Видеокурс от Megafon + курсовой проект

Александр Бабкин
25.06.2021

Предоставляемые данные

- Презентация с описанием решения
- 3 jupyter-notebook с обработкой данных и обучением модели готовой модели (course-project-[1-3].ipynb)
- Модель в формате pickle (model.pcl)
- Файл с предсказаниями (answers_test.csv)

Описание модели

- Модель построена на базе алгоритма бустинга CatBoost
- Финальная модель состоит из:
  - Функций начальной предобработки данных и добавления признаков к тестируемому датасету из файла features.csv
  - Подбора оптимального порога вероятности
  - Пайплайна:
    - Модели предобработки признаков (стандартизация численных признаков, кодирование категориальных признаков)
    - Классификатора
- Оценочный F1 macro score модели на базе кросс-валидации по 3 фолдам: ~0.6
- Параметры модели, подобранные в ходе оптимизации:
  - max_depth: 0.3
  - l2_leaf_reg: 5

План исследования

Работа была разбита на следующие подзадачи:

- Объединить данные из датасетов test с features
  Объединение происходило по правилу ближайшего по времени профиля к buy_date в тренировочных и тестовых датасетах
- Обработать признаки
  Выделить числовые и категориальные признаки, числовые – стандартизировать, категориальные – закодировать по методу OneHot
- Подготовить функцию корректной кросс-валидации
- Протестировать несколько моделей и выбрать модель для дальнейшего улучшения
- Проанализировать и сохранить результаты модели
- Построить предсказания

Дополнительные комментарии

- В качестве базовой модели была взята логистическая регрессия; f1-score=0.151
- Выбор модели происходил на основе оценки результатов нескольких моделей на кросс-валидации по 3 фолдам
- По финальным результатам, думаю, что в данной задаче необходимо больше времени посвятить исследованию признаков, несмотря на то, что они полностью анонимизированы.