# ФИНАЛЬНЫЙ ПРОЕКТ ПО КУРСУ "ВИДЕОКУРС ОТ MEGAFON + КУРСОВОЙ ПРОЕКТ"

КРИВОНОГОВ НИКОЛАЙ ВЛАДИМИРОВИЧ

# У НАС ПОЯВИЛСЯ ЗАПРОС ИЗ ОТДЕЛА ПРОДАЖ И МАРКЕТИНГА. КАК ВЫ ЗНАЕТЕ «МЕГАФОН» ПРЕДЛАГАЕТ ОБШИРНЫЙ НАБОР РАЗЛИЧНЫХ УСЛУГ СВОИМ АБОНЕНТАМ. ПРИ ЭТОМ РАЗНЫМ ПОЛЬЗОВАТЕЛЯМ ИНТЕРЕСНЫ РАЗНЫЕ УСЛУГИ. ПОЭТОМУ НЕОБХОДИМО ПОСТРОИТЬ АЛГОРИТМ, КОТОРЫЙ ДЛЯ КАЖДОЙ ПАРЫ ПОЛЬЗОВАТЕЛЬ-УСЛУГА ОПРЕДЕЛИТ ВЕРОЯТНОСТЬ ПОДКЛЮЧЕНИЯ УСЛУГИ.

- В качестве исходных данных вам будет доступна информация об отклике абонентов на предложение подключения одной из услуг. Каждому пользователю может быть сделано несколько предложений в разное время, каждое из которых он может или принять, или отклонить. Отдельным набором данных будет являться нормализованный анонимизированный набор признаков, характеризующий профиль потребления абонента. Эти данные привязаны к определенному времени, поскольку профиль абонента может меняться с течением времени.
- Данные train и test разбиты по периодам – на train доступно 4 месяцев, а на test отложен последующий месяц. Итого, в качестве входных данных будут представлены: ● data_train.csv: id, vas_id, buy_time, target ● features.csv.zip: id, feature_list
- И тестовый набор: ● data_test.csv: id, vas_id, buy_time target - целевая переменная, где 1 означает подключение услуги, 0 - абонент не подключил услугу соответственно. buy_time - время покупки, представлено в формате timestamp, для работы с этим столбцом понадобится функция datetime.fromtimestamp из модуля datetime. id - идентификатор абонента vas_id - подключаемая услуга Примечание: Размер файла features.csv в распакованном виде весит 20 гб, для работы с ним можно воспользоваться pandas.read_csv, либо можно воспользоваться библиотекой Dask.

# ДЛЯ АНАЛИЗА МЕТРИК БЫЛИ ИСПОЛЬЗОВАНЫ ТРИ МОДЕЛИ:

- Логистическая регрессия, lr: f1_macro = 0.4812 (+/- 0.0000)
- LightGBM, lgbm: f1_macro = 0.5299 (+/- 0.0058)
- CatBoost, cb: f1_macro = 0.7150 (+/- 0.0024)

# ЛУЧШУЮ МЕТРИКУ ПОКАЗАЛ CATBOOST

- После подбора гиперпараметров модель показала следующие метрики:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.99 | 0.87 | 0.93 | 254585 |
| 1.0 | 0.35 | 0.89 | 0.51 | 19861 |
| accuracy | | | 0.87 | 274446 |
| macro avg | 0.67 | 0.88 | 0.72 | 274446 |
| weighted avg | 0.94 | 0.87 | 0.90 | 274446 |

- Модель сохранена в файле model.pkl
- Вероятности подключения услуги сохранены в файле answers_test.csv

- Ссылка на проект на GitHub