Megafon: курсовой проект

Алекс Максаков

Задача:

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

Данные:

В качестве исходных данных предоставлена информация об отклике абонентов на предложение подключения одной из услуг **(data_train.csv/data_test.csv)**.

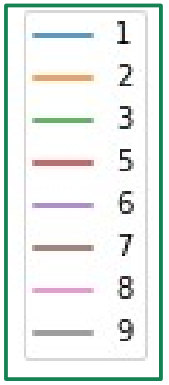
Каждому пользователю может быть сделано несколько предложений в разное время, каждое из которых он может или принять, или отклонить.

Отдельным набором данных будет являться нормализованный анонимизированный набор признаков, характеризующий профиль потребления абонента **(features.csv)**. Эти данные привязаны к определенному времени, поскольку профиль абонента может меняться с течением времени.

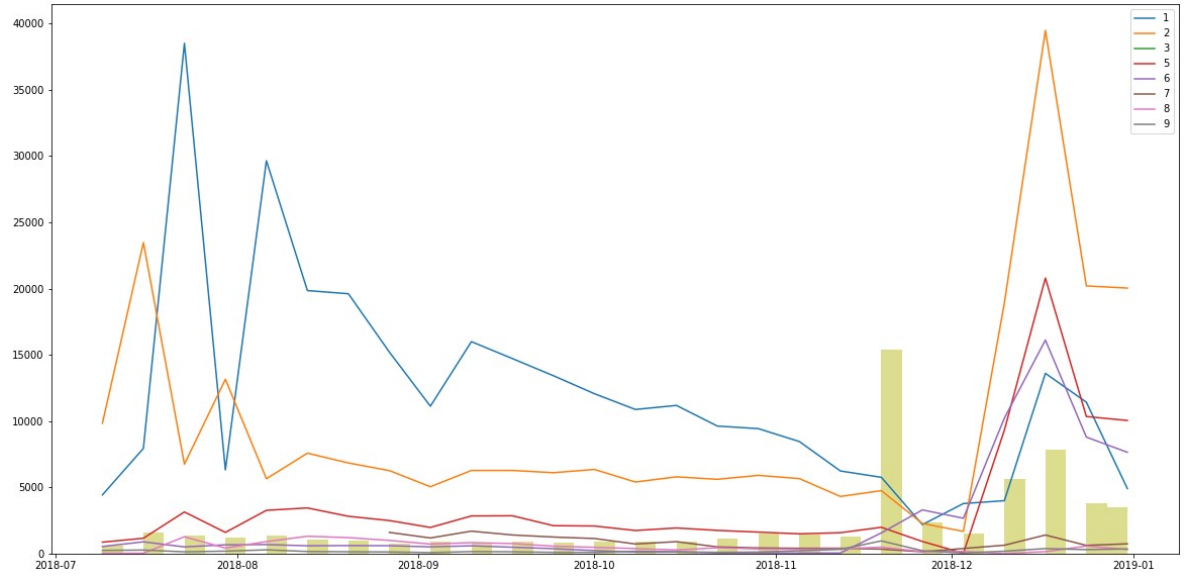
Анализ данных

Из графика следует, что:

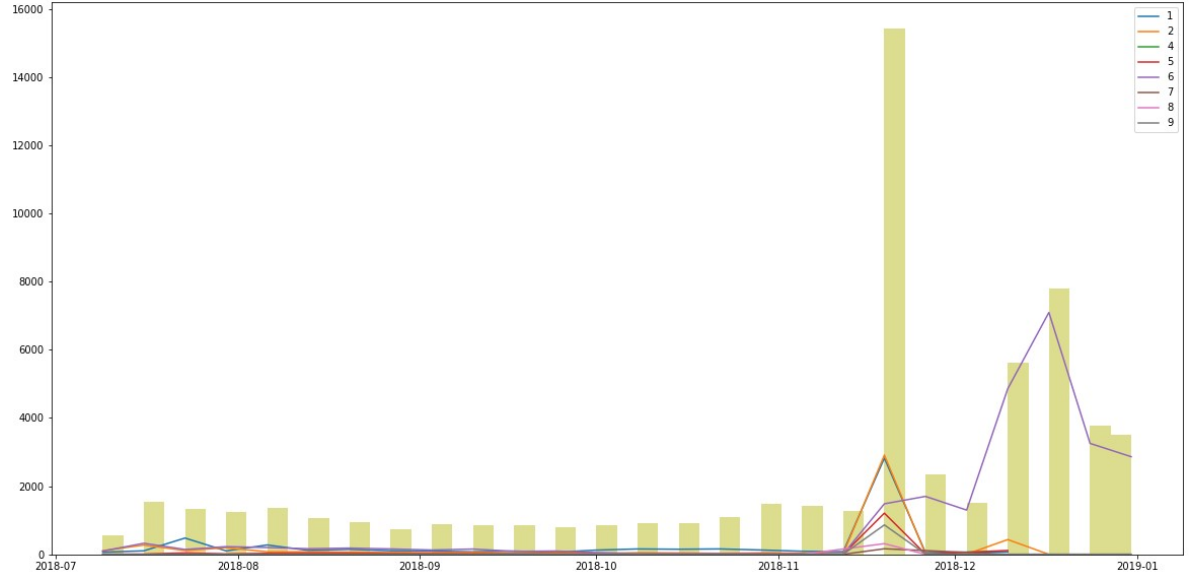
- Чаще всего предлагают услуги **1** и **2**.
- Количество взаимодействий с пользователями постепенно уменьшается, с июля по ноябрь.
- В середине декабря имеется всплеск активности, по всей видимости из-за новогодних акций.
- Количество положительных откликов вплоть до конца ноября практически не меняется.
- **19-10-2018** — большое к-во абонентов подключивших у слуги (видимо сработала акция).



Количество предложений услуг



Количество положительных откликов



Анализ данных

Популярность услуг

В процентах

	vas_id	0	1
0	1.0	0.982	0.018
1	2.0	0.981	0.019
2	4.0	0.746	0.254
3	5.0	0.982	0.018
4	6.0	0.573	0.427
5	7.0	0.986	0.014
6	8.0	0.974	0.026
7	9.0	0.817	0.183

Абсолютное значение

	vas_id	0	1
0	1.0	304511	5664
1	2.0	244708	4797
2	4.0	63991	21765
3	5.0	92393	1692
4	6.0	33174	24704
5	7.0	15219	213
6	8.0	13003	347
7	9.0	4468	1004

vas_id — код услуги

0 – Отклоненные предложения.

1 – Принятые предложения.

Вывод:

Из всего списка услуг наибольшей популярностью пользуются услуги **4** и **6**.

Подготовка данных

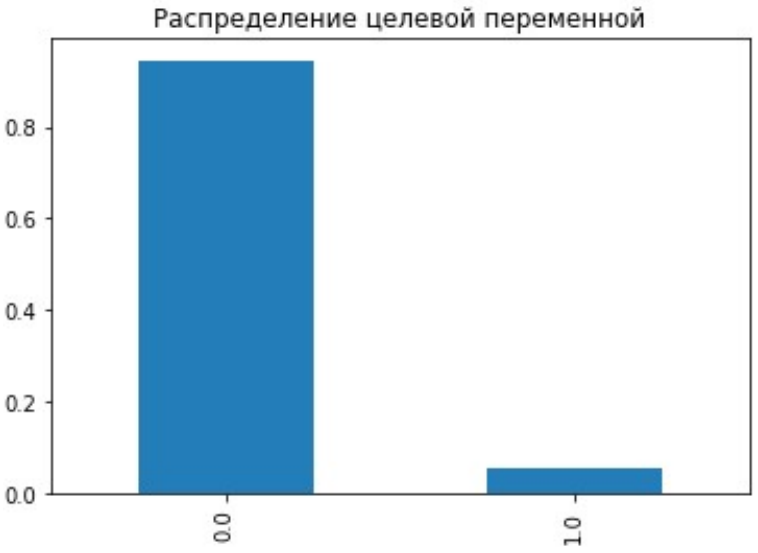
1. Из исходной таблицы с признаками абонентов (features), отсеяны записи по абонентам отсутствующим в data_train/data_test.

2. Выполнено объединение таблиц по id пользователей. При этом учитывался тот факт, в таблице features должна иметься запись о пользователе с датой меньше либо равной записи в таблице data_train/data_test.

3. Данные из обучающего набора разделены на тренировочную и тестовую выборки.

4. Выполнена балансировка данных (UnderSampling) на тренировочной выборке. Т.к. в исходном наборе имеется очень сильный дисбаланс классов.

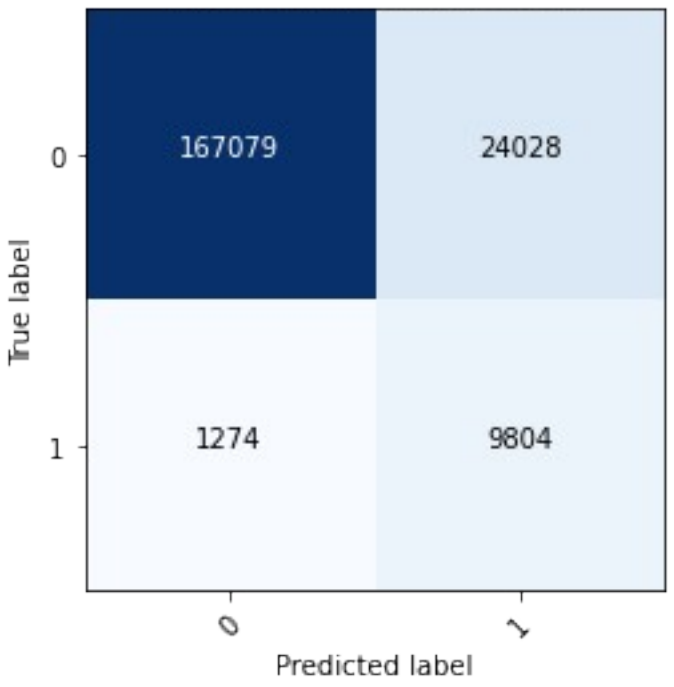
5. Собран pipLine для обучения модели.



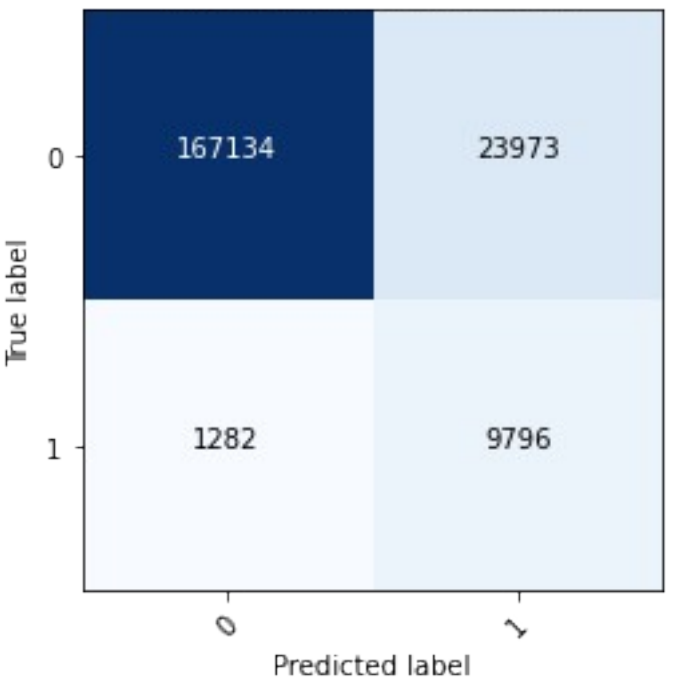
Выбор модели.

В рамках работы было опробовано две модели: Catboost и LigthAutoMl ('lgb_tuned', 'cb_tuned'). Обе модели показали очень близкое качество.

Lama: Confusion matrix



Catboost: Confusion matrix



Выбор модели.

В качестве модели для построения финального предсказания была выбрана catboost. Исходя из того, что при практически одинаковом с lightautoml качестве, catboost обучается быстрее.

Результирующее качество обучения.

	precision	recall	f1-score	support
0.0	0.99	0.87	0.93	191107
1.0	0.29	0.88	0.44	11078
accuracy			0.88	202185
macro avg	0.64	0.88	0.68	202185
weighted avg	0.95	0.88	0.90	202185

Принцип составления индивидуальных предложений для абонентов.

При составлении индивидуальных рекомендаций, к предсказаниям модели следует добавлять ряд бизнес-ограничений.

Например:

- Установить минимальный порог вероятности положительного отклика от клиента. Если вероятность положительного отклика ниже этого порога, то предлагать услугу не стоит.
- Проверить имеется ли уже услуга в списке подключенных услуг абонента.
- Проверить, когда было последнее взаимодействие с клиентом. Если с клиентом уже было неудачное взаимодействие в течение последнего месяца - двух, то предлагать ему новую услугу скорее всего не стоит, т.к. это может уменьшит лояльность клиента.