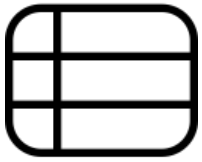
Прогнозирование маршрутов передвижения пассажиров Московского метрополитена

на основании данных о валидации транспортных карт

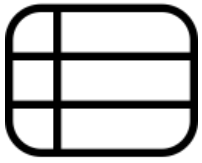
Старт обучения: август 2021
Дата защиты: 10 июня 2023

Описание задачи

- **время** повторной валидации (регрессия) → **R2** метрика
- **станция** повторной валидации (классификация на 276 классов) → **Recall** метрика



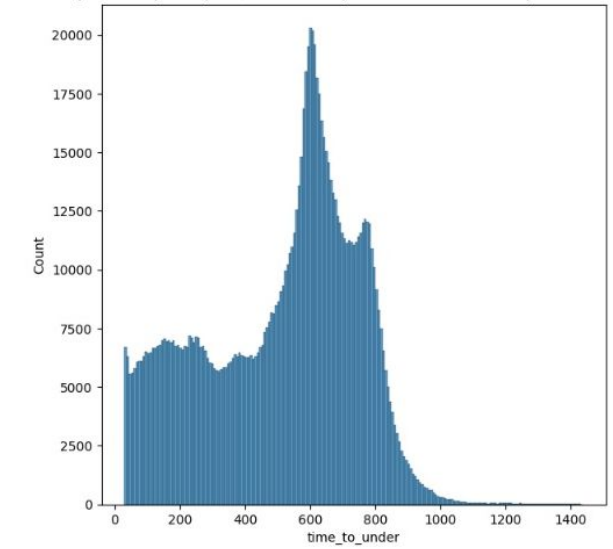
- основной файл (размер 1 091 021 x 12)



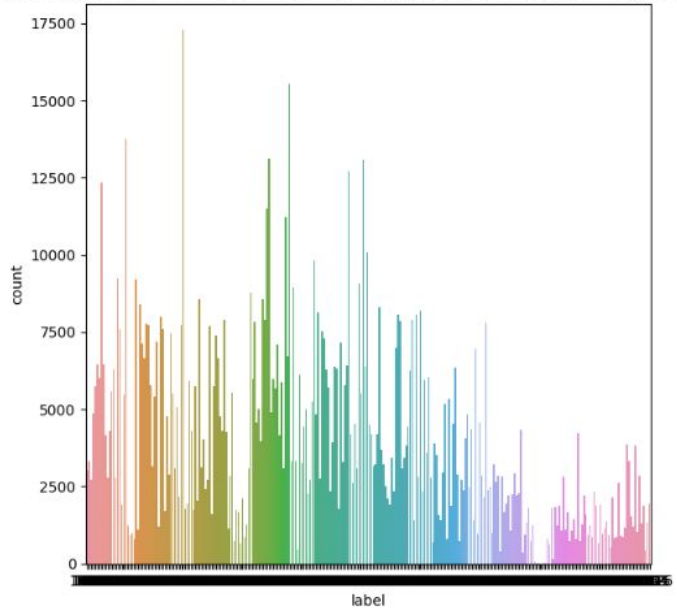
- дополнительный файл (размер 2 991 571 x 16)

Анализ таргетов

Гистограмма распределения времени до повторной валидации



Гистограмма распределения станций повторной валидации



Время между валидациями

- несбалансирован
- слабая корреляция с признаками
+ 5-11 часов между валидациями
+ точное число в минутах

Станция второй валидации

- несбалансирован
- многоклассовая классификация на 276 классов
- плохая связь с признаками
+ конкретная информация о месте назначения

Моделирование

DecisionTreeRegressor
+
DecisionTreeClassifier
+
DataFrame (label
encoder)

Поиск модели

- **AutoML:**
  - AutoKeras
  - Auto-sklearn
  - Optuna + LGBM
- **Decision Tree (sklearn)** ★

Подбор формата датасета

- **DataFrame:** ★
  - Преобразованные данные (label encoder)
  - Непреобразованные данные
- **Numpy:**
  - Преобразованные данные (normalize, ohe)

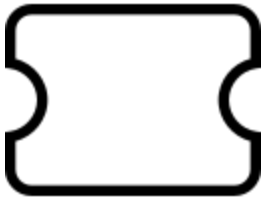
Генерация новых признаков

Источники для создания признаков:

- Фичи по ticket_id (каждый билет ездил ~ 3.25 раз.)
- Между валидациями 300 - 660 минут
- Типы билетов описывают категорию пассажира (резидент, студент, сотрудник МВД, персонал т.п.)
- Дата валидации
- Номер станции (label)
- Линия станции
- Дополнительный датасет

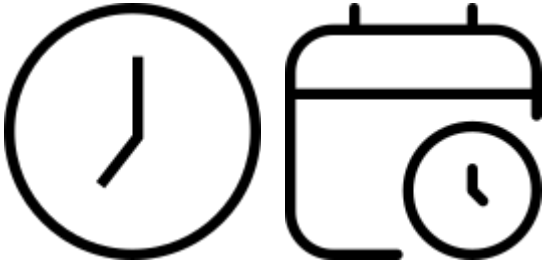
Генерация новых признаков

будний или выходной



фичи по id билета

линия станции назначения



час первой валидации и час второй валидации

средняя и максимальная продолжительность маршрута от станции начала

Частное от номеров станций (конец/начало)

категории пассажиров

‘class’ станции назначения

удаление избыточных фичей

Результаты

Модели регрессии:

1. Autokeras, r2 = 0.83

2. DecisionTreeRegressor, r2 = 0.99

Модели классификации:

1. Autokeras, recall < 0.1

2. LGBM + Optuna, recall < 0.1

3. Auto-sklean, recall < 0.1

4. DecisionTreeClassifier, recall = 0.99

Выводы

01 Генерация признаков, коррелирующих с таргетом

02 Всегда чистить данные (label encoder, ohe, normalize)

03 Использовать классические алгоритмы ML

04 С нейронными сетями использовать callbacks

05 Визуализация данных (обязательно хитмап, графики целевых переменных)

Ресурсы и инструменты

01 Collab (9.99$)

02 Библиотеки:

- pandas
- numpy
- matplotlib
- seaborn
- autokeras
- optuna
- lightgbm
- auto-sklearn
- sklean
- joblib
- keras

03 Материалы:

1. Документация библиотек
2. Общедоступные источники

Заключение

01 Ещё модели для стабильности:

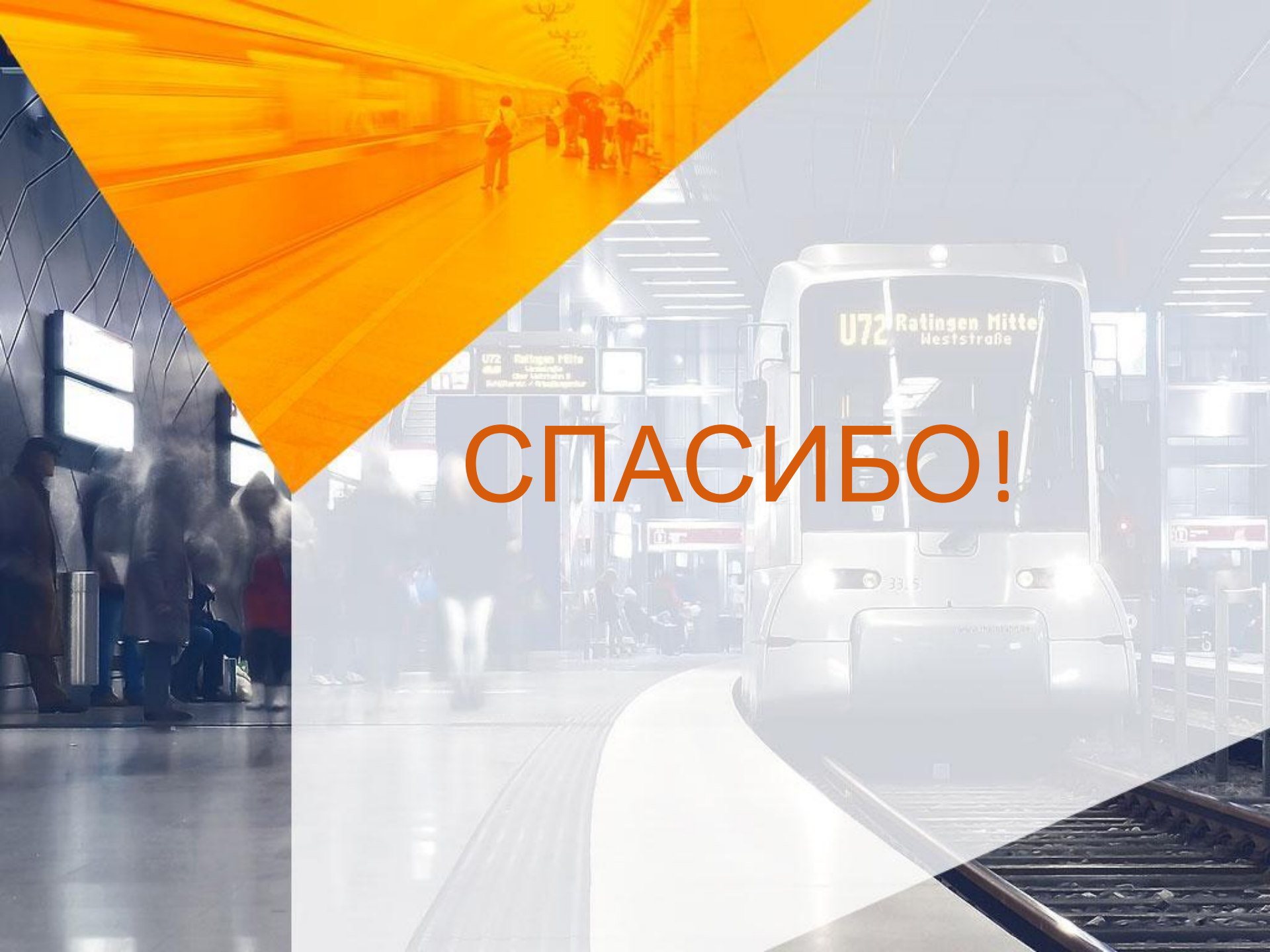
- LinearRegression
- RandomForestRegressor
- ExtraTreesRegressor
- GradientBoostingRegressor
- Ridge
- Lasso
- XGboost

02 Библиотеки:

- xgboost
- sklean

03 Дополнить датасет из [портала открытых данных]

04 Рассмотреть вариант решения через TimeSeries методы



СПАСИБО!

Контактные данные:

- Mobile: +7 903 951 8601
- Telegram: to_Yankovskaya
- E-mail.: airagent@mail.ru