КУРСОВОЙ ПРОЕКТ

Определение вероятности подключения услуги абонентом

Александрина Вяткина, 2022 г

МегаФон.
Начинается
с тебя

ПОСТАНОВКА ЗАДАЧИ

Исходные данные :

- Информация об отклике абонентов на предложение подключения одной из услуг. Каждому пользователю могут быть предложены несколько услуг в разное время, которые он мог принять или нет.
- Нормализованный анонимизированный набор признаков, описывающий профиль пользования абонента

Необходимо: для каждой пары пользователь-услуга определить вероятность подключения услуги и предложить принцип составления индивидуальных предложений для абонентов.

ЭТАПЫ ВЫПОЛНЕНИЯ ПРОЕКТА

- Составление общего набора данных из двух датасетов из исходных данных
- Изучение таргета и генерация новых признаков на основе признака времени
- Деление исходного набора данных на две части – тренировочный и валидационный (обучение на наблюдениях с июля до ноября включительно), валидация – на наблюдениях декабря
- Построение бейзлайна – модель Логистическая регрессия
- Оверсэмплинг для устранения дисбаланса классов, тестирование ансамблевых моделей
- Проведение отбора признаков на основе feature importance в целях понижения размерности, классификация признаков для последующего улучшения модели, подбор гиперпараметров
- Подбор порога вероятности, при вероятности выше которого объекты относим к 1 классу, нижу – к 0, определение величины метрики f1 macro
- Предсказание вероятностей подключения услуг пользователями на заданном тестовом датасете.

Генерация новых признаков

Были сгенерированы новые признаки на основе признака buy_time:

offer_year,

offer_month,

offer_day,

offer_weekday,

offer_hour

* Большая их часть не показала себя как полезные

МОДЕЛИ, УЧАСТВОВАВШИЕ В ЭКСПЕРИМЕНТАХ

- **Baseline** – ***Logistic Regression*** с небольшой обработкой признаков (удалены константные и добавлены новые на основе buy_time). *f1 macro score:*

До oversampling-а*: 0,58*

После oversampling-а: *0,59*

- ***CatBoost*** – после oversampling-а и обработки и отбора признаков *f1 macro score = 0.74*
- ***GradientBoosting*** – после oversampling-а, обработки и отбора признаков, подбора гиперпараметров *f1 macro score*= ***0,752***

Последняя модель показала наилучший результат

КОММЕНТАРИИ ПО ПОСТФИЛЬТРАЦИИ РЕЗУЛЬТАТА

На мой взгляд, одним из основных факторов для компании при составлении рекомендации абонентам является *экономическая* составляющая, а также общий показатель конверсии.

Мы должны быть уверены в своих предложениях абонентам, то есть в их наибольшей конверсии и, как следствие, оправданных издержках -> метрика точности имеет наибольшее значение

МОИ РЕКОМЕНДАЦИИ ПО СОСТАВЛЕНИЮ ИНДИВИДУАЛЬНЫХ ПРЕДЛОЖЕНИЙ

На мой взгляд, индивидуальные предложения должны строиться примерно следующим образом:

- Предсказание вероятности подключения каждой из услуг клиентом
- Сортировка по убыванию вероятности
- «Отсечение» услуг, которые ниже порога отнесения к 1 классу (помимо расчета по величине метрики, при определении порога важно учесть экономическую составляющую)
- Определение оптимальное количество услуг для рекомендаций и интервал. Критериями определения одновременно могут быть опыт, бюджет на рекомендации, коэффициент лояльности абонента и пр). Можно перестроить схему рекомендации последующих услуг, в зависимости от отклика на первое предложение. Ориентируясь на себя, могу сказать, что восприму за навязчивость более 2-3 предложений в месяц.
- Рекомендация выбранного количества услуг из тех, которые были отобраны по порогу вероятности.