**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

Федеральное государственное автономное образовательное учреждение высшего образования

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»

РУТ (МИИТ)

Выпускная квалификационная работа  
на тему

**«Оценка влияния времени работы машиниста на возникновение грубых нарушений безопасности движения и построение модели прогнозирования факта совершения нарушения в ближайшие 30 дней»**

ФИО: Кулагин Максим Алексеевич

Группа: ППП-111

Вариант: 1

Содержание

[1. Постановка задачи 3](#_Toc162335309)

[2. Анализ полученных данных о работе машиниста 5](#_Toc162335310)

[2.1. Первичный анализ данных 5](#_Toc162335311)

[2.2. Формирование набора данных для обучения моделей 7](#_Toc162335312)

[2.3. Анализ данных на предмет влияние признаков на целевую метку 8](#_Toc162335313)

[3. Обучение моделей и оценка результатов 12](#_Toc162335314)

[3.1. Разделение на обучающую и тестовую выборку 12](#_Toc162335315)

[3.2. Результат обучения нескольких моделей 12](#_Toc162335316)

[3.3. Сохранение и применение результатов работы моделей 13](#_Toc162335317)

[Заключение 15](#_Toc162335318)

# Постановка задачи

В железнодорожном транспорте обеспечение безопасности движения является приоритетной задачей. Роль машиниста как ключевого элемента в этом процессе нельзя переоценить. Однако, несмотря на все меры предосторожности, нарушения безопасности, происходящие в результате ошибок машинистов, остаются серьезной проблемой. Эти события не только представляют угрозу для пассажиров и грузов, но и влекут за собой финансовые потери для железнодорожных компаний.

В данной работе решается задача прогнозирования нарушений безопасности движения машинистом на железнодорожном транспорте. Сосредоточившись на анализе данных о прошлых инцидентах и факторах, приводящих к ним, следует разработать модель, способную предсказывать вероятность возникновения подобных ситуаций.

Целью данной работы является создание модели прогнозирования нарушений безопасности движения машинистом, которая позволит оперативно выявлять потенциально опасные ситуации и принимать меры для их предотвращения. Это включает в себя анализ данных о прошлых инцидентах, выявление закономерностей и факторов, а также разработку модели, способной предсказывать возможные нарушения безопасности на основе имеющихся данных.

В ходе работы будет проведен анализ данных о прошлых инцидентах и факторах, приводящих к нарушениям безопасности. На основе этих данных разработана модель прогнозирования, которая способна выявлять потенциально опасные ситуации и помогать принимать меры по их предотвращению.

Задача данной работы включают в себя:

1. Сбор и анализ данных о работе машиниста, а также данных о прошлых нарушениях безопасности движения.
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 нарушений безопасности движения на основе имеющихся данных.
4. Валидация и тестирование модели на независимых наборах данных для оценки ее точности.

# Анализ полученных данных о работе машиниста

## 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 машинистов.

В рамках данной квалификационной работы представлен один файл с данными о работе машинистов на каждый день для одной дороги (crews\_trips\_{ID\_ROAD}.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:

data = pd.read\_csv("crews\_trips\_1088292703.csv")

Первые 5 строк изображены ниже (Рисунок 1). Всего записей в crews\_trips\_1088292703.csv – 170598.

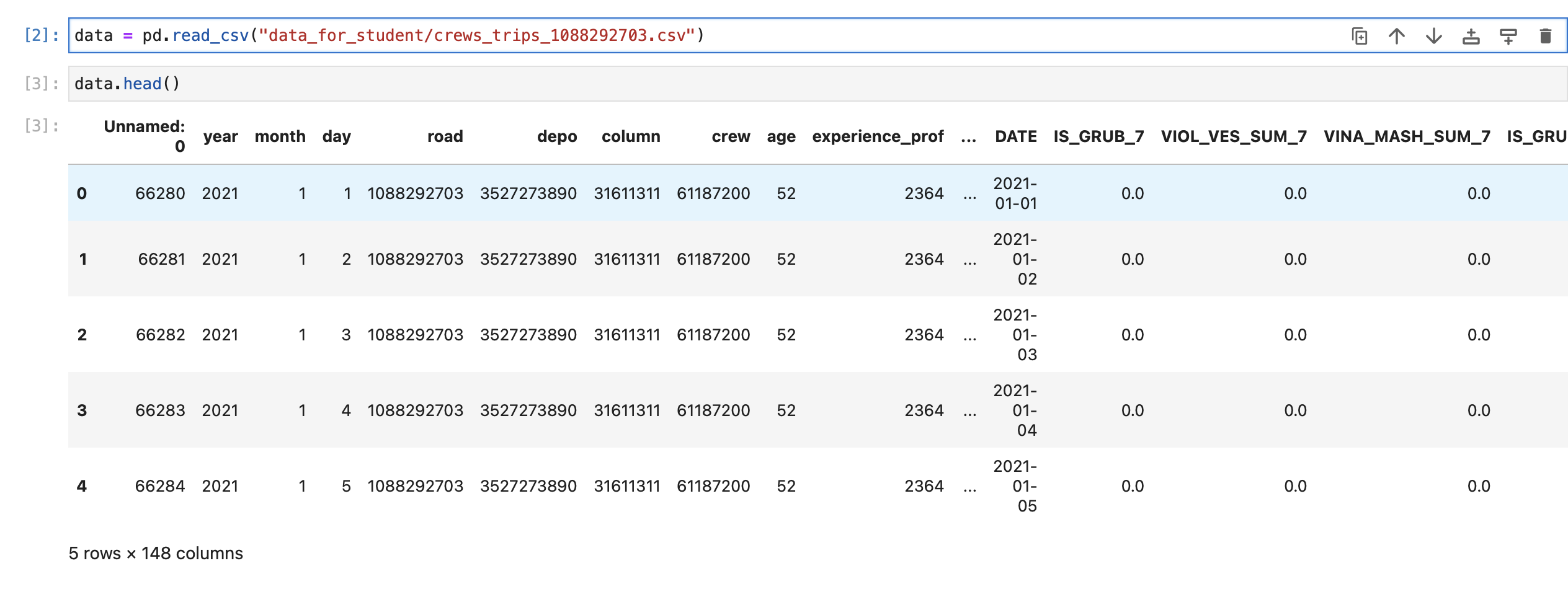


Рисунок 1 Первые строки набора данных ' crews\_trips\_1088292703.csv'

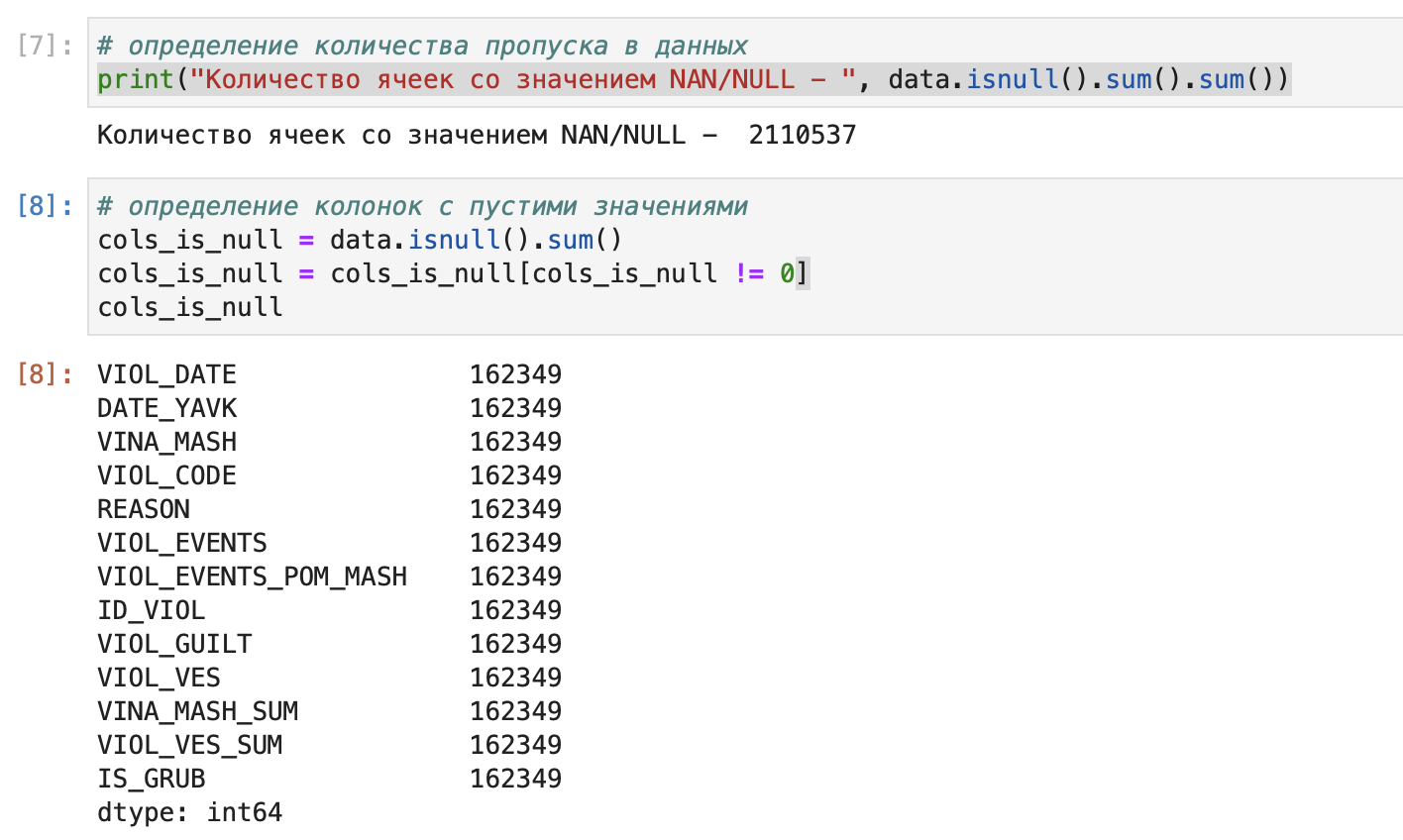
Далее с использованием Pandas проанализируем пропуски в данных и найдем столбцы, в которых есть значения null или nan (Рисунок 2). Это необходимо для того, чтобы подготовить корректные и очищенные данные для обучения модели.

print("Количество ячеек со значением NAN/NULL - ", data.isnull().sum().sum())

cols\_is\_null = data.isnull().sum()

cols\_is\_null = cols\_is\_null[cols\_is\_null != 0]

print(cols\_is\_null)



а)

Рисунок 2 Количество пропусков в данных

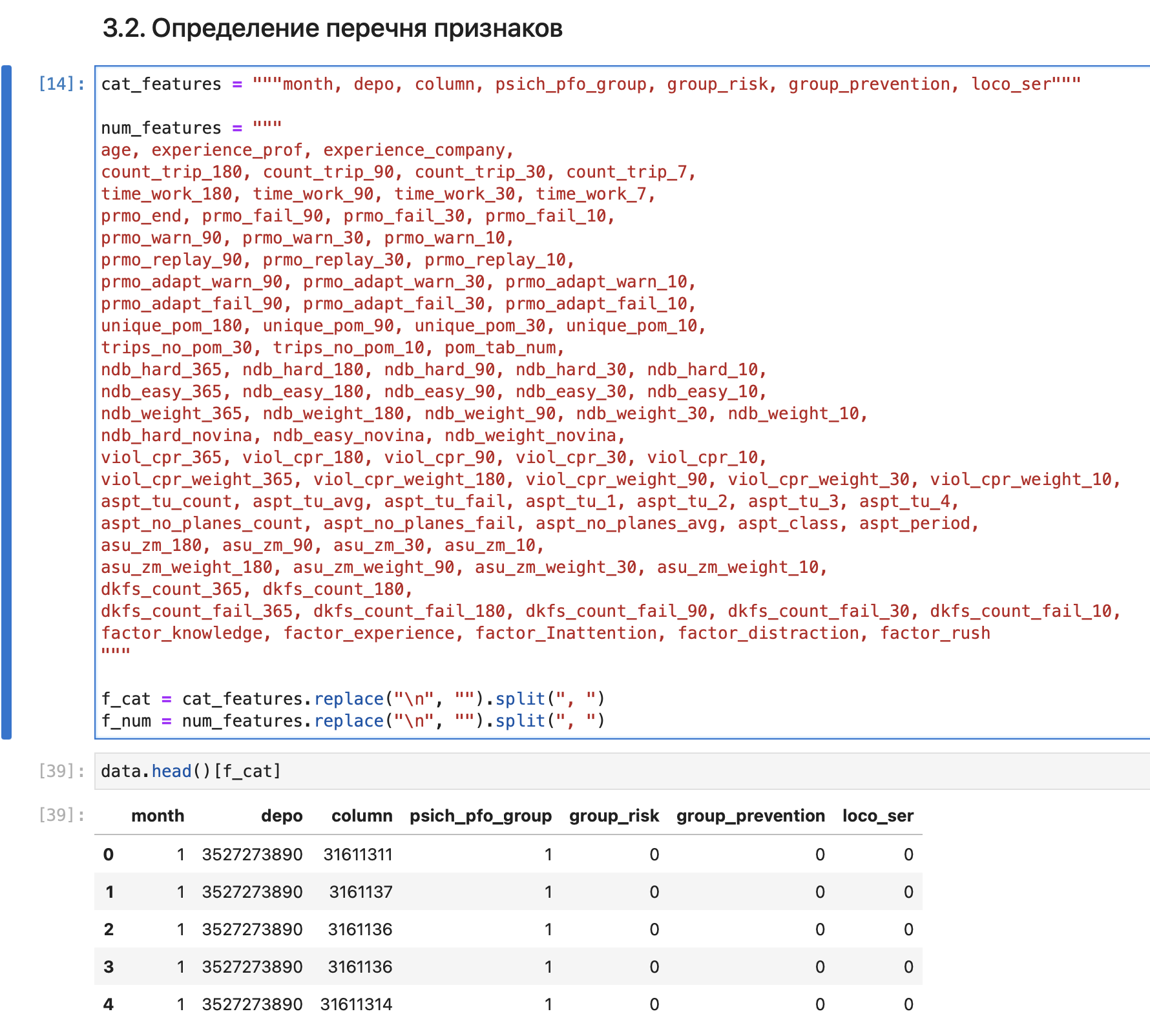
Из первичного анализа наблюдаем колонки «*Unnamed: 0»*, которые не были изначально заявлены в предоставленной спецификации (Таблица 1).

Таблица 1 Несколько примеров описания полей из предоставленных данных

|  |  |
| --- | --- |
| **Наименование столбца** | **Описание** |
| year | Год |
| month | Месяц |
| day | День |
| road | ID дороги |
| depo | ID депо |
| column | ID колонны |
| crew | ID бригады |
| age | Количество лет |
| experience\_prof | Опыт в профессии |
| experience\_company | Опыт в компании |
| talon\_btp | Наличие талона по безопасности |
| talon\_ot | Наличие талона по ОТ |
| is\_trip | Была ли поездка в день измерений |
| trip\_start | Начало поездки |
| trip\_end | Окончание поездки |
| count\_trip\_\* | Количество поездок за 7,30,90,180 дней |
| time\_work\_\* | Время работы за 7,30,90,180 дней |
| prmo\_\* | Агрегированные медицинские данные за 7,30,90,180 дней (fail, warn, replay, adapt\_warn, adapt\_fail, end) |
| unique\_pom\_\* | Количество смен уникальных помощников за 7,30,90,180 дней |
| trips\_no\_pom\_\* | Количество поездок без помощника за 10 и 30 дней |
| pom\_tab\_num | Табельный номер помощника |
| ndb\_\* | Количество нарушений за 10,30,90,180,365 дней для грубых и негрубых нарушений (hard, easy) |

## Формирование набора данных для обучения моделей

При работе с данными были выделены используемые признаки. Признаки были разделены на две категории: числовые и категориальные. А в качестве меток было взято поле, показывающее факт совершения грубого нарушения в ближайшие 30 дней (Рисунок 3).

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание

Рисунок 3 Выбранные признаки и метки для обучения

Первично проведем анализ меток. т.е. фактических совершенных нарушений безопасности. 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.

## Анализ данных на предмет влияние признаков на целевую метку

Рассмотрим детальнее полученные данные. Сначала необходимо распределение данных по годам. В результате анализа было выявлено, что данные по годам распределены равномерно, значит можно попробовать обучать на 2021 и 2022 году и тестировать на 2023 году (Рисунок 4).

Можно в качестве примера анализа признаков привести признак, который показывает количество часов работы за последние 7,30,90,180 дней (Рисунок 5). Из графика видно, что в основном данные о времени распределены корректно, так как укладываются в заданный диапазон 7, 30, 90, 180 дней. Например, для 7 и 30 дней в среднем машинисты в неделю работают около 35–40 часов, а в месяц около 145–160 часов, что соотносится с действительностью.

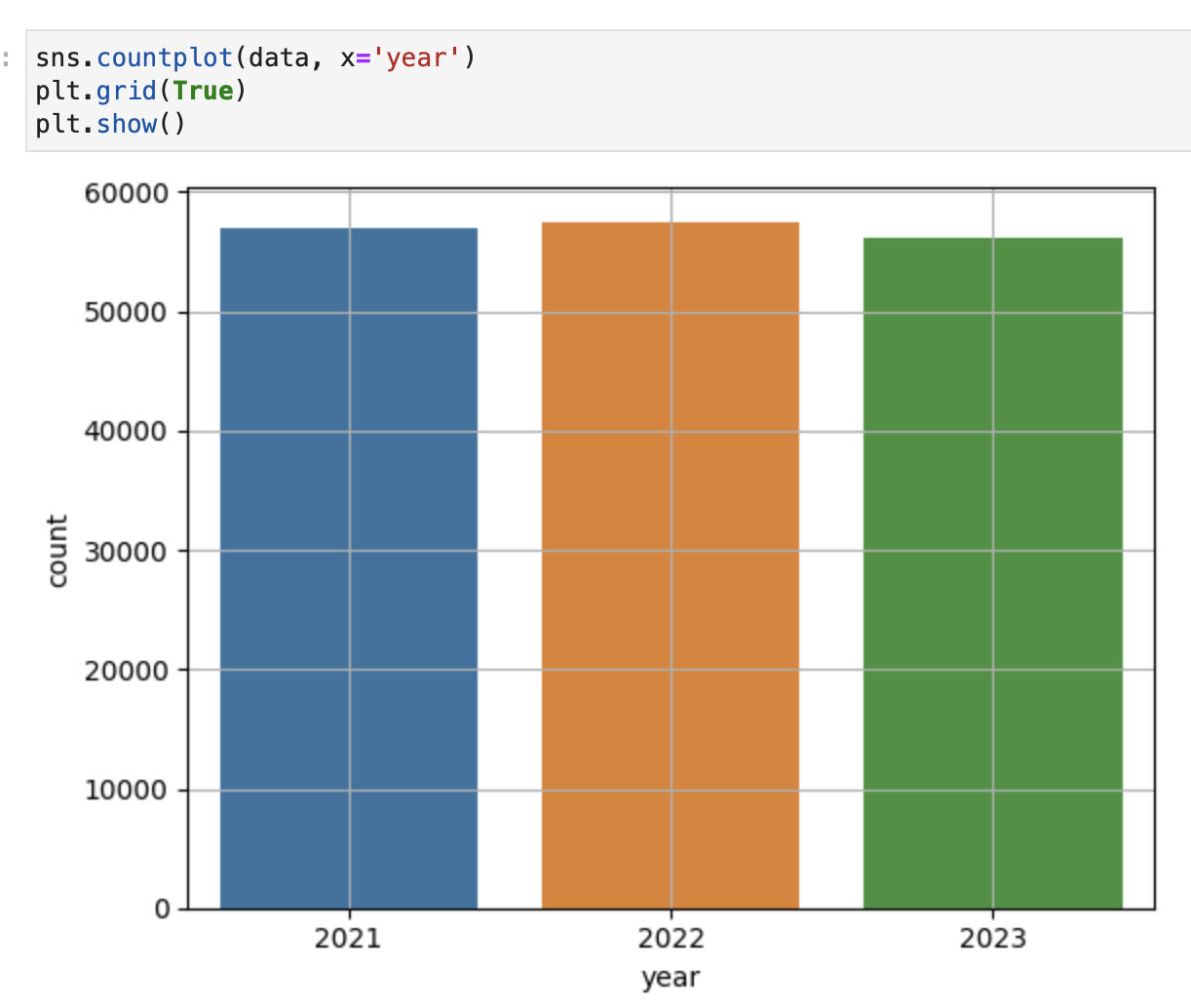


Рисунок 4 Распределение количества данных по годам

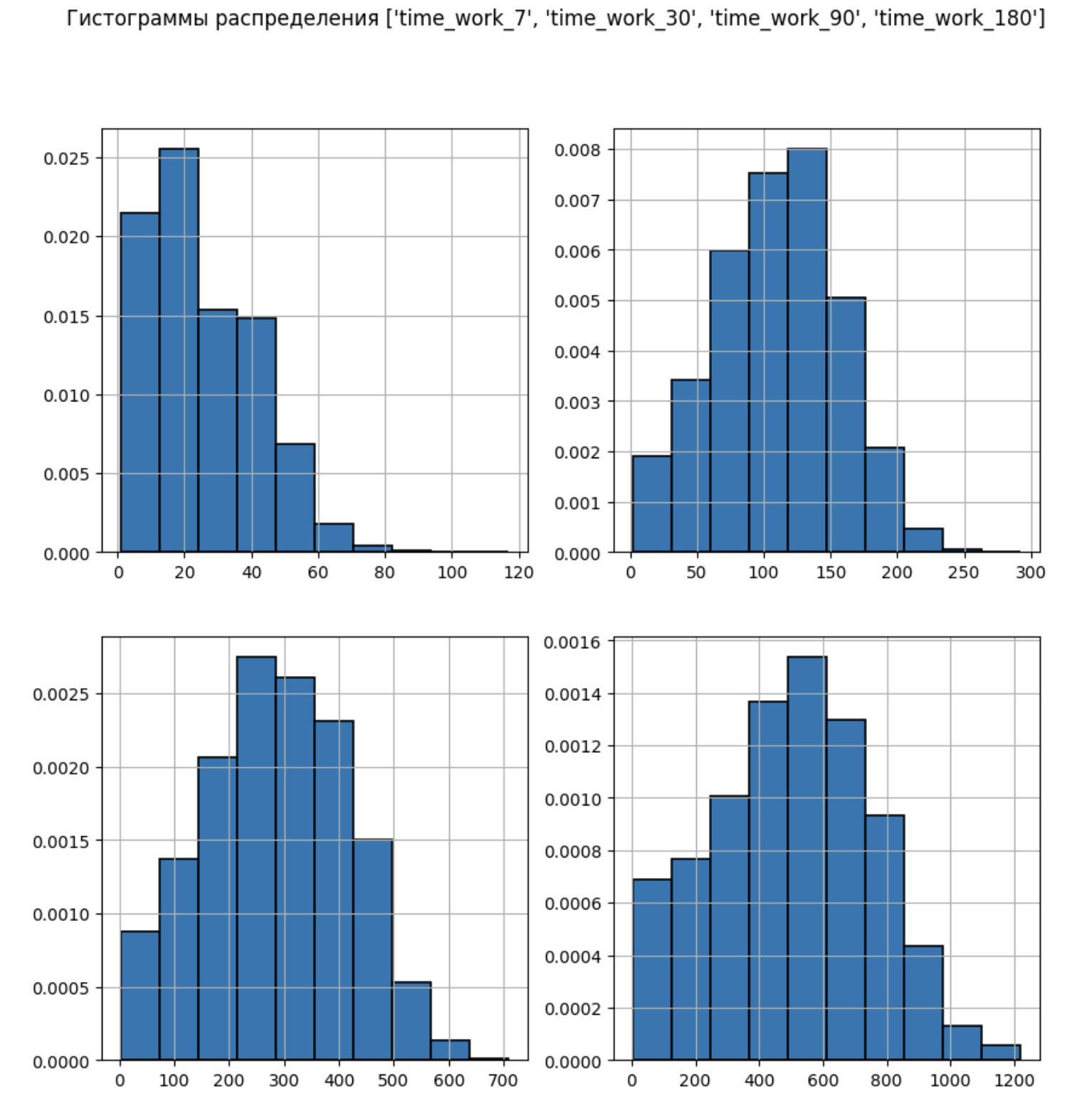
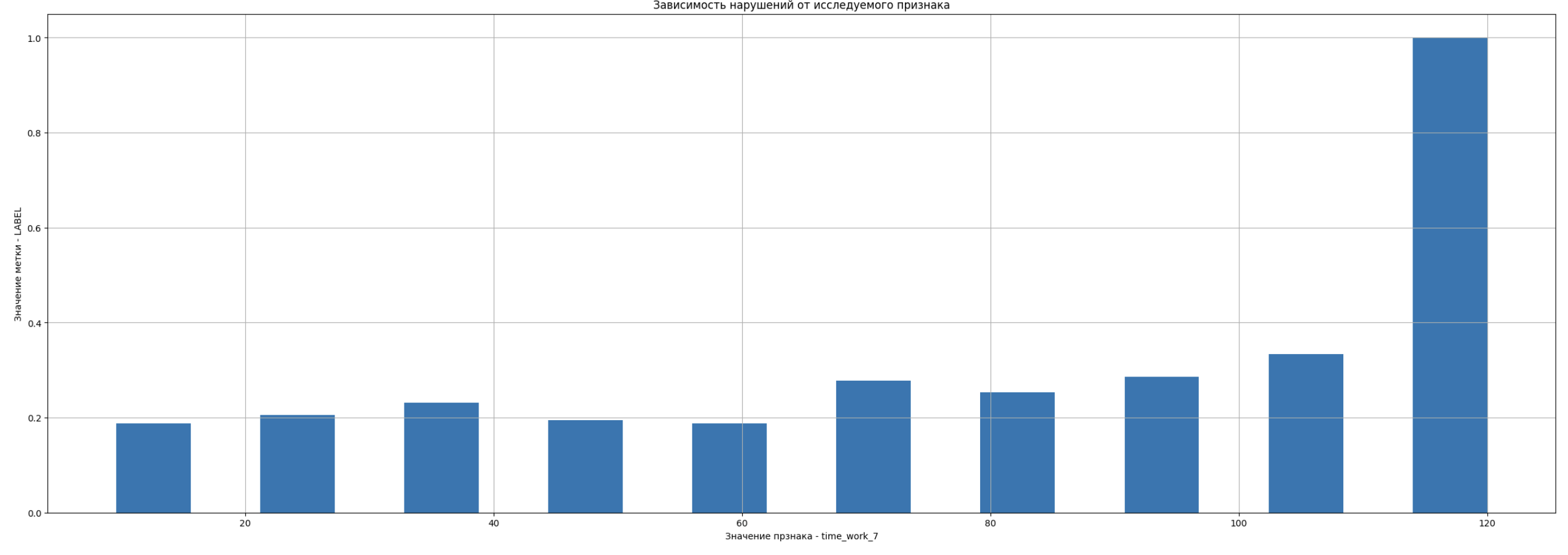


Рисунок 5 Распределение времени за 7,30,90,180 дней

Рассмотрим практическую зависимость между временем работы машиниста за последние N дней (7 и 30 дней) и фактом совершения грубых нарушений, которые совершил машинист.

plot\_helper.plotsDataDenepnds(data[data.time\_work\_7 != 0], "time\_work\_7", "LABEL", num\_interval=10, width=6)

plot\_helper.plotsDataDenepnds(data[data.time\_work\_30 != 0], "time\_work\_30", "LABEL", num\_interval=8, width=10)



а)

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

б)

Рисунок 6 Зависимость количества грубых нарушений от количества часов работы машиниста за: а) 7 дней и б) 30 дней

В результате анализа данных можно сделать вывод, что существует положительная корреляция между количеством часов работы машинистов за последние 7 и 30 дней и средним количеством нарушений безопасности движения. Это означает, что при увеличении времени работы машиниста наблюдается увеличение среднего количества нарушений безопасности.

Наши графики и статистические показатели подтверждают эту тенденцию. Мы видим, что при увеличении числа отработанных часов наблюдается увеличение среднего числа нарушений безопасности движения. Это может быть обусловлено различными факторами, такими как усталость машинистов, снижение концентрации внимания, а также возможное ухудшение физического и психического состояния.

Тем не менее, для более точного понимания влияния времени работы машинистов на нарушения безопасности движения может потребоваться дополнительный анализ других факторов, таких как опыт работы машинистов, условия работы и т.д. Это поможет разработать более эффективные стратегии управления безопасностью и снизить риски на железнодорожном транспорте.

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 следующим образом: 2021 и 2022 год добавляется в обучающую выборку, а 2023 год в тестовую выборку.

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*). 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 114525 и 56073 соответственно.

## Результат обучения нескольких моделей

В итоге сформированный в предыдущем разделе набор данных можно использовать для обучения модели градиентного бустинга над решающими деревьями (Рисунок 7).

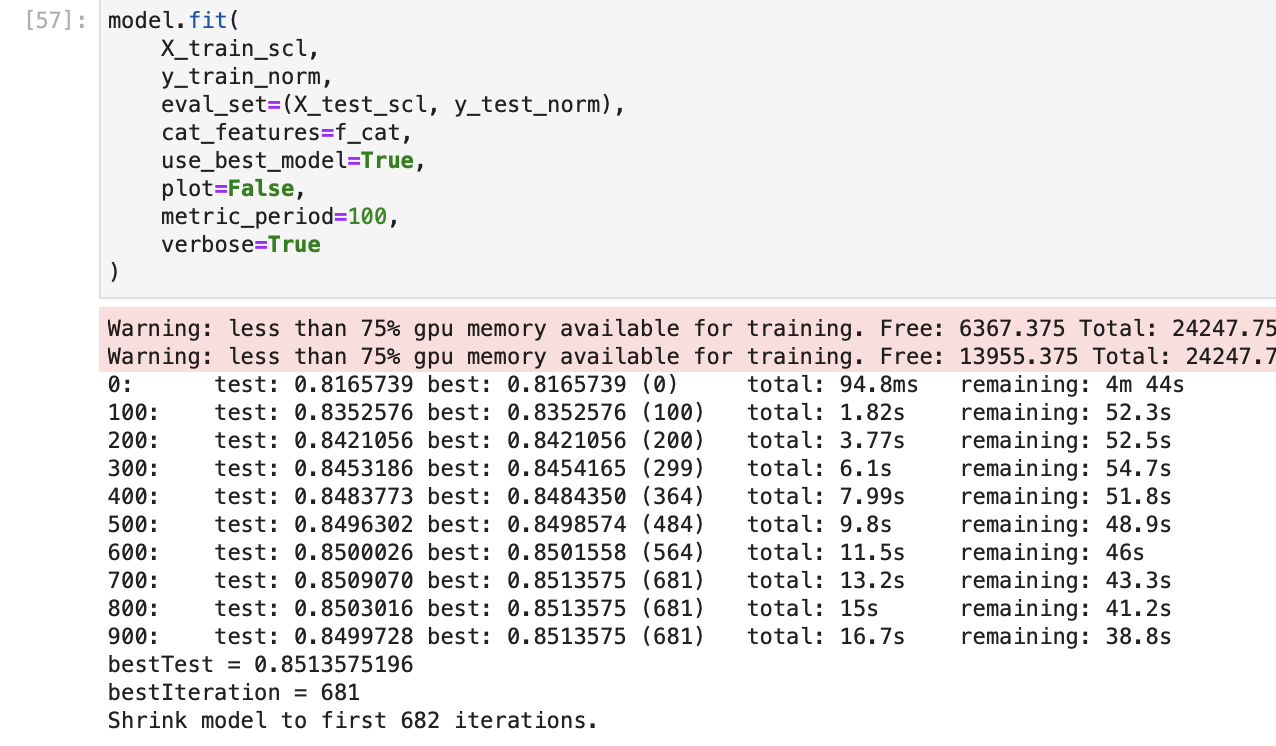


Рисунок 7 Вывод логов обучения модели

Качество работы модели оценивалось с использованием матрицы неточности и площадь под ROC-кривой (Рисунок 8).

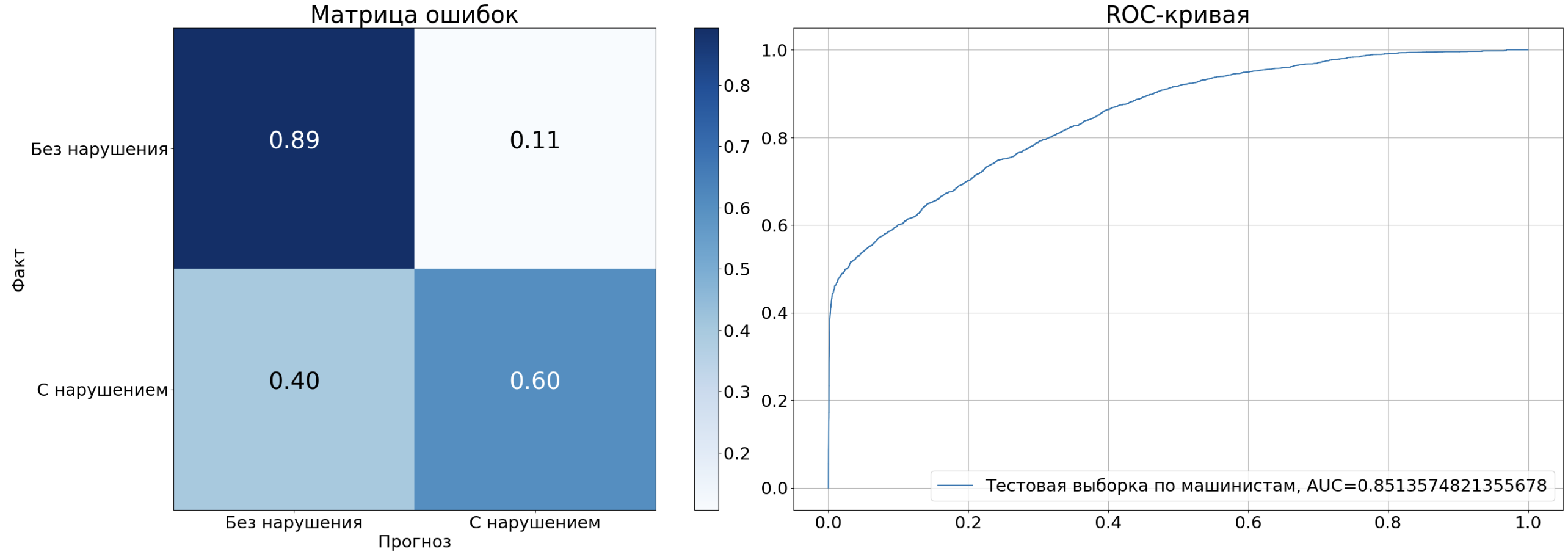


Рисунок 8 Матрица неточностей и ROC-кривая для алгоритма градиентного бустинга над решающими деревьями

## 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 прогнозы по работе машиниста на протяжении всего 2023 года (Рисунок 9).

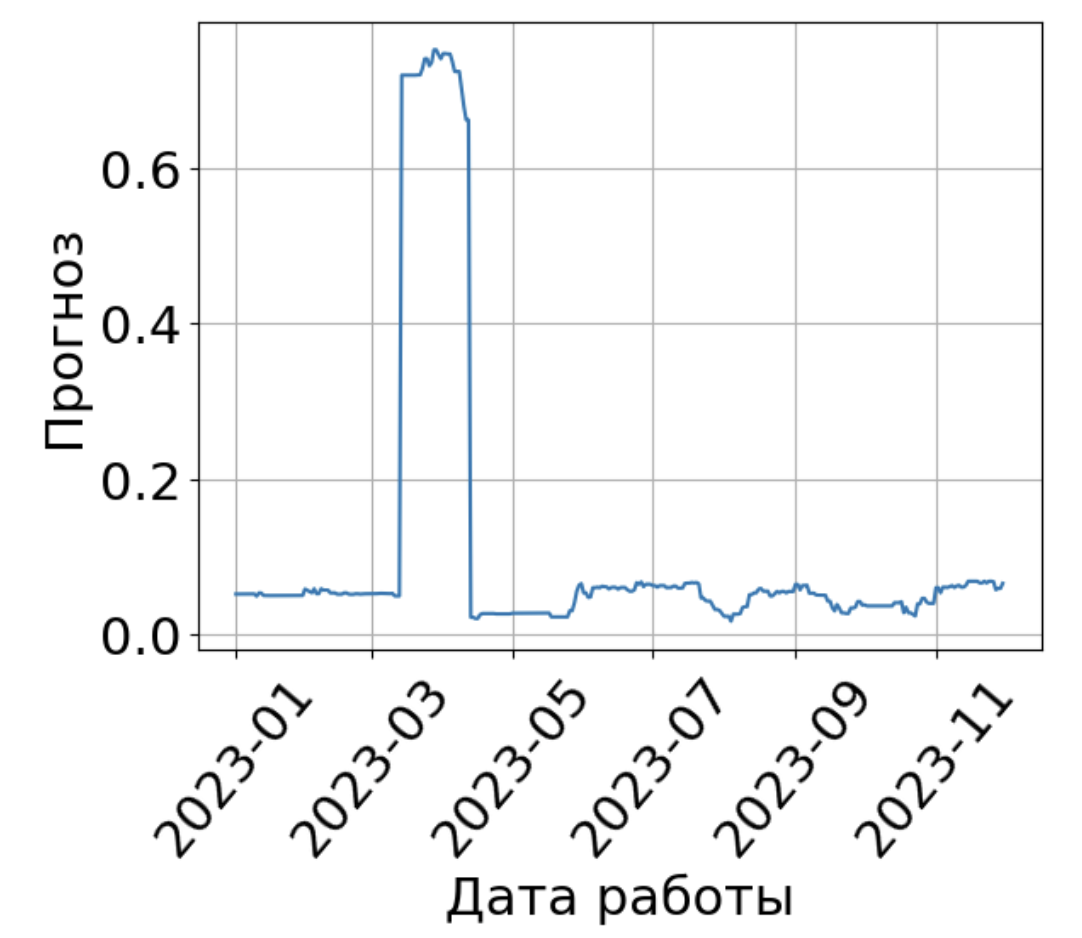


Рисунок 9 Применение модели на тестовом периоде машиниста

# Заключение

В рамках данного исследования разработана модель прогнозирования грубых нарушений машинистами на железнодорожном транспорте. Работа включала следующие этапы:

1. Сбор и предварительная обработка данных о прошлых грубых нарушениях и времени работы машинистов за последние периоды времени. Это включало в себя чистку данных, заполнение пропущенных значений и преобразование форматов для удобства анализа.
2. Анализ полученных данных, включающий изучение структуры информации, выявление закономерностей и определение взаимосвязей между временем работы машинистов и нарушениями безопасности движения.
3. На основе анализа данных разработана модель, способную оценить вероятность возникновения грубых нарушений в ближайшем будущем на основе времени работы машинистов.
4. Модель была протестирована на независимых наборах данных, и ее эффективность была оценена с использованием различных метрик, таких как точность и полнота.

В заключении можно отметить, что результаты исследования демонстрируют потенциал использования аналитических методов для предотвращения грубых нарушений на железнодорожном транспорте. Разработанная модель может стать ценным инструментом для железнодорожных предприятий в управлении безопасностью движения и планировании работы машинистов, что в конечном итоге способствует снижению рисков и повышению эффективности железнодорожного транспорта.