Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра системного программирования

Долголев Филипп Петрович

Повышение качества предсказания оттока абонентов оператора сотовой связи

Курсовая работа

Научный руководитель:
ведущий разработчик ООО "НМТ" Невоструев К. Н.

Санкт-Петербург
2016

Оглавление

Введение

Сегодня, вследствие конкуренции, перед компаниями стоит проблема оттока клиентов. Так как затраты на удержание клиента обычно меньше, чем затраты на привлечение нового, то становится выгодно заранее выявлять склонных к уходу клиентов и пытаться их удержать с помощью различных методов, например, делая специальные предложения. К текущему моменту многие компании накопили довольно много данных о своих клиентах, что делает актуальным применение методов машинного обучения для решения этой проблемы.

В этой работе рассматривалась конкретно проблема оттока абонентов российского оператора сотовой связи. По некоторым данным в России на каждого жителя в среднем приходится по 1.6 - 1.7 SIM-карты, а так же суммарный рост абонентской базы по всем операторам в последние годы практически не наблюдается, что свидетельствует о насыщении рынка сотовой связи, что в свою очередь влечёт крайне высокую конкуренцию между существующими операторами. Как результат, мы можем наблюдать очень высокие показатели оттока абонентов.

Рис. 1: Отток абонентов, % от абонентской базы

1. Терминология

1.1. Базовые определения

Образец - вектор вещественных чисел - **признаков**

Выборка - множество пар: образец, класс

Классификатор - отображение действующее из множества образцов в множество классов

Задача бинарной классификации - задача построения классификатора, сопоставляющего образцам два класса: положительный и отрицательный

1.2. Метрики

Множество образцов в результате классификации разбивается на четыре множества:

1. **True Positive** (**TP**) - образцы из положительного класса определенные в положительный класс
2. **True Negative** (**TN**) - образцы из отрицательного класса определенные в отрицательный класс
3. **False Positive** (**FP**) - образцы из отрицательного класса определенные в положительный класс
4. **False Negative** (**FN**) - образцы из положительного класса определенные в отрицательный класс

Определим уходящих абонентов как положительный класс, а остающихся как отрицательный.

На основе четырёх вышеперечисленных множеств определены следующие метрики:

- **Precision** $=\frac{\mathrm{TP}}{\mathrm{TP+FP}}$
- **Recall** $= \frac{TP}{TP+FN}$

- **ROC-кривая** - кривая зависимости $TPR = \frac{TP}{TP+FN}$ и $1 - FPR = \frac{FP}{FP+TN}$
- **ROC AUC** - численная метрика, равная площади под ROC-кривой
- **PR-кривая** - кривая зависимости Precision и Recall
- **PR AUC** - численная метрика, равная площади под RP-кривой

1.3. Оценка классификатора

Для проведения достоверной оценки производилась перекрёстная проверка[6]. Заключалась она в следующем, сначала данные разбивались на пять частей, далее каждая часть отдельно выбиралась для проверки предсказывающей способности классификатора, а на остальных четырёх проводилось обучение. Результирующее значение каждой метрики для классификатора считалось как среднее арифметическое всех значений полученных при различных выборах проверочной части.

2. Обзор

Существует множество работ по предсказанию оттока клиентов. После рассмотрения некоторых из них, были сделаны следующие выводы:

1. Универсального подхода нет
2. Результаты сильно зависят от предметной области и исходных данных
3. Комбинация классификаторов часто улучшает результат

Автор	Предметная область
Chih-Ping Wei, I-Tang Chiu[1],	Сотовая связь
Yaya Xie, Xiu Li и др.[12]	Банковское обслуживание
Shin-Yuan Hung и др.[5]	Сотовая связь

Таблица 1: Рассмотренные работы

Отдельного внимания требует работа Корыстова М.[14], так как мы в текущей работе будем повышать качество предсказания уходящих абонентов основываясь на результатах этой работы и на тех же данных что использовались в ней. Лучший результат в этой работе был получен при полу-автоматической кластеризации и объединении независимо обученных классификаторов XGBoost на каждом кластере.

Модель	Precision	Recall	AUC
Набор классификаторов	0.75	0.66	0.90

Таблица 2: Результаты работы[14]

3. Постановка задачи

3.1. Цель

Цель работы заключается в повышении качества предсказания оттока абонентов оператора сотовой связи по сравнению с работой Корыстова М.[14] Чтобы достичь поставленную цель в этой работе планировалось решить следующие задачи:

1. Проанализировать и подготовить данные для классификации
2. Провести кластеризацию абонентов
3. Построить бинарный классификатор на основе результатов кластеризации
4. Оценить оптимальное значение метрик для построенного классификатора, с точки зрения оператора сотовой связи

3.2. Инструменты

Следующие инструменты использовались для решения поставленных задач:

- Язык программирования Python3[4]
- Библиотека Pandas для обработки данных[2]
- Библиотеки scikit-learn[13] и XGBoost[11] для кластеризации, постройки и оценки классификаторов

4. Данные

4.1. Исходные данные

Исходные данные, предоставленные оператором сотовой связи, содержат информацию о помесячной активности абонентов на протяжении 15 месяцев, а так же некоторые персональные данные.

Персональные данные:

- Дата подключения
- Пол
- Дата рождения

Данные об активности:

- Количество минут входящих вызовов
- Количество отправленных SMS
- Объём интернет трафика в мегабайтах
- Количество минут исходящих международных вызовов
- Количество минут исходящих междугородних вызовов
- Количество минут исходящих вызовов внутри оператора внутри региона подключения
- Количество минут исходящих вызовов внутри оператора за пределы региона подключения
- Количество минут исходящих вызовов на других операторов внутри региона подключения
- Количество минут исходящих вызовов на других операторов за пределы региона подключения
- Количество минут исходящих вызовов на городские номера в пределах

4.2. Подготовка данных

Для начала необходимо сформировать выборку, а для этого нужно сначала определить когда считать абонента ушедшим. В рамках этой работы было решено поступить аналогичным образом как и в работе[14]. А именно, если рассматривая временной ряд, можно обнаружить четыре месяца подряд с активностью, а в пятом и последующем активности нет, то то тогда считать абонента ушедшим в четвертом месяце, в противном случае оставшимся.

На основе этого были сформированы образцы следующим образом:

- Сформированы последовательности активности по месяцам для каждого абонента (где в соответствующей позиции-месяце стоит 1 если была какая-нибудь активность, и 0 в противном случае)
- Исключены из рассмотрения те абоненты, у которых активность наблюдалась менее чем в 4 месяцах подряд
- Для каждого абонента выбраны подпоследовательности следующим образом:
  - Если в хвосте подпоследовательности было **N** нулей то отбрасывались **N-1** нулей
  - Взята подпоследовательность из конца длиной 5
  - Далее формировался образец с активность за первые 3 месяца из последовательности, и если была активность в пятом месяце, то образец помечался как оставшийся абонент, в противном случае как ушедший абонент.

В результате была получена выборка размером порядка 114000 образцов, где в каждом образце отображена активность на протяжении трёх месяцев. Однако, было замечено что в этой выборке всего 12000 ушедших абонентов, и 10000 из них - из Санкт-Петербурга, а остальные 2000 распределены по 7 другим городам. Дальнейшие эксперименты показали что поведение абонентов из разных городов различается, и по

этой причине абоненты, распределенные по разным городам, лишь мешали работе классификатора. В связи с чем было решено оставить в выборке только абонентов из Санкт-Петербурга.

Формирование новых признаков

Далее для полученных образцов были сформированы новые признаки на основе существующих следующим образом:

- Были подсчитаны в минутах: суммарное количество исходящих вызовов, суммарное количество всех вызовов
- Для каждой услуги связанной со звонками подсчитано какую долю от суммарного количество вызовов в минутах она составляет
- На основе всех исходных и полученных признаков касательно вызовов, а так же SMS и интернет-трафика были получены следующие признаки:
  - Разность и отношение для соответствующих признаков за 3 и 1 месяц в последовательности активности
  - Дисперсия, математическое ожидание, коэффициенты асимметрии и эксцесса для соответствующих признаков по всем трём месяцам в последовательности активности

Далее было сделано предположение, что среди всех признаков, содержащих информацию о вызовах, для построения классификатора достаточно оставить лишь те, которые основаны на долях вызовов (например: разница долей за 3 и 1 месяц), а остальные исключить. Это было подтверждено дальнейшими экспериментами, в результате которых классификатор, обученный на выборке, в которой было произведено данное исключение, давал результаты немного лучше. Из чего можно сделать вывод, что начиная с определённого момента добавление новых признаков может ухудшить результат, следовательно стоит обращать внимание на это.

Последним этапом было преобразование категориального признака "пол" в числовой. Существует два варианта такого преобразования.

Один из них заключается в том, чтобы просто перенумеровать категориальные значения в числовые. Например, значения признака пол "мужчина" и "женщина" превратились бы в 1 и 2 соответственно. Но такой вариант подходит в дальнейшем не для всех классификаторов, а так же плохо влияет на кластеризацию. Поэтому был выбран второй вариант, который на примере пола заключается в выделении двух новых признаков-столбцов, в каждый из которых для соответствующего образца выставляется либо 1, если образец является носителем этого признака, либо 0 в противном случае.

Итог: выборка получилась размером в 50000 образцов, каждый из которых имел 148 признаков. Ушедших абонентов в этой выборке порядка 10000.

5. Кластеризация

5.1. Выбор признаков для кластеризации

Существует множество алгоритмов и подходов к кластеризации. В связи с большими объёмами выборки и ограниченностью вычислительных ресурсов, было решено проводить кластеризацию с помощью алгоритма K-Means[9].

Большая размерность данных обычно плохо сказывается на результатах кластеризации. Из-за этого было решено сначала выбрать подмножество признаков, на основе которых строить кластеризацию, а затем отдельно кластеризовать по признакам, характеризующим персональные данные абонентов, и по признакам, характеризующим активность абонентов.

Из признаков, характеризующих персональные данные абонентов, были выбраны все.

Из признаков, характеризующих активность абонентов, были выбраны те, которые представляли собой отношение долей каждой услуги, связанной с вызовами за первый и третий месяц, а так же отношение количества интернет-трафика и SMS за первый и третий месяц.

Так как при кластеризации K-Means в качестве метрики расстояния между образцами используется евклидово[8] расстояние, то необходимо изначально преобразовать данные одним из следующих способов:

- Нормирование - для каждого x провести замену $x^* = \frac{x - min(x)}{max(x) - min(x)}$
- Стандартизация - для каждого x провести замену $x^* = \frac{x - \overline{x}}{S^*}$, где S^* - среднее квадратическое отклонение, $\overline{x}$ - среднее арифметическое.

Были проведены эксперименты и с нормированием, и со стандартизацией. Проведённая на нормированных данных кластеризация дала более чёткое разделение на кластеры, в результате чего было решено рассматривать только её.

На основе выбранных признаков была построена кластеризация.

5.2. Определение количества кластеров

Алгоритм K-Means требует указать количество кластеров для своей работы, что не всегда известно заранее. Для решения этой проблемы используется несколько подходов. Воспользуемся одним из этих подходов:

1. Построим кластеризации с количеством кластеров k, где $k \in [a, b]$, a - нижняя граница на число кластеров, b - верхняя граница на число кластеров.

2. Для каждой кластеризации просчитаем сумму сумм квадратов расстояний внутри каждого кластера, $\mathbf{W} = \sum_{i=1}^{k} \sum_{j \in [S_i]} ||x_i - x_j||^2$, где k - количество кластеров, S_i - i-ый кластер, x_i - центр кластера S_i, x_j - образец попавший в кластер S_i

3. Для подсчитанных $\mathbf{W}$ построим график зависимости от количества кластеров, такой график называется **графиком каменистой осыпи**.

4. Найдём на данном графике "излом". Другими словами говоря, нас интересует то количество кластеров, после увеличения которого $\mathbf{W}$ уменьшается заметно слабее.

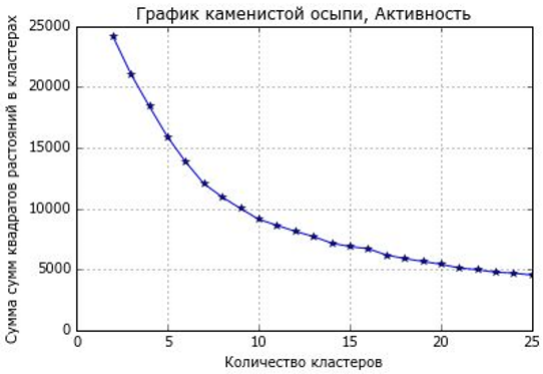

Рис. 2: График для кластеризации по активности

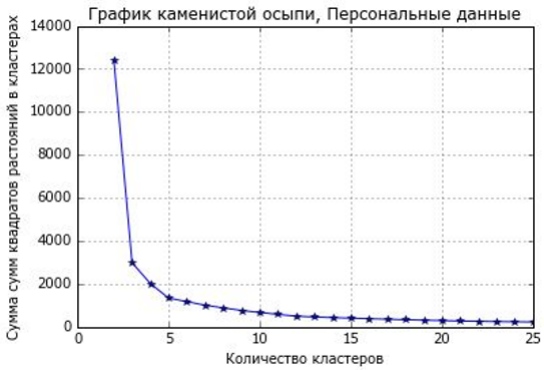

Рис. 3: График для кластеризации по персональным данным

Судя по графикам (Рис. 2, Рис. 3), стоит выделить 7 кластеров при кластеризации по активности, а так же 5 кластеров при кластеризации по персональным данным.

5.3. Анализ кластеризации

Чтобы оценить качество кластеризации, дополнительно был проведён анализ кластеризации по персональным данным. Для этого был применён метод описанный в статье[3]

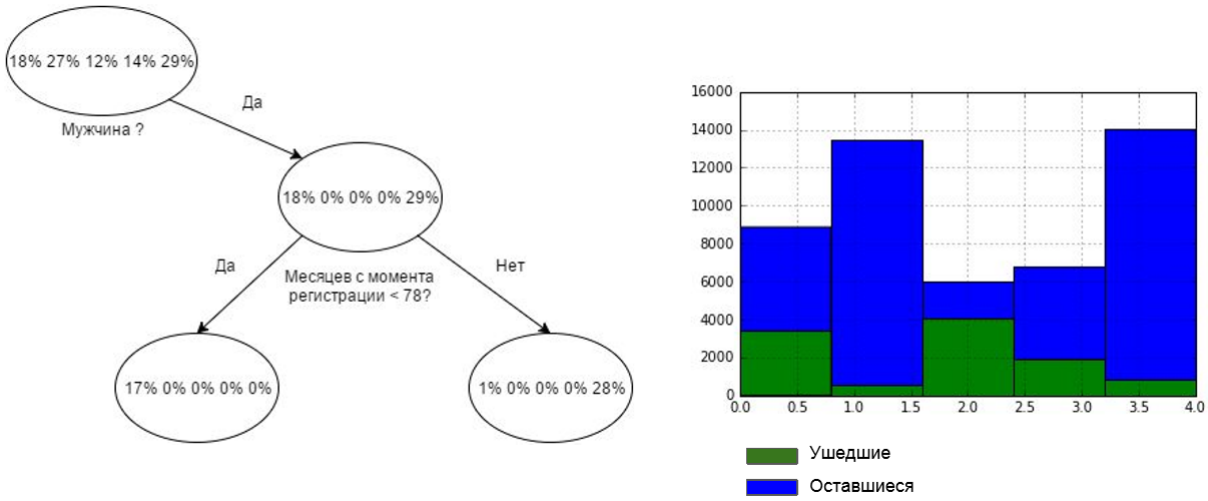

Рис. 4: Часть дерева решения и гистограмма ушедших/оставшихся абонентов

Данный метод заключается в том, что к образцам из исходной вы-

борки данных добавляется новый признак - номер кластера, вычисленный во время кластеризации, а после, на полученной выборке строится дерево решений[7], предсказывающее номер кластера. В узлах дерева указано распределение объектов по кластерам в процентах. Это позволит наглядно увидеть, чем различаются кластеры между собой.

Дополнительно к этому построим гистограммы ушедших и оставшихся абонентов в кластерах (Рис. 4).

Как видим, в кластер с номером 0 попали мужчины недавно перешедшие на услуги данного оператора, а в кластер с номером 5 попали мужчины давно пользующиеся услугами данного оператора.

Исходя из полного дерева решений и данной гистограммы было выяснено что:

- Мужчины и женщины пользующиеся услугами оператора менее 78 месяцев более склоны к оттоку
- Мужчины и женщины пользующиеся услугами оператора более 78 месяцев менее склоны к оттоку
- Отток среди юридических лиц является наибольшим

5.4. Построение ансамбля на кластерах

Построение ансамбля заключалось в следующем, на полученных кластерах независимо обучались классификаторы XGBoost[11], а на результатах их предсказаний была обучена логистическая регрессия[10].

Результаты

Модель	Precision	Recall	AUC	PR AUC
Набор классификаторов[14]	0.75	0.66	0.90	–
XGBoost	0.74	0.69	0.92	0.80
Ансамбль на кластерах	0.75	0.72	0.92	0.81

Таблица 3: Сравнение классификаторов

6. Выбор оптимальных значений метрик

Полученный классификатор выдаёт в качестве предсказания вероятность принадлежности к положительному классу (уходящие абоненты). Следовательно, задавая порог вероятности, выше которого мы считаем абонента уходящим, мы можем балансировать между Precision и Recall. Чем выше порог - тем больше Precision и ниже Recall.

Для оператора, как для конечного пользователя классификатором, важны в первую очередь не абстрактные для него метрики, а объёмы возможной прибыли за счёт использования классификатора. Представитель оператора сообщил, что 80% прибыли от удержанного абонента уходит на покрытие затрат на удержание. Исходя из этого, мы можем оценить в некоторых единицах прибыль оператора в зависимости от TP и FP. Была получена следующая формула: Выгода = $TP - 0.8 * (TP + FP)$ На основе этой формулы была оценена зависимость получаемой выгоды от Precision и Recall для построенного классификатора.

Эксперименты проводились на выборках в 10000 абонентов, среди которых порядка 20% ушедших.

Обозначим объём прибыли от одного абонента t. В результате экспериментов было выяснено, что наибольшая выгода составляет порядка $94.8t$ при Precision = 0.92 и Recall = 0.33

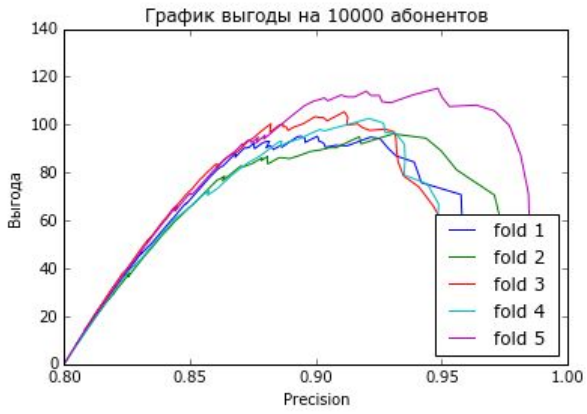

Рис. 5: Зависимость выгоды от Precision

Заключение

В рамках данной работы было получено повышение качества предсказания оттока абонентов оператора сотовой связи по сравнению с работой Корыстова М.[14].

В частности, были решены следующие задачи:

- Проанализированы и подготовлены данные для классификации
- Проведена кластеризация абонентов
- Построен бинарный классификатора на основе результатов кластеризации
- Оценены оптимальные значения метрик для построенного классификатора, с точки зрения оператора сотовой связи

Список литературы

[1] Chih-Ping Wei I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. — 2002. — URL: http://www.sciencedirect.com/science/article/pii/S0957417402000301 (дата обращения: 22.05.2016).

[2] Pandas. — 2016. — URL: http://pandas.pydata.org/.

[3] Parisot Olivier, Ghoniem Mohammad, Otjacques Benoıt. Decision Trees and Data Preprocessing to Help Clustering Interpretation. — 2014. — URL: https://www.researchgate.net/publication/263057551_Decision_Trees_and_Data_Preprocessing_to_Help_Clustering_Interpretation (дата обращения: 22.05.2016).

[4] Python3. — 2016. — URL: https://www.python.org/.

[5] Shin-Yuan Hung David C Yen, Wang Hsiu-Yu. Applying data mining to telecom churn management. — 2006. — URL: http://www.sciencedirect.com/science/article/pii/S0957417405002654 (дата обращения: 22.05.2016).

[6] Wikipedia. Cross-validation // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Cross-validation_(statistics).

[7] Wikipedia. Decision tree // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Decision_tree.

[8] Wikipedia. Euclidean distance // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Euclidean_distance (дата обращения: 22.05.2016).

[9] Wikipedia. K-Means // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/K-means_clustering.

[10] Wikipedia. Logistic regression // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Logistic_regression.

[11] XGBoost. — 2016. — URL: https://xgboost.readthedocs.io/.

[12] Yaya Xie Xiu Li EWT Ngai, Ying Weiyun. Customer churn prediction using improved balanced random forests. — 2009. — URL: http://www.sciencedirect.com/science/article/pii/S0957417408004326 (дата обращения: 22.05.2016).

[13] scikit-learn. — 2016. — URL: http://scikit-learn.org/.

[14] Корыстов Максим. Применение методов машинного обучения для предсказания поведения абонентов оператора сотовой связи. — 2015. — URL: http://se.math.spbu.ru/SE/diploma/2015/bmo/444-Korystov-report.pdf (дата обращения: 22.05.2016).