Курсовой проект от Megafon Зорин Илья Geek Brains, факультет Искусственного интеллекта 2023 год

Задача

- Построить алгоритм, который для каждой пары пользователь-услуга определить вероятность подключения услуги.
- Метрика :

  Sklearn.metrics.f1 score(avarage="macro")

: 0	id	vas_id	buy_time	target
0	540968	8.0	1537131600	0.0
1	1454121	4.0	1531688400	0.0
2	2458816	1.0	1534107600	0.0
3	3535012	5.0	1535922000	0.0
4	1693214	1.0	1535922000	0.0

Исходные данные

- ● data_train.csv и data_test.csv трейновый и тестовые датасеты с признаками

id, vas_id, buy_time, target.

Где:

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

Выбор модели

Для моделирования выбран Catboost, т.к. показал лучший результат с минимальной предобработкой.
Плюсы Catboost:

- поддержка графического процессора
- автоматическое кодирование категориальных функций
- хорошие результаты с дефолтными параметрами.

Были проверены 3 базовые модели:

- Логистическая регрессия

  f1_macro = 0.4812 (+/- 0.0000)
- Lightgbm

  f1_macro = 0.5296 (+/- 0.0072)
- Catboost

  f1_macro = 0.7143 (+/- 0.0019)

Составление индивидуальных предложений для выбранных абонентов

- Обработать данные выбраных абонентов через модель и на основании предикта предлагать конкретную услугу.
- Для большей уверенности что услуга будет приобретена нужно ограничить точность Precision (как пример >=85%) и максимизировать Recall
- Для расширения списка абонентов которые могут купить услугу необходимо ограничить полноту Recall (±80%) и максимизировать точность Precision

	precision	recall	f1-score	support
0.0	0.99	0.87	0.93	254585
1.0	0.35	0.90	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

Итоговые материалы

- Megafon_CP – Jupyter Notebook с кодом
- Model2023-04-02 - модель в формате pickle
- Answer_test - файл с предсказанными вероятностями
- Презентация