Отчет о проведенном исследовании по удержанию клиентов, склонных к оттоку.

Цели и задачи проекта

Предотвращение оттока пользователей – одна из самых важных задач на текущий момент в сфере телекоммуникационных услуг и иных сферах, где насыщенность рыка составляет 100%.

В конкретной задаче мы имеем данные о клиентах, значение которых мы не знаем, 230 категорийных и числовых признаков. Это задача бинарной классификации, где отток – класс 1, а не отток – класс 0. По результатам работы классификатора компания сможет определять клиентов, склонных к оттоку, и проводить мероприятия для их удержания.

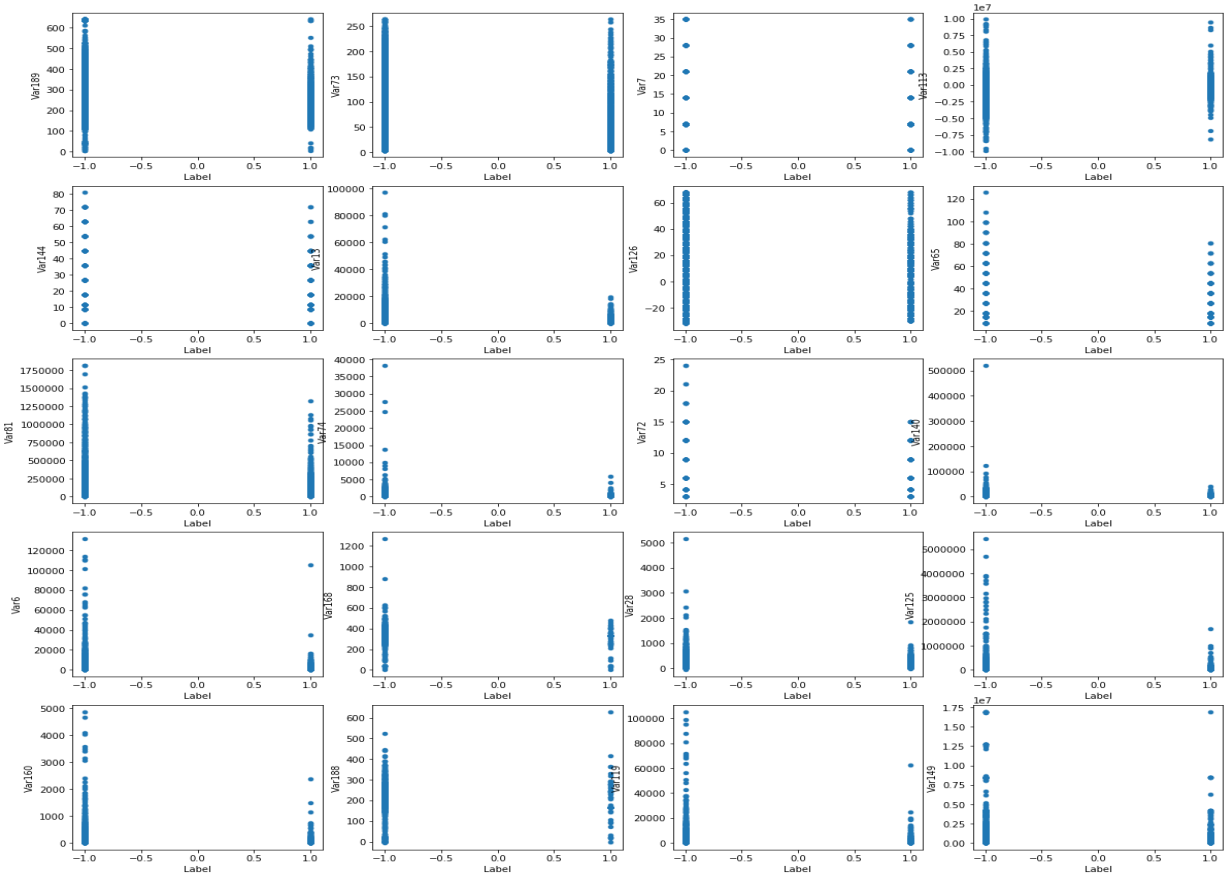
Выборка сильно несбалансированна – 7.44% объектов класса 1 и 92.56% объектов класса 0.

Методика измерения качества

Из-за дисбаланса классов основной метрикой качества будет ROC-AUC. Модель лучше тестировать на AB-тесте, и затем измерить экономический эффект от её внедрения. Считаю, что успешной её можно считать, если будет достигнут значительный экономический эффект (напр., 50% пользователей решило остаться), и ROC-AUC на тестовой выборке не ухудшится.

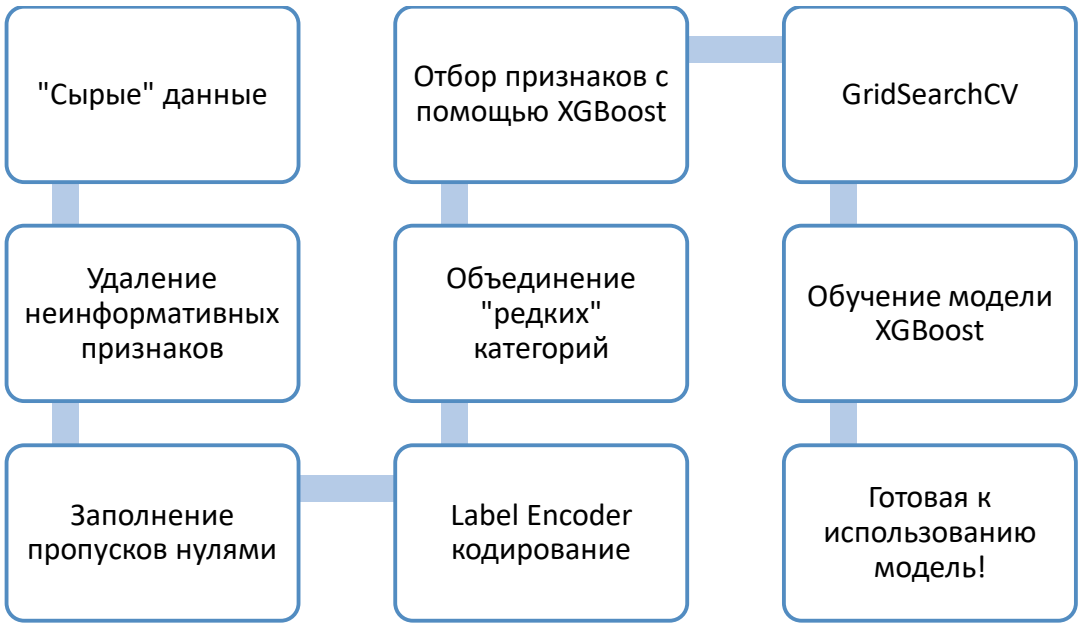
Техническое описание решения

Мы имеем большое количество признаков – 230 шт., и во многих категорийных признаках гигантское количество различных категорий. На рисунках сложно что-то понять, и даже при ближайшем рассмотрении не удалось выявить интересных закономерностей. Например, вот распределения топ-20 числовых переменных, наиболее сильно коррелирующих с целевой функцией:



Также в данных было большое количество пропусков. Было решено выкинуть признаки, в которых пропусков было более 75%, оставшиеся заполнить средними в числовых и новой категорией ‘no_value’ в категориальных. Также сократить количество категорий в категориальных признаках – если категория в признаке «редкая» - повторяется менее 50 раз – все такие категории объединяем в одну. Затем закодировали Label Encoder’ом категориальные признаки, а вещественные оставили как есть – в экспериментах такие преобразования показали себя лучше всего.

Далее провели отбор признаков с помощью XGBoost классификатора, в результате чего осталось 28 информативных признаков. И в конце концов на очищенных данных настроили с помощью GridSearchCv гиперпараметры GradientBoosting из библиотеки XGBoost, т.к. он показал наилучшие результаты.



Данный пайплайн показался наилучшим и был выбран из следующих возможных вариантов:

1. Балансировка выборки: undersampling, XGBoost веса
2. Заполнение пропущенных значений: среднее, медиана, нули
3. Обработка категориальных признаков: LabelEncoder, Dummy-кодирование
4. Регуляризация: Lasso, XGBoost
5. Классификаторы: Lasso, LogisticRegression, RandomForest, XGBoost

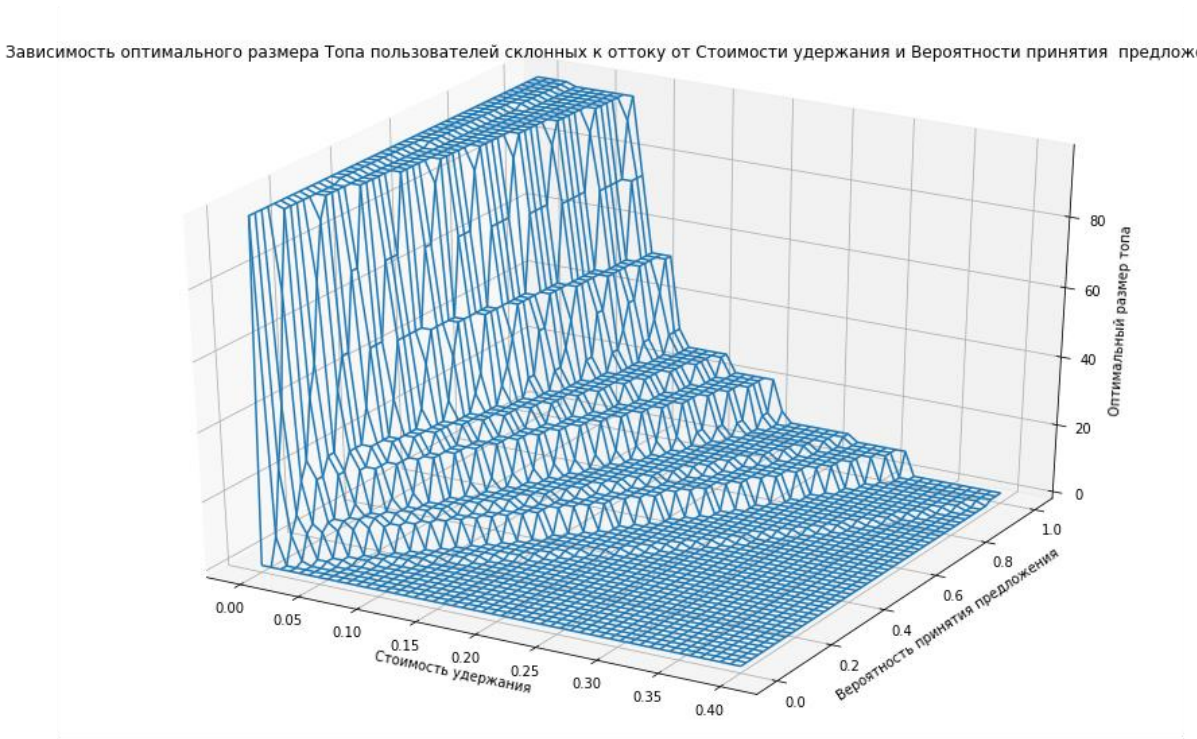
Качество модели

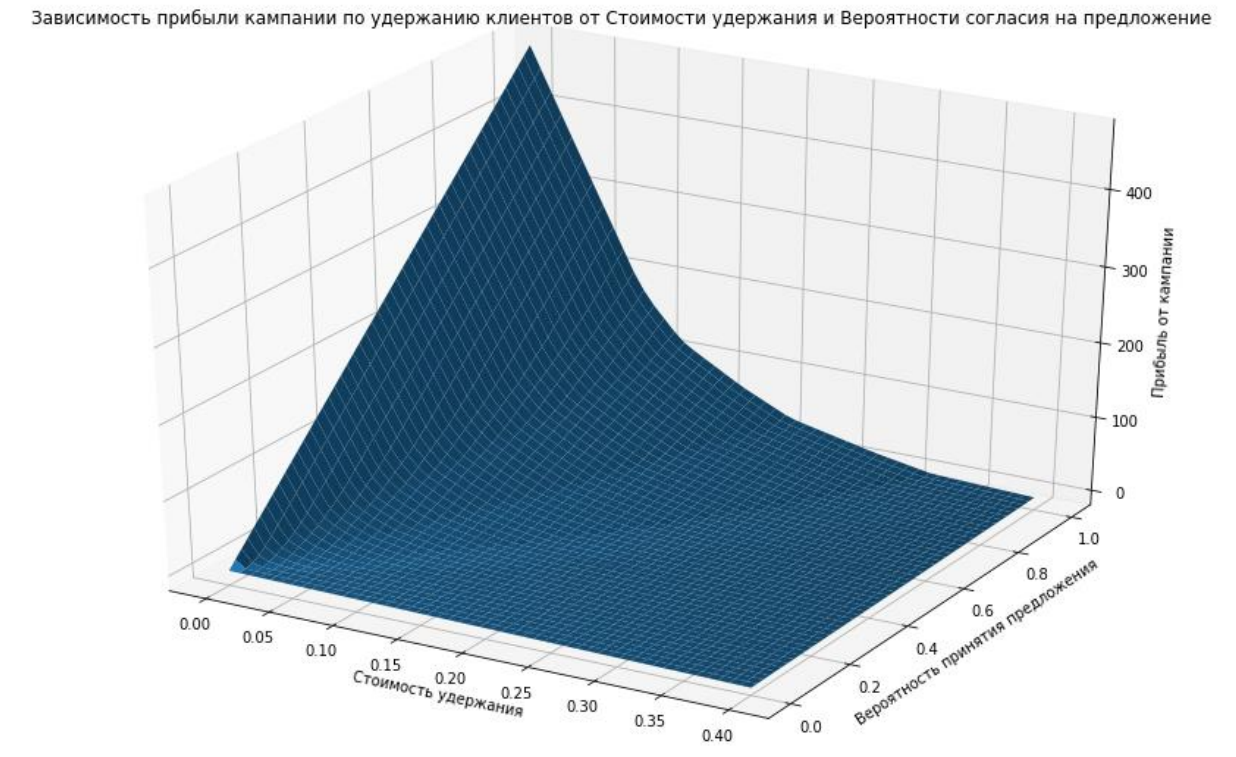
На отложенной выборке модель показала ROC-AUC = 0.74

Нижеуказанные признаки наиболее важны для построения модели:

```
('Var126', 0.21116506),
 ('Var218', 0.10436893),
 ('Var189', 0.07524272),
 ('Var73', 0.05582524),
 ('Var113', 0.050970875),
 ('Var81', 0.048543688),
 ('Var74', 0.03883495),
 ('Var205', 0.033980582),
 ('Var199', 0.029126214),
 ('Var192', 0.029126214)
```

Модель, несомненно, будет иметь экономический эффект. При её грамотном применении возможно найти более 92% пользователей, склонных уйти, и, проведя мероприятия по их удержанию, получить большую прибыль, чем при отсутствии каких-либо действий. Был проведен эксперимент, результаты которого видны на графиках:





Главные параметры при оценке экономического эффекта: стоимость удержания и вероятность принятия предложения клиентом.

Выводы

Применение данной модели экономически выгодно и позволит сохранить сотни тысяч, а может и миллионов рублей в условиях перенасыщенного рынка, даже при условии, что не все клиенты согласятся на то, чтобы их удерживали, привлечение одного нового стоит намного дороже, чем удержание старого. Данная модель весьма точна, охватывает почти всех клиентов склонных к оттоку, а при улучшении качества прогноза (recall) на 1% дает оценку увеличения прибыли кампании по удержанию клиентов более 2%.