Видеокурс от Megafon + курсов ой проект

Предоставляемые данные

- Презентация с описанием решения
- Jupyter-notebook с обработкой данных и обучением готовой модели
- Модель в формате pickle
- Файл с предсказаниями

Описание модели

- Модель построена на базе CatBoost
- Финальная модель состоит из:
  - Функций начальной предобработки данных и добавления признаков к тестируемому датасету из файла features.csv
  - Подбора оптимального порога вероятности
  - Pipeline:
  - Модели предобработки признаков (стандартизация численных признаков, кодирование категориальных признаков)
  - Классификатора
- Оценочный Macro F1-Score модели основанный на кросс-валидации: ~0.6
- Параметры модели, подобранные в ходе оптимизации:
  - max_depth: 0.3
  - l2_leaf_reg: 5

План исследования

- Объединение данных из датасетов test и features
- Объединение совершено по правилу ближайшего по времени профиля к buy_date в тренировочных и тестовых датасетах
- Обработка признаков
- Выделение числовых и категориальных признаков.

  Числовые – стандартизировать, категориальные – закодировать припомощи OneHot.
- Подготовить функцию корректной кросс-валидации
- Тестирование нескольких моделей и выбор модели для дальнейшего улучшения
- Анализ и сохранение данных
- Построение предсказания

Комментарии

- Базовая модель - Logistic Regression: F1-score = 0.151
- Выбор модели происходил на основе оценки результатов нескольких моделей на кросс-валидации
- Нужно больше времени посвятить исследованию признаков, несмотря на то, что они полностью анонимизированы.