В рамках курсового проекта решалась задача классификации: поиск пользователей, которые подключат некую услугу на основании их поведенческого профиля.

В качестве входных данных представлены:

- data_train.csv: id, vas_id, buy_time, target
- features.csv.zip: id, <feature_list>

И тестовый набор:

- data_test.csv: id, vas_id, buy_time

target - целевая переменная, где 1 означает подключение услуги, 0 - абонент не подключил услугу соответственно.

buy_time - время покупки, представлено в формате timestamp, для работы с этим столбцом понадобится функция datetime.fromtimestamp из модуля datetime.

id - идентификатор абонента

vas_id - подключаемая услуга

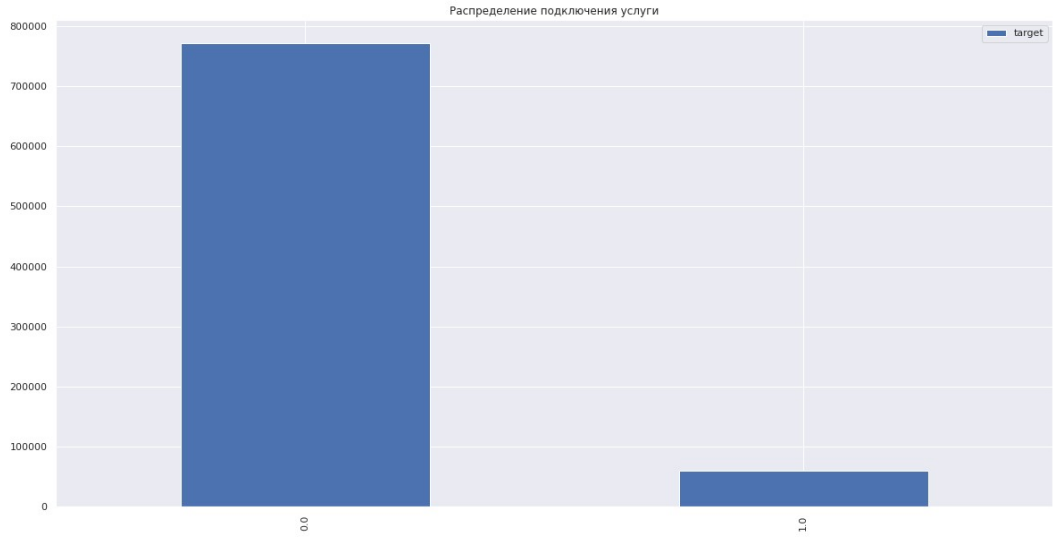
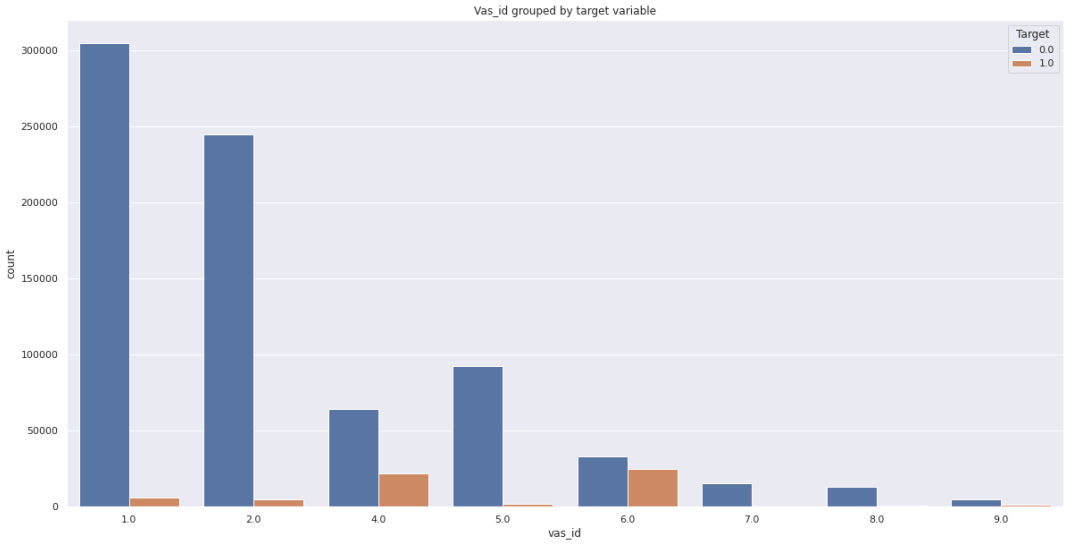
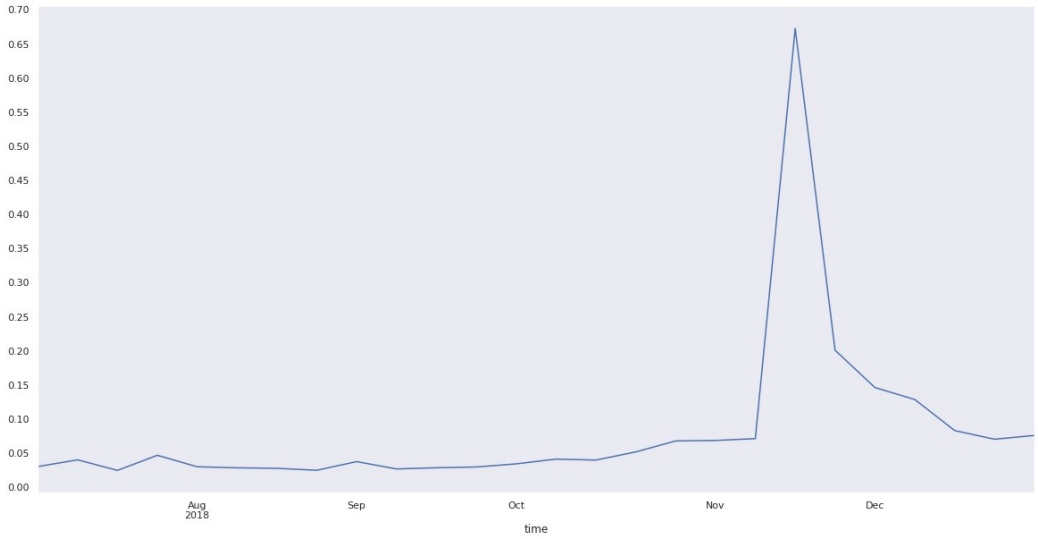


Диаграмма распределения показывает сильный дисбаланс по классам. Следующий график показывает распределение для каждой услуги в отдельности.

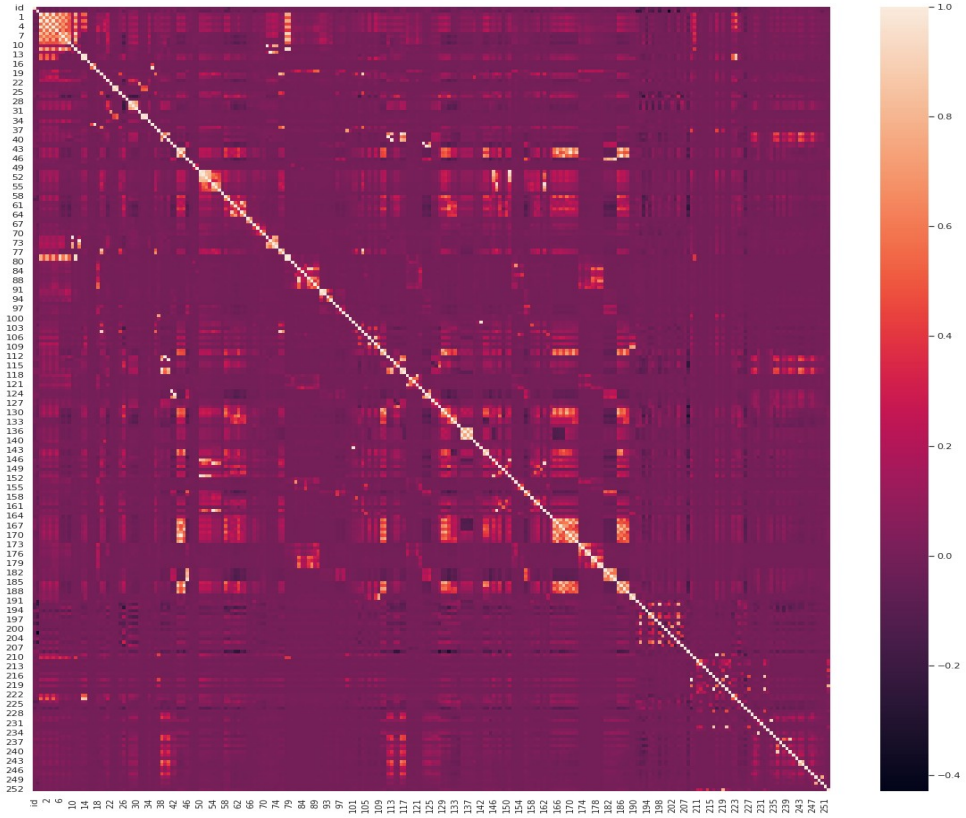
Услуги 1 и 2 предлагались максимальному количеству абонентов, но особым откликом не отличились. В то время как у 4 и, особенно, 6 - имеют значительно больше процент подключившихся. Понятно, что все услуги предлагают разную выгоду абонентам и издержки кампании, и для трактовки, вероятно, не хватает бизнес-информации.



Видно аномально большое количество подключений 19 ноября и плавный спад в последующие даты. Вероятно, была проведена некая акция.
В тестовом периоде предположительно, ничего подобного не было. Возможно, данный “пик” искажает реальную совокупность. Не имея дополнительной информации, не представляется возможным отличить абонентов, подключившихся по акции от тех, кто сделал это независимо. Вероятно, стоит исключить 19-е ноября из обучающей выборки.

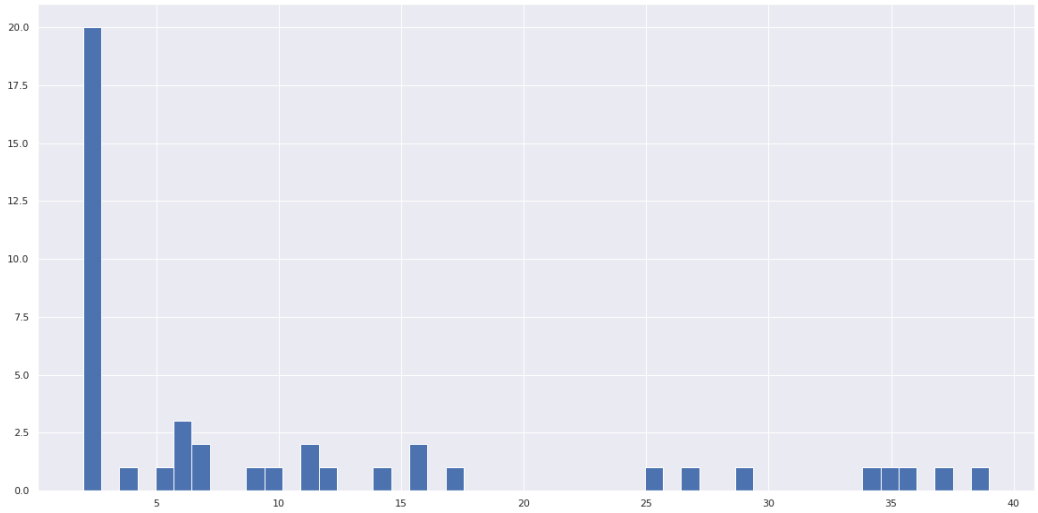


Так же, график показывает положительную динамику — с течением времени все больше пользователей подключали услуги.



По тепловой карте видно, что есть группы признаков с высокой корреляцией. Зачастую коррелирующие признаки идут подряд, что может говорить о том что они частично повторяются, но поскольку данные обезличены, невозможно отобрать наиболее характерные признаки из данных групп. Вероятно, модель будет переобучаться. Необходимо использовать регуляризацию.

Гистограмма по количеству значений получается разряженной и не информативной, поэтому следующий график ограничен справа 50-ю значениями максимум.



В качестве границы между категориальными и вещественными признаками выбрано 10 значений. Это число не окончательно и для более тонкой настройки стоит попробовать другие.

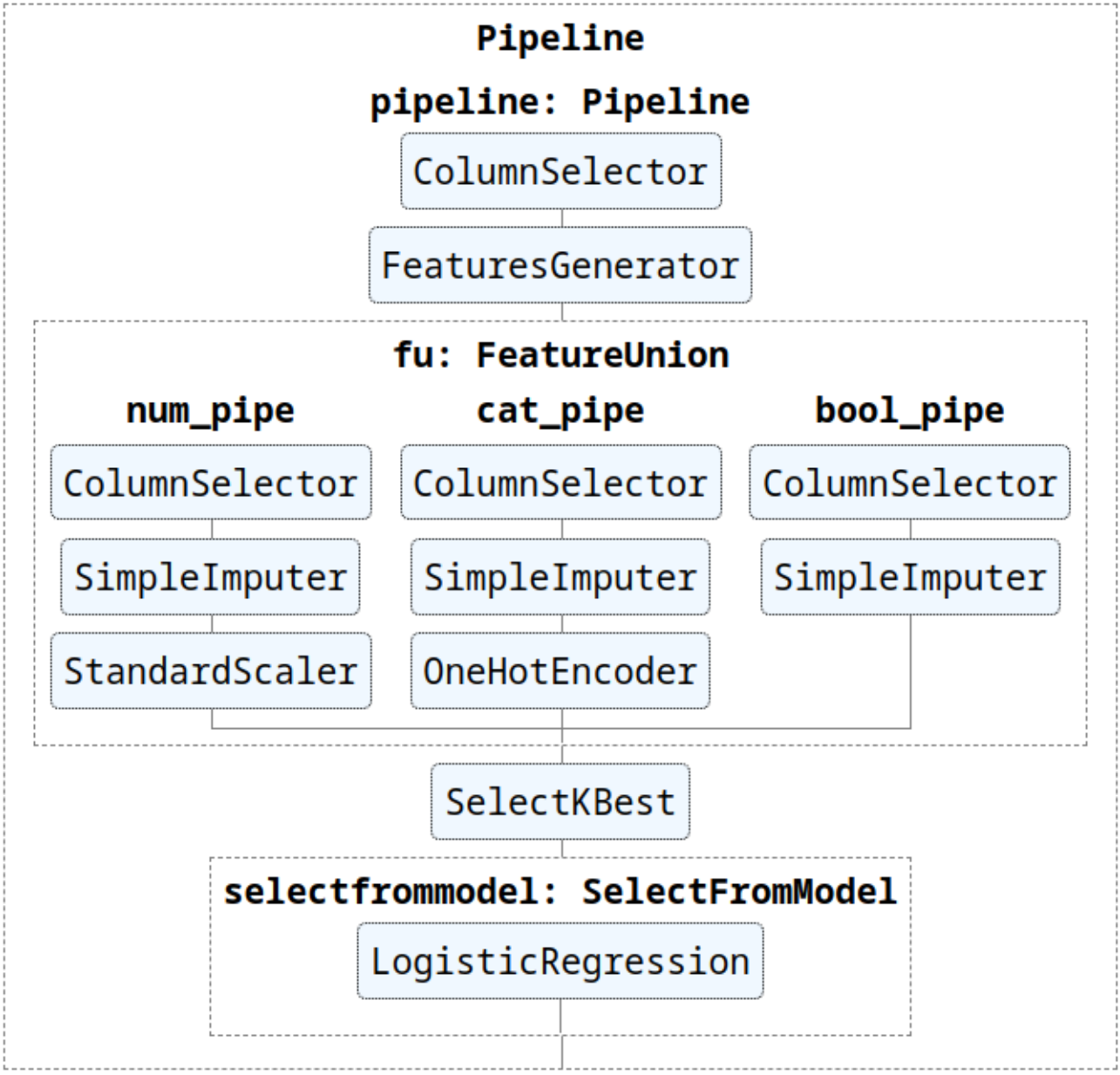
Для обработки булевых, категориальных и вещественных признаков использованы разные стратегии.

- Булевы — пропуски заполнены наиболее часто встречающимся значением.
- Категориальные - пропуски заполнены наиболее часто встречающимся значением, применено one hot кодирование.
- Вещественные — Пропуски заполнены средним. Данные стандартизированы.

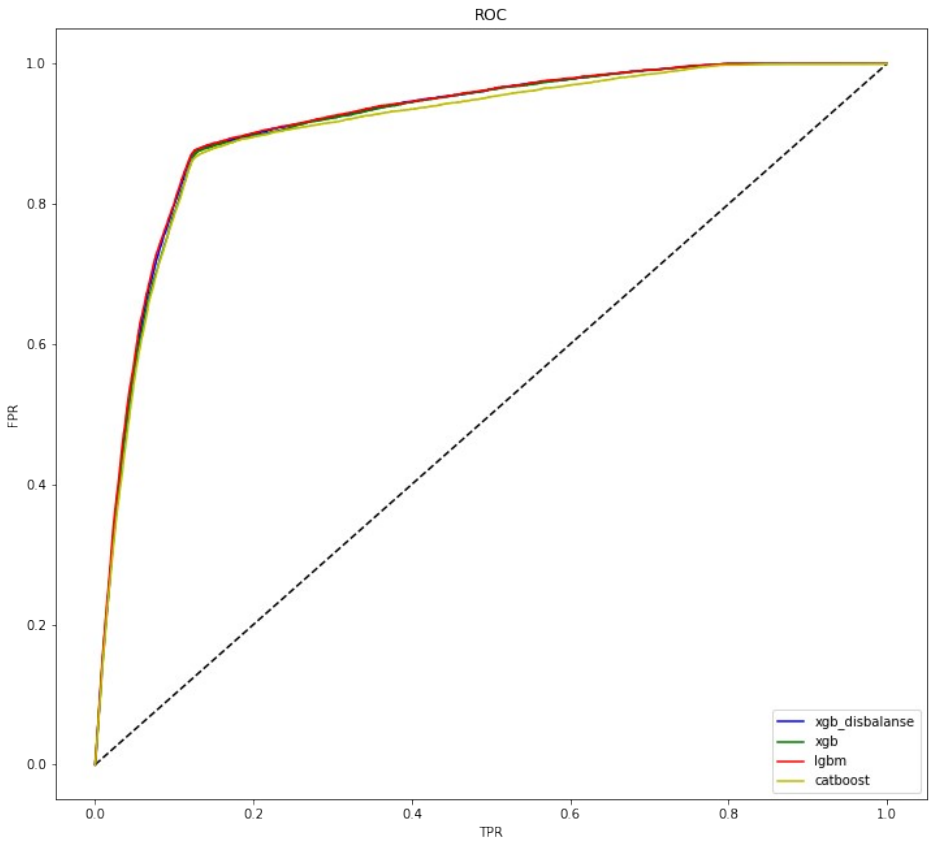
Поскольку вещественных признаков много больше, чем предполагалось оставить в модели, часть из них отсеяна на данном этапе.

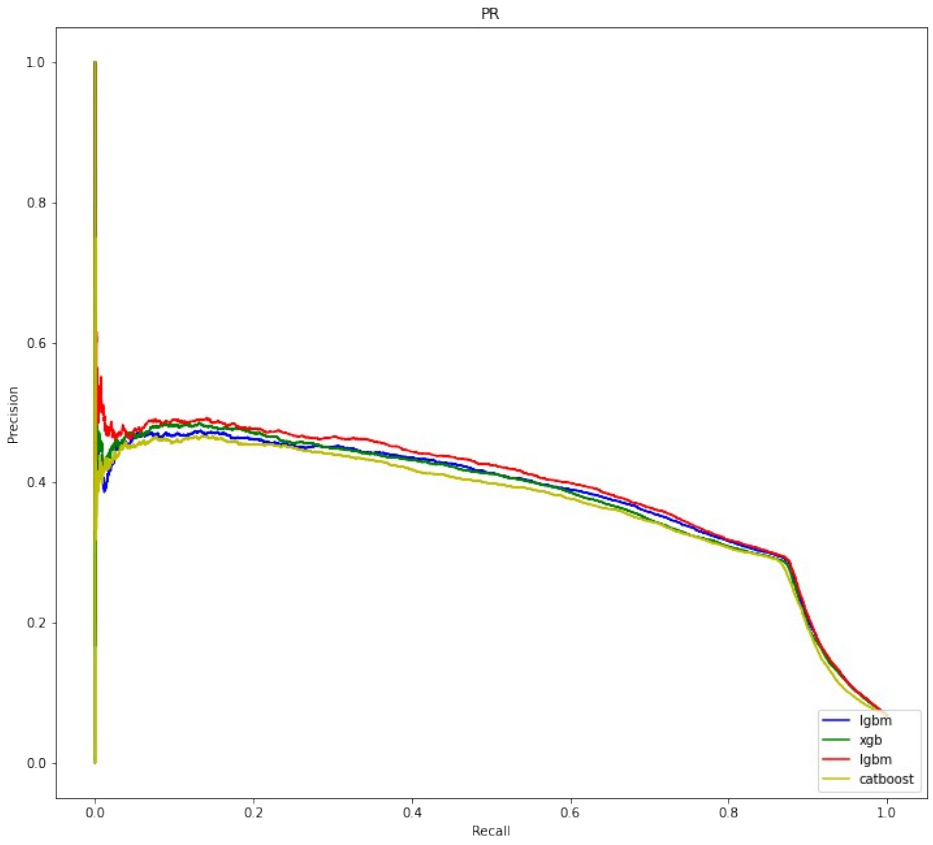
После FeatureUnion были в два этапа удалены малозначащие признаки.

- Первым шагом выбраны 64 признака с помощью SelectKBest
- Вторым — SelectFromModel (использована логистическая регрессия с L1 регуляризацией)



В качестве финальной модели тестировались разные алгоритмы. Лучше всего себя показали бустинговые алгоритмы, а среди них (на базовых настройках) CatBoost. Этот алгоритм был выбран для тонкой настройки, однако разница не велика, и вполне возможно, другие бустинги при тюнинге покажут себя не хуже.


ROC
FPR
TPR
0.0
0.2
0.4
0.6
0.8
1.0
xgb_disbalanse
xgb
lgbm
catboost


PR
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
lgbm
xgb
lgbm
catboost

Гиперпараметры перебирались по сетке. Лучшие значения использовались для обучения финальной модели на всем массиве данных.

Финальный F1 скор 0.684.

Для балансировки классов применялся оверсемплинг, что дало хорошие результаты.

- Во-первых, выросло значение f1.
- во-вторых после балансировки модель показывает значительно больший recall по первому классу.

Учитывая то, что ищем потенциальных клиентов, которые подключат услугу, FP ошибки менее страшны, а TP результаты более интересны для бизнеса. Другими словами, лучше попытаться предложить услугу тому, кто ее не подключит, нежели не предложить тому, кто в ней нуждается.

Согласно ТЗ, предсказания модели были добавлены к другим требуемым колонкам.
Модель показала неплохой результат, выявив (при разбиении на трейн-тест) подавляющее большинство потенциальных клиентов. Однако, есть потенциал для улучшения и некоторые тесты, которые не были проведены из-за срочности. Так, стоит попробовать изменять:

- Количество признаков и методы их отбора.
- Методы заполнения пропусков
- Генерацию новых фич (затруднено из-за обезличенности данных)
- Границу между «вещественными» и «категориальными» признаками.
- Другие алгоритмы (напр. LGBM и XGB) в качестве модели для подбора параметров.
- Другие гиперпараметры и (или) их значения при переборе по сетке.