Финальный проект. Прогнозирование оттока пользователей.

Суть задачи заключается в заблаговременном нахождении сегмента пользователей, склонных через некоторый промежуток времени отказаться от использования некоторого продукта или услуги. Точное и своевременное нахождение таких пользователей позволяет эффективно бороться с их оттоком, например, выявлять причины оттока и принимать меры по удержанию клиентов. Эта задача актуальна для большинства организаций, оказывающих услуги в сегменте B2C и вдвойне актуальна в областях, где распространение услуги близко к отметке 100%.

В задании используются исходные данные из соревнования KDD Cup: Customer relationship prediction (2009). Данные для соревнования были предоставлены французской телекоммуникационной компанией Orange. В задаче речь идет о клиентских данных, поэтому данные были предварительно обфусцированы и анонимизированны: из датасета убрана любая персональная информация, позволяющая идентифицировать пользователей, а также не представлены названия и описания переменных, предназначенных для построения прогнозов. Мы будем работать с набором данных orange small dataset. Он состоит из 50 тыс. объектов и включает 230 переменных, из которых первые 190 переменных - числовые, и оставшиеся 40 переменные - категориальные.

Качество модели оценивалось с помощью метрики ROC-AUC, так как данные несбалансированные, а эта метрика не чувствительна к дисбалансу классов и хорошо подходит для вероятностных моделей. Также отслеживались метрики Precision и Recall и F1 score. Для кросс-валидации была выбрана стратегия StratifiedKFold(n_splits=5, shuffle=True). Первая построенная модель плохо определяла целевой класс, но показала лучшую оценку на соревновании в Kaggle. Но так как мы пытаемся решить реальную, а не соревновательную задачу, было принято решение повысить метрику Precision, хороший результат был получен с использованием Undersampling’a. Далее будет отражена полная схема построения «пайплайна».

Обработку данных проводили следующим образом:

- Были удалены все вещественные признаки, для которых пропущено более 30% значений
    - Оставшиеся пропущенные значения заполнены средним значением
    - Стандартизация вещественных признаков
- Категориальные данные были разделены на 2 группы, в 1-й число уникальных значений признака менее 30, а во второй где больше ю
    - Пропущенные значения заполнены переменной «unknown»
    - Для первой группы OneHotEncoding, для второй – BinaryEncoder
- Undersampling. Сбалансируем данные откинув случайным образом наблюдения большего класса
- С помощью Lasso откинем незначимые признаки

Блок схема «пайплайна» показана на рисунке 1

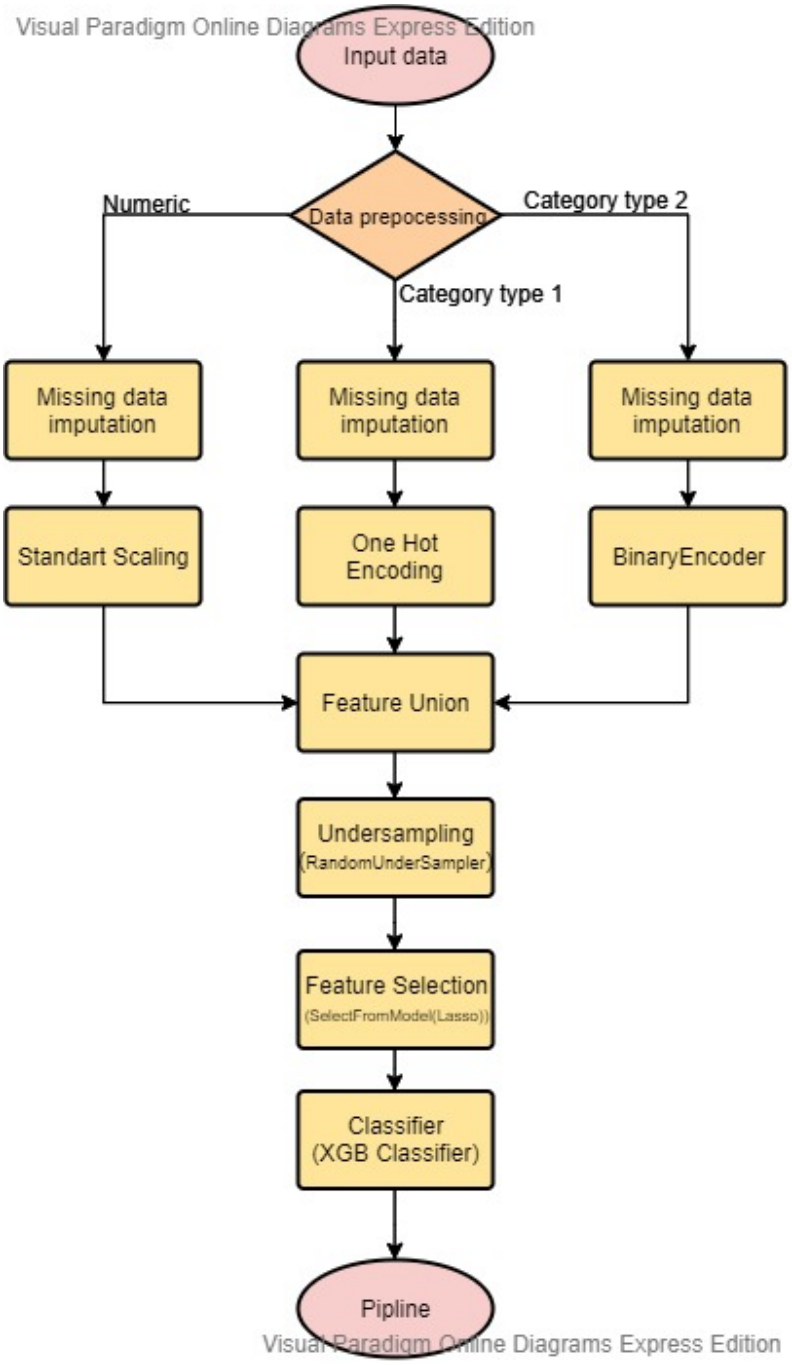

Рисунок 1

На последней неделе курса была построена экономическая модель для оценки эффекта от внедрения полученного решения на практике. Идея использования предсказательной модели пользователей склонных к оттоку заключается в предложении скидки тарифной оплаты пользователям, вероятность оттока которых, выше заданного порога. На рисунке 2 представлена зависимость чистой прибыли, от порога вероятности, с которого начинается скидочная программа.

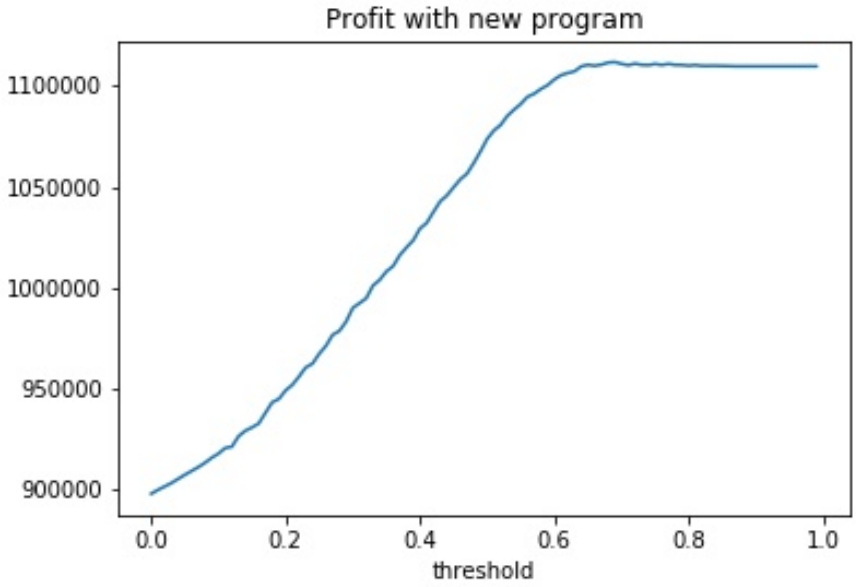

Рисунок 2

В работе был вычислен оптимальный порог, он равен 0.71. При таком пороге Общая прибыль компании увеличится на 0.2% от общей прибыли. При увеличении качества модели на 1% и 3%, рост прибыли окажется 0.07% и 0.29% соответственно. Но прежде чем инвестировать в улучшение модели, я бы предложил провести А/Б тестирование на 5% случайно выбранных пользователях с целью проверки предсказательной и экономической модели.

Сначала нужно определиться со временем теста, с какого времени можно начинать оценивать значимость изменений, приведенных моделью. В нашем проекте данные анонимизированны и нет никаких временных меток. В реальности скорее всего есть события с временной меткой, после действия которых происходит отток пользователя. Необходимо вычислить данное время и примерно через 5 таких временных интервалов после введения скидочной программы можно начать отслеживать среднюю ежемесячную прибыль в двух группах. Параллельно необходимо смотреть на адекватность экономической модели, выполняются ли принятые допущения. Например, в моей модели было выдвинуто предположение о том, что вероятность изменения решения прямо пропорциональна размеру скидки.