ЦИФРОВОЙ ПРОРЫВ 2022: АМУРСКАЯ ОБЛАСТЬ

”ПРОГНОЗИРОВАНИЕ МАРШРУТОВ ПЕРЕДВИЖЕНИЯ МОСКОВСКОГО МЕТРОПОЛИТЕНА”

Выполнил: Николаев Иван Витальевич

Суть поставленной задачи:

- На основании данных о пассажирах, которые воспользовались метро дважды за сутки, при наличии информации о первом заходе в метро, необходимо предсказать, на какой станции и через какой промежуток времени, этот пассажир воспользуется метро повторно.
- Таким образом, данное задание подразумевает в себе сразу два предсказания: одно (предсказание станции второго захода из ограниченного числа станций) - задача классификации, второе – задача регрессии (определение подходящего времени из непрерывного промежутка).

Данные:

- Строк в тренировочном датасете
- Признаков в тренировочном датасете, включая ID

Добавление внешних данных:

- Мною были проанализированы открытые источники информации о Московском метрополитене, которые могли бы добавить ясности к существующему датасету и помочь будущей модели более точно разделить данные на требуемые классы.
- Подходящий датасет был найден на портале открытых данных Правительства Москвы – Пассажиропоток по станциям Московского метрополитена: https://data.mos.ru/opendata/7704786030-passajiropotok-po-stantsiyam-moskovskogo-metropolitena
- Он был обработан (файл ”Пассажиропоток.ipynb” в репозитории), были выбраны самые новые данные и переименованы необходимые станции. После этого, его интегрировали в основной датасет.

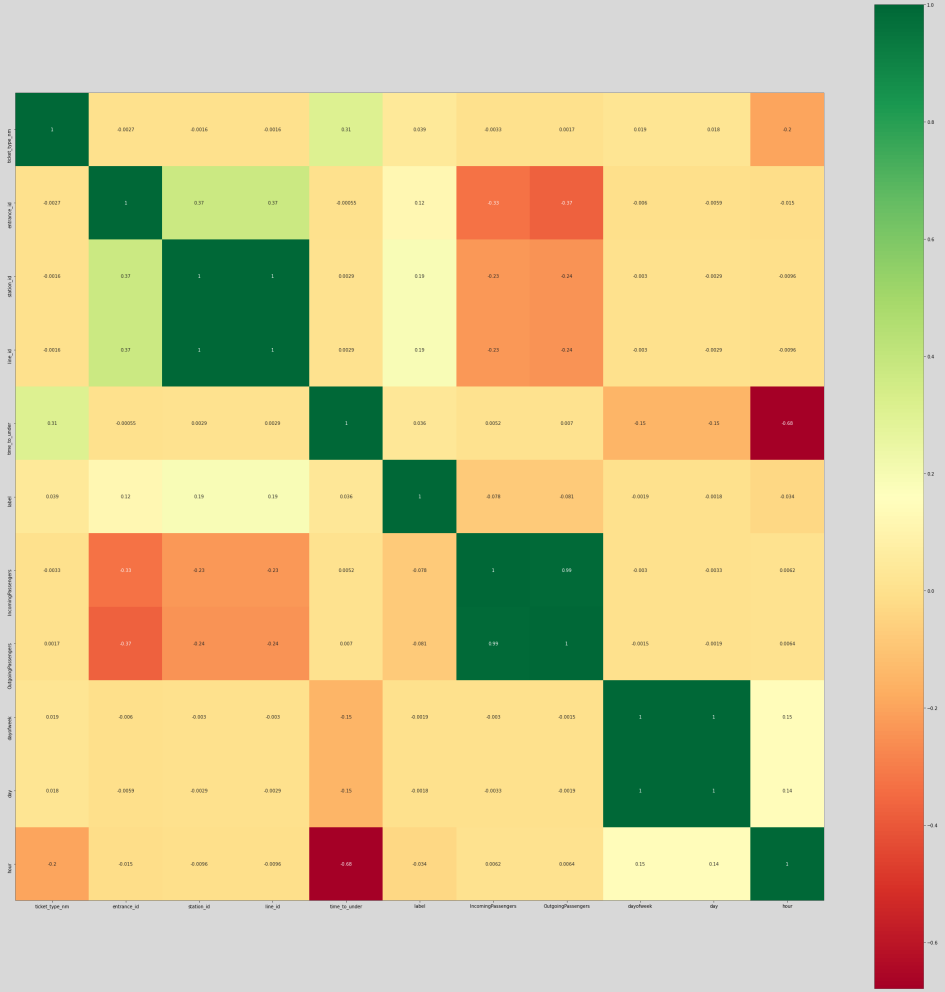
Гипотезы, инсайты и выводы по работе с данными

- Данные загромождены ненужной информацией. Поэтому производилась существенная предобработка:

1) Зачастую информация признаков дублировала друг друга: ID линии или Название линии – необходимо оставить что-то одно.

2) Содержалась лишняя информация. Например, признак ticket_id для каждого объекта разный и никак не может повлиять на исход предсказаний.

3) После отбора полностью нелогичных данных также была визуализирована корреляционная матрица и с помощью нее отобраны взаимозаменяемые признаки (коэф. корреляции ~ 1)



Описание выбранной модели

- Для задачи классификации оптимальным вариантом стала модель с использованием случайных лесов (Random Forest)
- Для задачи регрессии была выбрана модель CatBoost, разработанная компанией Яндекса. По утверждениям ее создателей она предсказывает значения лучше других похожих алгоритмов. И на наших данных, по результатам тестирования данный алгоритм предсказывает регрессионные значения лучше всего.

Обоснование точности решения

- В качестве метрики для оценки качества предсказаний была взята формула из задания:

$$result = 0.5 * Recall + 0.5 * R2$$

Таким образом, для задачи классификации у нас мы стремимся максимизировать recall, а для задачи регрессии максимизировать R2.

Контактные данные

- **Обо мне:**
- Студент Финансового университета при Правительстве РФ
- Факультет: Информационных Технологий и Анализа Больших Данных
- Курс: 3
- Направление: Бизнес-информатика
- Проходил несколько курсов по Машинному обучению от Samsung Innovation Campus, Stepik и Karpov Courses

- Контактный телефон: +79197626712
- Почта: ivannikolaev02@mail.ru