опорных векторов строит классифицирующую функцию f(x) следующим образом:

$$f(x) = sign(<w, s> + b)$$

где ⟨w, s⟩ – скалярное произведение; w – нормальный (перпендикулярный) вектор к разделяющей гиперплоскости; b – вспомогательный параметр, который равен по модулю расстоянию от гиперплоскости до начала координат. Если параметр b равен нулю, гиперплоскость проходит через начало координат. Объекты, для которых f(x) = 1, попадают в один класс, а объекты с f(x) = -1 – в другой. С точки зрения точности классификации лучше всего выбрать такую прямую, расстояние от которой до каждого класса максимально. Такая прямая (в общем случае – гиперплоскость) называется оптимальной разделяющей гиперплоскостью. Задача состоит в выборе w и b, максимизирующих это расстояние.

1.5. Выводы

В главе были рассмотрены различные типы механических повреждений, методы их предотвращения, а также подходы машинного обучения, применяемые для прогнозирования риска повреждений.

В главе 2 мы перейдем к анализу и предобработки данных, необходимых для разработки модели прогнозирования риска механических повреждений станков. Анализ данных играет ключевую роль в выявлении важных признаков и трендов, а предобработка позволяет подготовить данные к обучению модели.

ГЛАВА 2. АНАЛИЗ И ПРЕДОБРАБОТКА ДАННЫХ ДЛЯ МОДЕЛИ

В данной главе мы сосредоточимся на важном этапе исследования - анализе и предобработке данных, необходимых для разработки модели прогнозирования риска механических повреждений станков. Анализ данных позволит нам понять характеристики наших данных, выявить важные тренды и особенности, а предобработка поможет подготовить данные к обучению модели, устранить шумы и недостатки.

2.1. Описание параметров

Перед началом работы надо посмотреть на имеющийся набор данных, который мы получаем на вход в виде таблицы, и с которыми будем проводить дальнейшие преобразования. Фрагмент таких данных представлен в таблице ниже. В таблице

Таблица 2.1

Фрагмент исходных данных

$Rotationalspeed[rpm]$	$Torque[Nm]$	$Toolwear[min]$	$Target$	$FailureType$
1551	42.8	0	0	No Failure
1408	46.3	3	0	No Failure
1498	49.4	5	0	No Failure
1433	39.5	7	0	No Failure
1408	40.0	9	0	No Failure

UID	$ProductID$	$Type$	$Airtemperature[K]$	$Processtemperature[K]$
1	M14860	M	298.1	308.6
2	L47181	L	298.2	308.7
3	L47182	L	298.1	308.5
4	L47183	L	298.2	308.6
5	L47184	L	298.2	308.7

данных для обучения модели представлены 10000 записей о различных станках, описанных с помощью следующих столбцов- признаков:

- UID (Уникальный идентификатор): Данный параметр представляет собой уникальный номер для каждой записи в наборе данных. Мы можем использовать его для идентификации конкретных данных и для проверки наличия дубликатов.

- productID (Идентификатор продукта): Параметр содержит информацию о качестве продукта (низкое, среднее, высокое) и его серийный номер. Это важный

атрибут, который можно использовать для анализа влияния качества продукции на вероятность поломки оборудования.

- Air temperature и Process temperature: Эти параметры представляют собой значения температуры, измеренные в кельвинах. Исследование изменений температуры может помочь выявить возможные корреляции с риском механических повреждений.

- Rotational speed и Torque: Эти параметры отражают величины частоты вращения и момента, связанные с работой станка. Анализ изменений этих параметров может быть полезным для выявления аномалий или предполагаемых причин поломок.

- Tool wear: Этот параметр отражает время использования инструмента в минутах. Анализ износа инструмента может помочь понять, как долго оборудование работало без замены инструмента, что в свою очередь может быть связано с риском поломки.

- Метка Target: Эта метка указывает, произошел ли отказ оборудования (0 - нет, 1 - да). Мы будем анализировать этот параметр, чтобы выявить особенности данных перед отказом и попытаться предсказать вероятность отказа.

Итак, эти основные параметры будут подвергнуты анализу в рамках нашего исследования, с целью выявления паттернов и связей, которые могут помочь в прогнозировании риска механических повреждений оборудования.

2.2. Предварительный анализ и обработка данных

После анализа связи признаков с предметной областью начинается этап первичного анализа, включающий получение базовых метрик по данным и выполнение ряда простых, но крайне важных этапов обработки данных перед их подачей на вход модели машинного обучения.

Удаление дублирующихся записей

В набор данных по разным причинам могут попасть повторяющиеся наблюдения. Это негативно сказывается на дальнейшем анализе, создавая иллюзию увеличенной значимости определенных паттернов или трендов. В результате модель машинного обучения может неправильно обучиться, что также повлияет на вычисление статистических метрик, таких как среднее значение, медиана или стандартное отклонение, которые будут искусственно завышены или занижены. Это может привести к неправильным выводам и неверным решениям, основанным на этих

статистических показателях. В нашем случае, к счастью, в тренировочных данных не выявлено повторяющихся записей.

Удаление столбцов без данных

Часто в данных отсутствует информация об объекте в отдельных столбцах или во всей таблице, что может происходить из-за неисправностей датчиков или халатности сотрудников, ответственных за разметку данных. Важно на начальном этапе выявить такие пропуски, так как они сказываются на полноте данных и могут исказить информацию о распределении признака. Кроме того, большинство моделей машинного обучения не могут работать с пропущенными значениями. В таких случаях либо удаляют всю строку с пропущенными значениями, чтобы избежать неверных предположений, либо заменяют пропуски на характерные значения для конкретного столбца (например, среднее значение). В данном наборе данных отсутствуют пропуски, что упрощает задачу предобработки.

Так как исходные данные являются синтетическими, а не снятыми с реальных измерительных приборов, то в них нет дупликатов и пустых значений. Проверено с помощью методов класса pandas.Dataframe.

Детекция и обработка выбросов/аномалий

Выбросы обычно возникают из-за некорректной работы оборудования для сбора данных и отличаются аномальными значениями по сравнению с общим распределением. Выбросы могут искажать результаты анализа, приводить к ошибкам и неправильным выводам, а также влиять на надежность статистических метрик и моделей машинного обучения. Обработка выбросов помогает обеспечить точность и надежность анализа данных, улучшить качество моделирования и повысить точность прогнозов. Для обнаружения выбросов строят распределение конкретного признака и выявляют точки, сильно отличающиеся от большинства. Далее, со знанием предметной области, определяют, является ли данное наблюдение выбросом или его можно считать допустимым в рамках задачи.

Несмотря на то, что исходные данные являются синтетическими, в них присутствует выброс. А именно есть такие строки данных, в которых столбец "Target"говорит о том, что в станке поломка (значение 1), при этом в типе поломки указано "No Failure". Такие записи необходимо исключить.

Эти этапы предобработки данных являются критически важными для подготовки качественного набора данных, что в конечном итоге влияет на точность и эффективность моделей машинного обучения.

Убираем из данных параметры, которые не влияют на результаты анализа, такие как ’UID’ (уникальный идентификатор) и ’Product ID’ (идентификатор продукта). Эти параметры лишь обозначают некоторый идентификатор и не несут пользы при обучении модели.

В качестве дополнительного параметра возьмем мощность вращающегося объекта:

$$P = \tau * \omega$$

$$\omega = 2 * \pi * \frac{n}{60}$$

, где n - кол-во оборотов в минуту Запишем этот параметр в датасет, как power.

Следующим шагом анализируем данные, чтобы определить, сколько из всех записей в наборе данных соответствует станкам с поломками и сколько станков работает нормально. Это позволяет оценить баланс классов и определить необходимость применения стратегий балансировки классов в дальнейшем анализе.

С помощью языка программирования Python и библиотек (таких как pandas, numpy, seaborn), покажем количество неисправных станков, представленных в данных. Для этого воспользуемся методом countplot библиотеки seaborn.

На рисунке 2.1 видим, что из 10000 строк данных только порядка 250 станков

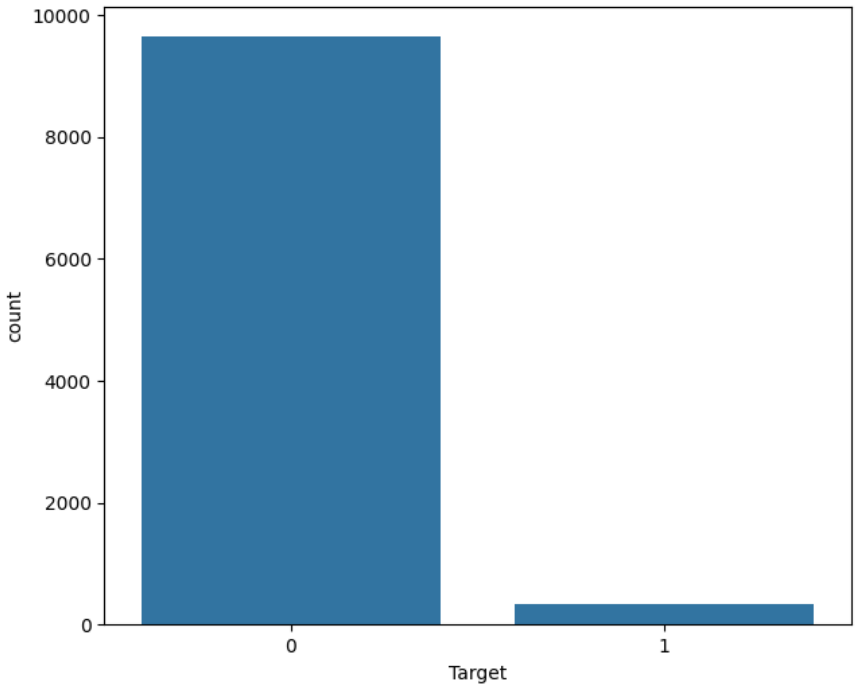

Рис.2.1. Количество неисправных и рабочих станков среди данных

с поломками. Можно сделать вывод, что класс "поломка"представлен в данных

недостаточно часто. Такое неравномерное распределение классов может повлиять на дальнейшую работу и подготовку модели к обучению следующим образом:

Дисбаланс классов: Неравномерное распределение классов может привести к переобучению модели на более представленный класс (например, "нормальная работа") и недообучению на менее представленный класс ("поломка"). Это приведет к снижению качества модели в прогнозировании риска поломок.[5]

Необходимость балансировки классов: Для корректного обучения модели и достижения более устойчивых результатов потребуется применение стратегий балансировки классов.

Затем анализируем количество различных типов поломок и их распределение. Эта информация помогает нам понять, какие типы поломок наиболее распространены, что может быть важно при выборе и настройке модели классификации. Опять же воспользуемся методом countplot

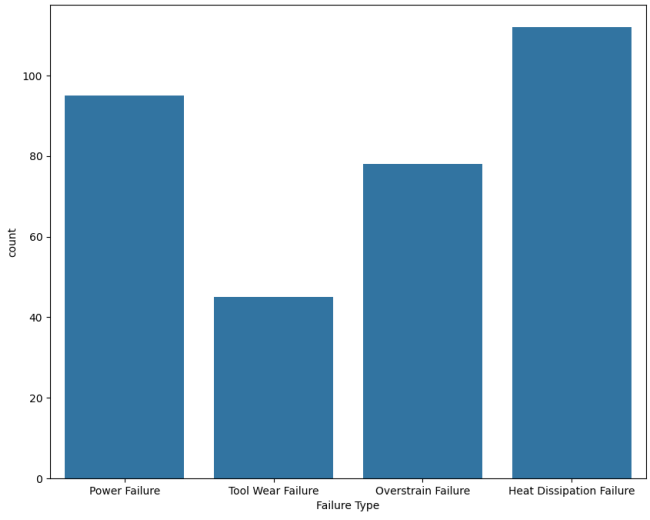

Рис.2.2. Распределние типов поломок

На рисунке 2.2 мы видим, что количество всех типов поломок близко друг к другу, за исключением типа поломки "Tool wear failure". Но так как в задачу классификации входят также данные о станках без поломок, то нам необходимо провести балансировку классов. Для этого воспользуемся SMOTE - это метод увеличения количества образцов в миноритарном классе путем создания синтетических образцов. В отличие от обычного метода повторного использования существующих данных, SMOTE создает новые примеры, которые являются линейными комбинациями ближайших соседей существующих образцов миноритарного класса. Это помогает улучшить качество обучения моделей машинного обучения,

обеспечивая более равномерное распределение данных.

На рисунке 2.3 можно наблюдать большой перевес в пользу класса "No Failure". После применения алгоритма балансировки на рисунке 2.4 можно увидеть, что кол-во данных увеличилось, но пропал дисбаланс классов.

Одним из ключевых методов в машинном обучении является корреляционный

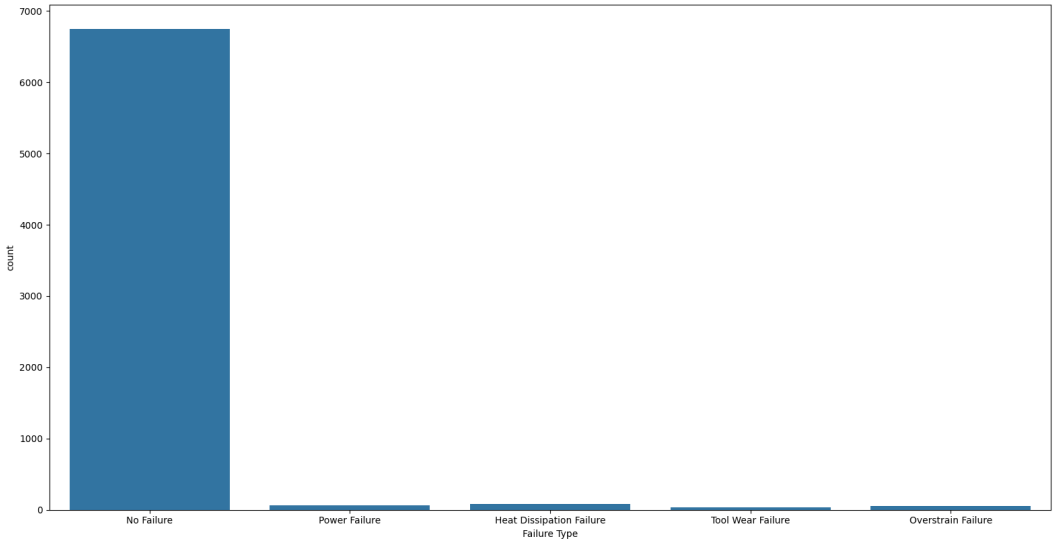

Рис.2.3. Распределение классов до балансировки

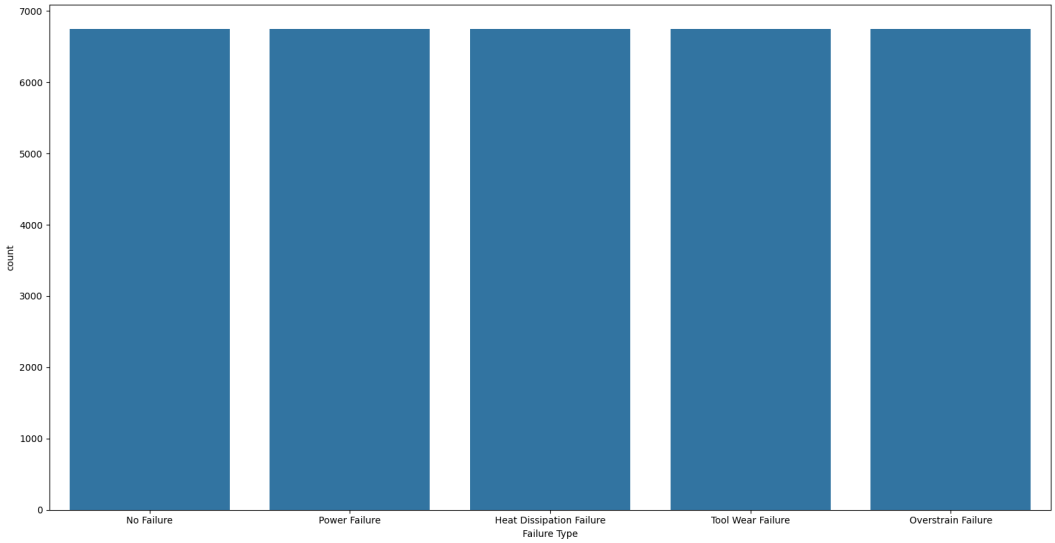

Рис.2.4. Распределение классов после балансировки

анализ. Этот статистический метод позволяет определить степень связи между двумя или более переменными в наборе данных, оценивая, насколько сильно и в

какую сторону (положительно или отрицательно) они связаны друг с другом[3, 2].

Цель корреляционного анализа заключается в изучении природы и силы связей между переменными. Метод позволяет построить численный показатель, известный как коэффициент корреляции, который измеряет степень линейной зависимости между переменными. Коэффициент корреляции принимает значения от -1 до +1. Чем ближе значение коэффициента к ±1, тем сильнее линейная зависимость между переменными. Знак коэффициента показывает направление зависимости: положительное значение указывает на прямую зависимость, а отрицательное – на обратную.

В нашем случае мы будем применять корреляционный анализ с вычислением корреляции Пирсона к столбцам-признакам, чтобы выявить степень линейной зависимости между ними. Предварительно убрав строковые типы данных, построим коррелограмму.

Исходя из представленной коррелограммы на рисунке 2.5 можно сделать

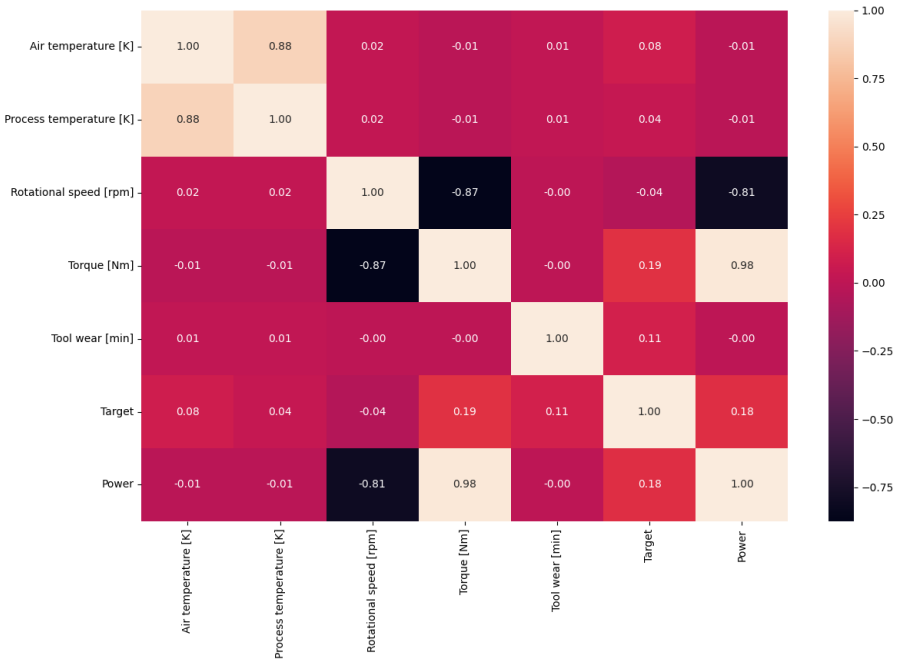

Рис.2.5. Коррелограмма

следующие выводы: общая матрица корреляций показывает отсутствие линейной зависимости между большей частью рассматриваемых признаков, что говорит о корретности их одновременного использования в модели машинного обучения. Но есть и скореллированные величины: Air temperature с Process temperature, а также Rotation speed с torque.

2.3. Выводы

Глава, посвященная анализу и предобработке данных, позволила выявить и устранить несколько ключевых проблем, что значительно повысило качество и надежность подготовленного набора данных для построения модели машинного обучения.

Обнаружение и удаление выбросов:

В ходе анализа был обнаружены выброс в данны, который мог искажать результаты моделирования. Выбросы был удален, что позволило обеспечить более достоверное и интерпретируемое представление данных.

Отсутствие дубликатов и пропущенных значений:

В наборе данных не было обнаружено дублирующихся записей, что избавило от необходимости их удаления и гарантировало, что все наблюдения уникальны. Также в данных не выявлено пропущенных значений, что исключило необходимость дополнительных методов их обработки и обеспечило полноту информации для каждого признака.

Коррелированные величины:

Анализ корреляционной матрицы показал, что некоторые величины сильно коррелируют друг с другом, например, air temperature и process temperature (0.88), torque и rotational speed (-0.88). Для предотвращения мультиколлинеарности и избыточности информации в модели было принято решение включить только один из коррелирующих признаков.

Дисбаланс классов:

В наборе данных наблюдается значительный дисбаланс классов: из 10,000 записей только около 250 относятся к случаям поломок. Этот дисбаланс необходимо учитывать при построении модели, применяя методы балансировки классов, чтобы модель могла эффективно распознавать случаи поломок и не переобучаться на более представленный класс.

Эти шаги по анализу и предобработке данных обеспечили создание качественного и надежного набора данных, готового для использования в построении и обучении модели машинного обучения для прогнозирования риска механических повреждений станков. В следующей главе мы перейдем к выбору и обучению модели на подготовленных данных, учитывая все выявленные особенности и проведенные преобразования.

ГЛАВА 3. ОБУЧЕНИЕ МОДЕЛЕЙ И АНАЛИЗ РЕЗУЛЬТАТОВ

3.1. Общий подход к обучению модели и оценки эффективности

Для начала необходимо поделить исходные данные на признаки и целевые значения, а также на данные для обучения и данные для оценки модели.

Так как ранее выяснили, что в исходных данных присутсвует дисбаланс классов - необходимо стратифицированное разбиение. Стратифицированное разбиение означает, что разбиение данных будет сделано таким образом, чтобы пропорции классов в обучающей и тестовой выборках оставались такими же, как в исходном наборе данных. Это особенно важно при работе с несбалансированными данными, где количество примеров одного класса значительно отличается от количества примеров другого класса.[7]

Перед обучением модели, так как некоторые признаки представлены в текстовом формате (Failure Type и Type), необходимо преобразовать их в бинарное представление с помощью:

OneHotEncoder используется для преобразования категориальных признаков (например, "цвет "тип "страна") в бинарное представление, где каждый уникальный категориальный уровень представляется отдельным бинарным столбцом. Это помогает алгоритмам машинного обучения работать с категориальными данными.

LabelEncoder используется для преобразования категориальных целевых значений (или меток) в числовые значения. Это полезно, когда целевая переменная в задаче машинного обучения представлена в виде категорий.

После чего необходимо создать модель и обучить на закондированных данных с помощью метода fit. Далее для каждого вида модели, будет одинаковый подход - Для оценки модели будем использовать отчет о классификации ($classification - report$) из библиотеки scikit-learn. Он предоставляет важные метрики для оценки качества модели классификации [14]. Эти метрики включают:

1. Precision (Точность) Определение: Доля истинных положительных предсказаний среди всех предсказаний положительного класса.

Формула:

$$Precision = \frac{TP}{TP + FP}$$

где:

TP (True Positives) — количество истинных положительных предсказаний.
FP (False Positives) — количество ложных положительных предсказаний.

Интерпретация: Точность показывает, насколько модель точна при предсказании положительного класса. Высокая точность означает, что ложных положительных предсказаний мало.

2. Recall (Полнота) Определение: Доля истинных положительных предсказаний среди всех фактических положительных примеров.

Формула:

$$Recall = \frac{TP}{TP + FN}$$

где:
TP (True Positives) — количество истинных положительных предсказаний.
FN (False Negatives) — количество ложных отрицательных предсказаний.

Интерпретация: Полнота показывает, насколько хорошо модель находит все положительные примеры. Высокая полнота означает, что ложных отрицательных предсказаний мало.

3. F1-score Определение: Гармоническое среднее точности и полноты.
Формула:

$$Recall = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Интерпретация: F1-score обеспечивает баланс между точностью и полнотой. Полезен, когда важно найти компромисс между точностью и полнотой, особенно при несбалансированных данных.

В отчете о классификации также присутствуют метрики accuracy, macro avg и weighted avg. Рассмотрим их подробнее:

1. Accuracy (Точность) Определение: Доля правильных предсказаний среди всех предсказаний.
Формула:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

где:
TP (True Positives) — количество истинных положительных предсказаний.

TN (True Negatives) — количество истинных отрицательных предсказаний.
FP (False Positives) — количество ложных положительных предсказаний.
FN (False Negatives) — количество ложных отрицательных предсказаний.

Интерпретация: Accuracy показывает общую долю правильных предсказаний. Однако, в случае несбалансированных данных, эта метрика может быть обманчивой, так как модель может иметь высокую точность просто за счет правильного предсказания преобладающего класса.

2. Macro Average (Макро Среднее) Определение: Среднее значение метрик (precision, recall, f1-score) по всем классам без учета дисбаланса классов. Каждый класс имеет равный вес при вычислении макро среднего.

Формулы:

$$MacroPrecision = \frac{1}{N}\sum_{i=1}^{N} Precision_i$$

$$MacroRecall = \frac{1}{N}\sum_{i=1}^{N} Recall_i$$

$$MacroF1-score = \frac{1}{N}\sum_{i=1}^{N} F1-score_i$$

где N — количество классов.

Интерпретация: Macro Average полезно, когда важно одинаково учитывать все классы, независимо от их частоты. Эта метрика не учитывает дисбаланс классов.

3. Weighted Average (Взвешенное Среднее) Определение: Среднее значение метрик (precision, recall, f1-score) по всем классам с учетом дисбаланса классов. Взвешивается по количеству примеров каждого класса (support).

Формулы:

$$WeightedPrecision = \frac{1}{\sum_{i=1}^{N} Support_i}\sum_{i=1}^{N}(Precision_i * Support_i)$$

$$WeightedRecall = \frac{1}{\sum_{i=1}^{N} Support_i}\sum_{i=1}^{N}(Recall_i * Support_i)$$

$$WeightedF1 - score = \frac{1}{\sum_{i=1}^{N} Support_i} \sum_{i=1}^{N} (F1 - score_i * Support_i)$$

где N — количество классов, $Support_i$— количество истинных примеров для класса i.

Интерпретация: Weighted Average учитывает дисбаланс классов, предоставляя более точное представление о производительности модели на несбалансированных данных.

3.2. Обучение на данных с дисбалансом классов

После обучения всех моделей на подготовленных данных, были получены следующие результаты по различным метрикам. Сначала обучение проводилось на данных с дисбалансом классов.

Высокая точность для класса "No failure". Это указывает на то, что модель

Таблица 3.1

Логистическая регрессия. Дисбаланс классов

	$Precision$	$Recall$	$F1 - score$
Heat Dissipation Failure	0	0	0
No Failure	0.97	1	0.98
Overstrain Failure	0	0	0
Power Failure	0	0	0
Tool Wear Failure	0	0	0
accuracy	0.97	0.97	0.97
macro avg	0.19	0.2	0.2
weighted avg	0.94	0.97	0.95

хорошо идентифицирует случаи, когда нет отказа, что связано с доминированием этого класса в данных.

Низкие показатели для других классов - Показывают, что модель иногда пра-

Таблица 3.2

Алгоритм K - ближайших соседей. Дисбаланс классов

	$Precision$	$Recall$	$F1-score$
Heat Dissipation Failure	0	0	0
No Failure	0.98	1	0.99
Overstrain Failure	0	0	0
Power Failure	1	1	1
Tool Wear Failure	0	0	0
accuracy	0.98	0.98	0.98
macro avg	0.4	0.4	0.4
weighted avg	0.95	0.98	0.97

Таблица 3.3

Алгоритм опорных векторов. Дисбаланс классов

	$Precision$	$Recall$	$F1-score$
Heat Dissipation Failure	0	0	0
No Failure	0.97	1	0.98
Overstrain Failure	0	0	0
Power Failure	1	0.34	0.51
Tool Wear Failure	0	0	0
accuracy	0.97	0.97	0.97
macro avg	0.39	0.27	0.3
weighted avg	0.95	0.97	0.96

Таблица 3.4

Случайный лес. Дисбаланс классов

	$Precision$	$Recall$	$F1-score$
Heat Dissipation Failure	0.91	0.85	0.85
No Failure	1	1	1
Overstrain Failure	0.83	0.83	0.83
Power Failure	0.97	0.97	0.97
Tool Wear Failure	0.8	0.92	0.86
accuracy	1	1	1
macro avg	0.9	0.91	0.91
weighted avg	1	1	1

Таблица 3.5

Дерево решений. Дисбаланс классов

	$Precision$	$Recall$	$F1-score$
Heat Dissipation Failure	0.88	0.82	0.85
No Failure	1	1	1
Overstrain Failure	0.76	0.83	0.79
Power Failure	0.89	0.86	0.88
Tool Wear Failure	0.79	0.85	0.81
accuracy	0.99	0.99	0.99
macro avg	0.72	0.73	0.72
weighted avg	0.99	0.99	0.99

вильно классифицирует случаи этого отказа, но пропускает большинство из них (очень низкий recall)

Нулевые значения для некоторых отказов - Это свидетельствует о том, что модель не распознаёт ни одного случая этих отказов, что указывает на проблемы с идентификацией этих редких событий.

Низкие значения macro average - Подчеркивает слабую производительность модели по всем классам отказов в целом

Но можно выделить два алгоритма, которые лучше всего справились с поставленной задачей:

Случайный лес и Дерево решений отлично справляются с задачей прогнозирования риска механических повреждений станков по нескольким причинам:

Устойчивость к дисбалансу классов: Случайный лес и Дерево решений могут обрабатывать дисбаланс классов, что часто встречается в реальных данных. Они имеют механизмы, такие как взвешивание классов или использование критерия Джини, которые позволяют эффективно работать с неравномерными распределениями классов.

Способность к обработке нелинейных зависимостей: В задачах, связанных с прогнозированием отказов станков, зависимости между признаками и целевой переменной могут быть сложными и нелинейными. Случайный лес и Дерево решений способны обнаруживать и учитывать такие зависимости, благодаря их древовидной структуре и возможности строить нелинейные разделяющие гиперплоскости.

Гибкость и масштабируемость: Оба алгоритма гибки и могут быть легко настроены под конкретные требования задачи. Кроме того, они масштабируются хорошо на большие объемы данных и могут обрабатывать большое количество признаков без значительного увеличения времени обучения.

Интерпретируемость: Дерево решений предоставляет понятные правила принятия решений, которые могут быть легко интерпретированы человеком. Это позволяет понять, какие признаки вносят наибольший вклад в прогнозирование, что важно для понимания процесса и выявления потенциальных причин отказов.

3.3. Обучение на данных без дисбаланса классов

Следующим этапом были обучены модели с помощью тех же алгоритмов, но уже на данных, к которым был применен алгоритм SMOTE для того, чтобы убрать дисбаланс классов путем добавление синтетических данных.

Полученные результаты показали, что модели логистической регрессии, алго-

Таблица 3.6

Логистическая регрессия. Отсутствие дисбаланса кдассов

	Precision	*Recall*	$F1 - score$
Heat Dissipation Failure	0.028	0.24	0.05
No Failure	1	0.58	0.73
Overstrain Failure	0.22	0.87	0.35
Power Failure	0	0	0
Tool Wear Failure	0.04	0.92	0.076
accuracy	0.58	0.58	0.58
macro avg	0.26	0.52	0.24
weighted avg	0.96	0.58	0.71

Таблица 3.7

Алгоритм K - ближайших соседей. Отсутствие дисбаланса кдассов

	Precision	*Recall*	$F1 - score$
Heat Dissipation Failure	0.019	0.38	0.037
No Failure	0.99	0.54	0.7
Overstrain Failure	0.13	0.83	0.22
Power Failure	1	1	1
Tool Wear Failure	0.016	0.69	0.031
accuracy	0.54	0.54	0.954
macro avg	0.43	0.69	0.4
weighted avg	0.97	0.54	0.69

ритма опорных векторов и K-ближайших соседей продемонстрировали улучшение в предсказании некоторых классов, которые ранее не распознавались. Однако, для других классов, которые ранее угадывались, точность предсказаний варьировалась: для некоторых она улучшилась, а для других ухудшилась.

Таблица 3.8

Алгоритм опорных векторов. Отсутствие дисбаланса кдассов

	$Precision$	$Recall$	$F1-score$
Heat Dissipation Failure	0.022	0.47	0.042
No Failure	0	0	0
Overstrain Failure	0.061	0.87	0.11
Power Failure	0.48	1	0.65
Tool Wear Failure	0.0048	0.69	0.0095
accuracy	0.025	0.025	0.025
macro avg	0.11	0.61	0.16
weighted avg	0.0054	0.025	0.0077

Таблица 3.9

Случайный лес. Отсутствие дисбаланса кдассов

	$Precision$	$Recall$	$F1-score$
Heat Dissipation Failure	0.93	0.82	0.88
No Failure	1	1	1
Overstrain Failure	0.85	0.74	0.79
Power Failure	0.97	0.97	0.97
Tool Wear Failure	0.65	1	0.79
accuracy	1	1	1
macro avg	0.88	0.91	0.88
weighted avg	1	1	1

Таблица 3.10

Дерево решений. Отсутствие дисбаланса кдассов

	$Precision$	$Recall$	$F1-score$
Heat Dissipation Failure	0.83	0.88	0.86
No Failure	1	1	1
Overstrain Failure	0.81	0.74	0.77
Power Failure	1	0.93	0.96
Tool Wear Failure	0.67	0.77	0.71
accuracy	0.99	0.99	0.99
macro avg	0.86	0.86	0.86
weighted avg	1	0.99	1

В то же время, модели случайного леса и дерева решений не продемонстрировали значительного улучшения, а в некоторых случаях даже показали ухудшение результатов.

Эти наблюдения можно объяснить особенностями работы алгоритма SMOTE. Алгоритм SMOTE увеличивает плотность данных за счет генерации синтетических образцов миноритарных классов. Эти синтетические образцы создаются путем интерполяции между существующими образцами, что позволяет увеличить количество данных для миноритарных классов и улучшить их представление в обучающей выборке. Из-за увеличения плотности объектов некоторым алгоритмам стало сложнее определять принадлежность объекта к одному определенному классу.

3.4. Выводы

Наилучшие результаты:

Случайный лес и Дерево решений продемонстрировали высокие значения метрик precision, recall и F1-score для большинства классов отказов, особенно для классов "Heat dissipation failure"; "Overstrain failure"; "Power failure"и "Tool wear failure". Эти модели показали высокую точность (accuracy 0.99) и сбалансированные результаты по большинству метрик.

Средние результаты:

K-ближайших соседей и Логистическая регрессия показали умеренные результаты, с точностью около 0.97. Однако они не справляются с классификацией некоторых классов отказов (например, "Random failure"и "Tool wear failure"), что указывает на ограниченные возможности этих моделей в данной задаче. Наихудшие результаты:

Алгоритм опорных векторов показал самые слабые результаты, не справляясь с классификацией большинства классов отказов (например, "Heat dissipation failure"; "Overstrain failure"; "Random failure"и "Tool wear failure"). Несмотря на высокую точность для класса "No failure". Эта модель не подходит для задач, требующих более точной классификации редких событий. Важность метрик оценки:

Accuracy: Показатель точности демонстрирует общее качество модели, однако при наличии дисбаланса классов, как в данном случае, он может вводить в заблуждение. Высокая точность может быть достигнута за счет правильной классификации наиболее частого класса ("No failure"). Precision, Recall и F1-score:

Precision указывает на долю правильных положительных предсказаний среди всех предсказанных положительных случаев.

Recall показывает, какую долю всех положительных случаев модель правильно классифицировала.

F1-score представляет собой гармоническое среднее precision и recall и является важной метрикой в условиях дисбаланса классов, так как учитывает как ложноположительные, так и ложноотрицательные предсказания.

Решение дисбаланса классов путем синтезирования данных частично улучшило результаты для тех моделей, которые вовсе не прогнозировали определенный класс, но из-за увеличения плотности объектов также и ухудшило результаты для моделей случайного леса и дерева решений.

ЗАКЛЮЧЕНИЕ

В данной дипломной работе был проведен комплексный анализ данных для прогнозирования риска механических повреждений станков с использованием методов машинного обучения. Основная цель заключалась в разработке модели, способной точно предсказывать вероятные поломки станков, что позволит предприятию принимать своевременные меры для предотвращения простоев и оптимизации производственных процессов.

Первоначальный этап включал тщательный анализ исходных данных, что позволило выявить ключевые признаки, влияющие на риск механических повреждений станков. Этот анализ показал высокую корреляцию между такими признаками, как температура воздуха и температура процесса, а также момент силы и скорость вращения.

Затем была проведена подготовка данных для обучения модели, включающая предобработку данных, удаление выбросов, заполнение пропущенных значений и масштабирование признаков. Это улучшило качество данных и позволило исключить мультиколлинеарность, что является важным для повышения точности модели.

В процессе выбора и обучения модели были рассмотрены и протестированы различные алгоритмы машинного обучения, включая логистическую регрессию, K-ближайших соседей, алгоритм опорных векторов, дерево решений и случайный лес. Наилучшие результаты показали модели на основе случайного леса и дерева решений благодаря своей способности обрабатывать сложные, нелинейные зависимости и устойчивости к дисбалансу классов.

Также была проведена оценка качества модели с использованием различных метрик, таких как точность, полнота и F1-мера. Модели случайного леса и дерева решений продемонстрировали высокие показатели по сравнению с другими моделями. Логистическая регрессия и алгоритм опорных векторов показали низкую эффективность из-за недостаточной способности справляться с нелинейными зависимостями и дисбалансом классов, в то время как модели K-ближайших соседей также продемонстрировали относительно низкую точность из-за влияния дисбаланса классов.

Был применен алгоритм SMOTE для балансировки классов. Результаты показали, что модели логистической регрессии, алгоритма опорных векторов и K-ближайших соседей улучшили предсказание миноритарных классов, которые

ранее не распознавались. Однако, для моделей случайного леса и дерева решений результаты либо остались на прежнем уровне, либо ухудшились из-за увеличения плотности объектов.

Для практического внедрения можно рекомендовать модель на основе случайного леса или дерева решений. Интеграция разработанной модели в систему мониторинга оборудования позволит своевременно выявлять и предотвращать потенциальные поломки, что значительно повысит надёжность и безопасность производственных процессов.

Также можно продолжить работу по улучшению качества данных, включая регулярный анализ и устранение выбросов, а также сбор дополнительных данных для повышения точности прогнозов. Это позволит адаптировать модель к новым данным и поддерживать её актуальность и точность.

В будущем следует исследовать возможности применения более сложных моделей машинного обучения, таких как глубокие нейронные сети, для дальнейшего повышения точности прогнозов и проводить эксперименты с различными методами балансировки классов и повышения эффективности моделей.

Результаты данного исследования подтверждают высокую актуальность и перспективность применения методов машинного обучения для задач прогнозирования рисков в промышленности. Разработанные модели способны значительно повысить надёжность и безопасность производственных процессов, минимизировать время простоя оборудования и связанные с этим финансовые потери. Продолжение работы в этом направлении позволит достигнуть ещё более значимых результатов и улучшить функционирование исследуемого объекта.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Биргер, И.А. Техническая диагностика. – М.: «Машиностроение», 1978. 240 с.

2. Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. – СанктПетербург: Питер, 2017. – 336 с.

3. Воронина, В. В. Теория и практика машинного обучения : учебное пособие / В. В. Воронина. — Ульяновск : УлГТУ, 2017. — 290 с. — ISBN 978-5-9795-1712-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/165053 (дата обращения: 30.06.2024)

4. Григорьев, С.Н. Диагностика автоматизированного производства / С.Н. Григорьев, В.Д. Гурин, М.П. Козочкин и др. – М.: Машиностроение, 2011. 600 с.

5. Демидова, Л. А., Шаршатов, М. А., Шыхыев, А. А. методы решения проблемы дисбаланса классов в задаче бинарной классификации [Текст] / Л. А. Демидова, М. А. Шаршатов, А. А. Шыхыев // «ИТ-Стандарт». — 2023. — No 1. — С. 22- 33.

6. Кабалдин Ю.Г., Шатагин Д.А., Кузьмишина А.М. Управление технологическим оборудованием предприятия в условиях цифровых производств на основе искусственного интеллекта и облачных технологий. Итоги 2017 года: научные исследования и разработки. Сб. мат. междунар. науч.-практ. конф., 20 января 2018, Иркутск, Научное партнерство «Апекс», 2017, с. 94–107.

7. Миронов А.М. Машинное обучение. Часть 1: учебник для вузов. – М.: МАКС Пресс, 2018. – 90 с.

8. Учебник по машинному обучению. Линейные модели / Яндекс. — 2023. — URL: https://education.yandex.ru/handbook/ml/article/linear-models (дата обращения: 24.04.2024).

9. Cherif, W. Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis. Procedia Comput. Sci. 127, 293–299 (2018).

10. Classification and Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen, C.T. Stone // Wadsworth. Belmont. California. 1984.

11. Random Forest Regression in Python / Geeks for geeks. — 2023. — URL: https://www.geeksforgeeks.org/ random- forest- regression-in- python/ (дата обращения: 29.05.2024).

12. Linear Regression Masterclass - ML / Kaggle. — 2023. — URL: https://www. kaggle.com/code/auxeno/linear-regression-masterclass-ml (дата об-

ращения: 31.05.2024).

13. McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.

14. Müller, A.C. Introduction to Machine Learning with Python / A.C. Müller, S. Guido. – O'Reilly Media, 2016. – 394 p. – ISBN 978-1-449-36941-5.

15. Stehman, S.V. Selecting and interpreting measures of thematic classification accuracy // Remote Sensing of Environment. 1997. Vol. 62, No. 1. P. 77–89. DOI: 10.1016/S0034-4257(97)00083-7.