**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**

**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**

**(РУТ (МИИТ)**

Институт/факультет «Академия “Высшая Инженерная Школа”, АВИШ»

Специальность/Направление подготовки Информатика и вычислительная техника

**КУРСОВАЯ РАБОТА**

**по дисциплине:** Машинное обучение и анализ данных.

|  |  |
| --- | --- |
| **на тему:** | "Применение машинного обучения для прогнозирования времени доставки грузов" |

|  |  |
| --- | --- |
| **Студент группы ШАД-211** | **(Рябов С.В.)** |
| **Научный руководитель** | **(Атамасов Н.В.)** |

**Москва 2023 г.**

**Оглавление**

[**Основные термины** 3](#_Toc152703299)

[**Введение** 4](#_Toc152703300)

[**Основные шаги выполнения. EDA** 5](#_Toc152703301)

[**Основные шаги выполнения. Разделение на выборки.** 11](#_Toc152703302)

[**Основные шаги выполнения. Интерпретация лучшей моделей** 17](#_Toc152703303)

[**Заключение** 19](#_Toc152703304)

[**Источники** 20](#_Toc152703305)

[**Приложение** 21](#_Toc152703306)

## **Основные термины**

– EDA (Исследовательский анализ данных) – это процесс изучения и анализа данных для получения информации о их структуре, характеристиках и взаимосвязей между различными переменными;

– Data Profiling – процесс сбора и анализа данных, включающий в себя определение количество и тип данных, обнаружение аномалий и повторяющихся значений, а также оценка качества данных;

– Encoding – преобразование исходных данных в форму, пригодную для использования в моделях машинного обучения;

– Gradient Boosting Regressor – алгоритм машинного обучения для прогнозирования непрерывных значений на основе набора данных;  
– CatBoost Regressor, Random Forest Regressor и Linear Regression – методы машинного обучения и статистического анализа для прогнозирования цен автомобилей на основе различных характеристик;

– Auto\_ML – автоматизированная система для подбора оптимальных алгоритмов и гиперпараметров моделей машинного обучения;

– SHAP – библиотека для визуализации и интерпретации результатов

моделей машинного обучения.

## **Введение**

В данной курсовой работе проводится анализ датасета из варианта №1, содержащего информацию об отправках грузов. Основные задачи: реализация анализа данных (EDA); обучение моделей машинного обучения; сравнение эффективности моделей с использованием библиотеки Auto\_ML и SHAP.

Целью работы является выявление ключевых факторов, влияющих на процесс отправки грузов, а также предсказание эффективности и точности моделей машинного обучения. Также требуется обучить модель машинного обучения, превзойти результаты Auto\_ML на наборе данных о грузах.

Целевая переменная: время доставки в днях.

Характеристики набора данных: в датасете содержится 36 тысяч строк, 48 столбцов.

Оценка моделей: в данном примере, так как решаем задачу регрессии, мы будем использовать R2-квадрат (или коэффициент детерминации) - это мера того, насколько хорошо регрессионная модель соответствует данным. Он указывает, какую часть дисперсии для зависимой переменной объясняет модель. R2-квадрат принимает значения от 0 до 1, где 1 - идеальное соответствие модели данным.

Задачи:

1. Анализ данных.

2. Исследование данных с помощью EDA и графиков.

3. Преобразование данных для обучения моделей машинного обучения.

4. Обучение различных моделей машинного обучения.

5. Сравнение лучшей модели с Auto\_ML и визуализация результатов с использованием SHAP.

## **Основные шаги выполнения. EDA**

В ходе курсовой работы был проанализирован датасет, состоящий из 36 тыс. строк и 48 столбцов, содержащий информацию о перевозках грузов.

В результате были просмотрены данные, которые включает в себя датасет.

Загрузка датасета. Просмотр данных, которые содержатся в столбцах.

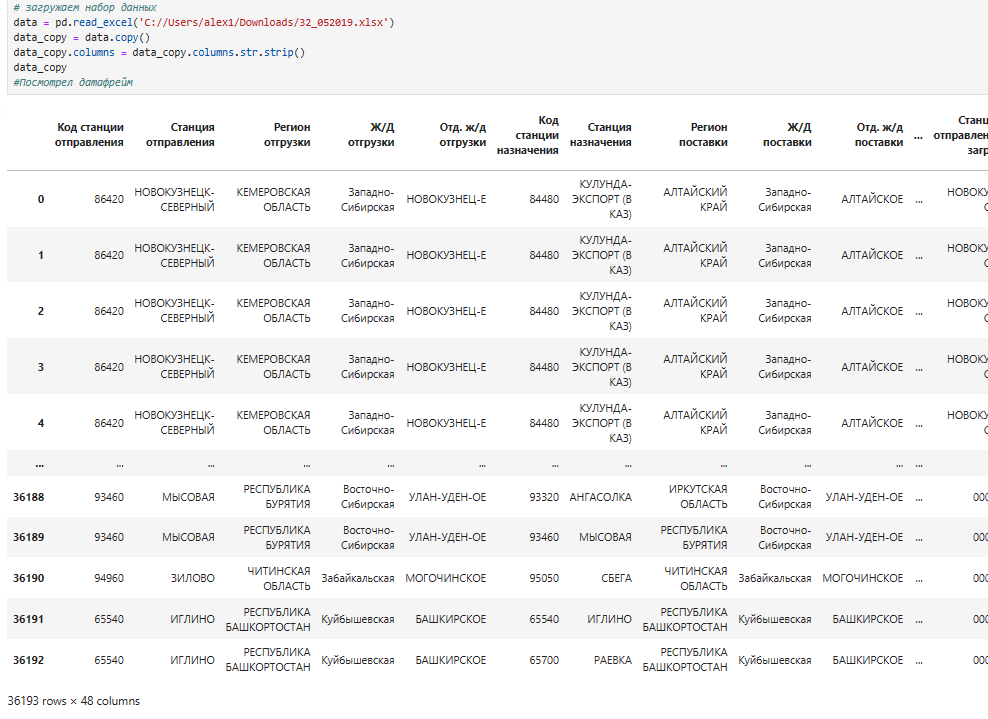


Рис 1. Загрузка данных.

1. Проверка на дубликаты и их удаление.

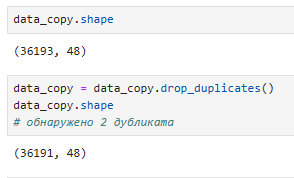


Рис 2. Проверка на дубликаты.

На основе гипотез и предположений были реализованы следующие визуализации:

1. Было интересно узнать, как распределено количество вагонов в зависимости от станции.

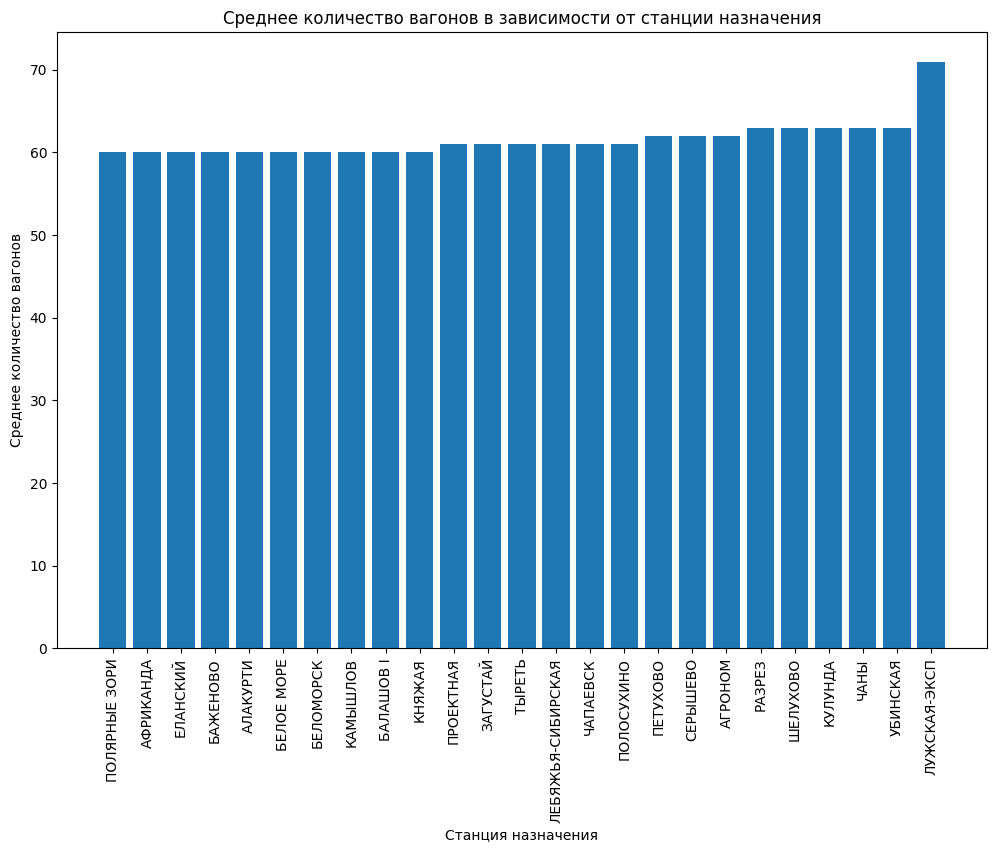


Рис 3. Среднее количество вагонов в зависимости от станции назначения.

1. Был интересен ответ на вопрос «У какой именно станции ж/д отгрузки больше всего поставок?». Впоследствии для нахождения ответа реализована визуализация: Доля различных видов ж/д отгрузки.

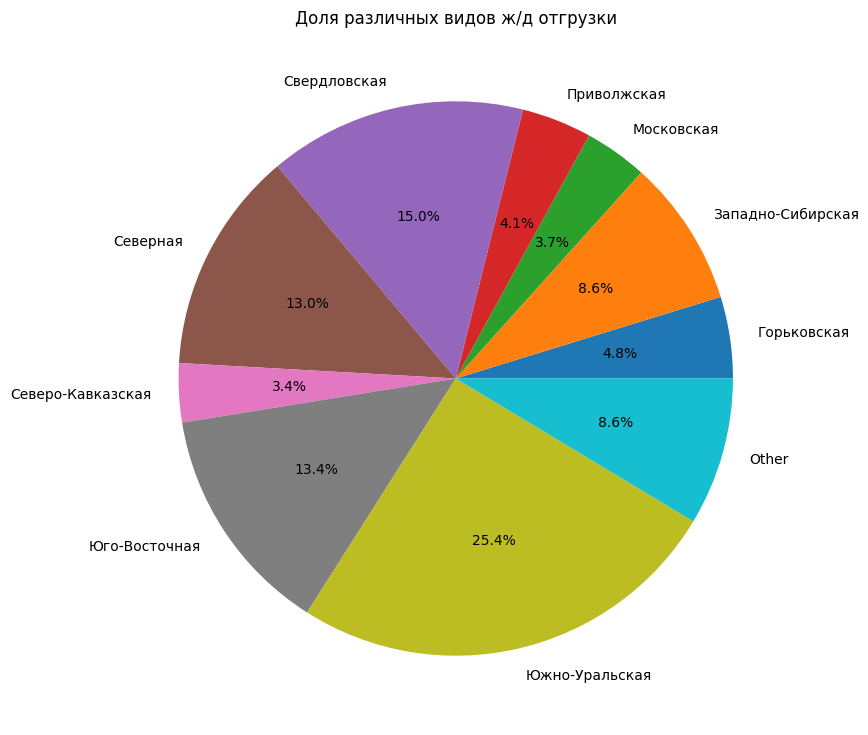


Рис. 4 Доля различных видов ж/д отгрузки.

1. Далее была составлена визуализацию со сравнением количества грузов по странам. По ней можно увидеть, что самое большое кол-во грузов в России.

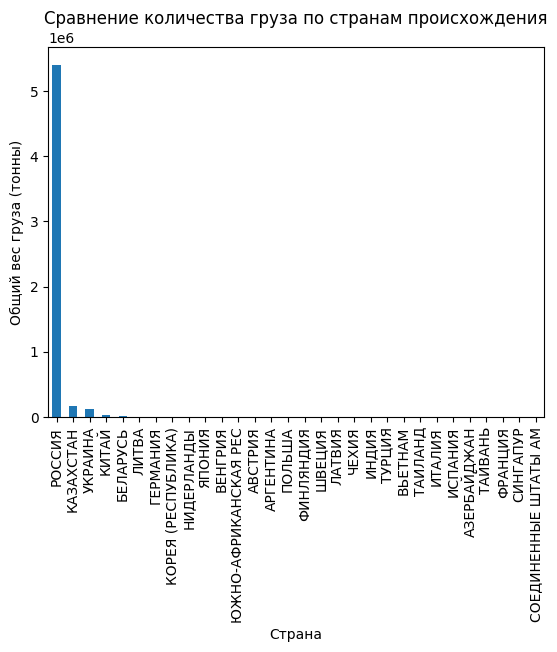


Рис. 5 Сравнение количества грузов по странам.

1. Выдвинута гипотеза о том, что самый легкий груз доставляется дольше. Для её проверки реализована визуализация.

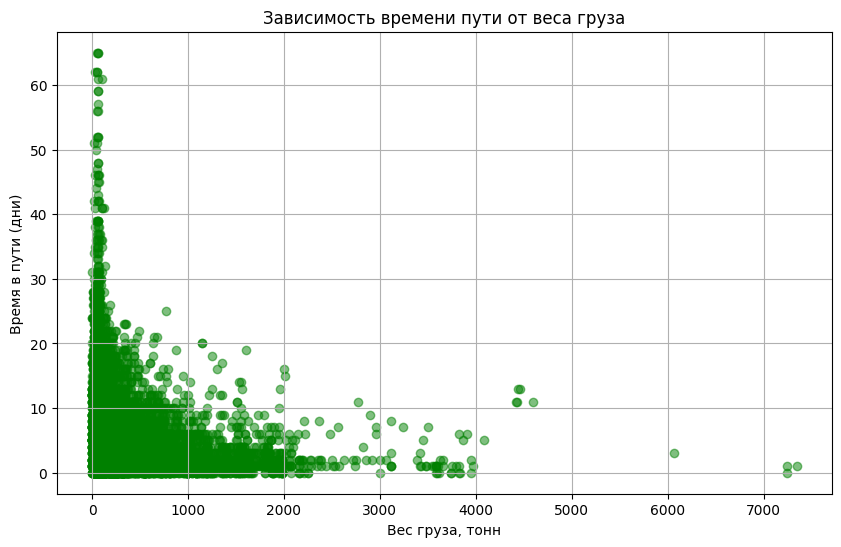


Рис. 6 Зависимость веса груза от времени в пути.

1. Было интересно узнать, как распределено количество грузов по разнице в днях. Реализована визуализация: Анализ разницы в днях между отправкой и прибытием. По графику видно, что большинство грузов были доставлены в границах 10 дней.

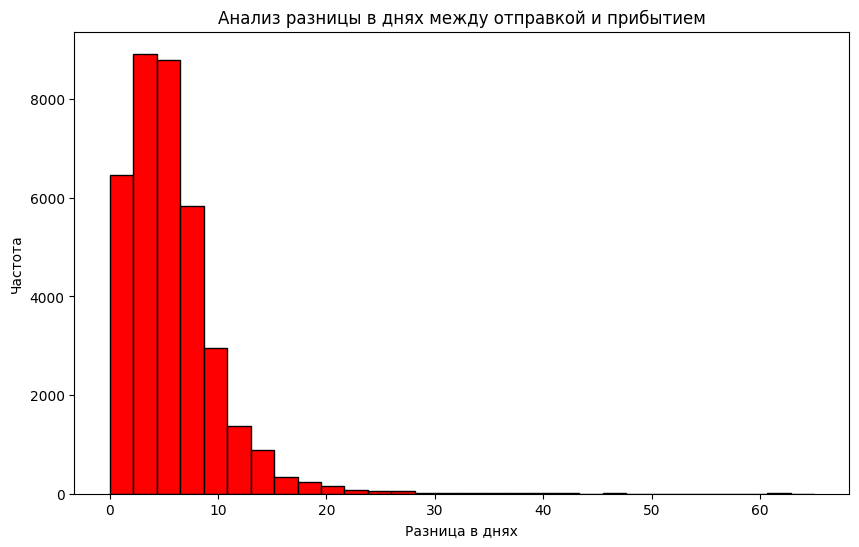


Рис. 7 Анализ разницы в днях между отправкой и прибытием.

**Основные шаги выполнения. Подготовка данных.**

1. Матрица пустых значений.

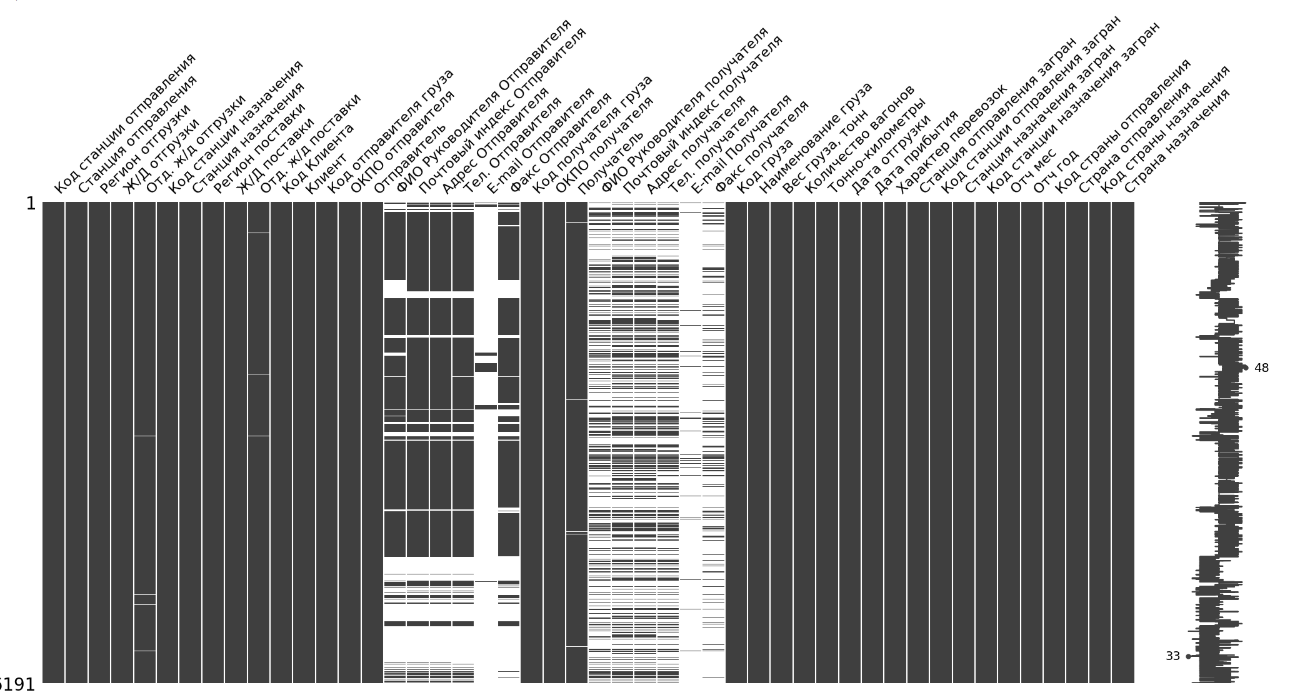


Рис 8. Матрица пустых значений.

1. Вывод статистики по столбцам. Краткая статистика по каждому столбцу, который состоит из числовых значений. По графику можно увидеть максимальное и минимальное значения.



Рис 9. Статистика по столбцам.

* 1. Преобразование столбцов: 'Дата отгрузки', 'Дата прибытия'. Т.к. они неправильно читались.

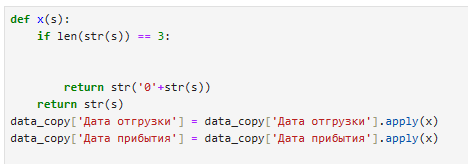


Рис. 10 Преобразование столбцов.

1. Добавление столбцов для удаления.

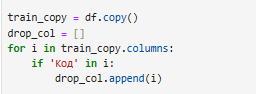


Рис 11. Добавление в список для удаления.

1. Удаление столбцов, влияющих на ответ и удаление на основе логики. Некоторые столбцы не влияют на целевую переменную.

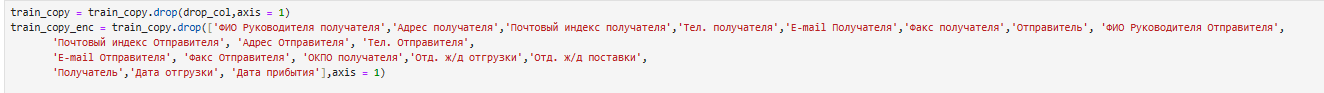


Рис 12. Удаление столбов.

1. Выполнение условия из варианта. Оставил только грузы, в названии которых присутствует ‘Сталь’.

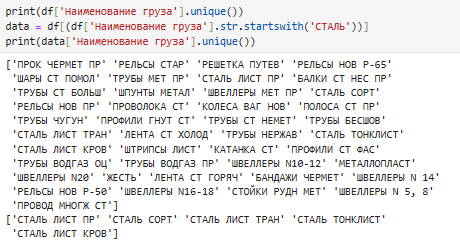


Рис. 13. Выполнение условия.

1. Преобразование столбцов в закодированную форму, потому что модели машинного обучения лучше работают с числовой формой данных.

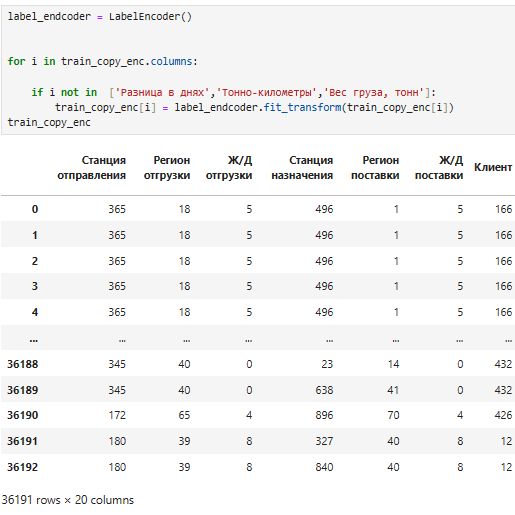


Рис. 14 Кодирование переменных.

## **Основные шаги выполнения. Разделение на выборки.**

Сначала было сделано разделение на две выборки: тренировочную и тестовую. Также здесь выделен таргет, то есть тот столбец, который мы должны предсказывать.

Разделение данных на категориальные признаки и целевую переменную

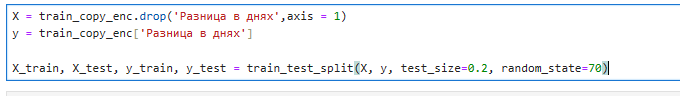


Рис 15. Разделение на выборки.

**Основные шаги выполнения. Построение модели.**

На следующем этапе работы были разработаны различные модели машинного обучения для прогнозирования времени доставки грузов.

После обучения каждой модели были определены оптимальные параметры и их обучение было оценено с использованием R2-квадрата, который является главным параметром для моделей регрессии.   
 Для создания базовой модели был выбран TPOT AutoML, удобная и эффективная библиотека, показывающая хорошие результаты при обучении.

**TPOTRegressor** - это инструмент машинного обучения, который использует генетическое программирование для автоматизации выбора моделей и настройки гиперпараметров. Его преимущества включают автоматизацию выбора моделей, но он требует значительных вычислительных ресурсов и может быть медленным на больших наборах данных.

Trot выбрала лучшей моделью – XGBRegressor.

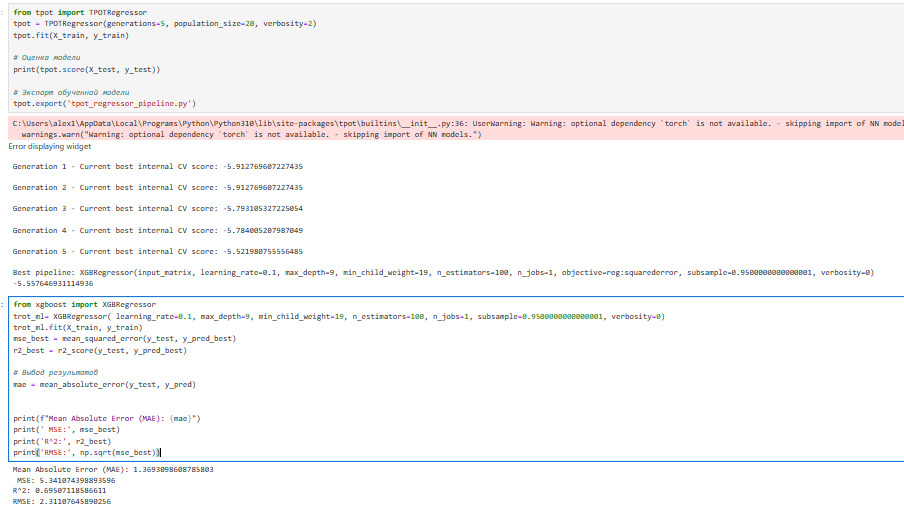


Рис 16. Обучение auto\_ml. Вывод R2 модели trot.

Модели которые использовались в работе:

1. RandomForestRegressor:

- Описание: Этот метод создает ансамбль деревьев решений и комбинирует их прогнозы для уменьшения переобучения.

- Преимущества: Устойчивость к переобучению, хорошая работа с большим числом признаков.

- Недостатки: Возможна медленная работа на больших наборах данных.

2. LinearRegression:

- Описание: Модель, предсказывающая зависимую переменную как линейную комбинацию независимых переменных.

- Преимущества: Простота, интерпретируемость.

- Недостатки: Предполагает линейную зависимость, не учитывает нелинейные отношения.

3. Ridge и Lasso:

- Описание: Регуляризованные линейные регрессионные модели с добавлением штрафов за большие коэффициенты.

- Преимущества: Уменьшение переобучения, устойчивость к мультиколлинеарности.

- Недостатки: Могут занулять коэффициенты, не учитывая важность всех признаков.

4. DecisionTreeRegressor:

- Описание: Модель, предсказывающая целевую переменную путем разбиения данных на подмножества в зависимости от значений признаков.

- Преимущества: Простота визуализации и понимания, подходит для небольших наборов данных.

- Недостатки: Склонность к переобучению, неустойчивость к изменениям в данных. Параметры ко всем перечисленным моделям, подбирались с помощью библиотеки GridSearchCV.

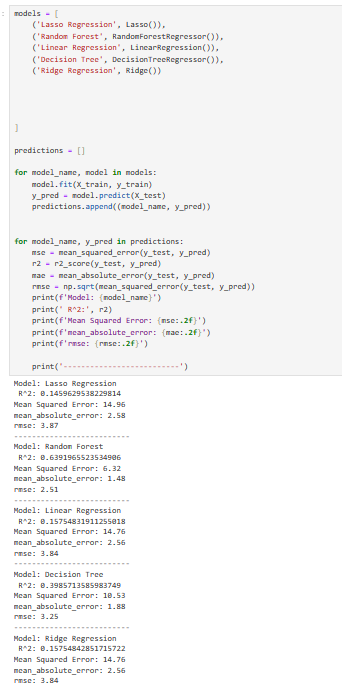


Рис 17. Обучение сразу нескольких моделей с помощью цикла. Вывод результатов.

**Основные шаги выполнения. Лучшая модель.**

GradientBoostingRegressor:

- Алгоритм работы: Это метод градиентного бустинга, который строит прогностическую модель в форме ансамбля слабых прогностических моделей, обычно решающих деревьев.

- Плюсы: Высокая точность, хорошо работает на больших наборах данных, не требует предварительной обработки данных.

- Минусы: Медленное обучение из-за последовательной природы алгоритма, склонность к переобучению.

Параметры для данной модели подбирались с помощью GridSearchCV.

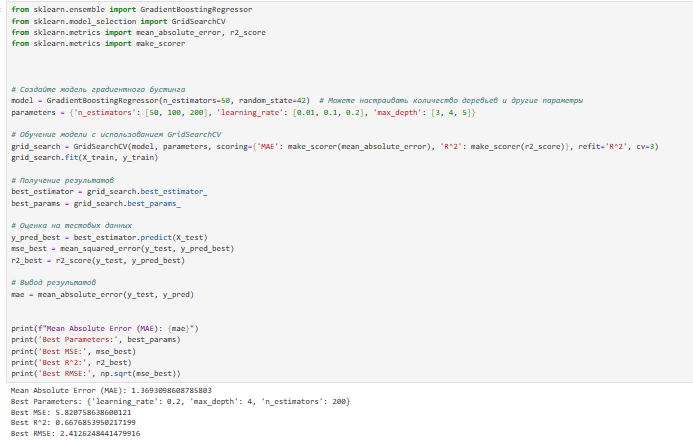


Рис 24. Результаты GradientBoostingRegressor.

CatBoostRegressor:

- Работа: Градиентный бустинг, оптимизированный для работы с категориальными признаками.

- Плюсы: Хорошо обрабатывает категориальные данные, автоматически обрабатывает пропущенные значения.

- Минусы: Может быть медленным на больших наборах данных.

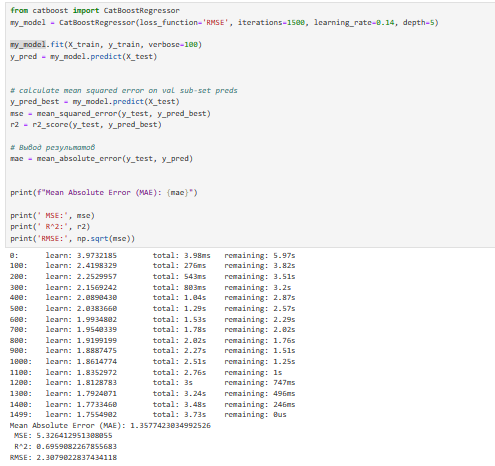


Рис 25. Результаты CatBoostRegressor.

## **Основные шаги выполнения. Интерпретация лучшей моделей**

После завершения обучения каждой модели определяются оптимальные параметры, и оценивается их качество по главному параметру - R^2.

На последнем этапе производится сравнение различных моделей. Лучшие модели для каждого алгоритма выбираются с использованием модели, обученной непосредственно, и модели, обученной автоматически с использованием библиотеки TPOT. Затем эти модели сравниваются с использованием SHAP.

График SHAP для собственной модели.

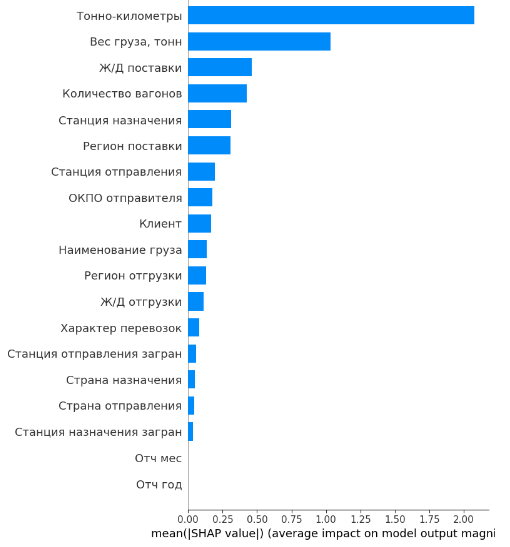


Рис 26. Важность параметров для модели CatBoostRegressor

График SHAP для модели Auto\_ML.

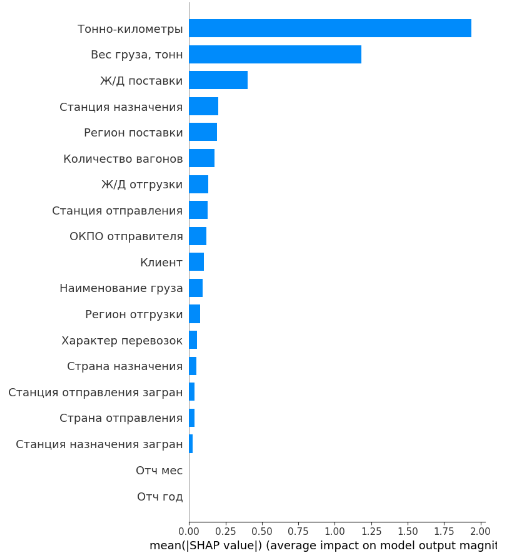


Рис 28. Важность параметров для модели TPOT.

Также построены 2 локальных графика SHAP для самого большого и маленького грузов.

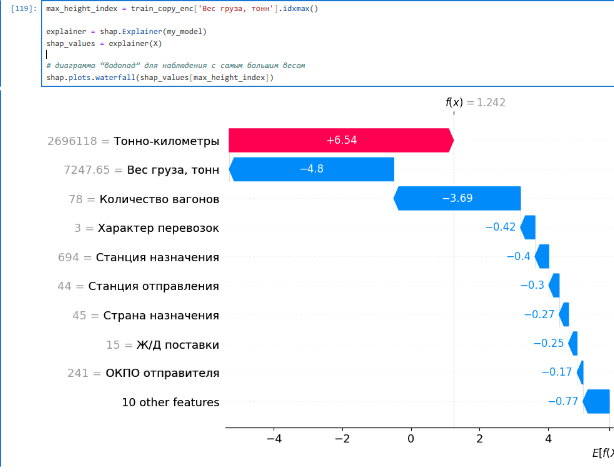


Рис 29. Важность фичей для самого большого груза

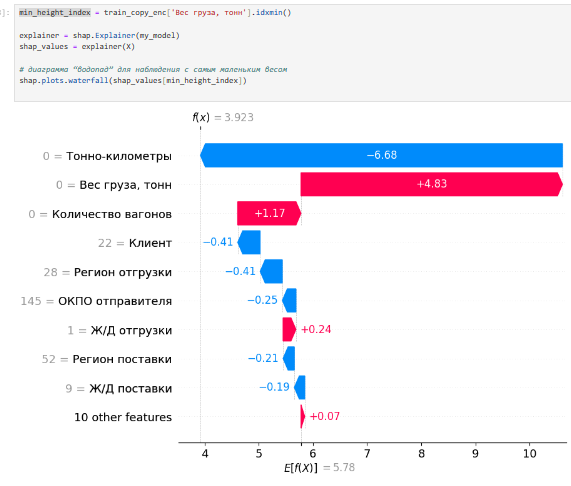


Рис 30. Важность фичей для самого маленького груза.

## **Заключение**

В процессе выполнения курсовой работы были достигнуты все поставленные цели. Проведен анализ датафрейма с применением различных методов, включая исследовательский анализ данных (EDA) и построение корреляционных матриц. Также была осуществлена проверка типов данных и поиск дубликатов записей.

В заключении отмечается, что анализ данных о перевозках позволил провести исследование на высоком уровне. Данные были обработаны и подготовлены для обучения моделей машинного обучения, а также для сравнения с моделью автоматического машинного обучения.

В результате работы были разработаны модели машинного обучения с высокой точностью предсказаний, превосходящей модель Auto\_ML. Это подтверждается визуализацией моделей с использованием SHAP.

В заключении можно сделать вывод, что лучшей моделью оказалась CatBoostRegressor, обогнав XGBRegressor, выбранную моделью Auto\_ML (TPOT). Для сравнения моделей использовалась метрика – R^2

## **Источники**

Данные были взяты из Варианта 1.

Версии библиотек находятся в файле requirements.txt

## **Приложение**

Ссылка на github, с выложенным проектом:

https://github.com/drunkjedii/MLcargoCoursework