Курсовой проект

Вероятность подключения услуги абонентом

Роман Космодемьянский, декабрь 2021

Задача

В качестве исходных данных доступна информация об отклике абонентов на предложения подключения разных услуг.

Каждому пользователю может быть сделано несколько предложений в разное время, каждое из которых он может или принять, или отклонить.

Также доступен анонимизированный набор признаков, характеризующий профиль потребления абонента. С течением времени профиль абонента может изменяться.

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

Этапы решения задачи

1. Трейн датасет делим на две части: обучающая выборка (наблюдения с июля по ноябрь 2018 года) и валидационная выборка (декабрь 2018 года).

2. Генерация новых признаков.

3. Построение бейзлайн модели.

4. Классификация признаков для последующего построения модели, позволяющей добиться лучшего качества предсказаний + подбор гиперпараметров.

5. Подбор порога вероятности при превышении которого наблюдение относим к позитивному классу. Важно, так как присутствует дисбаланс классов!

6. Оценка вероятностей подключения услуг пользователями на заданном тестовом датасете.

Признаки

Кроме признаков присутствующих в датасетах сгенерировал дополнительные, однако в большинстве они оказалось малоинформативными:

- ***buy_time*** преобразовал к формату даты, а затем на основе этого признака создал новые - год, месяц, день, день недели, час предложения и, аналогично, для датасета с профилями;

- ***vas_count*** - количество предложений соответствующего типа услуг в обучающей выборке;

- ***id_count*** - количество предложений соответствующий пользователь получил в обучающей выборке.

Модели

- *Бейзлайн:* **логистическая регрессия** без предобработки признаков. Метрика f1_macro = **0.6344**

Далее на основе изменчивости признаки были разделены на категориальные и вещественные. Оставил только информативные признаки и обучил две модели на их основе:

- **CatBoostClassifier**: признаки, принимающие не более 10 уникальных значений, были отнесены к категориальным. Метрика f1_macro = **0.6968**

- **GradientBoostingClassifier** + OneHotEncoder для категориальных признаков с подобранным гиперпараметром max_features = 70. Метрика f1_macro = **0.7462**

Финальные комментарии

- На мой взгляд показатель точности при решении подобной задачи должен иметь больший вес, т.к. предложение услуги клиенту связано с издержками, следовательно мы хотим быть уверены, что мы делаем предложения только таким клиентам, которые готовы покупать и рекомендуем им те услуги, которые они купят;

- Схема построения индивидуальных рекомендаций пользователю по моему мнению должна выглядеть следующим образом: оценка вероятности приобретения каждой из услуг с использованием модели ⟶ сортировка услуг по рассчитанной вероятности ⟶ определяем сколько из услуг превышает выбранный порог для отнесения к позитивному классу (при определении этого порога нужно учитывать финансовые издержки рекомендации пользователю) ⟶ рекомендуем отобранные услуги. Дополнительным ограничением может выступать общий бюджет на продвижение услуг.

Использованные библиотеки

- Pandas + NumPy;
- Sklearn;
- CatBoost;
- Seaborn;
- Pickle.