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

Итого, в качестве входных данных будут 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

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

• Булевы — пропуски заполнены наиболее часто встречающимся значением.

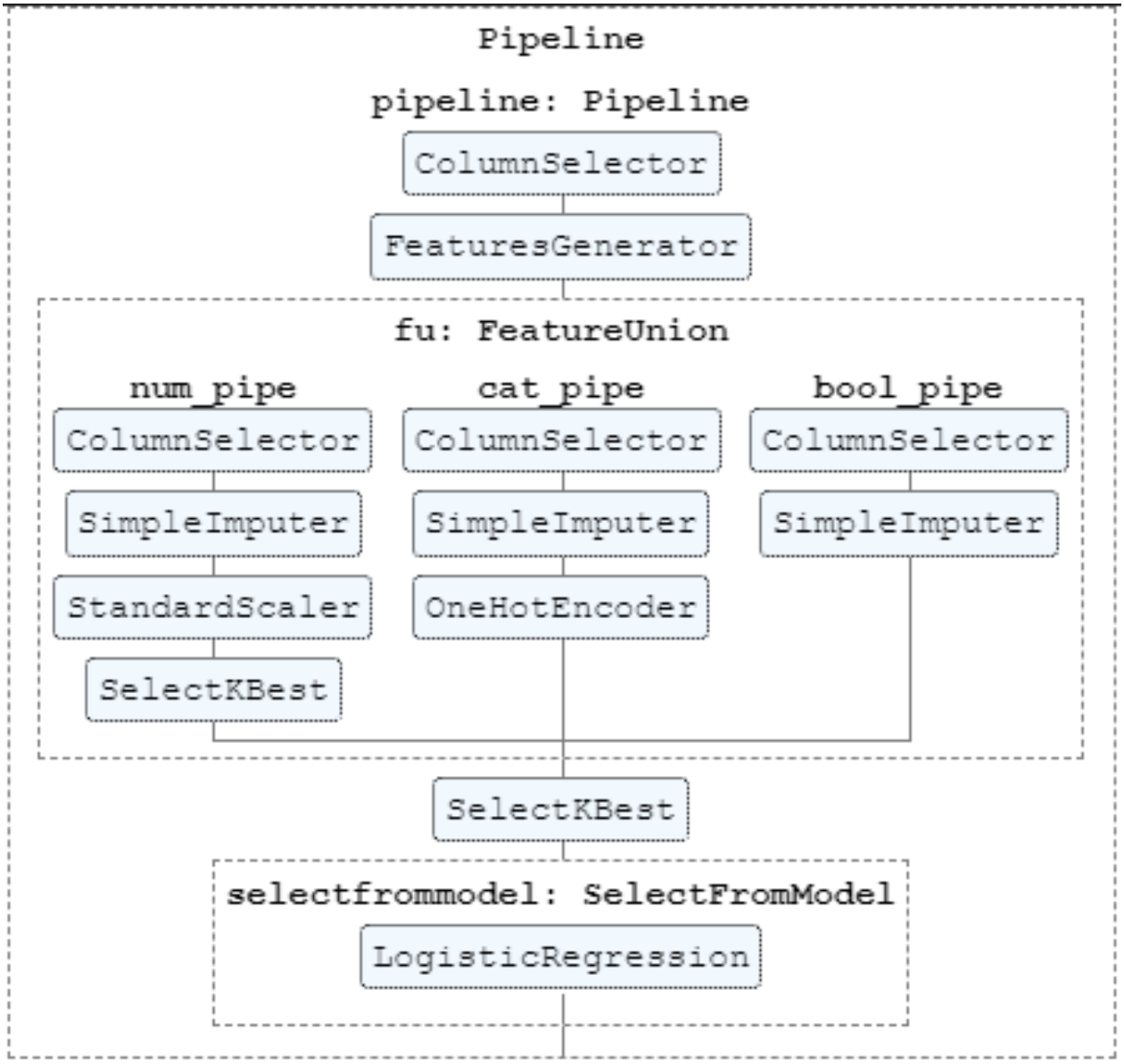
• Категориальные - пропуски заполнены наиболее часто встречающимся значением, применено one hot кодирование.

• Вещественные — Пропуски заполнены средним. Данные стандартизированы. Поскольку вещественных признаков много больше, чем предполагалось оставить в модели, часть из них отсеяна.

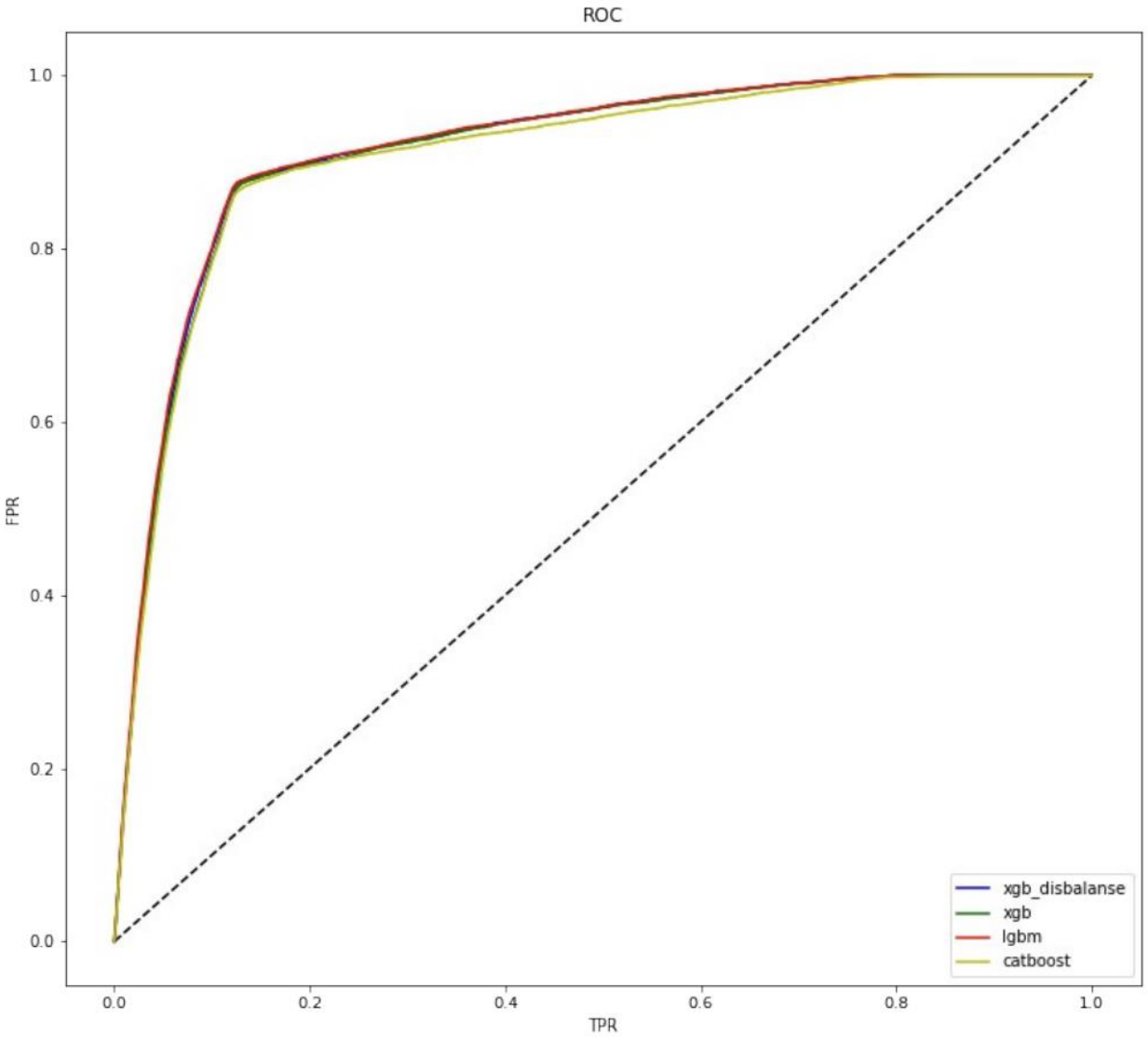
После FeatureUnion были в два этапа удалены малозначащие признаки.

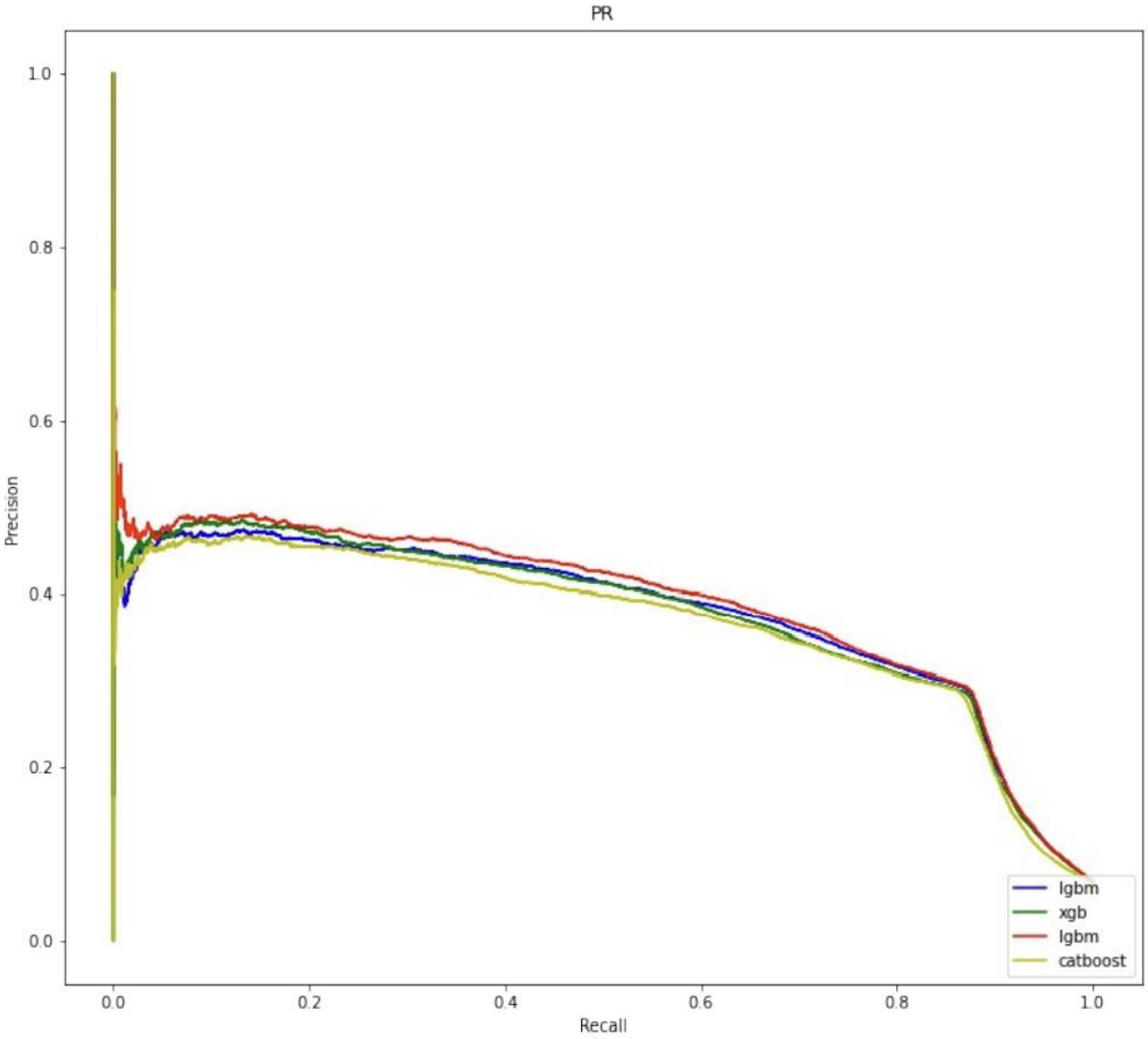
• Первым шагом выбраны 64 признака с помощью SelectKBest

• 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
1.0
0.8
0.6
0.4
0.2
0.0
FPR
0.0
0.2
0.4
0.6
0.8
1.0
TPR
xgb_disbalanse
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.