Прогнозирование оттока пользователей провайдера телекоммуникационных услуг

•••

ЧемпИИонат по ML и Big Data Республики Карелия

05.09 - 26.09.2022

Дата проведения исследования: сентябрь 2022
Исполнитель: Ганева М. В.
Контактный телефон: +79116658587
E-mail: ainslie.red@gmail.com

Задача исследования

Сфера телекоммуникаций — высококонкурентная динамичная среда, в которой уровень осведомленности пользователей позволяет легко получить информацию о других поставщиках услуг с аналогичным или наиболее высоким качеством по более выгодной цене. Отсюда и возникает проблема оттока пользователей.

Участникам чемпионата предлагалось разработать модель прогнозирования оттока пользователей с использованием данных о них, опираясь на информацию о запросах пользователей к другим сайтам данной отрасли и историю обращений в компанию.

Цели чемпионата:

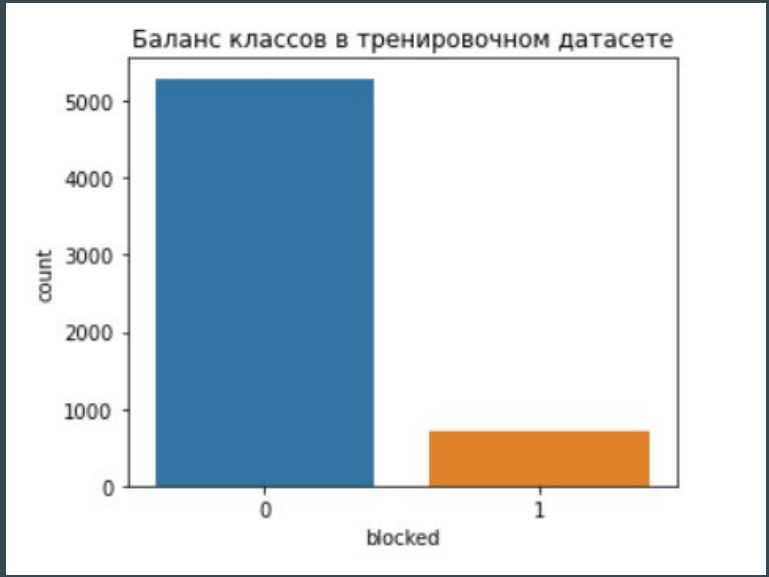
Участникам чемпионата необходимо было создать свое решение по прогнозированию оттока с использованием данных о клиентах. В доступе находится информация о запросах пользователей к сайтам конкурентов и история обращений в компанию.

Загрузка данных

Для решения поставленной задачи было предоставлено в свободное пользование пять наборов данных:

1. train.csv — файл содержащий данные пользователей для тренировки. Где: 1 - клиент ушел, 0 - остался.
2. log.csv — содержит данные обращения пользователей.;
3. named.csv — лог днс-запросов к доменам конкурентов (rt.ru и sampo.ru).
4. type contract.csv - тип списания у пользователей, где: 1 - посуточная, 0 - помесячная.
5. submission.csv — пример файла для отправки.

Баланс классов целевой переменной в тренировочном датасете



- 0.88% пользователей продолжают пользоваться услугами компании
- 0.12% пользователей отказались от услуг компании

Вывод: классы целевой переменной не сбалансированы.

Добавление новых признаков

1. В процессе исследования и группировки данных для каждого пользователя были добавлены признаки:
    a. day_or_month_contract - тип списания (ежедневно, помесячно)
    b. cnt_contacting - общее число обращений за услугами
    c. cnt_domen - число доменов конкурентов
    d. avg_support - среднее число обращений пользователя к услугам компании
    e. и другие

2. Было изучено распределение данных признаков в тренировочном датасете и последующая фильтрация пользователей, у которых значения признаков были аномальными.

Добавление признаков (категориальные данные)

Далее были проанализированы датасеты с данными обращений пользователей и логом днс-запросов к доменам конкурентов. Для каждого запроса и обращения пользователя было подсчитано общее число обращений. Были получены таблицы следующего вида:

```
In [38]: df_log4[df_log4['blocked']==0].sort_values(by='contract_id', ascending=False).head(7)
```

Out[38]:

	event_type	blocked	contract_id
66	Информер ВК. Показ	0.0	5414
28	Включение интернета на 20 минут	0.0	1857
71	Обращение в службу заботы о клиентах	0.0	1649
48	Гарантированный платеж за деньги	0.0	1192
136	Турбокнопка бесплатно	0.0	193
129	Смена тарифа	0.0	186
131	Состояние клиентского оборудования	0.0	169

Добавление признаков (категориальные данные)

Для каждой группы пользователей по целевой переменной были выделены топ-5 обращений пользователей за услугами и топ-5 днс-запросов к логам конкурентов (часто они совпадали у активных и отвалившихся пользователей). Таким образом, в тренировочный и тестовый датасеты было добавлено порядка 12 новых признаков, которые фиксировали обращался ли пользователь по конкретному запросу или нет.

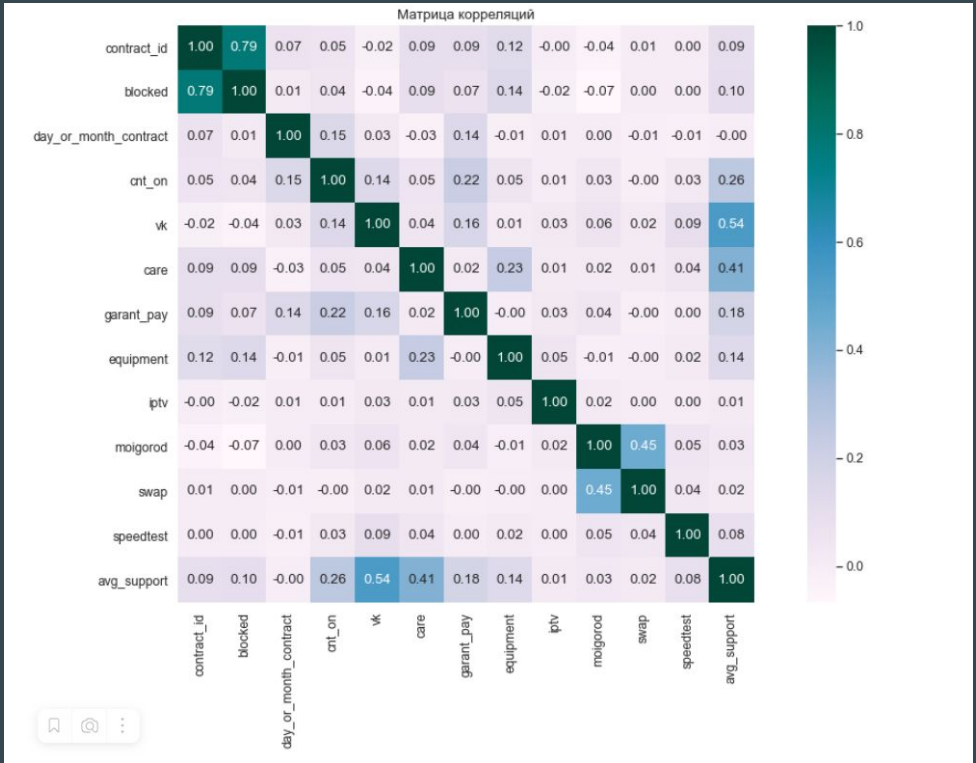
Анализ обращений пользователей

Примеры таблиц, получившихся в результате группировки данных.

	event_type	blocked	contract_id
67	Информер ВК. Показ	1.0	558
72	Обращение в службу заботы о клиентах	1.0	374
29	Включение интернета на 20 минут	1.0	315
49	Гарантированный платеж за деньги	1.0	206
70	Оборудование	1.0	66
41	Внутреннее сообщение	1.0	61
31	Включение интернета на 20 минут с IVR	1.0	56

	event_type	blocked	contract_id
66	Информер ВК. Показ	0.0	5414
28	Включение интернета на 20 минут	0.0	1857
71	Обращение в службу заботы о клиентах	0.0	1649
48	Гарантированный платеж за деньги	0.0	1192
136	Турбокнопка бесплатно	0.0	193
129	Смена тарифа	0.0	186
131	Состояние клиентского оборудования	0.0	169

Итоговая матрица корреляций



В процессе изучения матрицы корреляций получившихся признаков, часть из них была удалена из датасетов во избежание проявления мультиколлинеарности.

Для числовых признаков было проведено масштабирование.

Алгоритм машинного обучения

В качестве модели машинного обучения был выбран градиентный бустинг CatBoost - это алгоритм машинного обучения с расширенным деревом решений, разработанный Яндексом.

Он использует небрежные (oblivious) деревья решений, чтобы вырастить сбалансированное дерево. Также CatBoost позволяет работать категориальными переменными, но в данном проекте они не использовались.

Процесс обучения модели

1. Данные были разделены на обучающую и валидационную выборки. Тестовая выборка формировалась в отдельном датасете изначально.
2. Для подбора оптимальных гиперпараметров модели был использован GridSearchCV.
3. Проблема несбалансированных классов целевой переменной была решена посредством class_weight (из пакета sklearn.utils). Для каждого класса был рассчитан вес, затем эти данные были указаны в модели.

Процесс обучения модели

4. Не все признаки были использованы для обучения финального варианта модели, часть была удалена из датасета после изучения графика model.feature_importances_.

5. Затем была обучена финальная модель и сделано предсказание для тестовой выборки.

6. Метрика recall финальной модели составила 0.730519.