Домашнее задание
ML kurs 2022. 1 семестр.

Условие 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.

Описание входных значений (https://drive.google.com/drive/folders/1rZXJbc1gd--BP9Z1n6qOp_xUV-bmoe78?usp=sharing)

В данных присутствуют только те люди, которые совершили ровно две поездки в день, при этом статистика валидаций взята за несколько дней.

- **train.csv** — файл, содержащий данные о валидациях для обучения;
- **test.csv** — файл, содержащий данные для предсказания;
- **subway. csv** — вспомогательный файл содержащий информацию о всех возможных способах попасть со станции «А» на станцию «Б»;

Описание столбцов для **train и test:**

- id - уникальный идентификатор столбца;
- ticket_id - уникальный идентификатор билета, считается, что у одного

  билета один владелец
- ticket_type_nm - тип билета
- entrance_id - уникальный id входа в станцию
- entrance_nm - название
- station_id - уникальное id станции захода
- station_nm - наименование станции захода
- line_id - уникальный id ветки на, которой находится станция
- line_nm - наименование ветки, на которой находится станция
- pass_dttm - дата валидации
- **time_to_under (столбец для предсказания) -** сколько времени прошло

между первой и второй валидацией

- **label (столбец для предсказания)** - id второй станции, на которой

произошла валидация

Метрика

В качестве метрики сумма $Recall$ по столбцу label и $R2$ по time_to_under.

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

$R2$ считается как:

$$R2 = 1 - \frac{SSres}{SStot}$$

$SSres$ - сумма квадратов остаточных ошибок. $SStot$ - общая сумма ошибок.

$Recall$ считается как: $recall = \frac{TP}{TP+FN}$

TP (True Positive) — количество верно угаданных значений одного класса FN (False Negative) — количество неправильно угаданных значений класса

Форма сдачи:

1. Готовый ноутбук с исследованием
2. Ноутбук нужно показать и ответить на вопросы по нему
3. Оформить pipeline – по желанию (при реализации отдельно также показать ноутбук с исследованием)

4. **Также нужно сделать fork на гитхабе с репозитория https://github.com/kurmakaevAlsu/ML_kurs2022 и залить заранее ноутбук туда**