**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:** | "Прогнозирование необходимого количества дней для доставки груза" |

|  |  |
| --- | --- |
| **Студент группы ШАД 212** | **Петров А.Г** |
| **Научный руководитель** | **Атамасов Н.В.** |

**Оглавление**

[**Введение** 3](#_Toc153577657)

[**Основные шаги выполнения. EDA** 4](#_Toc153577658)

[**Основные шаги выполнения. Подготовка данных** 10](#_Toc153577659)

[**Основные шаги выполнения. Разделение на выборки** 14](#_Toc153577660)

[**Основные шаги выполнения. Построение модели.** 15](#_Toc153577661)

[**Основные шаги выполнения. Лучшая модель** 17](#_Toc153577662)

[**Основные шаги выполнения. Интерпретация лучшей модели** 20](#_Toc153577663)

[**Выводы** 26](#_Toc153577664)

# **Введение**

В современном мире эффективная логистика и грузоперевозки имеют решающее значение для экономики. В работе создана модель прогнозирования с использованием набора данных «32\_052019.xlsx», чтобы предсказать, сколько дней будет доставляться груз.

**Характеристики набора данных:**

В наборе данных содержится 36193 строк и 48 столбцов.

**Оценка:**

В данной курсовой работе, так как решаем задачу классификации, мы будем использовать Accuracy как метрику оценки.

# **Основные шаги выполнения. EDA**

В данной курсовой работе были данные импортированы из Excel, и была создана их копия.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 1. Загрузка датасета

Просмотрена статистика датасета.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис 2. Знакомство с данными

Просмотр общей информации о наборе данных, выявление основных характеристик.

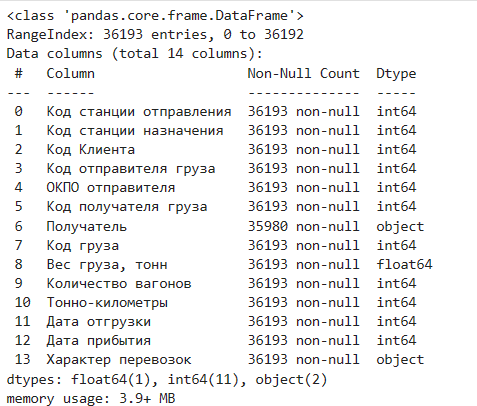


Рис 3. Просмотр информации о датасете

Просмотр данных, типов данных, выявление некорректных типов.

Изображение выглядит как текст, снимок экрана, Параллельный, число

Автоматически созданное описание

Рис 4. Матрица корреляции

Построена корреляционная матрицa, она отображает степень связи между переменными в наборе данных

Изображение выглядит как текст, снимок экрана, дисплей, Прямоугольник

Автоматически созданное описание

Рис 5. График веса груза

Данная гистограмма показывает количество записей в наборе данных по весу.

Изображение выглядит как текст, диаграмма, круг, снимок экрана

Автоматически созданное описание

Рис 6. Круговая диаграмма

Затем построен график зависимости среднего веса груза от его характера перевозок.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рис 7. Средний вес груза по каждой категории

# **Основные шаги выполнения. Подготовка данных**

Уберем ненужные пробелы

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 8. Удаление пробелов справа

Изображение выглядит как текст, снимок экрана, Шрифт, программное обеспечение

Автоматически созданное описание

Рис 9. Оставлены столбцы где наименование груза начинается на СТАЛЬ

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 10. Удаление колонок с пропусками

Решено было удалить пропуски столбцы с пропусками, потому что они никак не будут влиять на модель.

После сделано кодирование нужного столбца для обучения модели.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 11. Кодирование столбца

Сделана функция для правильной работы с датой.

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

Рис 12. Преобразование столбца времени

Создание целевого столбца из разницы даты отправки и даты прибытия.

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Рис 13. Создание целевого столбца

После было удаление столбцов, которые связаны с целевым столбцом.

Изображение выглядит как текст, электроника, снимок экрана, программное обеспечение

Автоматически созданное описание

Рис 15. Удаление не коррелирующих столбцов

Удаление столбцов, которые не повлияют на обучение, потому что корреляция почти отсутствует.

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

Разделили данные на два набора для обучения модели.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 16. Разделение на выборки

# **Основные шаги выполнения. Построение модели.**

Для создания базовой модели была выбрана библиотека automl - tpot.

Результатом работы с этой библиотекой стала модель, которая показала наилучший показатель Accuracy, а также оптимальные параметры для ее работы.

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

Рис 17. Вывод tpot

Для оценки была взята MAE. После получения модели и параметров для нее, она была запущена и построены метрики для её оценки.

Изображение выглядит как текст, снимок экрана, Шрифт, программное обеспечение

Автоматически созданное описание

Рис 19. Вывод метрик для tpot

# **Основные шаги выполнения. Лучшая модель**

Следующим шагом стало построение собственных моделей, которые превзойдут Accuracy у модели automl.

Изначально проверено:

1. KNeighborsRegressor

Метрики очень отличались, поэтому была взята другая модель.

Изображение выглядит как текст, снимок экрана, Шрифт, программное обеспечение

Автоматически созданное описание

Рис 20. KNeighborsRegressor

1. RandomForestRegressor

Данная модель показала неплохие результаты на нашем наборе данных.

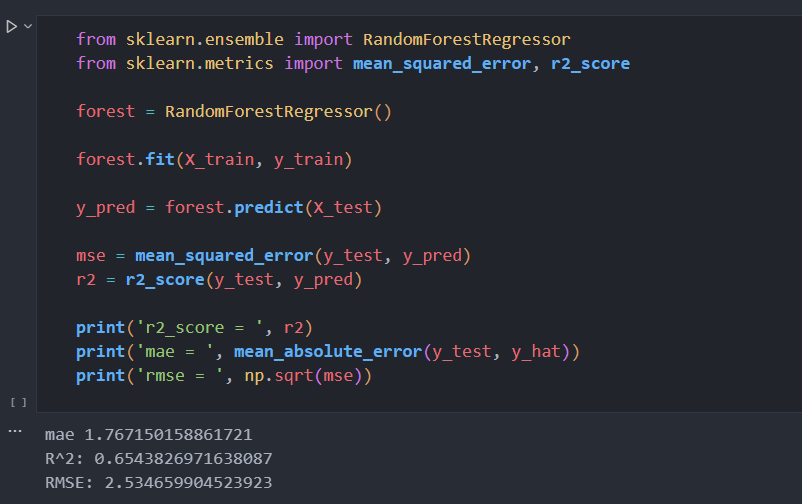


Рис 21. RandomForestRegressor

1. GradientBoostingRegressor

Следующей неплохой моделью стал градиентный бустинг, для которого подбор параметров осуществлялся через GridSearch.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 22 GradientBoostingRegressor

1. CatBoostClassifier

Последней и лучшей стала модель разработанная Яндексом - CatBoost, превосходящая модель из tpot.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 22. CatBoostRegressor

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

Лучшей моделью является CatBoost, а моделью от tpot – GradientBoostingRegressor. Для того, чтобы понять какие фичи сильно влияют на модели, была выполнена настройка SHAP b интерпретация для каждой модели.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис 23. Настройка SHAP

После настройки стала визуализация, сначала был сделан график для моей лучшей модели, а после для модели из tpot.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Рис 24. Код для создания SHAP визуализации

Визуализация для лучшей модели.

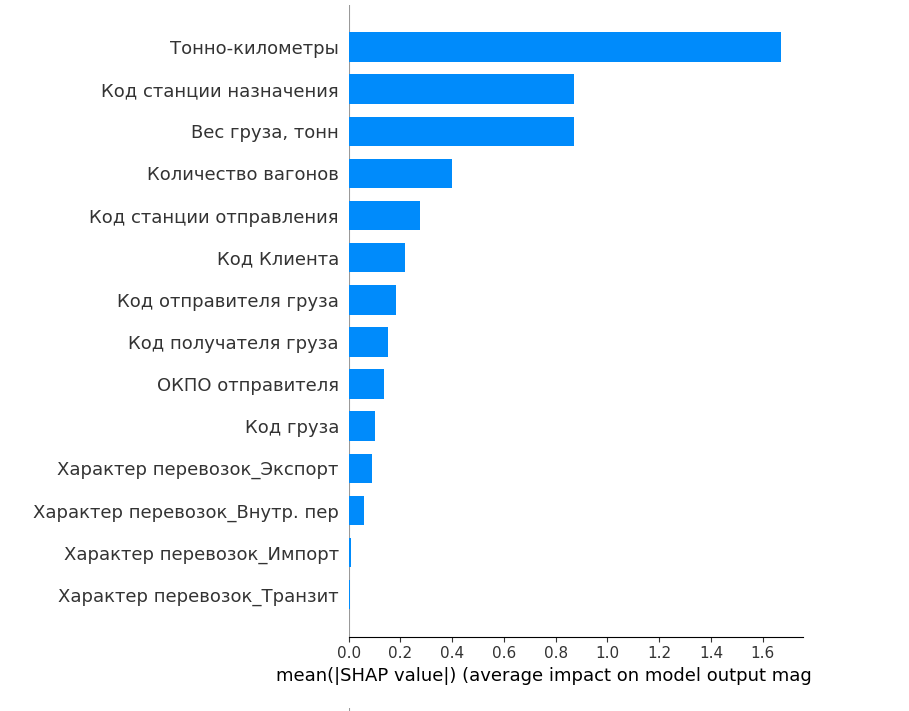


Рис 25. Визуализация лучшей модели

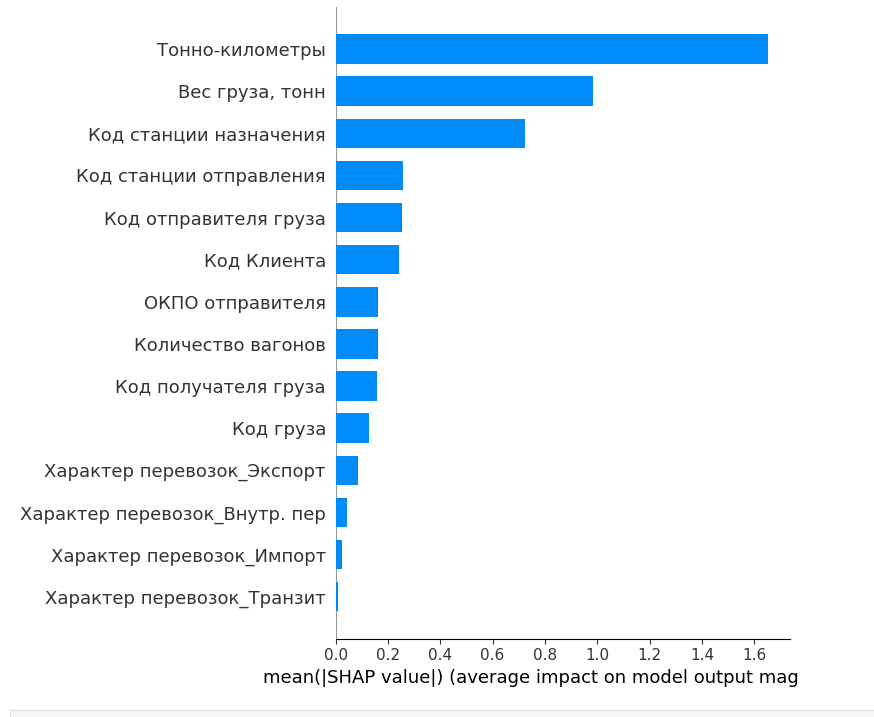


Рис 26. Визуализация для модели из tpot

Потом сделана локальная интерпретация и выбрано две строчки для проверки важности фичей для данных строчек данных.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

Рис 27. Интерпретация модели CatBoost

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

Рис 28. Интерпретация модели CatBoost

После этого была сделана визуализация Partial Dependency для трех самых важных столбцов, на основании визуализации SHAP.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Рис 29. Partial Dependency

# **Выводы**

В заключении можно сказать, что CatBoostClassifier проявил себя эффективнее и лучше по сравнению с базовой моделью.

Была использована визуализация SHAP для лучшего понимания влияния различных признаков на само обучение модели. Было подробно изучено воздействие конкретных наблюдений на предсказания модели.

Для дальнейшего повышения производительности модели рекомендуется более тщательно изучить различные параметры и оценить их влияние на конечные результаты.