Правительство Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Кафедра Системного Программирования

Корыстов Максим Андреевич

# Применение методов машинного обучения для предсказания поведения абонентов оператора сотовой связи

Бакалаврская работа

Допущена к защите.
Зав. кафедрой:
д. ф.-м. н., профессор Терехов А. Н.

Научный руководитель:
д. ф.-м. н., профессор Терехов А. Н.

Рецензент:
ведущий разработчик ООО «НМТ» Невоструев К. Н.

Санкт-Петербург
2015

SAINT-PETERSBURG STATE UNIVERSITY
Software Engineering Chair

Maxim Korystov

# Machine learning methods for predicting behavior of mobile network operator customers

Bachelor's Thesis

Admitted for defence.
Head of the chair:
Professor Andrey Terekhov

Scientific supervisor:
Professor Andrey Terekhov

Reviewer:
Senior developer at "NMT" LLC Constantin Nevostruev

Saint-Petersburg
2015

# Содержание

# Введение

Ежегодно провайдеры телекоммуникационных услуг терпят большие убытки из-за оттока абонентов. Сейчас у клиентов телекоммуникационных компаний есть возможность выбора среди большого количества поставщиков услуг. В этом конкурентном рынке клиенты предпочитают высокое качество за меньшую цену, в то время как провайдеры сосредоточены над созданием выгодных предложений. В этой отрасли годовой отток клиентов достигает 25-50% (Рис. 1). Причем, привлечение новых клиентов стоит в несколько раз дороже чем удержание старых. Это делает проблему оттока абонентов особо привлекательной для изучения. Сама задача предотвращения оттока клиентов - это и статистическая задача и задача маркетинга. Телекоммуникационные компании используют различные стратегии для удержания клиентов. Достаточно точные предсказания оттока клиентов позволяют менеджерам применять эти стратегии целенаправленно, экономя наибольшее количество средств.

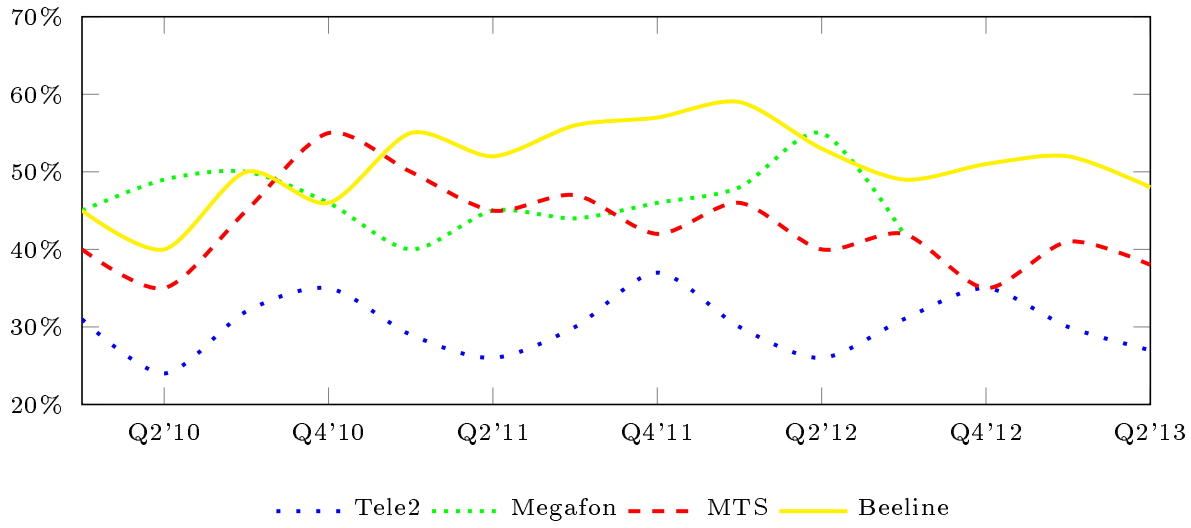


Рис. 1: Отток абонентов мобильных операторов в России [1]

Эти факты дали толчок для разработки множества интеллектуальных систем предсказывающих риск ухода абонентов. Эта задача классификационного анализа, то есть, разделения абонентов на две группы: удержанные и имеющие риск ухода. Для решения задач классификации применяются

[1] По данным ежеквартальных отчетов мобильных операторов.

популярные методы машинного обучения, такие как логическая регрессия, нейронные сети, метод опорных векторов. Также, важным этапом решения задачи является понимание и измерение точности предсказания полученной модели машинного обучения. Абонентам имеющим риск ухода рассылаются специальные предложения: скидки, бонусы, которые позволяют увеличить вероятность удержания абонента. Необходимо чтобы классификатор достаточно точно указывал на абонентов собирающихся сменить провайдера, в противном случае мобильный оператор может потерпеть убытки делая скидки не собирающимся уходить абонентам.

В представленной работе описан процесс разработки классификатора основанного на кластеризации, также приводятся сравнения полученного классификатора с другими методами машинного обучения. Для обучения и тестирования классификатора были использованы реальные данные абонентов крупной российской телекоммуникационной компании. Лучшая полученная модель достигает 0.90 AUC на проверочной выборке.

# 1 Предметная область и постановка задачи

## 1.1 Термины и определения

*Образец* – вектор $\langle x_1^i, x_2^i, \ldots, x_n^i \rangle$ вещественных чисел – *характеристик*.
*Выборка* – множество пар: образец, ответ.
*Классификатор* – отображение $C : S \to \{0, 1, \ldots, n\}$, действующая из множества образцов в множество классов.
*Задача бинарной классификации* – задача построения классификатора разделяющего множество образцов на два класса: положительный, отрицательный
**Оценка эффективности работы бинарного классификатора.** Все множество образцов разбивается классификатором на четыре части:

- TP (True positive) – образцы верно определенные классификатором в положительный класс.
- FP (False positive) – образцы неверно определенные классификатором в положительный класс.
- TN (True negative) – образцы верно определенные классификатором в отрицательный класс.
- FN (False negative) – образцы неверно определенные классификатором в отрицательный класс.

Из размеров этих частей определяются precision (точность) и recall (полнота):

$$\text{precision} = \frac{TP}{TP + FP}, \quad \text{recall} = \frac{TP}{TP + FN}$$

Откуда определяется функция используемая для оценки эффективности бинарных классификаторов:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Основной величиной оценки является AUC (Area under curve), определяе-

мая как площадь под ROC-кривой (кривая зависимости $TPR = \frac{TP}{TP+FN}$ от $FPR = \frac{FP}{FP+TN}$).

## 1.2 Исходные данные

Одна из крупнейших российских телекоммуникационных компаний предоставила данные о своих абонентах. Для каждого из абонентов известна его активность в течение 15 месяцев (весь 2013 год и начало 2014). Каждый месяц активности абонента описывается десятью показателями:

- количество минут входящих вызовов (INCOMING)
- количество минут исходящих вызовов на городские номера в пределах области подключения (PSTN)
- количество минут исходящих вызовов на мобильные номера прочих мобильных операторов в за пределы области подключения (ALIEN)
- количество минут исходящих вызовов на мобильные номера прочих мобильных операторов в пределах области подключения (REGION)
- количество минут исходящих вызовов на данного оператора за пределы области подключения (OTHER)
- количество минут исходящих вызовов на данного оператора в пределах области подключения (AREA)
- количество минут исходящих вызовов на междугородние вызовы (LONG)
- количество минут исходящих вызовов на международные вызовы (INTERNATIONAL)
- количество мегабайт потребленного интернет трафика (GPRS)
- количество отправленных СМС (SMS)

Так же известна дата регистрации абонента (REG), его пол (SEX), корпоративный ли он клиент (COMPANY), возраст (AGE), модель используемого устройства связи. Всего в выборке присутствуют данные о 150 тысячах абонентов.

## 1.3 Постановка задачи

**Цель работы:** разработать алгоритм предсказания ухода абонентов по представленной выборке.

**Задачи:**

1. обработать исходные данные
2. проанализировать характеристики представленной выборки
3. провести эксперименты сравнения моделей
4. выбрать модель машинного обучения с лучшим результатом

# 2 Обзор существующих работ

Решения схожей задачи уже были описаны в существующих работах. В этом разделе будут перечислены работы и методы машинного обучения применявшиеся в них, рассмотренные в процессе исследования (см. Таблица 1). Все эти работы имеют общую область исследования - отток абонентов. Целью обзора этих работ был выбор перспективных моделей машинного обучения.

| **Автор** | **Источник данных** | **Модели** |
|---|---|---|
| V. Umayaparvathi, K. Iyakutti [1] | Сингапурский оператор сотовой связи | Деревья решений, Нейронные сети |
| Mozer MC, Wolniewicz R и др. [2] | Провайдер услуг беспроводной связи | Логистическая регрессия, Нейронные сети |
| Chih-Ping Wei, I-Tang Chiu [3] | Тайваньский оператор сотовый связи | Деревья решений |
| Yaya Xie, Xiu Li и др. [4] | База клиентов китайского банка | Random forest |
| Hung, Shin-Yuan and Yen и др [5] | База клиентов тайваньской телеком-муникационной компании | Нейронные сети, Деревья решений, K-Means – кластеризация |
| Coussement, Kristof and Van den Poel, Dirk [6] | Подписчики бумажных изданий бельгийской компании | Логистическая регрессия |

Таблица 1: Рассмотренные работы

В большинстве рассмотренных работ сравниваются несколько моделей машинного обучения на одном наборе данных. Это зарекомендованная практика построения хорошего классификатора. В некоторых работах показывает хорошие результаты логическая регрессия – базовый алгоритм клас-

сификации. В нескольких работах перед обучением проводится обработка данных. В работах [5] [6] исходные данные содержат множество характеристик, поэтому перед обучением применяется алгоритм уменьшения размерности, PCA (метод главных компонент). В работах [1], [4], [5] лидирующие результаты показывает модель машинного обучения – деревья решений. Нейронные сети в рассмотренных работах [1], [2], [5] показали результаты близкие к лидирующим. В работе [5] лучший результат показала комбинация методов машинного обучения – кластеризация методом K-Means и деревья решений. Результаты исследований в рассмотренных работах сложно сравнить между собой, потому что во многом итоговый результат зависит от характеристик исходных данных и от выбора алгоритма машинного обучения. Однако, можно выделить список перспективных моделей, которые показывают лучшие результаты на данном классе задач:

- Деревья решений
- Random forest
- Логистическая регрессия
- Нейронные сети

# 3 Обработка данных

Важным этапом построения модели машинного обучения является обработка исходных данных и правильный выбор параметров для обучающей выборки. Во многом результат обучения классификатора зависит от набора представленных характеристик.

## 3.1 Метод скользящего окна

Для превращения задачи предсказания временного ряда в классическую задачу машинного обучения применяется метод скользящего окна [7]. Метод скользящего окна заключается в следующем: для каждого исходного временного ряда $x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots$ выбирается соответсвующее окно $w^{(j)} = \langle x_{t+1}^{(i)}, x_{t+2}^{(i)}, \ldots, x_{t+l-1}^{(i)}, x_{t+l}^{(i)} \rangle$ и $y^{(j)} = x_{t+l+1}^{i}$ – следующий за окном элемент временного ряда, где $t$ – сдвиг окна, $l$ – фиксированная ширина окна. Обычно для каждого исходного временного ряда выбирается несколько последовательных окон (окно скользит по временному ряду) чтобы увеличить количество образцов в итоговой выборке. Тогда новая задача звучит так: построить модель машинного обучения которая по $w^{(j)}$ предсказывает следующий за окном элемент $y^{(j)}$.

В исходном наборе данных для каждого абонента известно 10 показателей его активности в течение 15 месяцев. Если абонент в данный месяц еще не был зарегистрирован клиентом мобильного оператора, то его активность в этот месяц равна нулю. Для исходных данных была выбрана ширина окна 4 месяца. Для каждого временного ряда $x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots, x_{15}^{(i)}$, где $x_j^{(i)}$ – вектор из 10-и показателей активности за $j$-ый месяц для $i$-го абонента, было зафиксировано одно окно $w^{(i)} = \langle x_{t+1}^{(i)}, x_{t+2}^{(i)}, x_{t+3}^{(i)}, x_{t+4}^{(i)} \rangle$ и суммарная по всем показателям активность абонента в следующем после окна месяце $AS(x_{t+5}^{(i)}) = \sum_{k=1}^{10} x_{t+5}^{(i)}[k]$. Абонент считается ушедшим в месяце $t+4$ если $AS(x_{t+5}^{(i)}) = 0$, иначе абонент считается удержанным. Если для данного абонента возможно выбрать $t$ такое, что он будет считаться ушедшим, то выбирается именно такое $t$. Причем, в итоговую выборку попадает только часть окна: все элементы окна кроме последнего, потому что для ушедших абонентов активность в месяц ухода $t+4$ гораздо меньше, чем в предыду-

щие месяцы – это может стать причиной переобучения классификатора.

## 3.2 Дополнительные характеристики

Помимо использования исходных характеристик, улучшения модели машинного обучения можно добиться добавляя дополнительные характеристики, полученные из уже имеющихся. Распространенные дополнительные характеристики дающие наибольший результат при обучении для временных рядов [8]:

- арифметическое отношение (G)
- геометрическое отношение (A)
- математическое ожидание (E)
- дисперсия (D)
- коэффициент асимметрии (SKEW)
- коэффициент эксцесса (KURT)

**Арифметическое отношение.** Для соседних элементов временного ряда $x_{i-1}, x_i$ берется их разница $a_i$:

$$a_i = x_i - x_{i-1}$$

**Геометрическое отношение.** Для соседних элементов временного ряда $x_{i-1}, x_i$ берется их отношение $g_i$ по следующему правилу:

$$g_i = \begin{cases} \frac{x_{i-1}}{x_i}, & x_i \neq 0 \\ 1.0, & x_i = 0 \ \& \ x_{i-1} = 0 \\ \text{INF}, & x_i = 0 \ \& \ x_i \neq 0 \end{cases}$$

Где INF - число заведомо большее всех возможных в выборке геометрических отношений.

**Математическое ожидание.** Для всех элементов временного ряда $j$-го показателя активности $x_1^j, x_2^j, \ldots, x_k^j$ берется их оценка математического ожидания $e^j$:

$$e^j = \frac{1}{k} \sum_{i=1}^{k} x_i^j$$

**Дисперсия.** Для всех элементов временного ряда $j$-го показателя активности $x_1^j, x_2^j, \ldots, x_k^j$ берется их оценка дисперсии $d^j$:

$$d^j = \frac{1}{k} \sum_{i=1}^{k} (x_i^j - e^j)^2$$

**Коэффициент асимметрии.** Для всех элементов временного ряда $j$-го показателя активности $x_1^j, x_2^j, \ldots, x_k^j$ берется их оценка коэффициента асимметрии $s^j$:

$$s^j = \frac{\frac{1}{k} \sum_{i=1}^{k} (x_i^j - e^j)^3}{{d^j}^{\frac{3}{2}}}$$

**Коэффициент эксцесса.** Для всех элементов временного ряда $j$-го показателя активности $x_1^j, x_2^j, \ldots, x_k^j$ берется их оценка коэффициента асимметрии $k^j$:

$$k^j = \frac{\frac{1}{k} \sum_{i=1}^{k} (x_i^j - e^j)^4}{{d^j}^2}$$

## 3.3 Модели устройств связи

Для каждого активного месяца абонента в исходных данных представлен TAC (Type Allocation Code) устройства – первые 8 цифр IMEI, описывающие модель и место происхождения устройства. Для каждого абонента из выборки был зафиксирован один код TAC - встречающийся наибольшее количество раз в рассматриваемом окне, если таких кодов несколько, то выбирался самый последний в хронологическом порядке. По этому коду TAC были получены следующие дополнительные характеристики об устройстве связи:

- работает под управлением Android (бинарная характеристика)
- работает под управлением iOS (бинарная характеристика)

- работает под управлением Windows Phone (бинарная характеристика)
- поддерживает передачу данных GPRS (бинарная характеристика)
- поддерживает передачу данных EDGE (бинарная характеристика)
- поддерживает передачу данных 3G (бинарная характеристика)
- поддерживает передачу данных LTE (бинарная характеристика)
- имеет модуль WiFi (бинарная характеристика)
- год выпуска модели
- коэффициент стоимости от 1 до 10
- диагональ экрана

Данные о моделях устройств брались с сайта содержащего крупную базу мобильных устройств связи [9].

# 4 Обучение

## 4.1 Выбор классификаторов

Исходя из анализа публикаций по теме, были выделены несколько перспективных методов машинного обучения:

- Логистическая регрессия
- Нейронные сети
- Random forest
- Градиентный бустинг над решающими деревьями

## 4.2 Подбор параметров

У всех моделей машинного обучения, перед началом самого обучения должны быть определены параметры модели. Для логической регрессии это параметр регуляризации. Для нейронных сетей это количество нейронов в скрытом слое. В Random forest это количество деревьев. У градиентного бустинга – это количество деревьев и другие. Возникает вопрос:

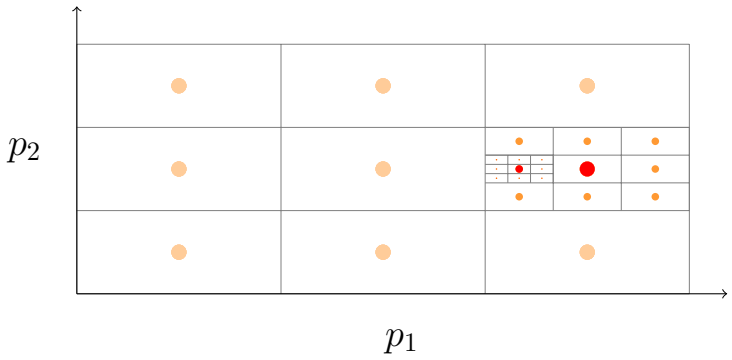


Рис. 2: Иллюстрация работы алгоритма поиска в сетке при n = 2.

как выбрать параметры модели машинного обучения для получения наилучших результатов? Для решения этого вопроса используется *алгоритм поиска в сетке (Grid search)* [10]. Для каждого из $n$ параметров модели определяются интервалы поиска $[a_i, b_i], 1 \leqslant i \leqslant n$, в которых находятся все допустимые значения. Представим фиксированные $n$ значений как точку

в $n$-мерном кубе. Суть алгоритма состоит в том чтобы разбить этот куб на кубы меньшего размера, в каждом из них выбрать центральную точку $p_j$. Обучить модель с параметрами $p_j$ для каждого из $j$ кубов. Посчитать метрику качества обучения в каждом из $j$ точек и выбрать лучшую точку. С соответствующим лучшей точке кубом проделать тоже самое и так далее, до тех пор, пока не найдем удовлетворяющую точку. Этот алгоритм не всегда находит оптимальные параметры модели машинного обучения, в некоторых случаях он находит только лишь локально оптимальные параметры. Чтобы избежать этого используется более мелкое разбиение куба.

## 4.3 Подготовка выборки

После этапа обработки данных была получена выборка из 96460 образцов с 84 характеристиками. Она была разделена на три части:

- Тренировочная выборка (60%)
- Тестовая выборка (20%)
- Кросс-тестовая выборка (20%)

Тренировочная выборка используется для обучения моделей. На тестовой выборке модели делают предсказания и оцениваются по заданным метрикам. Лучшая модель выбирается по результатам предсказаний по тестовой выборке. Кросс-тестовая выборка используется для независимой оценки моделей.

## 4.4 Обучение

Выборка была случайным образом перемешена и разделена на три части (60%, 20%, 20%) 10 раз, после каждого раза производилось обучение моделей и проверка модели на тестовой выборке. Причем для моделей: логическая регрессия и нейронные сети, характеристики полученной выборки были смасштабированны на отрезок $[0, 1]$. Результаты экспериментов представлены в таблице 2.

| Модель | precision | recall | F1 | F0.5 | **AUC** |
|---|---|---|---|---|---|
| Градиентный бустинг | 0.70 | 0.64 | 0.66 | 0.68 | **0.842**±0.005 |
| Random forest | 0.72 | 0.60 | 0.65 | 0.69 | **0.832**±0.008 |
| Нейронные сети | 0.69 | 0.59 | 0.63 | 0.66 | **0.825**±0.007 |
| Логистическая регрессия | 0.63 | 0.37 | 0.46 | 0.55 | **0.842**±0.002 |

Таблица 2: Сравнение моделей на тестовой выборке.

**Интерпретация сравнений.** Итоговые сравнения классификаторов производились с помощью AUC (площадь по ROC кривой, см. Рисунок 3) метрики. Лучшие результаты показал градиентный бустинг над решающими деревьями, схожие с ним результаты показал классификатор random forest. Однако, в задаче оттока абонентов также важной метрикой является precision. Предсказание ухода абонента делается с целью предложить клиенту более выгодные условия, бонусы и таким образом удержать его. Именно precision показывает насколько точно будут попадать эти предложения нужному абоненту. При низкой точности классификатора мобильный оператор будет часто отсылать предложения клиентам не собирающимся уходить, что может привести к финансовым потерям.

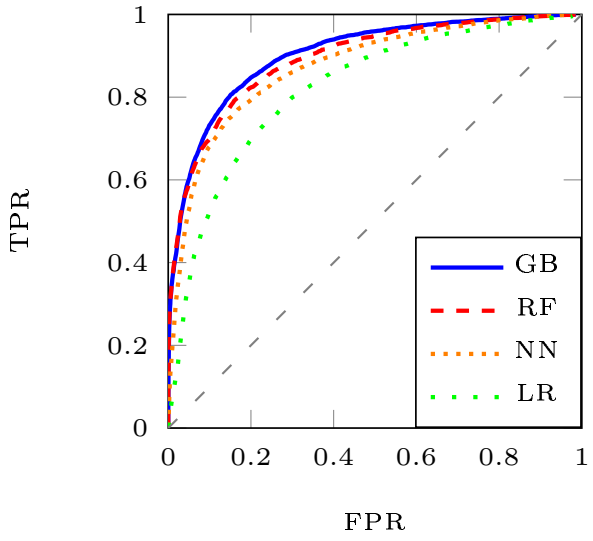


Рис. 3: ROC кривые полученных моделей

## 4.5 Важность характеристик

После полученных результатов сравнений возникает вопрос: какие именно из представленных 84 характеристик оказали наибольшее воздействие на результат. Рассмотрим важность характеристик полученной выборки с точки зрения алгоритма показавшего лучший результат в обучении – градиентного бустинга над решающими деревьями. Как видно из сравне-

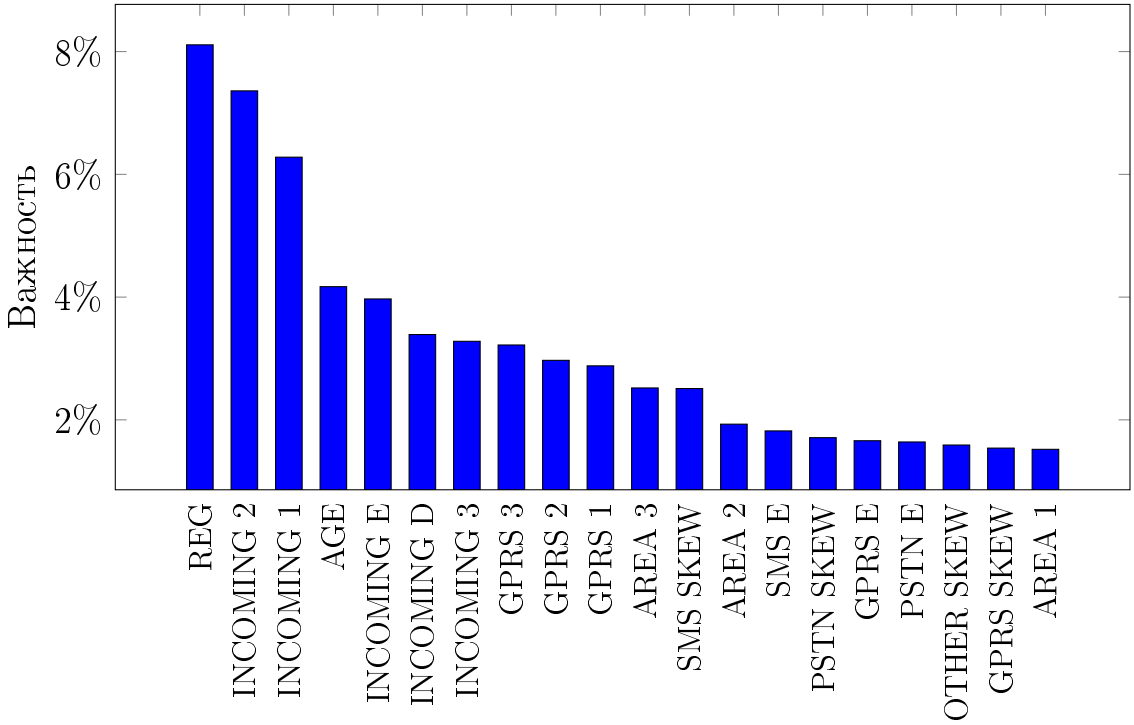


Рис. 4: Список информативных характеристик.

ний важности характеристик (см. Рисунок 4): наиболее значимый вклад в результат вносят характеристики: дата регистрации, количество входящих минут за месяц, возраст абонента, количество потребленного за месяц интернет трафика. Рассматривая полученные данные о важности характеристик можно заметить, что для некоторых характеристик (INCOMING, SMS, GPRS, OTHER) оценки временного ряда, такие как: математическое ожидание, дисперсия имеют, коэффициент асимметрии, почти такую же важность, а иногда и большую важность, как и исходные показатели активности. Это интересный факт – в некоторых случаях дополнительные характеристики временного ряда играют большую роль чем сами элементы временного ряда. В исходной выборке это можно объяснить посмотрев

на матрицу ковариации исходных показателей активности абонента за 3 последовательных месяца (см. Рисунок 5). Как видно из рисунка одни и те же

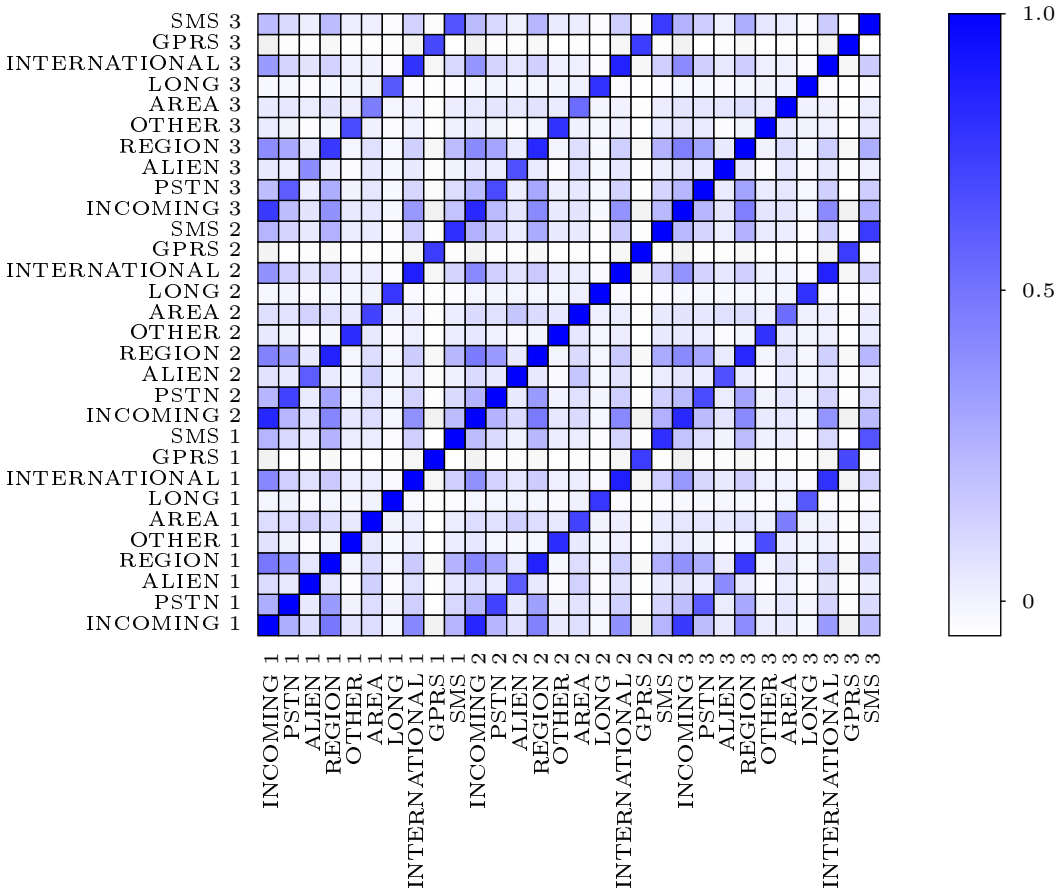


Рис. 5: Матрица ковариации показателей активности.

показатели активности за разные месяцы коррелируют между собой. Это говорит о схожих значениях показателей активности одного абонента за соседние месяцы. Наиболее коррелирующими показателями активности абонента оказались: INTERNATIONAL(0.85 средняя ковариация по всем месяцам), REGION(0.83), INCOMING(0.82), OTHER(0.77), SMS(0.75). Наименее коррелирующими оказались показатели: LONG(0.75), GPRS(0.75), PSTN(0.7), AREA(0.61), ALIEN(0.59).

Некоторые характеристики не внесли существенный вклад в полученный классификатором результат. Среди этих характеристик: данные о модели устройства связи абонента, пол абонента, некоторые показатели исходящих вызовов. Это свидетельствует о том, что отток абонента почти не зависит от этих характеристик. Опираясь на эти факты можно существен-

но уменьшить количество характеристик в итоговой выборке, тем самым достигнув более быстрой скорости обучения модели. Однако, при исключении этих характеристик возможно незначительное ухудшение качества классификации.

## 4.6 Кластеризация

Одним из подходов обработки данных перед обучением модели является кластеризация. Идея кластеризации заключается в разделении выборки на несколько частей – кластеров и обучения независимых классификаторов на каждой из частей. Часто с помощью этого метода можно добиться лучших результатов, чем с помощью стандартного подхода. Это улучшение является следствием следующего факта: в разных кластерах одни и те же характеристики имеют разную ценность. Поэтому классификаторы в различных кластерах работают по-разному, основываясь на разных характеристиках. Возникает вопрос: каким образом и на сколько кластеров разделить выборку? Для решения этих вопросов применяется две техники: автоматическая кластеризация и полуавтоматическая кластеризация.

**Автоматическая кластеризация.** Метод автоматической кластеризации заключается в разделении выборки на кластеры не основываясь на особенностях задачи и предметной области. Единственным существенным параметром автоматической кластеризации является количество кластеров. Один из популярных алгоритмов автоматической кластеризации является алгоритм K-Means [11]. Алгоритм заключается в выборе центров для каждого кластера и последовательном их приближении к центрам имеющим минимальное суммарное квадратичное отклонение от всех точек кластеров соответствующих центрам. Иногда используется отличная от евклидовой метрика. Однако, применение K-Means к несбалансированным выборкам, как в данной задаче, редко дает удовлетворительные результаты.

**Полуавтоматическая кластеризация.** Этот метод требует заранее зафиксировать набор характеристик для кластеризации. Эти характеристики выбираются вручную, часто в качестве этих характеристик исполь-

зуют самые важные для классификации. После фиксации характеристик, для каждой из характеристик перебирается набор границ. После того как были зафиксированы характеристики и соответствующие наборы границ, исходная выборка разбивается по границам на несколько кластеров и в каждом кластере обучается независимый классификатор. Для данной задачи были выбраны следующие характеристики для кластеризации:

- Пол (Мужской, Женский, Юридическое лицо)
- Дата регистрации ($>$ 2 лет, $>$ 1 года, $>$ 3 месяцев, $<$ 3 месяцев)
- Использует только интернет (телефон, модем)

| | | Мужской | | | Женский | | | Юр. лицо | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | prec | recall | AUC | prec | recall | AUC | prec | recall | AUC |
| $<$ 3 месяцев | телефон | 0.76 | 0.79 | 0.82 | 0.75 | 0.85 | 0.80 | 0.89 | 0.88 | 0.95 |
| | модем | 0.68 | 0.76 | 0.64 | 0.72 | 0.90 | 0.70 | 0.86 | 0.41 | 0.80 |
| $>$ 3 месяцев | телефон | 0.73 | 0.52 | 0.84 | 0.76 | 0.58 | 0.87 | 0.59 | 0.61 | 0.86 |
| | модем | 0.66 | 0.52 | 0.75 | 0.66 | 0.66 | 0.83 | 0.70 | 0.63 | 0.80 |
| $>$ 1 года | телефон | 0.63 | 0.60 | 0.88 | 0.78 | 0.65 | 0.86 | 0.75 | 0.69 | 0.85 |
| | модем | 0.80 | 0.53 | 0.77 | 0.74 | 0.74 | 0.84 | 0.63 | 0.70 | 0.79 |
| $>$ 2 лет | телефон | 0.83 | 0.52 | 0.92 | 0.86 | 0.57 | 0.94 | 0.86 | 0.60 | 0.92 |
| | модем | 0.83 | 0.53 | 0.85 | 0.72 | 0.71 | 0.85 | 0.84 | 0.64 | 0.92 |

Таблица 3: Результаты классификации в кластерах.

В итоге выборка была разделена по данным характеристикам на 24 кластера и в каждом из них был обучен независимый классификатор – градиентный бустинг над решающими деревьями, который показал наилучшие результаты в ходе сравнений. Результаты классификации на тестовой выборке в каждом из кластеров представлены в таблице 3. Лучшие результаты показал классификатор в кластере $<$ 3 месяцев, телефон, юридическое лицо. Худший результат показал классификатор в кластере $<$ 3 месяцев, модем, пол мужской. Как видно из сравнений (см. Таблицу 4): классификатор основанный на кластеризации показал результаты лучше, чем рассмотренные ранее модели машинного обучения. Это говорит о том что в

разных кластерах одни и те же характеристики имеют разную важность, иначе говоря: в разных кластерах разные причины оттока абонентов.

| Модель | precision | recall | F1 | F0.5 | **AUC** |
|---|---|---|---|---|---|
| **Набор классификаторов** | 0.75 | 0.66 | 0.70 | 0.73 | **0.904**±0.004 |
| Градиентный бустинг | 0.70 | 0.64 | 0.66 | 0.68 | **0.842**±0.005 |
| Random forest | 0.72 | 0.60 | 0.65 | 0.69 | **0.832**±0.008 |
| Нейронные сети | 0.69 | 0.59 | 0.63 | 0.66 | **0.825**±0.007 |
| Логистическая регрессия | 0.63 | 0.37 | 0.46 | 0.55 | **0.842**±0.002 |

Таблица 4: Сравнение полученной модели с предыдущими.

# Заключение

В рамках данной работы была проведена обработка исходных данных, содержащих информацию об абонентах крупного российского мобильного оператора, было проведено сравнение результатов классификации моделей машинного обучения на обработанных данных. Была установлена степень влияния дополнительных и исходных характеристик на результат обучения. Был построен классификатор на основе кластеризации, достигающий 0.90 AUC на тестовой выборке.

# Список литературы


[1] V Umayaparvathi and K Iyakutti. Applications of data mining techniques in telecom churn prediction. *International Journal of Computer Applications*, 42(20):5–9, 2012.

[2] Michael C Mozer, Richard Wolniewicz, David B Grimes, Eric Johnson, and Howard Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *Neural Networks, IEEE Transactions on*, 11(3):690–696, 2000.

[3] Chih-Ping Wei and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, 23(2):103–112, 2002.

[4] Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3):5445–5449, 2009.

[5] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, 2006.

[6] Kristof Coussement and Dirk Van den Poel. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34(1):313–327, 2008.

[7] Thomas G Dietterich. Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition*, pages 15–30. Springer, 2002.

[8] Jon Scott Armstrong. *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer Science & Business Media, 2001.

[9] База данных устройств связи. `http://gsmarena.com` (Дата обращения 15 мая 2015).

[10] PM Lerman. Fitting segmented regression models by grid search. *Applied Statistics*, pages 77–84, 1980.

[11] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.