Комментарии по заданию

Формулировка задачи: необходимо идентифицировать отток в понятии “клиент НЕ совершит хоть одну транзакцию в течение 60 дней с момента последней известной”. Если не совершит, считаем что это отток. Надо предсказать 0 (ноль если отток), 1 если НЕ отток. Если совершит, надо предсказать через сколько дней (0-60).

Для решения вам выдаются файлы hackaton2023\_train и hackaton2023\_test

***Описание полей:***

**customer\_id** - int id клиента

(присутствует только в трейн) **date\_diff\_post** - int кол-во дней от максимальльной даты в pre периоде до минимальной покупки в пост

(присутствует только в трейн) **buy\_post** - bool метка о том, что в пост периоде покупка имеется

**group\_name** - cat test или train

**revenue** - money выручка от продажи позиции заказа

**startdatetime** - datetime дата\время продажи

**dish\_name** - str название продукта

**ownareaall\_sqm** - float площадь ресторана

**format\_name** - cat формат ресторана

***Комментарий по датасету:***

По сути, если разбирать что есть в трейне, это транзакции клиентов, которые можно привязать к клиенту по customer\_id, а одна транзакция характеризуется одним startdatetime. Можно понять, что именно (какие блюда) и когда он покупал и по какой цене. И есть признак что после последней даты он еще сделает какую-то неизвестную нам транзакцию (buy\_post), и мы знаем через сколько дней после крайнего известного заказа (date\_diff\_post).

Файлы сохранены в структуру apache parquet, открыть можно командой: pd.read\_parquet в питоне

в колонке buy\_post наблюдается дисбалланс классов. На тесте какой-то другой баланс классов, чем в трейне.

***Формат сабмита:***

csv файл с колонками: customer\_id, date\_diff\_post, buy\_post

один customer\_id - одна запись в сабмите (совершит ли этот клиент транзакцию в течение 60ти дней с момента последней известной транзакции, и если да, то через сколько дней)

Сабмит грузится на платформу в csv файле (разделитель ;)

***Метрики:***

Главная метрика для сабмита: F1 (измеряется по колонке buy\_post)

Доп метрика, за которую также можно набрать до трех баллов: RMSE (измеряется по колонке date\_diff\_post)

***Требования к презентации:***

В презентации обязательно надо нарисовать ROC кривую для задачи классификации оттока, рассказать как пришли к решению, что позволило получить хороший результат, как соотносятся основные метрики в лидерборде (на тесте) и на трейне.