**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

Федеральное государственное автономное образовательное учреждение высшего образования

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»

РУТ (МИИТ)

Выпускная квалификационная работа  
на тему

**«Анализ влияния количества критических сообщений за 7 дней ранее на появление непланового ремонта и построение модели прогнозирования возникновение непланового ремонта в ближайший 1 день»**

ФИО: Кулагин Максим Алексеевич

Группа: ППП-111

Вариант: 1

Содержание

[1. Постановка задачи 3](#_Toc161610898)

[2. Анализ полученных данных о работе локомотивов 5](#_Toc161610899)

[2.1. Первичный анализ данных 5](#_Toc161610900)

[2.2. Формирование набора данных для обучение моделей 8](#_Toc161610901)

[3. Обучение моделей и оценка результатов 12](#_Toc161610902)

[3.1. Разделение на обучающую и тестовую выборку 12](#_Toc161610903)

[3.2. Результат обучения нескольких моделей 12](#_Toc161610904)

[3.3. Сохранение и применение результатов работы моделей 13](#_Toc161610905)

[4. Заключение 15](#_Toc161610906)

# Постановка задачи

В современном мире эффективное функционирование железнодорожного транспорта играет ключевую роль в обеспечении экономического развития и социальной стабильности. Локомотивы являются сердцем железнодорожной системы, обеспечивая перевозку грузов и пассажиров. Однако, как и любая техническая система, локомотивы подвержены износу и потребности в регулярном обслуживании и ремонте.

Неплановые ремонты локомотивов являются серьезной проблемой для российских железных дорог. Они не только приводят к значительным финансовым затратам, но и могут вызывать простои в работе, что в свою очередь отрицательно сказывается на качестве обслуживания пассажиров и грузов.

В данной работе решается задача прогнозирования неплановых ремонтов локомотивов серии 2ЭС6. Данная серия локомотивов широко применяется на железнодорожных дорогах. Прогнозирование неплановых ремонтов данной серии локомотивов позволит оптимизировать расходы на техническое обслуживание и ремонт, а также минимизировать вероятность простоев в работе, что в итоге способствует повышению эффективности железнодорожного транспорта.

В ходе работы проведен анализ данных о состоянии локомотивов серии 2ЭС6 и построена модель прогнозирования неплановых ремонтов.

Целью данной работы является разработка модели прогнозирования неплановых ремонтов локомотивов серии 2ЭС6, которую можно использовать для повышения оперативности принятия решений по допуску локомотивов на инфраструктуру, что в свою очередь позволит оптимизировать ресурсное обеспечение и обслуживание железнодорожного транспорта.

Задача данной работы состоит в разработке модели прогнозирования неплановых ремонтов локомотивов серии 2ЭС6 с целью повышения оперативности принятия решений по допуску локомотивов на инфраструктуру железнодорожных путей. Это включает в себя:

1. Сбор и анализ данных о состоянии локомотивов серии 2ЭС6, а также данных о прошлых неплановых ремонтах и причинах их возникновения.
2. Выбор и применение подходящих методов анализа данных и машинного обучения для разработки модели прогнозирования.
3. Создание и настройка модели прогнозирования, способной предсказывать вероятность возникновения неплановых ремонтов на основе имеющихся данных.
4. Валидация и тестирование модели на независимых наборах данных для оценки ее точности.

# Анализ полученных данных о работе локомотивов

## Первичный анализ данных

Для обучения и использования любой модели машинного обучения сначала требуется подготовить данные. В рамках данной главы описано решение 1 и 2 задачи, а именно изучить и проанализировать данные о работе локомотивов серии 2ЭС6.

В рамках данной квалификационной работы представлены в двух файлах: диагностические сообщения с локомотива (diagnostic\_message.csv) и неплановые ремонты (unplanned\_repairs.csv).

На первом шаге в анализе данных произведем изучение структуры и содержания обоих наборов данных. Это включает в себя ознакомление с переменными, их типами, а также общими характеристиками данных, такими как количество записей, диапазоны значений и наличие пропущенных данных.

Сначала мы загрузим данные с использованием следующих строк кода:

diagnostic\_messages = pd.read\_csv('data\_for\_student/1/diagnostic\_messages.csv', sep=';')

unplanned\_repairs = pd.read\_csv('data\_for\_student/1/unplanned\_repairs.csv', sep=';')

Первые 5 строк изображены ниже (Рисунок 1 и Рисунок 2). Всего записей в diagnostic\_message – 212435; unplanned\_repairs – 466.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 1 Первые строки набора данных 'diagnostic\_message.csv'

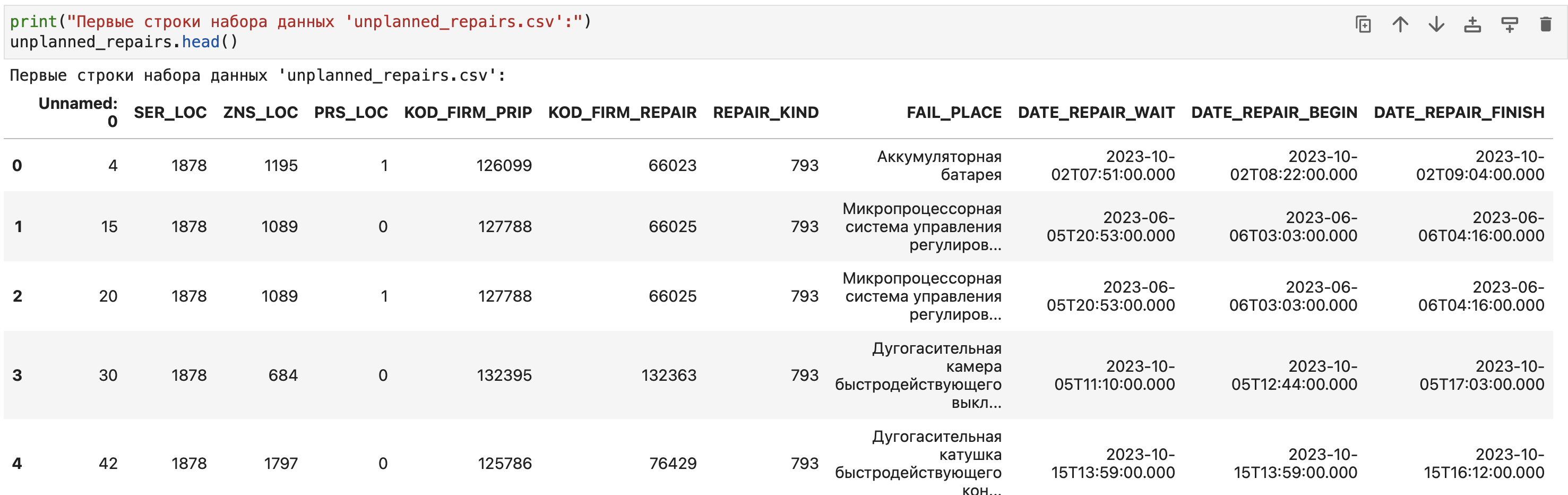
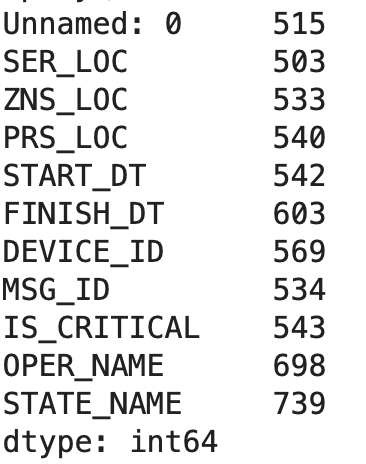
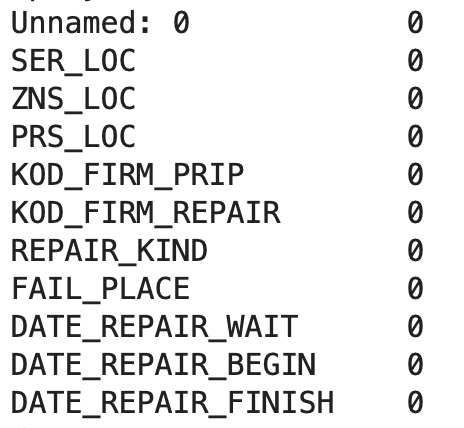


Рисунок 2 Первые строки набора данных 'unplanned\_repairs.csv'

Далее с использованием Pandas проанализируем пропуски в данных и найдем столбцы, в которых есть значения null или nan (Рисунок 3). Это необходимо для того, чтобы подготовить корректные и очищенные данные для обучения модели.

print(diagnostic\_messages.isnull().sum())

print(unplanned\_repairs.isnull().sum())

а) б)

Рисунок 3 Представлено количество пропусков в данных

Из первичного анализа наблюдаем колонки «*Unnamed: 0»*, которые не были изначально заявлены в предоставленной спецификации (Таблица 1).

Таблица 1 Спецификация на предоставленные данные

|  |  |  |
| --- | --- | --- |
| **Наименование столбца** | **Описание** | **Исходная таблица** |
| SER\_LOC | Серия локомотива | diagnostic\_message  unplanned\_repairs |
| ZNS\_LOC | Номер локомотива | diagnostic\_message  unplanned\_repairs |
| PRS\_LOC | Секция локомотива | diagnostic\_message  unplanned\_repairs |
| START\_DT | Дата начала диагностического сообщения | diagnostic\_message |
| FINISH\_DT | Дата окончания диагностического сообщения | diagnostic\_message |
| DEVICE\_ID | ID устройства | diagnostic\_message |
| MSG\_ID | ID сообщения | diagnostic\_message |
| IS\_CRITICAL | Критиченость сообщения (1 - критичное, 0 - не критичное) | diagnostic\_message |
| OPER\_NAME | Наименование операции | diagnostic\_message |
| STATE\_NAME | Наименование состояния | diagnostic\_message |
| KOD\_FIRM\_PRIP | Код предприятия приписки локомотива | unplanned\_repairs |
| KOD\_FIRM\_REPAIR | Код предприятия ремонта локомотива | unplanned\_repairs |
| REPAIR\_KIND | Вид ремонта | unplanned\_repairs |
| FAIL\_PLACE | Место ремонта | unplanned\_repairs |
| DATE\_REPAIR\_WAIT | Дата постановки в ожидание на ремонт | unplanned\_repairs |
| DATE\_REPAIR\_BEGIN | Дата начало ремонта | unplanned\_repairs |
| DATE\_REPAIR\_FINISH | Дата окончание ремонта | unplanned\_repairs |

После удаления неиспользуемых столбцов все данные были проверены на наличие дубликатов в данных. В таблице с диагностическими сообщениями выявлено 15743, а в таблице с неплановыми ремонтами выявлено 2 строчки дубликатов. С использованием следующих команды дубликаты были удалены из исходных данных:

diagnostic\_messages.drop\_duplicates(inplace=True)

unplanned\_repairs.drop\_duplicates(inplace=True)

После устранения дубликатов и устранения пропусков в данных (используемый способ устранения пропусков представлен в приложенном коде) количество записей в рассматриваемых наборах сократилось до 191118 (diagnostic\_message) и 464 (unplanned\_repairs).

## Формирование набора данных для обучение моделей

Перед формирование итогового набора данных для обучения были произведены следующие операции: преобразование типов данных, формирование новых признаков, соединение таблиц в одну таблицу, разделение на обучающую и тестовую выборки.

В качестве преобразования типов выделить можно выделить преобразование дат и времени, пример:

diagnostic\_messages["START\_DT"] = pd.to\_datetime(diagnostic\_messages["START\_DT"])

diagnostic\_messages["FINISH\_DT"] = pd.to\_datetime(diagnostic\_messages["FINISH\_DT"])

В качестве признаков было добавлена длительность диагностических сообщений и длительность ремонта.

diagnostic\_messages["DURATION"] = (diagnostic\_messages.FINISH\_DT - diagnostic\_messages.START\_DT).dt.total\_seconds()

unplanned\_repairs['repair\_duration'] = (unplanned\_repairs['DATE\_REPAIR\_FINISH'] - unplanned\_repairs['DATE\_REPAIR\_BEGIN']).dt.total\_seconds()

Алгоритм объединения таблиц, который был реализован в приложенном коде, предназначен для анализа данных о неплановых ремонтах локомотивов и диагностических сообщениях с локомотивов с целью объединения их. Для каждой секции локомотива алгоритм проходит через следующие шаги:

1. Из набора данных о неплановых ремонтах выбираются записи, соответствующие текущей секции локомотива.
2. Из набора данных о диагностических сообщениях выбираются записи, соответствующие текущей секции локомотива.
3. Для каждого непланового ремонта анализируются диагностические сообщения, полученные в определенный период до ремонта. Это включает подсчет количества критических сообщений за последний день, последние 7 дней и последние 30 дней до начала ремонта. Производится анализ уникальных сообщений за последние 30 дней.
4. Для каждого непланового ремонта создается запись, содержащая информацию о текущей секции локомотива, дате ремонта, количестве критических сообщений и уникальных сообщений в различные периоды перед ремонтом. Затем эти записи объединяются в один набор данных.
5. Для каждой секции локомотива создается также набор данных, содержащий информацию о диагностических сообщениях за последние 30 дней до случайного будущего дня (более чем 30 дней до начала ремонта). Этот набор данных также содержит информацию о критических сообщениях и уникальных сообщениях за аналогичные периоды.
6. Полученные наборы данных объединяются в два общих списка: один для случаев с ремонтами (с указанием метки 1) и один для случаев без ремонтов (с указанием метки 0).

Этот алгоритм формирует набор данных для обучения моделей и позволяет выявить связи между диагностическими сообщениями и неплановыми ремонтами локомотивов, а также использовать эти данные для анализа и прогнозирования возможных ремонтов в будущем.

В результирующем наборе данных получилось 12712 строк и 24 столбца.

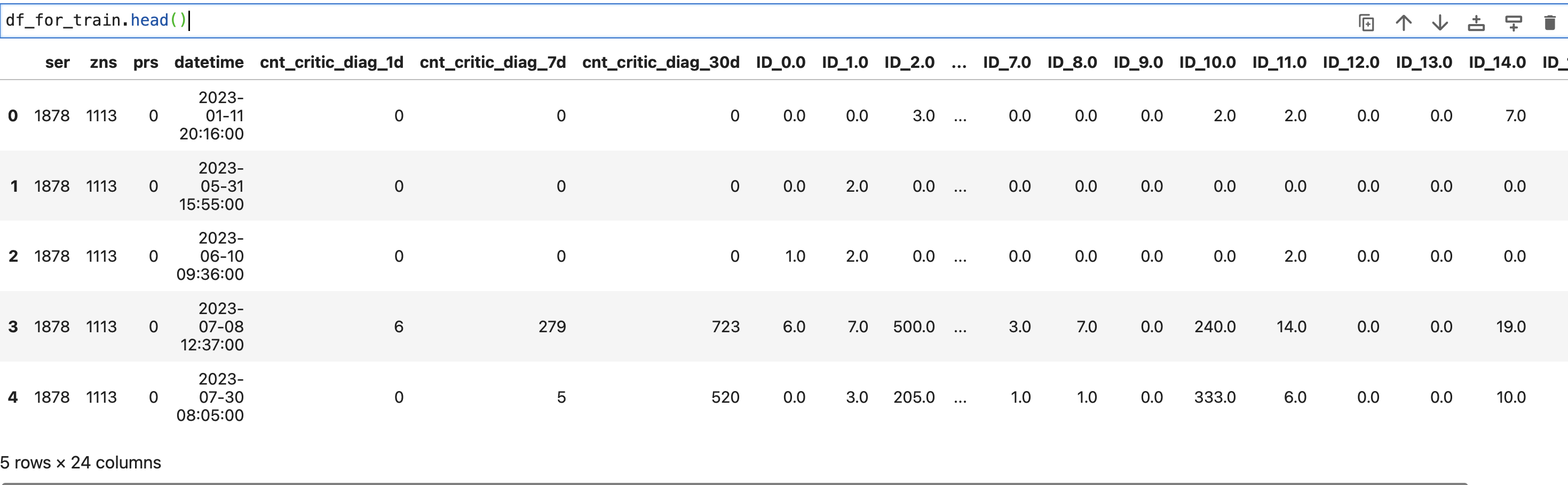


Рисунок 4 Пример сформированных признаков

Первично проведем анализ меток. т.е. фактических неплановых ремонтах (Рисунок 5). В рамках анализа меток можно сделать, что выборка несбалансирована и имеет явное смещение в сторону количества 0. Данную информацию обязательно нужно учесть при обучении моделей.

Можно в качестве примера анализа признаков привести признак, который показывает количество критических диагностических сообщений за 7 дней (Рисунок 6). На данном графике видно, что в случаях, когда возникал неплановый ремонт количество критических сообщений было в среднем больше, чем в тех случаях, когда непланового ремонта не было.

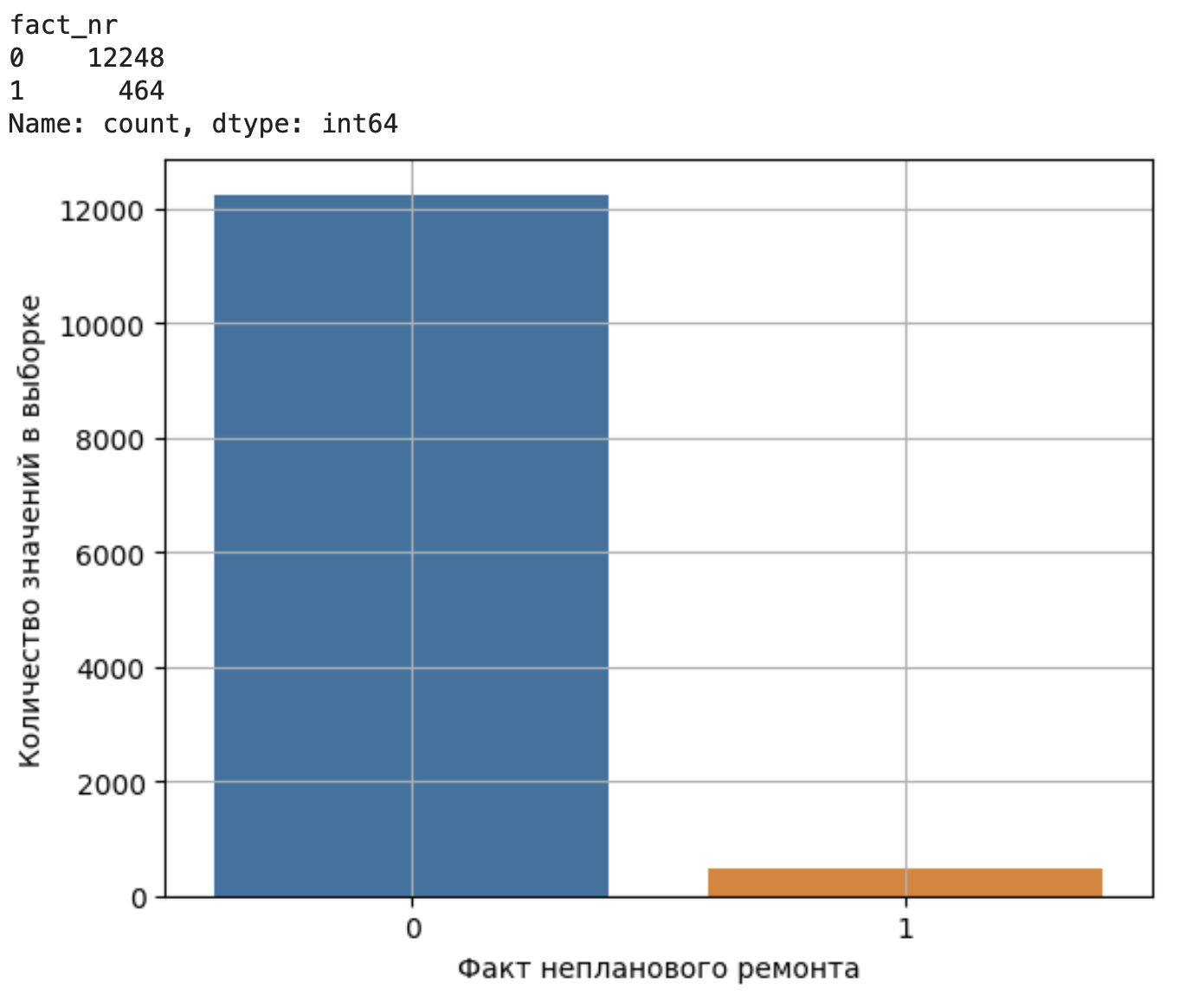


Рисунок 5 Распределение количество меток со значением 0 и 1 в подготовленном наборе данных

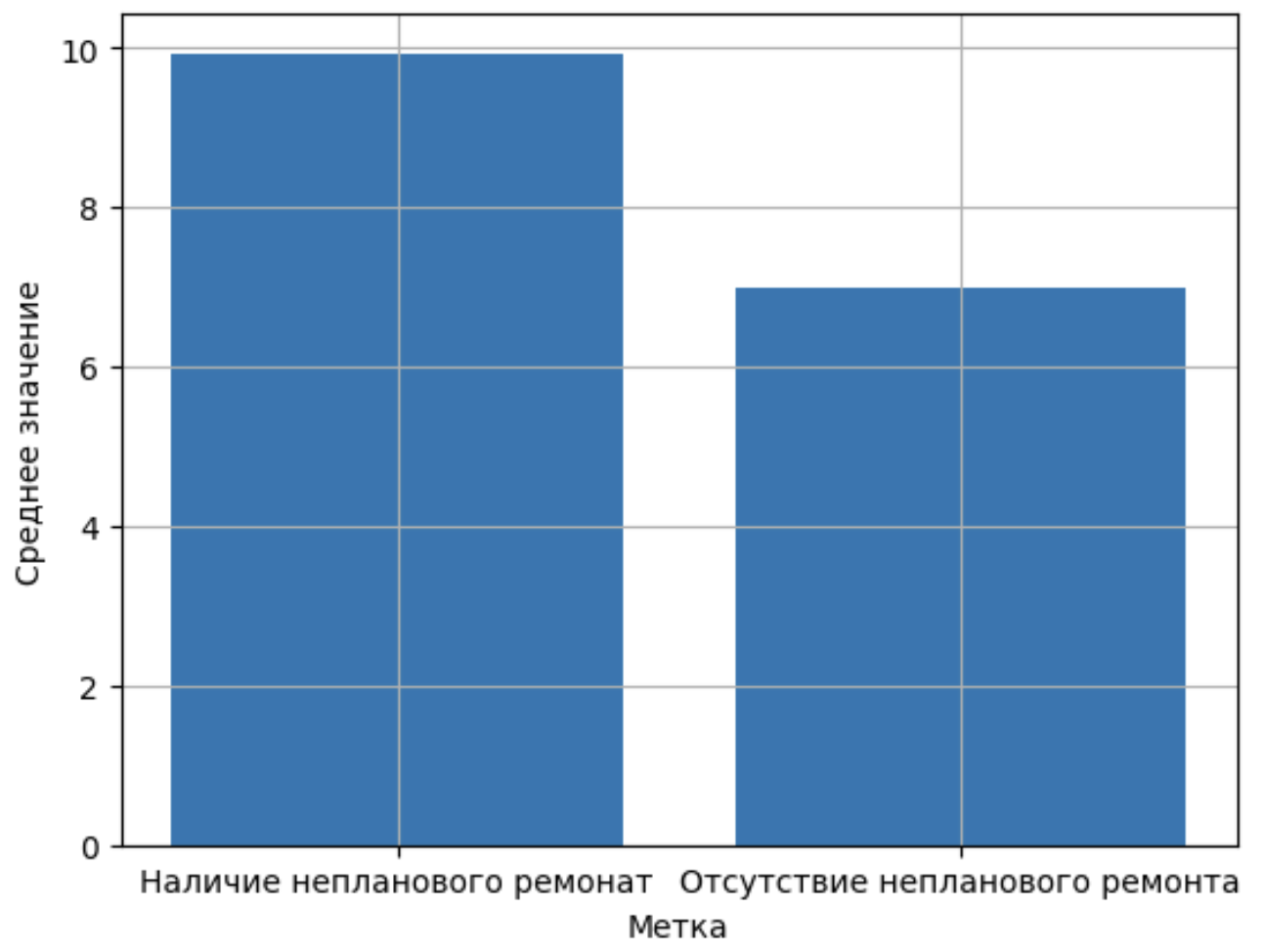


Рисунок 6 Средние значение количества критических сообщений за 7 дней

Анализ других признаков проводился перед построением модели и представлены в приложенном коде.

# Обучение моделей и оценка результатов

## Разделение на обучающую и тестовую выборку

После первичной подготовки данных и формирование итогового набора данных для обучения необходимо его разбить на обучающую и тестовую выборку.

Исходный набор данных разделяется случайным образом на два набора данных: обучающий и тестовый. Доля тестовых данных составляет 30% от общего объема данных.

Затем выбираются признаки и целевая переменная (метка), т.е. обучающие и тестовые данные разделяются на матрицы признаков и векторы целевых переменных. Далее выполняется масштабирование данных с использованием стандартного масштабирования (*StandardScaler*). Наконец, колонки данных переименовываются в соответствии с выбранными признаками. Реализованная функция в ноутбуке подготавливает данные, делая их готовыми для передачи на вход алгоритмам обучения. Он включает в себя такие этапы, как разделение данных на обучающий и тестовый наборы, масштабирование признаков и удаление пропущенных значений.

В результате получается обучающая и тестовая выборка размерами 8898 и 3814 соответственно.

## Результат обучения нескольких моделей

В итоге сформированный в предыдущем разделе набор данных можно использовать для обучения нескольких базовых моделей, чтобы потом выбрать наилучшую. Результаты обучения представлены ниже (Таблица 2).

Таблица 2 Результаты обучения моделей

|  |  |  |
| --- | --- | --- |
| Наименование модели | *F1-score* | *AUC-ROC* |
| Логистическая регрессия | 0.18 | 0.74 |
| Решающие деревья | 0.54 | 0.73 |
| Градиентный бустинг над решающими деревьями | 0.31 | 0.88 |

Качество работы модели оценивалось также с использованием матрицы неточности.

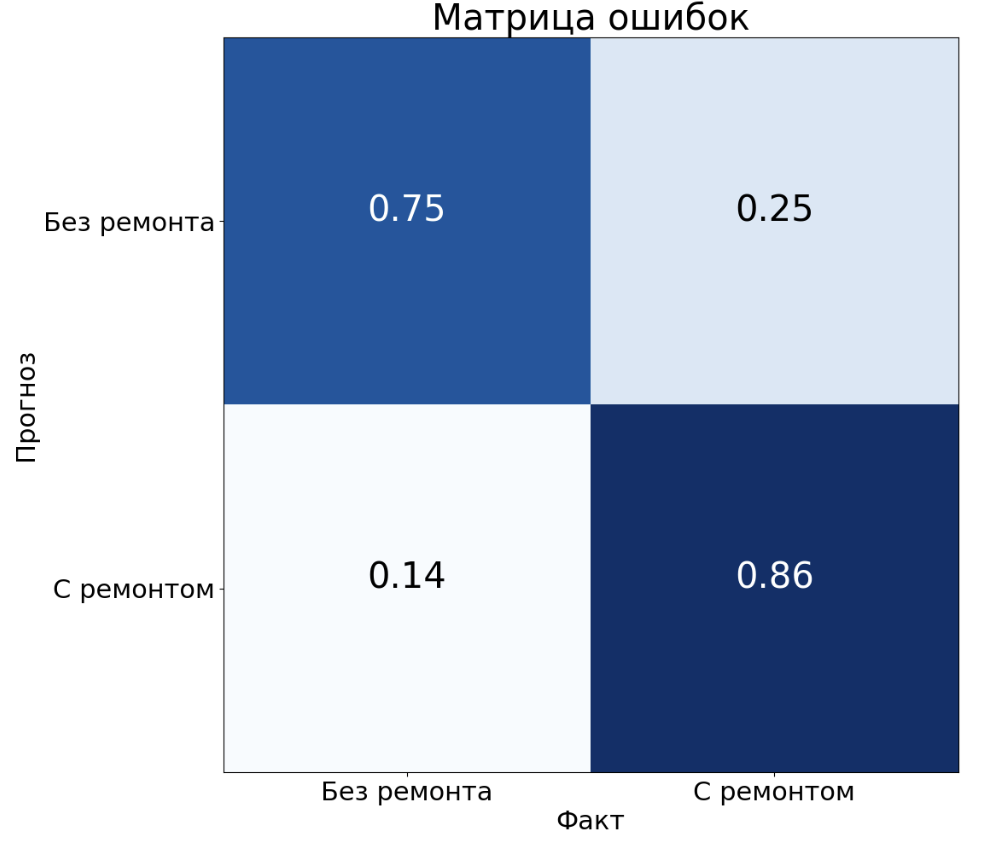


Рисунок 7 Матрица неточностей для алгоритма градиентного бустинга над решающими деревьями

## Сохранение и применение результатов работы моделей

После того, как модель была обучена необходимо сохранить веса модели в файл, чтобы потом ее использовать в информационных и автоматизированных системах. Для этого, в качестве примера, можно использовать следующую команду:

model.save\_model("models/name\_model")

После сохранения модели ее можно использовать для практических целей, внедряя в другие системы. Пример, как с помощью модели можно определить топ 7 локомотивов, вероятность выхода на неплановый ремонт у которых – высокая (Рисунок 8).

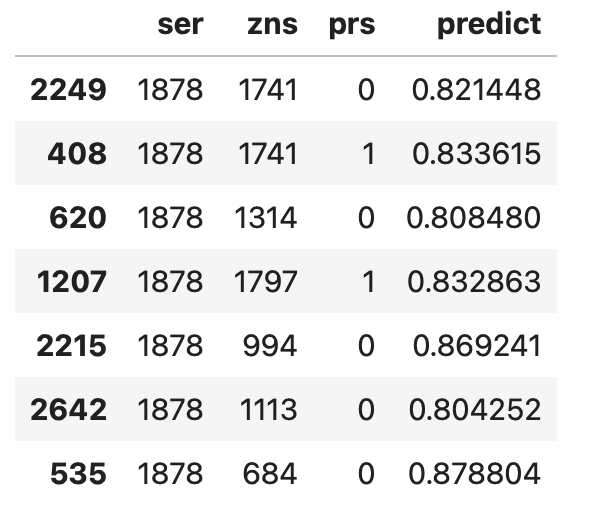


Рисунок 8 Применение модели на тестовых секциях

# Заключение

В рамках проведенного исследования была разработана модель прогнозирования неплановых ремонтов локомотивов серии 2ЭС6 с использованием данных о диагностических сообщениях и неплановых ремонтах. Работа включала в себя несколько этапов:

1. Собраны и предварительно обработаны данные о диагностических сообщениях с локомотивов и о неплановых ремонтах. Этот этап включал в себя чистку данных, заполнение пропущенных значений и преобразование форматов данных для удобства анализа.
2. Проведен анализ полученных данных о работе локомотивов, включающий изучение структуры данных, выявление закономерностей и анализ взаимосвязей между диагностическими сообщениями и неплановыми ремонтами.
3. Разработана модель, которая позволяет рассчитать вероятность возникновения неплановых ремонтов на основе данных о диагностических сообщениях.
4. Модель была протестирована на независимых наборах данных, и ее эффективность была оценена с помощью различных метрик.

В целом, результаты исследования демонстрируют потенциал использования аналитических методов для оптимизации обслуживания локомотивов. Разработанная модель может быть полезна для железнодорожных предприятий в управлении ресурсным обеспечением и планировании технического обслуживания, что в конечном итоге может привести к снижению операционных расходов и повышению эффективности использования ресурсов железнодорожного транспорта.