КУРСОВОЙ ПРОЕКТ К ВИДЕОКУРСУ ОТ MEGAFON

ЗАНОРИН Д.Ю., 2022 Г.

ЗАДАЧА И ДАННЫЕ

Задача:

- Построить алгоритм, который для каждой пары пользователь-услуга определит вероятность подключения услуги.

Данные:

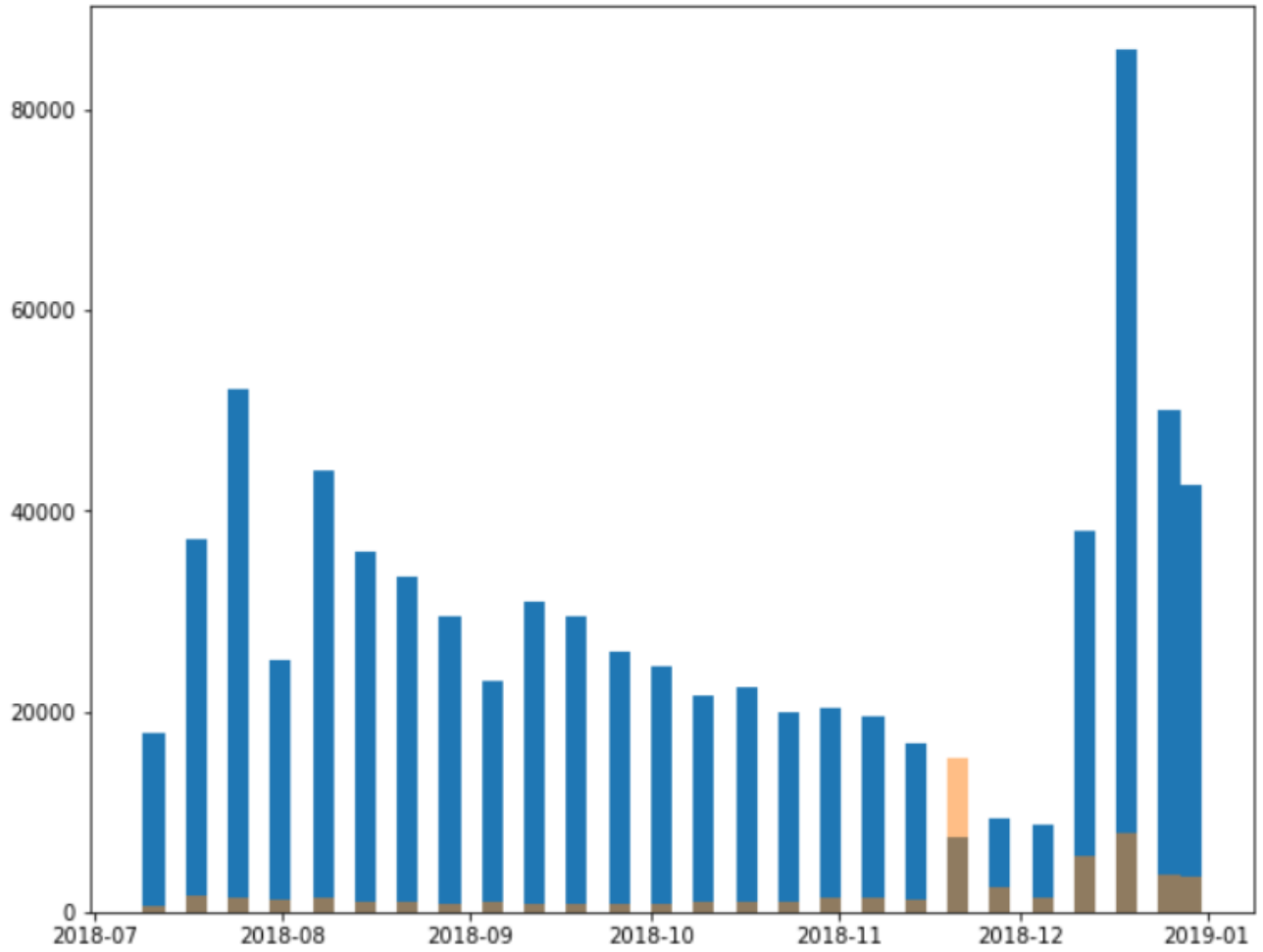
- В качестве исходных данных доступна информация об отклике абонентов на предложение подключения одной из услуг **(файлы data_train.csv и data_test.csv)**.
- Каждому пользователю может быть сделано несколько предложений в разное время, каждое из которых он может или принять, или отклонить.
- Отдельным набором данных будет являться нормализованный анонимизированный набор признаков, характеризующий профиль потребления абонента **(файл features.csv)**. Эти данные привязаны к определенному времени, поскольку профиль абонента может меняться с течением времени.
- Данные train и test разбиты по периодам – на train доступно 4 месяцев, а на test отложен последующий месяц.

АНАЛИЗ ДАННЫХ

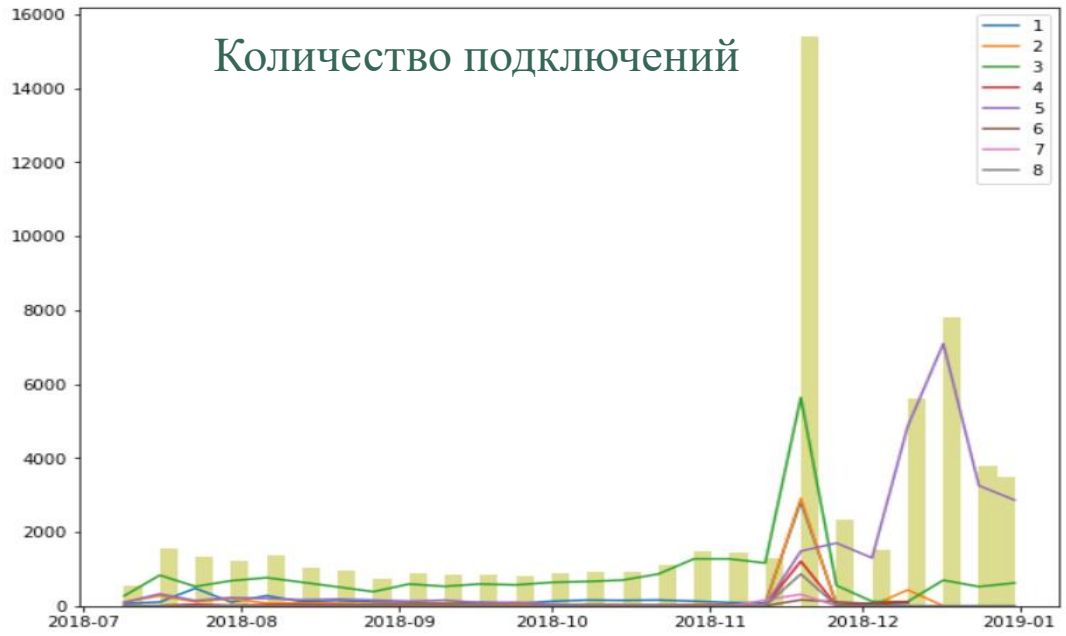
- Загрузили тренировочный датасет data_train.csv при помощи pandas, а датасет features.csv через dask.dataframe, так его размер слишком велик.
- Создали признак date, в котором дата из признака buy_time конвертирована в удобном формате.
- Убедились, что нет пропущенных значений.
- Узнали, за какой период предлагались услуги, а также ID этих услуг.
- Увидели, что услуг 8, пронумерованы по порядку, начиная с 1, но под номером 3 нет. Заменили значения на более корректные.
- Проверили, что есть пользователи, которым поступало предложение о подключении услуги несколько раз.
- Построили графики и посмотрели динамику предложений.

АНАЛИЗ ДАННЫХ ГРАФИКИ

- По этому графику видно, что в ноябре 2018 количество пользователей, которые приобрели предложенную услугу, оказалось намного больше, чем отказавшихся. Это было 19 ноября 2018. Этот день в дальнейшем удалили из выборки.



АНАЛИЗ ДАННЫХ ГРАФИКИ



- Видим, что количество предложений по услугам 1 и 2 постепенно уменьшалось, но к середине декабря произошёл резкий скачок (вместе с услугами 4 и 5). Особенно сильно подскочила услуга № 2. Вероятно, такой скачок связан с новогодними праздниками, поскольку в этот период часто бывают акции. Услуги 3, 6, 7 и 8 такого эффекта не имели. Кроме того, к декабрю услуга № 3 пошла на спад по количеству предложений. Лидерами по подключениям являются услуги 3 и 5. Услуга № 3 имела успех в ноябре, затем компания сделала акцент на других услугах перед новогодними праздниками, и уже в этот период самой эффективной оказалась услуга № 5. Кроме того, у услуги № 5 количество предложений практически совпадает с количеством подключений.

АНАЛИЗ ДАННЫХ
ПОПУЛЯРНОСТЬ УСЛУГ В ЦИФРАХ

- vas_id – ID услуги
- 0 – предложения, которые были отклонены
- 1 – предложения, которые были приняты

- Видим, что наиболее популярные услуги – 3 и 5.

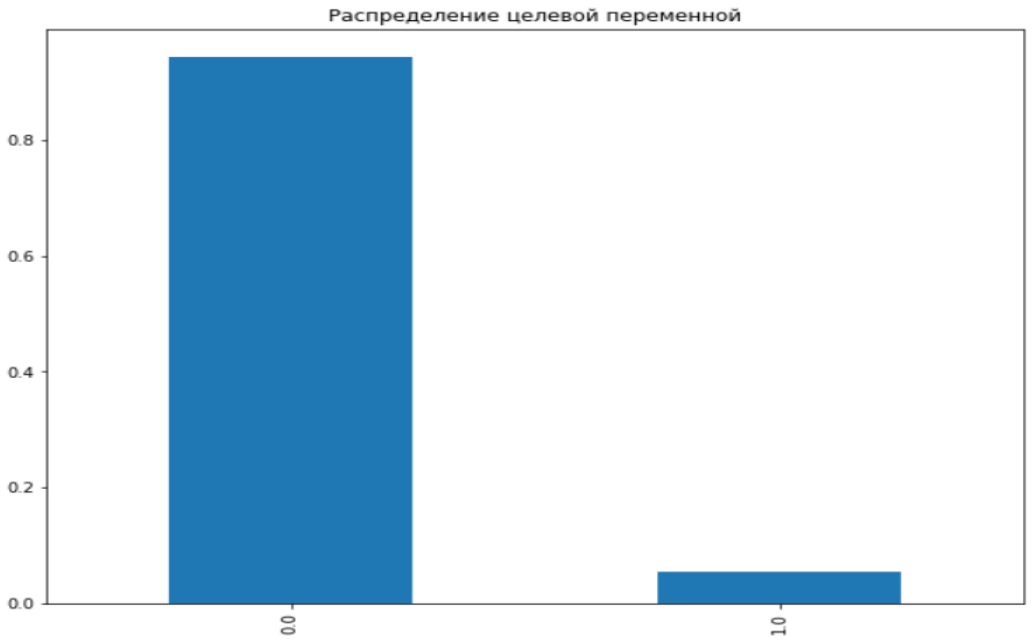
	vas_id	0	1
0	1	304511	5664
1	2	244708	4797
2	3	63991	21765
3	4	92393	1692
4	5	33174	24704
5	6	15219	213
6	7	13003	347
7	8	4468	1004

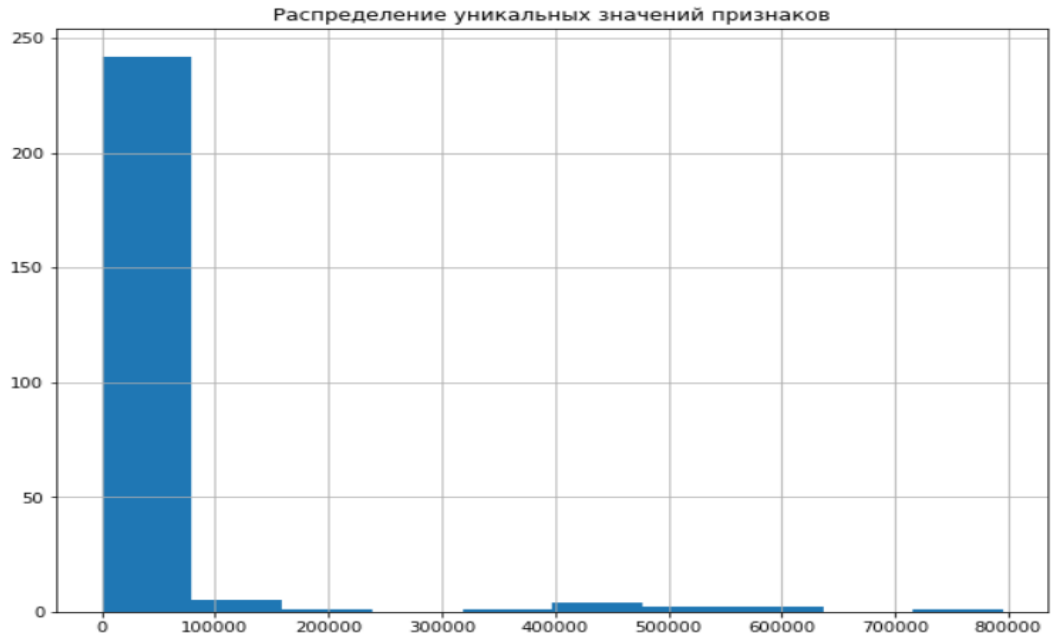
	vas_id	0	1
0	1	0.98	0.02
1	2	0.98	0.02
2	3	0.75	0.25
3	4	0.98	0.02
4	5	0.57	0.43
5	6	0.99	0.01
6	7	0.97	0.03
7	8	0.82	0.18

ПРЕДОБРАБОТКА ДАННЫХ

- Убрали из features тех пользователей, которых нет в data_train и data_test.
- Отсортировали датасеты по времени (признаку buy_time). Учитывается, что в датасете features есть строки с пользователями, у которых дата либо меньше, либо равна аналогичному пользователю из датасетов data_train и data_test.
- Объединили датасеты по id пользователей.
- Добавили в датасеты train_data и train_test 3 новых признака (день, неделю и месяц). Признаки buy_time и date удалили.
- Признаки типа float64 сконвертированы в тип float32, чтобы уменьшить вес датасетов.
- Сохранили выборки в файлы train_data_features.csv и train_test_features.csv.

РАБОТА С ПРИЗНАКАМИ

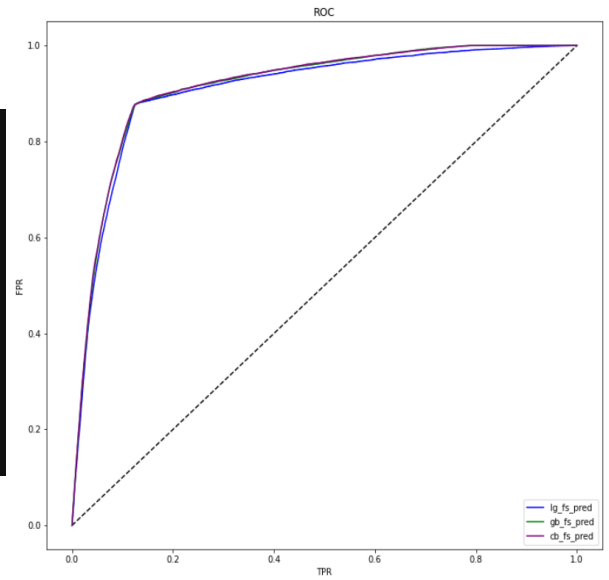


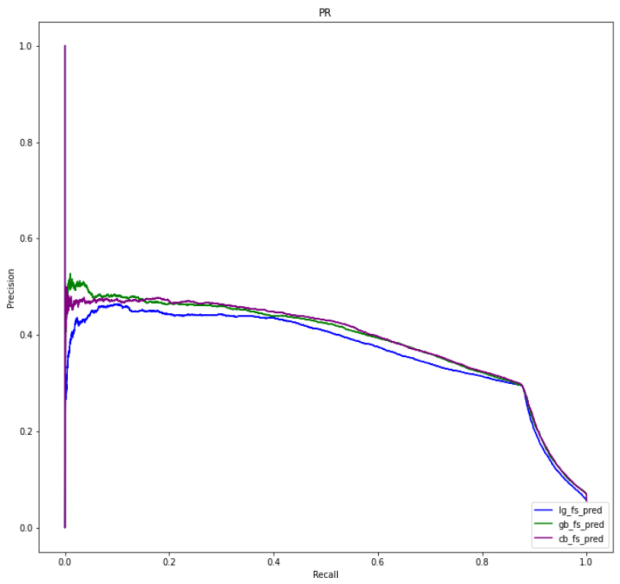


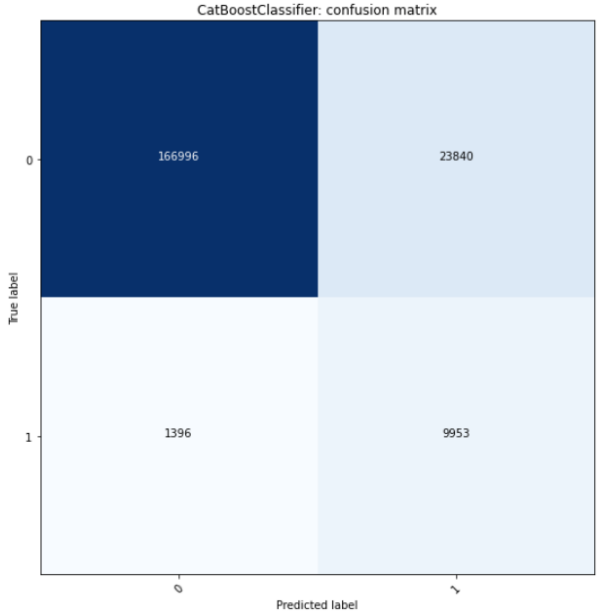
- Разделили признаки и целевую переменную. Посмотрели на распределение целевой переменной и распределение уникальных значений признаков.
- Отобрали признаки по категориям.
- Разбили обучающий набор на train и test.
- Выполнили балансировку данных (RandomUnderSampler), так как есть дисбаланс в распределении целевой переменной.

ПОСТРОЕНИЕ ПАЙПЛАЙНОВ
ОБУЧЕНИЕ МОДЕЛЕЙ И ВЫБОР ЛУЧШЕЙ

	precision	recall	f1-score	support
0.0	0.99	0.88	0.93	190836
1.0	0.29	0.88	0.44	11349
accuracy			0.88	202185
macro avg	0.64	0.88	0.69	202185
weighted avg	0.95	0.88	0.90	202185







- Собрали Pipeline.
- Использовали GridSearchCV для автоматического подбора параметров.
- Обучили несколько моделей, а именно 3: LogisticRegression, GradientBoostingClassifier и CatBoostClassifier.
- Предсказали на тестовой выборке и оценили качество моделей.
- Увидели, что все модели примерно похожи по качеству, результаты метрик почти одинаковые.
- Выбранная модель – CatBoostClassifier, так как она отработала быстрее, а также у неё самое высокое значение интересующей нас метрики f1 (average=`macro`): 0.69.
- Признали значение 0.5 оптимальным порогом вероятности для отнесения к положительному классу и предложения абоненту услуги.

ПРИНЦИП СОСТАВЛЕНИЯ ИНДИВИДУАЛЬНЫХ ПРЕДЛОЖЕНИЙ ДЛЯ ВЫБРАННЫХ АБОНЕНТОВ

- Можно установить минимальный порог вероятности того, подключит ли клиент услугу (цифра 1 в признаке target). Если вероятность ниже порога, то услугу не предлагать (как в данном случае использовали порог 0.5).
- Можно учесть лояльность клиента. Например, если клиент часто подключал предлагаемые услуги, то можно предложить и новую. Но если в течение последнего времени с клиентом были неудачные взаимодействия (услуга предлагалась, а клиент отказывался), то пока что новые услуги не предлагать.
- Желательно проверить, подключена ли предлагаемая услуга у клиента. Предложение об услуге, которая уже подключена у клиента, может сбить с толку и повлиять на лояльность.
- Можно учесть время последнего взаимодействия с клиентом, а также самые популярные услуги на текущий период. Если клиенту давно не предлагались услуги, то можно предложить ту, которая сейчас является самой подключаемой.