МегаФон

Финальный проект от компании МегаФон

Формирование персональных предложений абонентам компании "Мегафон"

Давыдов Александр

МегаФон

Что нужно было сделать?

Необходимо на основе анонимизированных профилей потребления абонентов предсказать вероятность подключения услуг.

Исходные данные

- Датасет с id абонентов, id услуг, временной меткой и целевой переменной – откликом на предложенную услугу (трейн)
- Датасет с id абонентов, id услуг и временной меткой (тест)
- Датасет с анонимизированными профилями потребления абонентов (id абонента, метка времени и 255 закодированных признаков)

Ожидаемый результат

- Модель в формате pickle, предсказывающая вероятность подключения услуг
- Сами предсказания в формате csv
- Код обучения модели в формате ipynb

Метрика

f1, невзвешенная (`sklearn.metrics.f1_score(…, average='macro')`

МегаФон

Из каких этапов состоит моё решение?

- Формализация задачи, загрузка данных
- EDA
- Преобразования данных
- Оптимизация features.csv – исключение лишних строк
- Объединение датасетов (merge_asof)
- Бейзлайн – дерево решений
- Балансировка классов
- Более продвинутые модели – бустинги, LAMA
- CNN
- Подведение итогов, выводы по работе моделей
- Обучение модели на всем датасете и сохранение прогноза
- Бонус – рекомендации для абонентов, выводы по рекомендациям

МегаФон

Какую модель я выбрал для финальных прогнозов?

xgboost.XGBClassifier, обученный на сбалансированном датасете, с параметрами:

- booster: gbtree (он более склонен к переобучению, но если за этим следить, зачастую показывает лучшие результаты, чем dart)
- learning rate: 0.02
- max_depth: 20

Параметры предобработки:

- Тип балансировки – oversampling
- Тип мерджа датасетов – nearest
- Из обучения исключены метки времени, они нужны были только для мерджа датасетов

МегаФон

Что я еще пробовал из того, что не вошло в финальное решение?

- Библиотеку autoEDA - не дала мне нужного представления о данных
- Фреймворк LightAutoML (LAMA) - выступил хуже XGBoost
- Полноценную кросс-валидацию - слишком долго на моих мощностях, а учитывая, что наверняка придется менять гиперпараметры - совсем нереалистично
- Стэкинг - выступил хуже XGBoost
- Другие методы мержа датасетов (кроме nearest, можно еще forward и backward) - не очень влияет на метрику
- Другие методы работы с дисбалансом классов (например, SMOTE и tomek) - метрика чуть ухудшилась

Что осталось в ноутбуке, но не участвует в формировании финальных предсказаний?

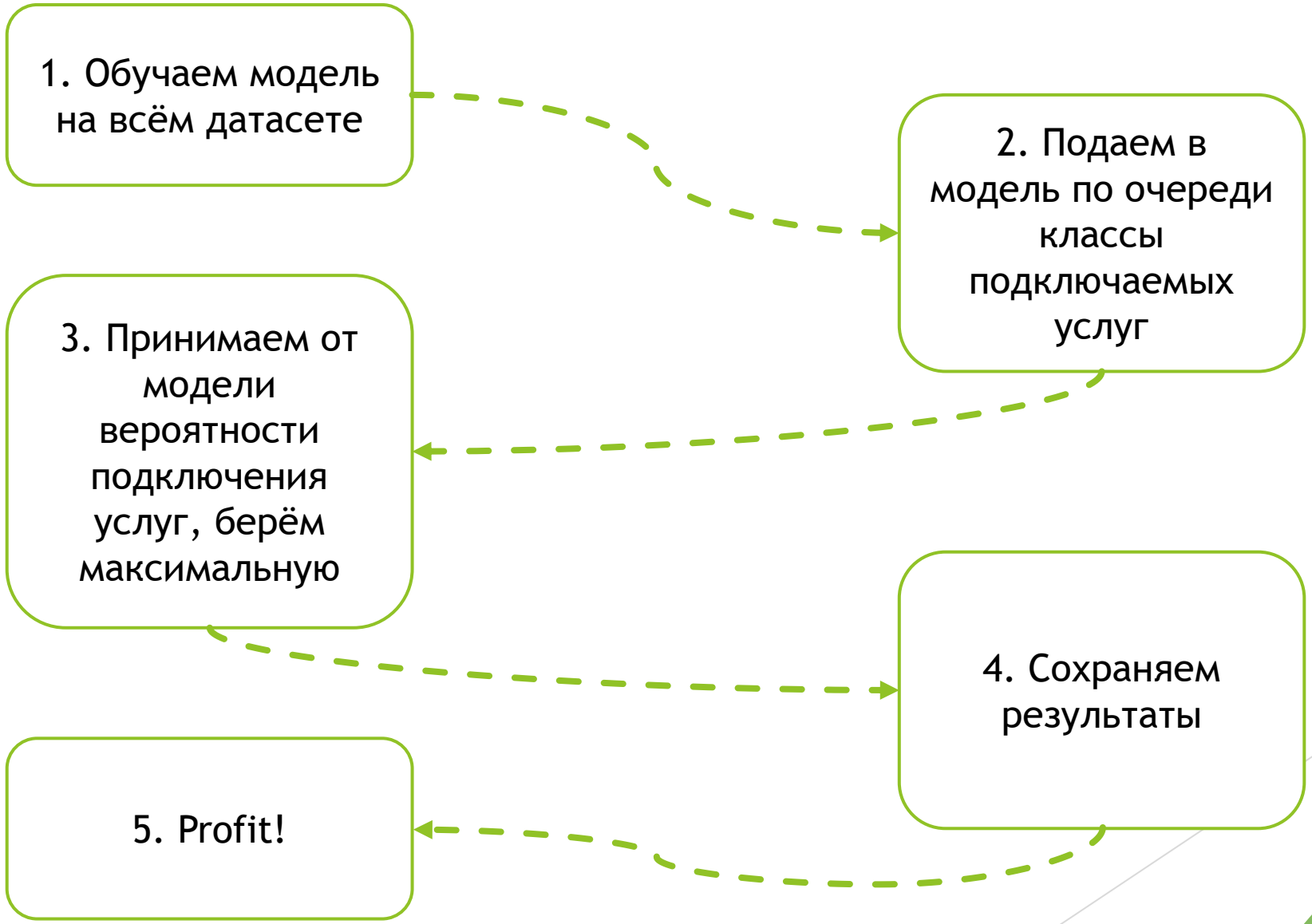
- Бейзлайн
- CNN
- Кросс-валидация

Что осталось невостребованным?

- Фреймворк Luigi
- Feature engineering

МегаФон

Как именно устроена работа модели для пользователя?



МегаФон

Вместо итогов

- Предсказуемо, лучший вариант показал бустинг с предварительным оверсемплингом (Train-score: 0.9439, Test-score: 0.9289)
- Неплохое приближение дает бейзлайн - дерево решений. Оно, конечно, слишком простое для таких сложных данных, но справляется неплохо без необходимости предобрабатывать данные.

МегаФон

И отдельно пару слов по рекомендациям

- Любопытно, что распределение кардинально отличается от базового распределения в тестовом датасете. Возможно, стоит попробовать внести изменения в существующую (да, я понимаю, что скорее существовавшую, курс то 2019 года) рекомендательную систему после тщательного анализа уже не ml, а бизнес-метрик.
- Кроме того, возможно, стоит выдавать рекомендации только пользователям с вероятностью подключения топовой услуги начиная с определенного трешхолда. Это напоминает задачу аплифт-моделирования - можно разделить пользователей на группы:

1. Пользователи, которые совершат нужное действие независимо от коммуникации
2. Те, кто совершат действие, если будет коммуникация
3. Те, кто не совершит действие независимо от коммуникации
4. Не совершат действие, если коммуникация будет

- И наша задача - не беспокоить пользователей из 4 группы. Это отдельная задача, в которой могут быть использованы результаты моей работы.

МегаФон

Ссылки

- Гитхаб проекта
- Ноутбук с решением (Колаб)