Прогнозирование оттока клиентов

Отчет по проекту

Марковский Олег, 5 сентября 2017

Цели и задачи проекта

Выстраивание взаимоотношений с клиентами — чрезвычайно важная задача для любых коммерческих фирм. Особенно актуален этот вопрос в условиях жёсткой конкуренции. Примером может служить телекоммуникационная отрасль, предоставляющая долгосрочные услуги своим клиентам. Зачастую, гораздо дешевле, оказывается, проводить ряд акций по удержанию клиентов реклама , новые тарифы, персональные услуги. Для этого компания должна определять ,какие из клиентов, приносящих большую прибыль, скорее всего откажутся от услуг. Для того, чтобы находить пользователей, склонных к оттоку, строят прогнозные модели - модели, позволяющие прогнозировать вероятность того, что пользователь покинет сервис. В классической постановке строятся вероятностные модели бинарной классификации, где целевой класс представляют собой пользователи, покидающие сервис. Вероятность того, что пользователь принадлежит целевому классу и есть целевая величина - вероятность оттока. Цель проекта - научиться находить пользователей, склонных к оттоку.

В рамках проекта требуется по исходным данным построить модель, позволяющую предсказывать клиентов склонных к оттоку с определенной точностью, а также предложить метод ее использования и оценить экономический эффект.

Описание исходных данных:
- Данные предоставлены французской телекоммуникационной компанией Orange
- Датасет состоит из 50 тыс. объектов и включает 230 переменных, из которых первые 190 переменных - числовые, и оставшиеся 40 переменные - категориальные. Названия и описания переменных, предназначенных для построения прогнозов отсутствуют.
- Метки классов: 1 соответствует классу отток, -1 - классу не отток.
- Признаки обозначены как Var1, Var2, … , Var230.
- Количество объектов в наборе данных: 40000
- Доля класса «отток»: 0.0744
-Доля класса «не отток»: 0.9256

Методика измерения качества и критерий успеха.

Основной методикой измерения качества будем считать площадь под roc-кривой (AUC-ROC)
В качестве вспомогательных метрик используем
-точность;
-полноту;
- f-меру.
Т.к. выборка несбалансирована, будем ориентироваться на AUC-ROC .Критерием успеха будем считать , если на тесте удалось добиться AUC-ROC >0,7, также будем одновременно пытаться максимизировать F-меру.
Оценивать качество модели будем по кросс-валидации. Взято 70% данных для проведения кросс-валидации, а оставшиеся 30% зарезервированы в качестве отложенной выборки. Для кросс-валидации были использована стратегия stratified k-fold, чтобы обеспечить в каждом фолде такую же долю классов, как и в исходном наборе данных. Количество фолдов = 3. В качестве оценочного значения каждой из метрик применялось мат. ожидание метрики на всех фолдах.
Качество алгоритмов, хорошо показавших себя во время кросс-валидации, дополнительно проверялось на отложенной выборке. После сравнения основной метрики проводилось дополнительная корректировка параметров модели.
Такая стратегия позволила компенсировать переобучение и взвешенно оценивать качество алгоритма. Указанная ниже модель классификатора обладает лучшим качеством согласно данной стратегии.

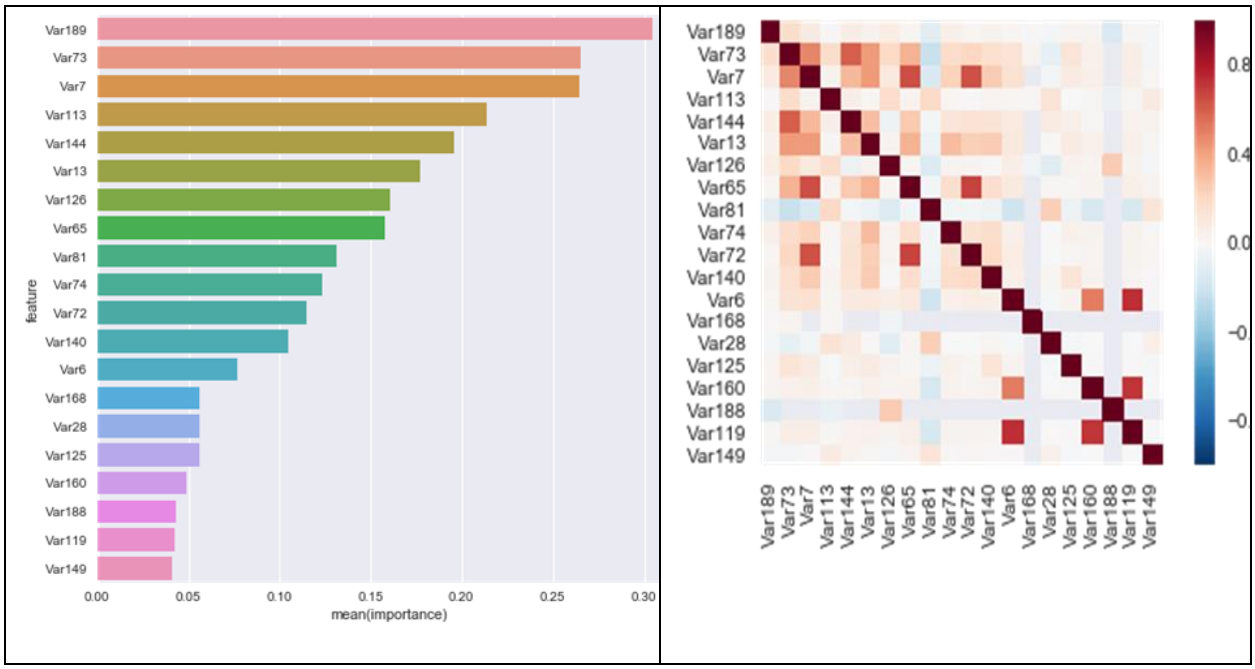
Техническое описание решения.

Отбор признаков
Вещественные

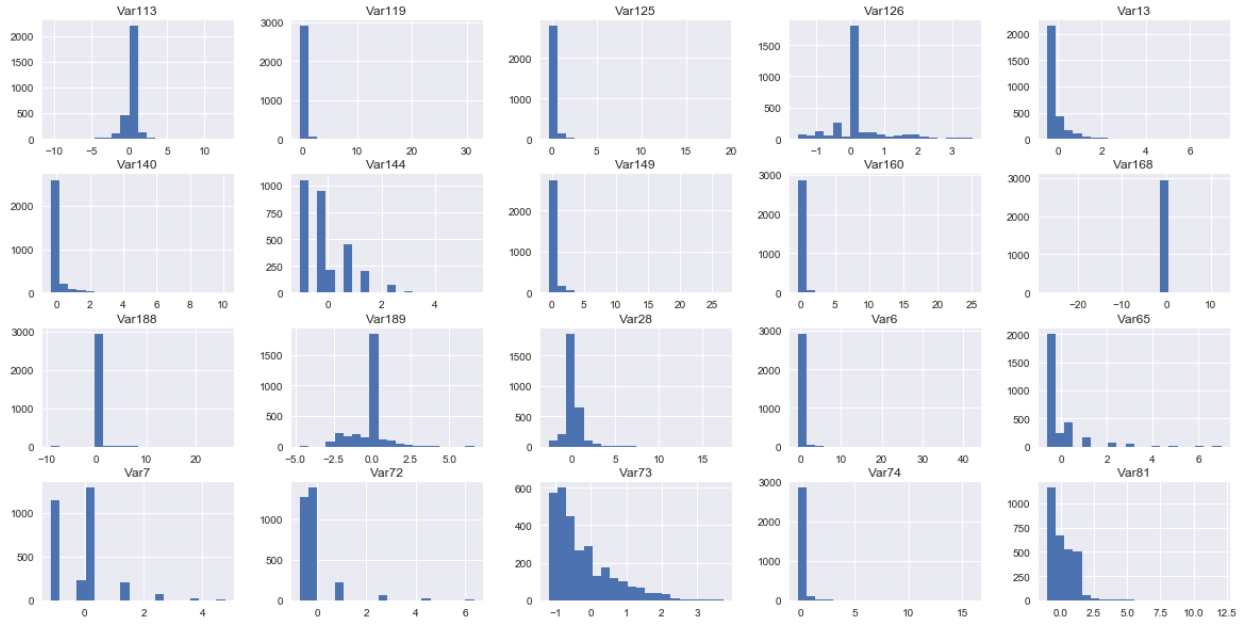
В первую очередь был проведен анализ признаков на долю NaN значений в каждом из них. По результатам этого анализа были сразу отброшены 18 признаков:
['Var8', 'Var15', 'Var20', 'Var31', 'Var32', 'Var39', 'Var42', 'Var48', 'Var52',
'Var55', 'Var79', 'Var141', 'Var167', 'Var169', 'Var175', 'Var185', 'Var209', 'Var230']

Далее был проведен анализ признакового пространства на предмет формы и характера распределений признаков и их корреляций с целевой переменной и между собой

20 наиболее коррелируемых с меткой класса числовых признаков и их взаимные корреляции:



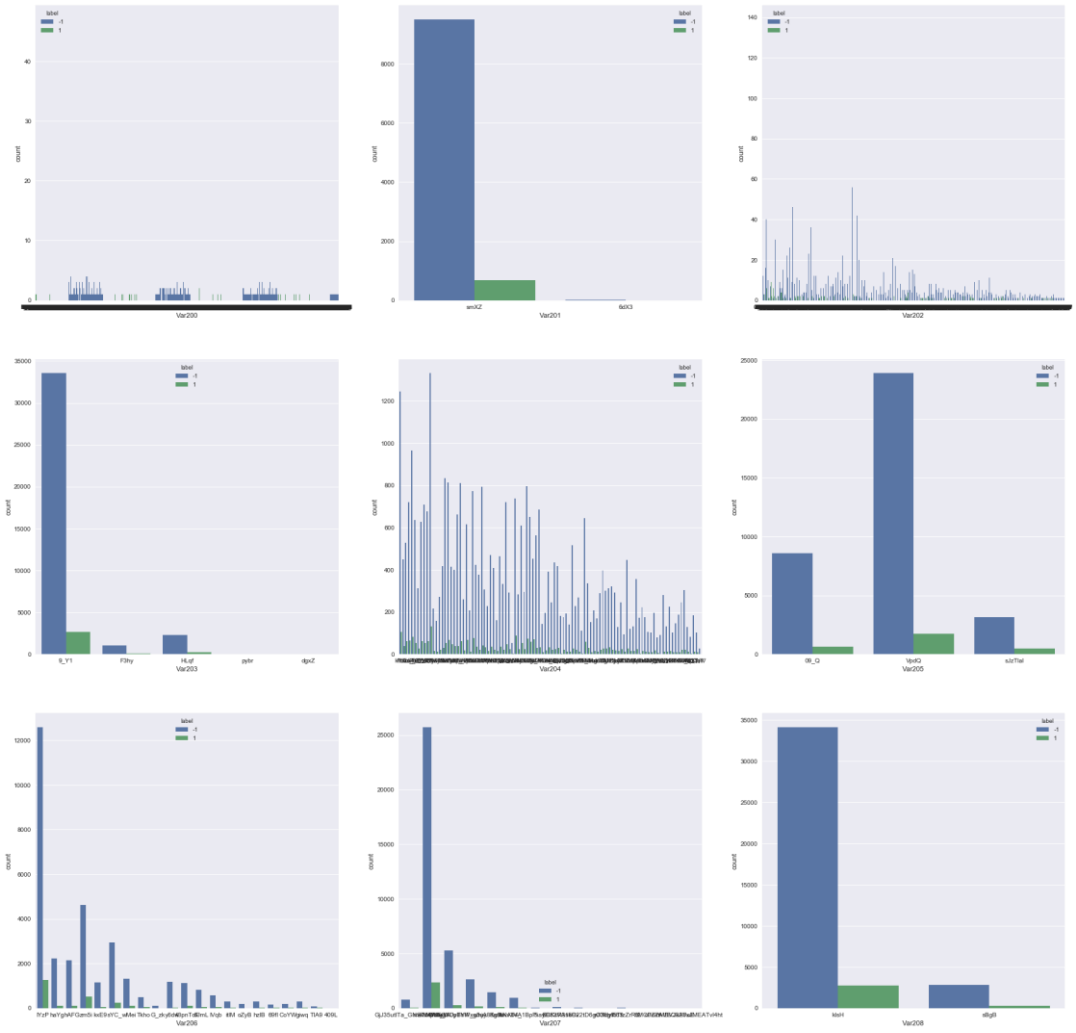
Распределения этих признаков :



Категориальные

При подробном рассмотрении категориальных признаков выяснилось, что есть признаки с количеством категорий более 10000.При этом большая часть этих категорий может быть в обучении и отсутствовать при тесте!

Пример распределений категориальных признаков



Для преобразования категорий в числовые данные я использовал стандартные процедуры кодирования Label и OneHotEncoder.

Работа с NaN-значениями

Анализируемый набор данных содержит очень много NaN-значений, поэтому один из ключевых моментов - надо было решить что с ними делать . При детальном рассмотрении можно увидеть, что есть две больших группы признаков: с долей NaN-значений > 90% и с долей категорий более 10 000.

По моему мнению «пустые» признаки , особенно для вещественных, бессмысленны в обучении, также большое кол-во категорий (например возможно это фамилии или персональные данные, которые нам будут только мешать).В пред процессинге решено было отбросить признаки ,которые на 90 % пустые и признаки с кол-вом категорий более 1000.В результате в обучении осталось 67 признаков для объектов(42 вещественных и 25 категориальных)

Выбор классификатора и подбор параметров

Для балансировки классов были проведены эксперименты с подбором весов классов и under и oversampling . В результате данные были досэмплированы объектами класса 1 .

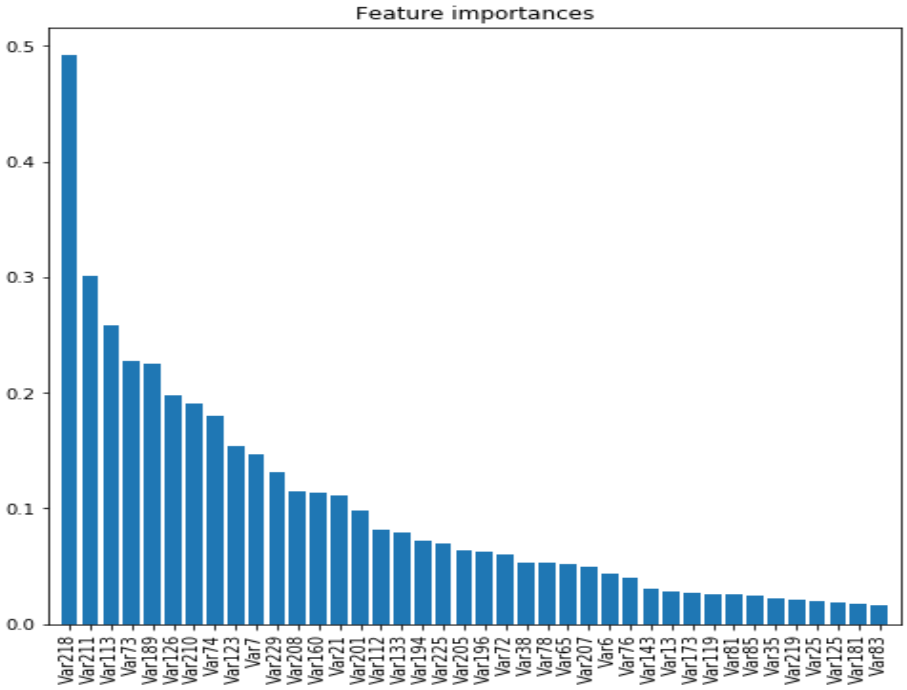
После некоторого подбора классификатора (выбирал между логистической регрессией и градиентным бустингом) остановился на градиентном бустинге GradientBoostingClassifier из пакета sklearn для Python.

Сравнение метрик качества классификаторов на кросс-валидации:

метрика	*LogisticRegression*	*GradientBoostingClassifier*
ROC-AUC	0.669220931852	0.790215765455
average precision	0.153772985737	0.771418425742
F1	0.196519829832	0.720759506436

Для модели был составлен Pipeline с предварительным отбором признаков с помощью L1-регуляризации с использованием метода SelectFromModel (LogisticRegression) и дальнейшим использованием их в GradientBoostingClassifier.

ТОП-40 наиболее важных признаков метода SelectFromModel (LogisticRegression)



Путем исследования были подобраны оптимальные параметры для классификатора:

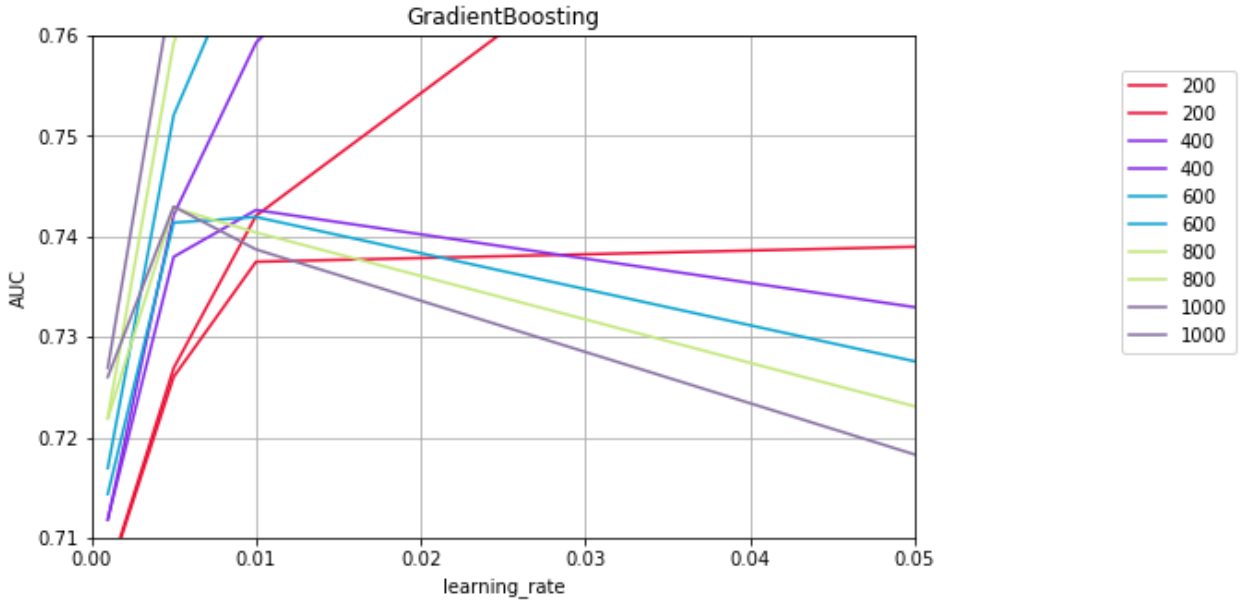
```
n_estimators=400,
max_depth=1,
min_samples_split=1,
subsample=1.0
```

Однако на тесте по метрикам был показан сильно отличающийся результат:

```
f1                  0.225441376188
average_precision 0.260855889724
roc_auc             0.692282572323
```

В результате чтобы компенсировать переобучение я дополнительно попробовал подобрать learning_rate=0.01 на сравнении результата на отложенной выборке

```
Качество модели с разными n_estimators
```



Результаты работы классификатора

Полученная модель дает следующие метрики качества:

метрика	*значение кросс-вал.*	*значение на отл. выб.*
ROC-AUC	0.75932086224	0.74455558075

Оценка ожидаемого экономического эффекта

Попытаемся оценить ожидаемый экономический эффект от использования данной модели в кампании по удержанию пользователей. Под экономическим эффектом будем понимать сколько денег мы получили (или наоборот потеряли) от проведения кампании по удержанию с использование нашей модели.

Введем следующие параметры:

- Средний ежемесячный доход от клиента (ARPU) = 300 руб.
- Средняя ежемесячная скидка на услуги для удержания клиента (DPU) = 15% ARPU
- Вероятность удержания пользователя предлагаемой скидкой (AP) = 0.5
- Общее кол-во клиентов в тестовой выборке -N
- topU -процент от общего числа клиентов, участвующий в компании
- Кол-во в topU верно классифицированных клиентов , склонных к оттоку - Nchurn

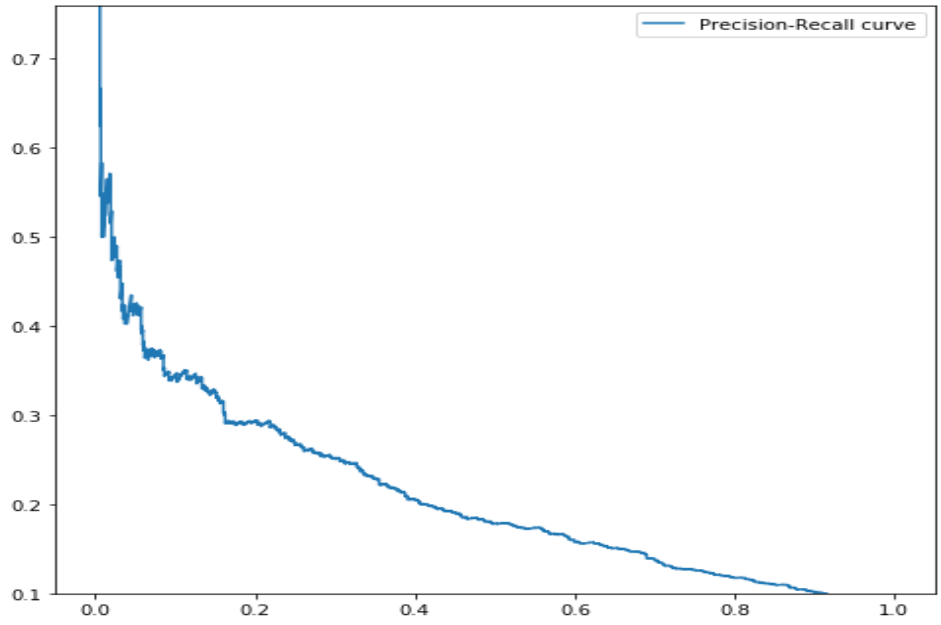
Попробуем оценить сколько классифицированных алгоритмом пользователей будет участвовать в кампании (top%) так чтобы мы оказались в прибыли и попробуем ее максимизировать.
Будем использовать ради простоты следующую формулу:

PROFIT = REVENUE-COST, где REVENUE = ARPU*AP*Nchurn
COST = DPU*N*topU

Прибыль -убыток при простом подборе топ %

Если мы внимательно посмотрим на формулу профита , то при неизменных ARPU,DPU,AP на прибыль влияет точность модели p и пороги вероятности при которых она достигается. У нас есть метрика precision_recall_curve, которая как раз выдает набор presicion и thresholds.Точка на кривой, определяющая порог и точность - соответствующий максимум прибыли, однозначно определяет топ-% пользователей текущей модели.

Кривая обучения



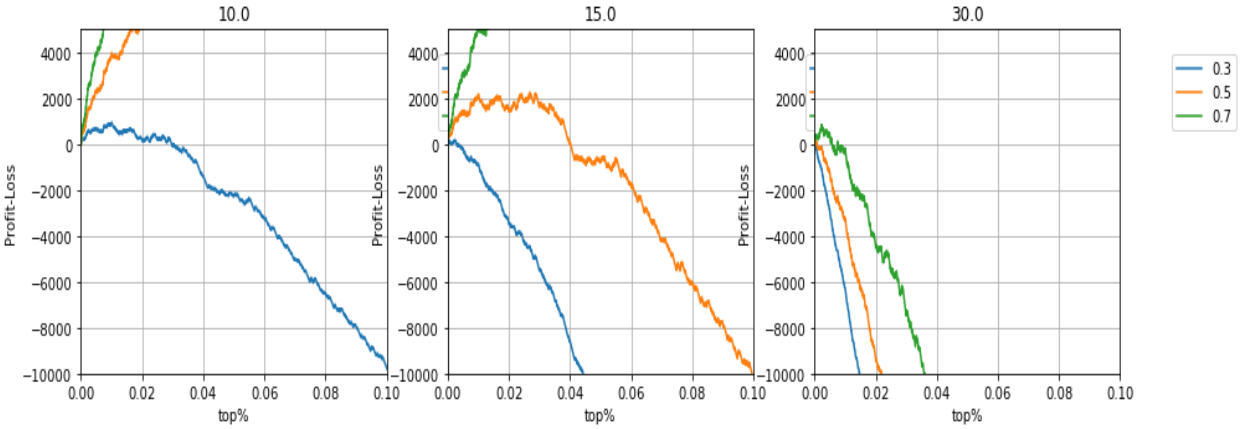
Тогда применив набор порогов и точности для выбранных нами параметров получим:

```
Лучший эффект достигается при использовании топ-% пользователей 2.68
Соотв. precision: 0.346749226006
Соотв. recall:    0.125419932811
Соотв. порог кл.: 0.77160946837
Соотв. эффект   : 2257.9876161 руб.
```

Если поэкспериментируем с затратами на удержание в % от ARPU и вероятностью принятия скидки, то :

```
Скидка, % 10.0 Вероятность принятия 0.7
Лучший эффект достигается при использовании топ-% пользователей 9.87
Соотв. precision: 0.24641350211
Соотв. recall:    0.326987681971
Соотв. порог кл.: 0.663915011173
Соотв. эффект   : 25748.2531646 руб.
```

Варианты для затрат 10%, 15%, 30 % и вероятностью 0.3, 0.5, 0.7



Чем меньше скидка и чем с большей вероятностью принята, тем больше эффект. Лучший результат скидка 10% и вероятность ее принятия 0.7. Также модель теряет смысл при скидках от 30% или вероятности принятия меньше 0,3.

Предложения по дальнейшему использованию

- Качество модели растет с увеличением объема исходных данных, поэтому, после внедрения рассматриваемого решения, необходимо сохранять историю оттока пользователей, и через какое-то время заново обучать модель с добавлением новых данных

-Как было показано выше -вероятность принятия скидки сильно влияет на прибыльность , поэтому точно надо проводить AB-тестирование кампаний по удержанию.

- Также очень желательно получить дополнительную информацию о признаках, чтобы провести их дополнительный отбор и таким образом улучшить качество.

- Для более сложных моделей расчетов надо изменить формулу расчета введя понятие Lifetime Value — пожизненной ценности клиента. Как вариант :

LTV = (Monthly Revenue per Customer * GrossMargin per Customer) / Monthly ChurnRate
ChurnRate = Q / Nt, где
Q — число ушедших пользователей на конец периода
Nt — общее число оставшихся на конец периода

Итог работы

В результате проделанной работы мы получили работающую модель классификации пользователей склонных к оттоку. Также мы вывели простую формулу оценки экономической эффективности классификатора, и оценили по ней рассматриваемую модель, получив наглядные показатели. В процессе работы над моделью, мы определили наиболее полезные признаки. Были построены экономические модели для оценки эффективности модели по удержанию. Описаны практические советы по внедрению и использованию модели.