Видеокурс от Megafon + курсовой проект

Итоговый проект

Курков И.В

Задача.

Поэтому необходимо построить алгоритм, который для каждой пары пользователь-услуга определит вероятность подключения услуги.

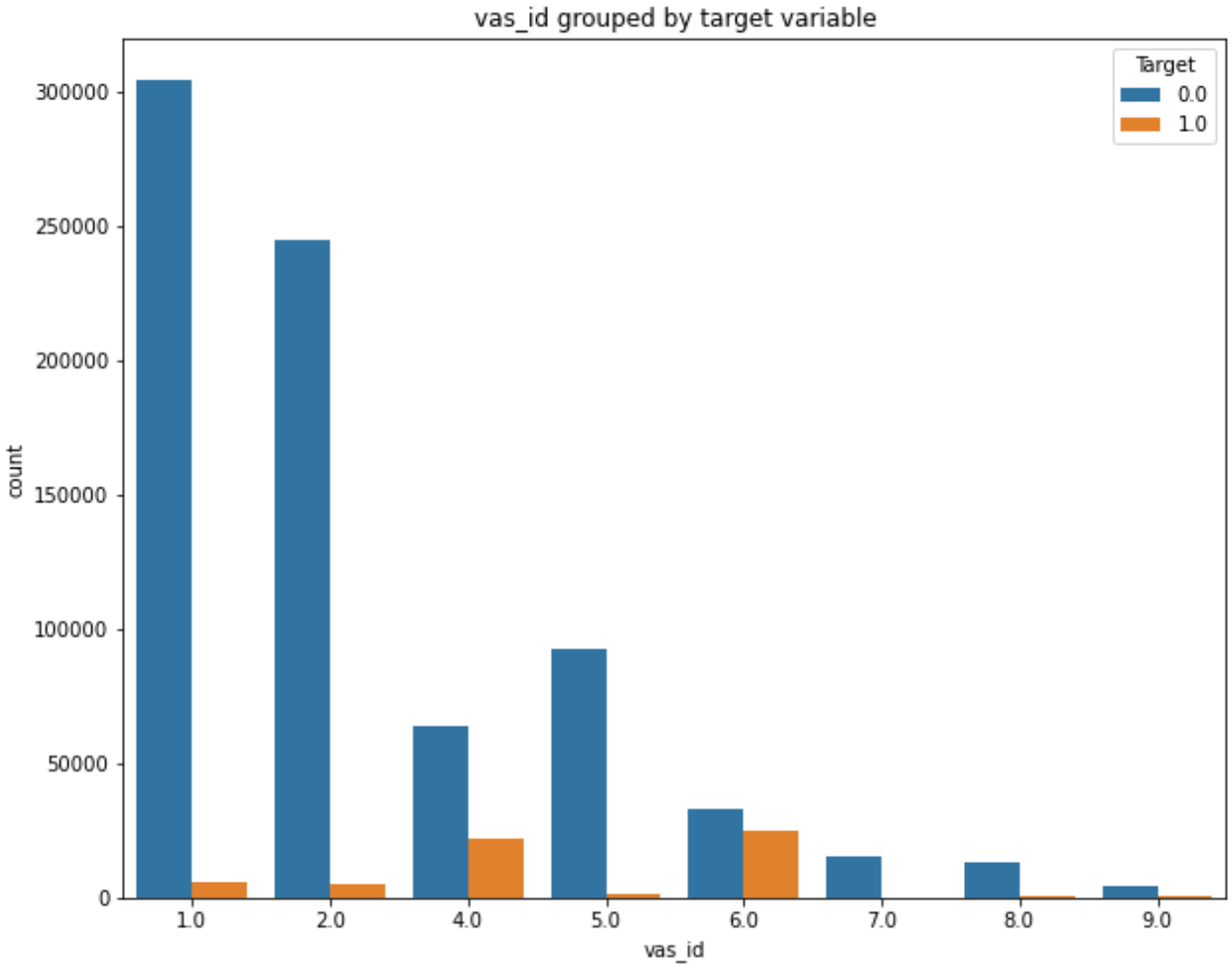
Данные.

В качестве исходных данных вам будет доступна информация об отклике абонентов на предложение подключения одной из услуг. 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. 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.

Метрика.

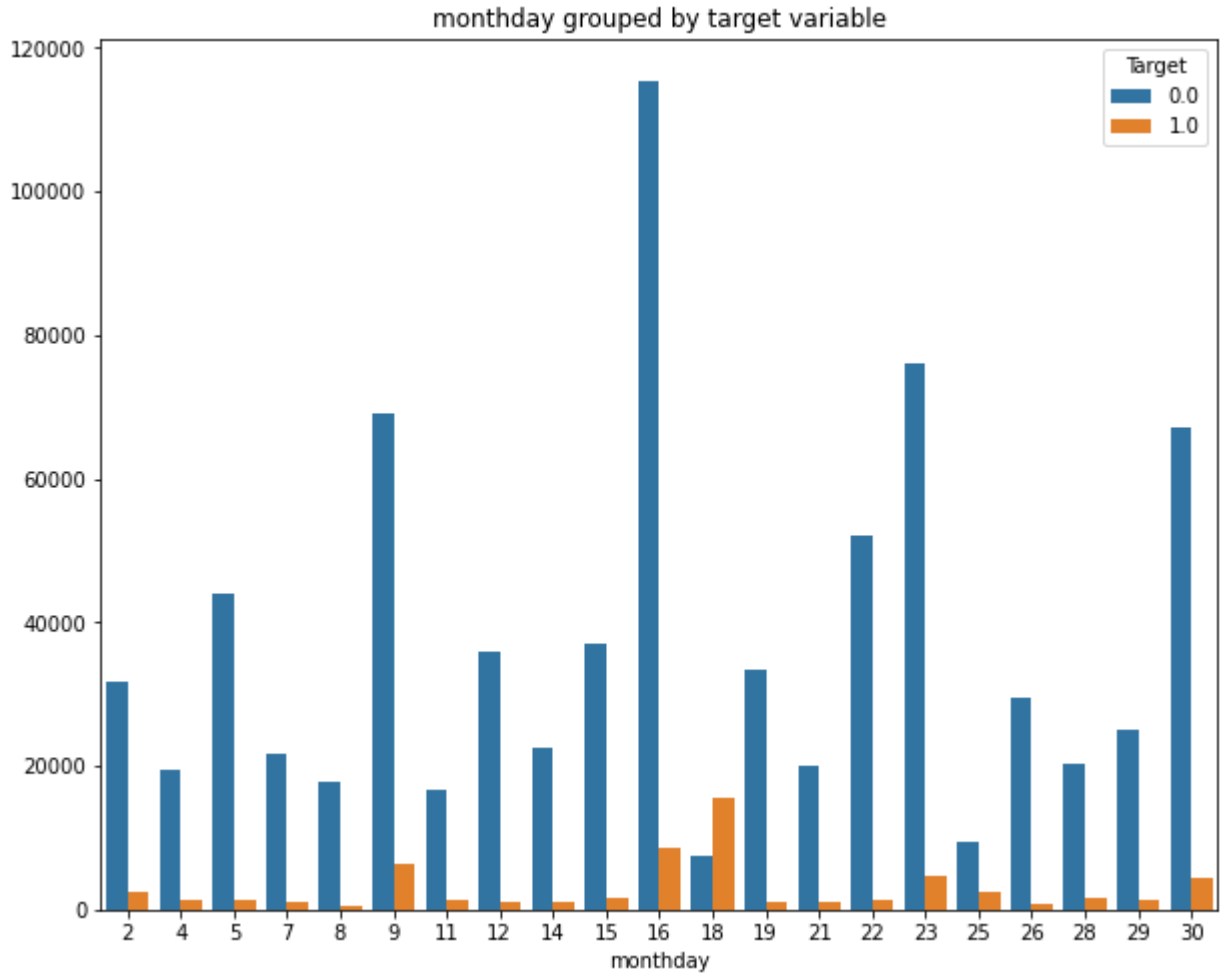
Скоринг будет осуществляться функцией f1, невзвешенным образом, как например делает функция sklearn.metrics.f1_score(..., average='macro').

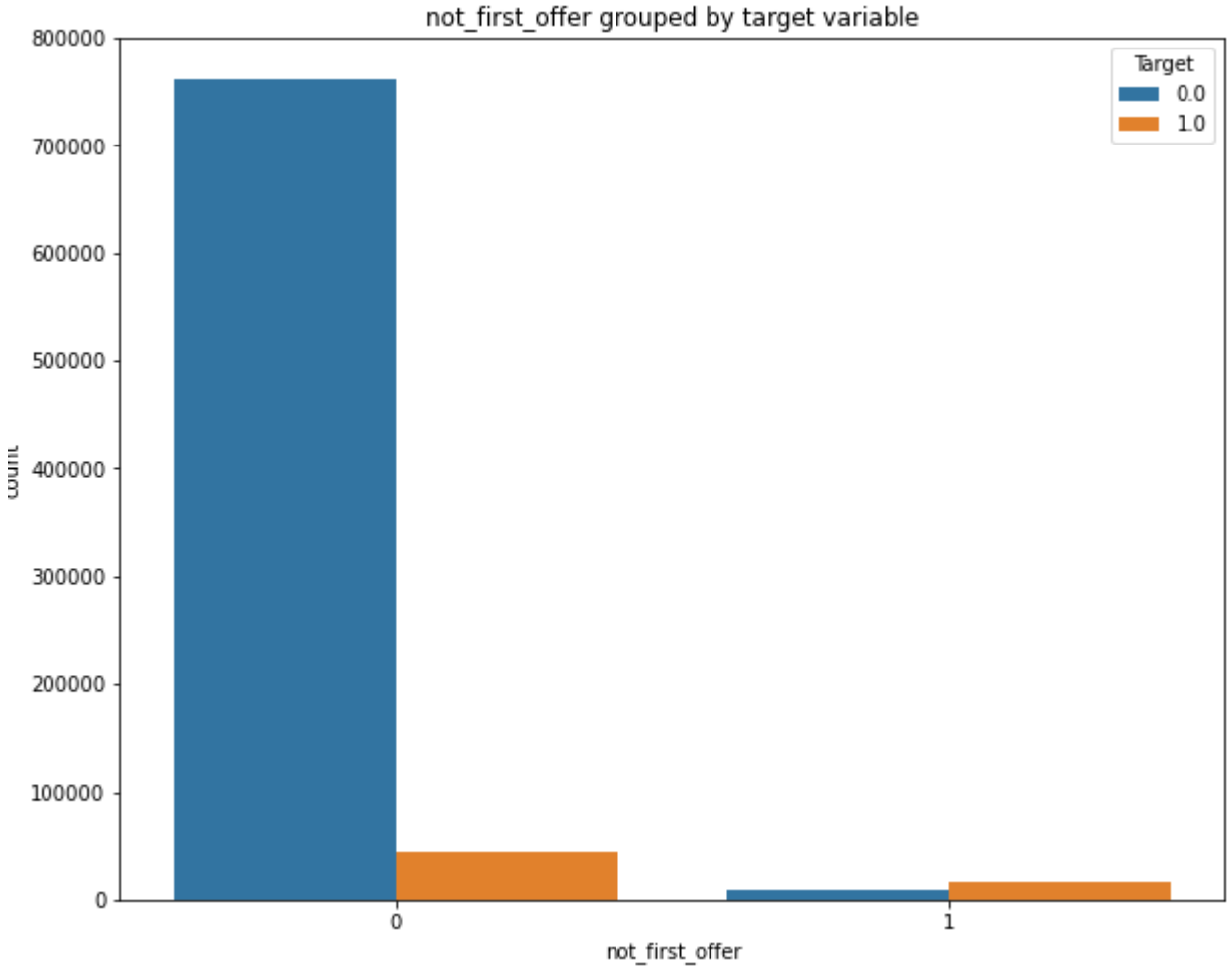
Анализ данных.



Видно, что услуги 4 и 6 подключают с большей охотой. Вероятно эти услуги имеют какие-то выгодные условия для абонента, но не выгодные для компании, может что-то безлимитное, потому как их предлагают в несколько раз реже, чем к примеру услуги 1 и 2, которые не пользуются спросом.

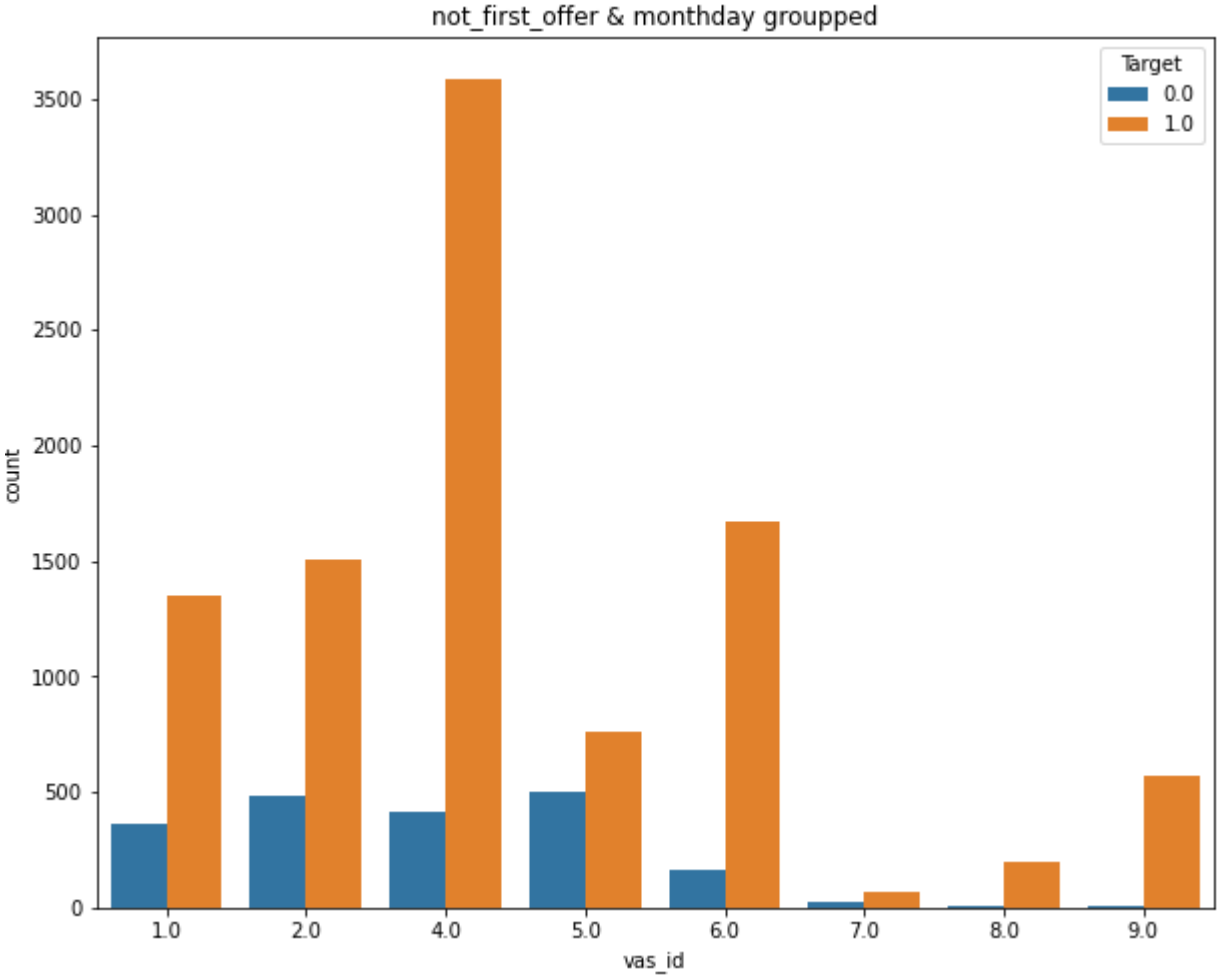
Если выделить из даты день месяца, когда был звонок, то можно увидеть, что в середине месяца видно сильное смещение соотношения отклика, по сравнению с другими днями. Можно сделать предположение, что в середине месяца у людей зарплата, и они более лояльны к приобретению доп. услуг. Т.е. можно сказать, что время звонка имеет большое значение. Также видно, что все звонки были сделаны в 21 час воскресенья. Вероятно это было уже выявлено, как самое удачное время для звонка, но было бы не плохо посмотреть, как например меняется отклик, если звонить в разное время и в разные дни.



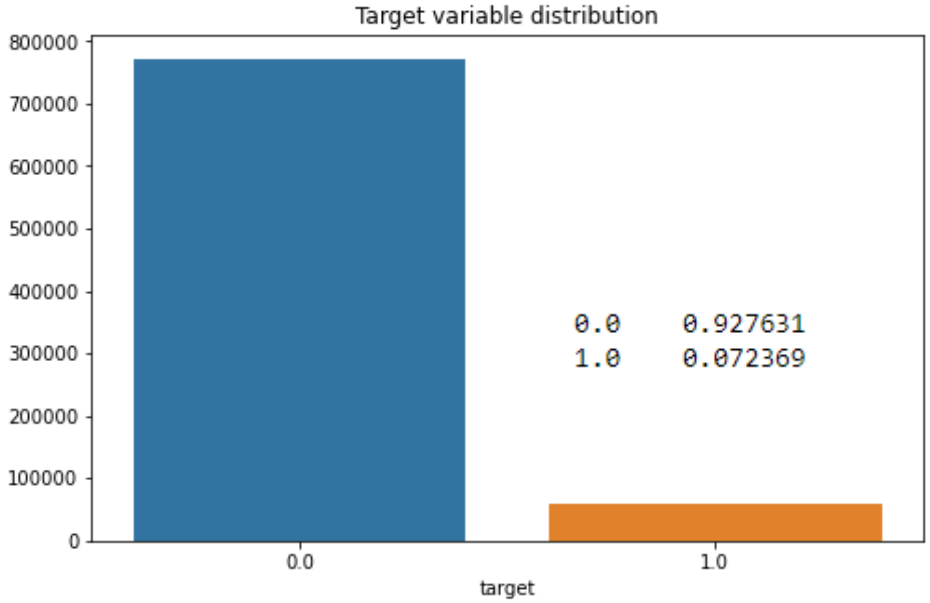


В наборе данных имеются абоненты, которым звонили более 1 раза. Таким образом, если мы пометим те наблюдения, в которых людям звонят второй и более раз, и рассмотрим отдельно группу, которой звонили повторно, то видим, что такие люди подключают услугу намного охотнее. Т.е. за время, пока абонент обдумал нашем предложение, вероятность положительного отклика от повторного звонка выросла. Наблюдение дает повод, в будущем провести исследование на предмет оптимальной паузы между звонками, чтобы вероятность положительного отклика повторного звонка была выше.

И так, если мы выделим в отдельную группу наблюдения с повторными звонками, сделанными в середине месяца, то увидим, что наблюдается большое смещение в сторону положительного отклика, по всем предлагаемым услугам. Из чего можно сделать базовую рекомендацию, всегда звонить повторно в середине месяца.



Выборка наблюдений крайне не сбалансирована. Так как данных много я принял решение использовать undersampling для балансировки классов на трейне. Для разделения на трейн\валид выборки я разделял по времени, и для валидации брал последний месяц из всего сета.



Также посмотрев корреляцию описательных признков, я увидел, что все признаки имеют корреляцию с целевым признаком меньше abs(+-0.01), из чего сделал вывод, что либо у нас не линейная связь, либо признаки – шум. По этому на этапе определения типа признаков я делал идентификацию признаков как категориальные с уникальными значениями в диапазоне 0–30

	target
vas_id	0.262972
target	1.000000
monthday	0.007250
not_first_offer	0.372296
0	0.001181
...	...
241	-0.004290
243	-0.001036
245	0.003425
247	-0.001163
248	-0.001394

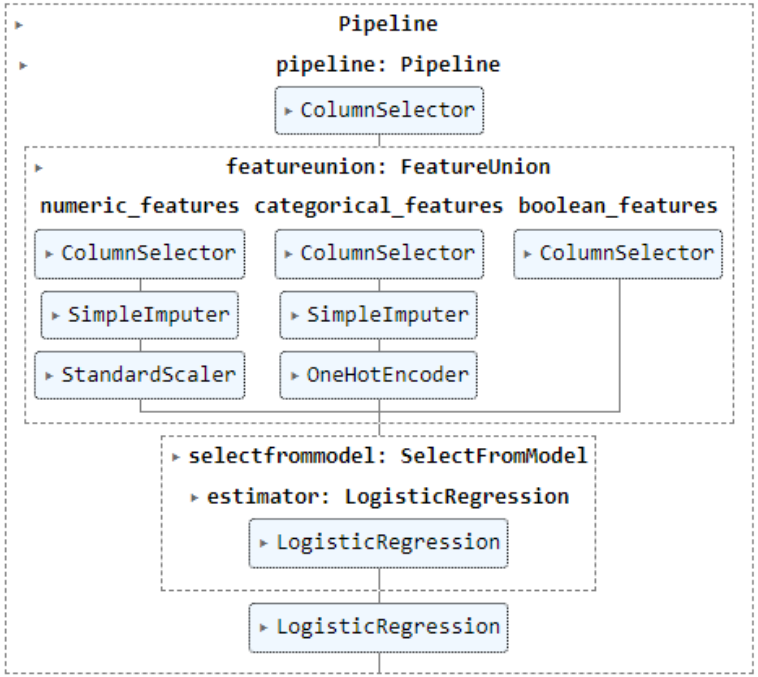
Отбор признаков и выбор модели.

На этапе отбора признаков я попробовал 2 метода – с помощью L1 регуляризации и вычисление уменьшение энтропии. В итоге GS показал лучшие параметры для L1 регуляризации, отобрав 29 признаков.
В процессе подбора моделей я испробовал RandomForestClassifier, LogisticRegression, XGBClassifier, LGBMClassifier, как представителей разных подходов. При этом в процессе тестов, модели показывали очень близкий результат, но из всех простая модель LogisticRegression меньше всех переобучалась, что позволило ей на валидационных данных показывать немного лучший результат. В итоге она и стала фаворитом, в виду еще и более быстрой скорости работы.

Итоговая точность модели на валидационной выборке
F1_macro: 0.7488

Лучшая отсечка : 0.5, Метрика F1_macro: 0.748843780243

	precision	recall	f1-score	support
0.0	1.00	0.87	0.93	178557
1.0	0.40	1.00	0.57	15083
accuracy			0.88	193640
macro avg	0.70	0.93	0.75	193640
weighted avg	0.95	0.88	0.90	193640



Принцип составления индивидуальных предложений для выбранных абонентов

- Звоните в середине месяца.

- Ищите абонентов, которым уже звонили не менее месяца назад.

- Предлагайте другую услугу, нежели ту, что уже предлагали.

- Предлагайте услугу в пакете с 4 или 6 услугой.