Видеокурс от Megafon, курсовой проект

Задача:

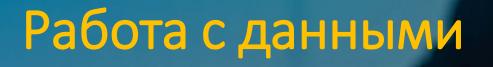
• построить алгоритм, определяющий вероятность подключения услуги пользователем.

В работе использовались файлы:

data_train.csv, data_test.csv, features.csv

Метрика:

• f1, невзвешенным образом, аналогично функции sklearn.metrics.f1_score



- Файл features.csv был слишком большим, поэтому была использована библиотека dask.
- Файлы data_train.csv и features.csv были соединены через inner join по полям id и buy_time.
- Итог работы с данными файл X_train.
- Была проведена балансировка классов для предотвращения негативных последствий явного перекоса по целевой переменной.
- Из X_train удалено поле target, что в итоге позволило создать итоговый набор данных: X_train и Y_train.

Pipeline

После разделения признаков на константные, вещественные, бинарные и категориальные, были обработаны пропуски:

- вещественные признаки замена на среднее значение;
- категориальные на моду.

Модель

- Из использованных моделей лучший результат показал **градиентный бустинг**, который и был использован в качестве финальной модели для предсказаний.
- Порог значений был выбран 0.5, т.е. значения >= 0.5 относилось к целевой переменной 1, а < 0.5 к переменной 0.