САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Пономарёв Артемий Александрович

Выпускная квалификационная работа аспиранта

Сегментация пользователей мобильных операторов с помощью моделей Больших Данных

Направление 09.06.01

«Информатика и вычислительная техника»

Заведующий кафедрой,
доктор физ.-мат. наук,
профессор Терехов А.Н.

Научный руководитель,
доктор физ.-мат. наук,
профессор Терехов А.Н.

Рецензент,
кандидат физ.-мат. наук Тихонов А.Б.

Санкт-Петербург

2018

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

За последнее время словосочетание «большие данные» набирает все большую популярность. Это словосочетание используют по поводу и без повода в книгах и публикациях о потенциальных возможностях маркетинга, промышленности, сервисных и обслуживающих организациях, телекоммуникационных услуг, и в целом любой отрасли, где предприятия и игроки рынка каким-либо образом собирают и накапливают информацию о своей клиентской базе. Однако, используя этот термин, авторы не всегда достаточно полно представляют, что на самом деле кроется за этим словосочетанием. Обычно имеются ввиду неструктурированные массивы сырых данных, которыми обладают компании. Но как такие массивы обрабатывать, и как потом интерпретировать результаты этой обработки — представляют немногие. За границей подобная обработка, анализ и в целом использование компаниями и корпорациями данных о своей клиентской базе — дело совсем не новое, но развитие направление Big Data получило лишь в последние несколько лет, когда на рынке у компаний появились технические возможности для накопления, обмена и обработки больших объемов сырых данных и информации. Широкое применение «Big Data» получила в государственных структурах Соединенных Штатов, в медицине, финансовой сфере и телекоммуникационной сфере, история развития данного направления подробно рассматривается в исследовании компании McKinsey [1]. Аналитика, структурирование и последовавшее за этим целевое применение информации о гражданах страны и/или о клиентах компании позволяет корпорациям экономить миллионы долларов на логистике, затратах на персонал и рентабельности.

В России на рынке телекоммуникационных компаний в данное время все еще существует сильный разрыв между достаточно большим объёмом информации об абоненте, накопленной за многие годы, и между тем, как данная информация используется для внутренних целей компании. Для решения проблем, которые появляются в ходе планирования радиосети или которые появляются во

время «развития» абонента используется лишь незначимый процент данных о своих абонентах.

Однако, имея данные о трафиковом потреблении, денежных тратах и платежах абонента, представляя его геосоциальные признаки, и имея информацию о том, в каких местах чаще бывает клиент, спектр задач, которые может решать Big Data можно рассматривать очень широко.

Как именно сотовые операторы могут использовать данные о своих абонентах, нам подсказывает зарубежный опыт. Первые пробы пера по работе с большими данными были использованы в Соединенных Штатах как раз для упомянутого выше планирования радиосети. Уже тогда на основе текущей загрузки базовых станций и ее динамики операторы стали делать выводы о том, как быстро растёт население в определенных районах городов, где ведётся активная застройка и заселение. Исходя из этих данных, операторы развивали радиочастотную сеть. В дальнейшем интерес сдвинулся из технологической и инфраструктурной области в коммерческую, а именно — в сторону развития абонента. Индивидуальный подход к работе с клиентом был поставлен во главу угла работы с большими данными. С появлением и ростом проникновения смартфонов, вовлечением людей в социальные сети персонализация и индивидуализация только усилились. Это позволяло и позволяет компаниям, в том числе и мобильным операторам, вести работу с каждым клиентом практически индивидуально, направляя ему целевые сегментированные предложения. Известен случай, который я описал и в своей работе «Варианты использования Больших Данных в телекоммуникационном бизнесе» [2], когда одна из торговых сетей в США направила SMS своему клиенту — молодой девушке с рекламой товаров для беременных [3]. Такой вывод система этой сети сделала, проанализировав покупки клиента за последние месяцы. Возмущённый отец девушки посчитал оскорблённым себя и свою, как ему казалось, невинную дочь и подал на сеть в суд. Однако вскоре дело было закрыто, поскольку девушка действительно оказалась беременной. Это один из многих примеров правильного

(хотя и не очень удачного с эмоциональной точки зрения) таргетинга. Возможно, Вы и сами обращали внимание, что стоит Вам посетить сайт определённой тематики, и в следующую секунду Вы видите рекламу соответствующего товара или услуги в браузерах и, возможно, получаете SMS с тематикой сайта, который посещали еще вчера. Пока подобный опыт таргетированных предложений завязан в основном на трафик клиента в сети интернет и его использовании мобильных приложений в телефоне. С ростом интеграции компаний типа Yandex или Mail в сервисные организации на подобии мобильных операторов уровень персонализации работы с клиентов только вырастет.

1 Обзор литературы по моделям больших данных, используемых в телекоммуникационном и ИТ-бизнесе

За последние три года усилия в исследовательской работе были направлены на решение задачи с сокращением оттока активной клиентской базы Оператора и выявлением популярных маршрутов клиентов для оптимизации рекламных расходов Оператора. Надо отметить, что задачи по предсказыванию и прогнозированию оттока абонентов в телекоммуникациях в частности и в сфере услуг в целом не нова. Решению задачи выявления склонных к оттоку абонентов посвящено достаточно большое количество научных работ, которые имеют весьма практический смысл. Дело в том, что привлечение нового клиента в практически любой сфере стоит на порядок дороже, чем удержание старого клиента. Так вот, одна из первых работ, посвященных предсказанию оттока, датируется 1999 годом [4]. Работа была проведена учеными университета Колорадо и называлась сокращение оттока в беспроводной индустрии. Проблему, которую они обозначили — переток склонных к оттоку абонентов к оператору-конкуренту. Исследователи изучили текущий отток, поведение абонента в части пополнения счета, использования трафика, использования приложения операторов и количества и характера жалоб и обращений в службу поддержки. Далее они делали выводы относительно того, с кем должна быть проведена работа по лояльности с целью увеличения сохранения абонентов в базе.

Но, конечно, большинство работ по данной теме приходится на последние годы, самый расцвет наступил после 2010-х годов, когда у компаний появились технологические инструменты обработки и анализа накопленных сырых данных. Заслуживает внимания работа, проведенная учеными Пакистана — «Customer Churn Prediction in Telecommunication A Decade Review and Classification» [5]. В этой работе группа исследователей сделала обзор по шестидесяти одной работе и статье, в которых рассматриваются техники дата майнинга, используемые для предсказания оттока в области телекоммуникаций, обеспечивая таким образом дорожную карту для маркетинговых исследований. В 2012 году опубликовано

исследование задачи предсказания оттока клиентов всех представителей телекоммуникационной индустрии в Ирландии [6]. Было выбрано случайным образом 827 124 клиентов с реальным потреблением. В обучающей выборке было проверено 400 тысяч клиентов, которые не уходили из компании и 13 562 клиентов, которые ушли в отток. В проверочной базе было такое же количество клиентов, несклонных к оттоку и такое же количество ушедших абонентов. У каждого клиента рассматривалось 738 характеристик. Эти характеристики включали в себя демографический профиль: возраст, пол, социальный статус, пользовательская информация: тип тарифа, начисления, трафик, жалобы и обращения в колл-центр. К выборкам применялись различные варианты кластеризации и в итоге после сравнения работы алгоритмов машинного обучения самые лучшие результаты были продемонстрированы методами деревья решений и метод опорных векторов (метрики AUC от 0.85 до 0.90). Необходимо отметить, что из-за специфики конкретных задач каждого оператора следует использовать разные алгоритмы классификации и методов обработки данных. Эта мысль так же описана и приведена в рассматриваемой работе.

Проблема оттока стоит остро не только у мобильных операторов связи. В перенасыщенном банковском секторе банки уже достаточно давно живут только за счет «старой», уже привлеченной базы. Привлечение нового клиента — задача дорогая, поэтому отток «старых» клиентов представляет собой угрозу финансовой стабильности банков. Поэтому все рассматриваемые работы имеют схожу область исследования — отток абонентов, а цель данных работ — помочь коммерческим службам в подготовке данных и сокращении этого оттока.

V Umayaparvathi и K Iyakutti в работе «Applications of data mining techniques in telecom churn prediction» [7] рассматривают методы деревья решений и нейронных сетей в работе сингапурского оператора сотовой связи. Акцент в работе сделан на применении методов дата майнинга в предсказании оттока и выборе характеристик абонента с влиянием на отток каждого элемента. В итоге авторы пришли к выводу, что в решаемой ими задаче дерево решений является более предпочтительным вариантом относительно метода нейронных сетей в

части точности предсказания, и, дополнительно оказалось, что дерево решений построить оказалось легче. Но задача определения влияний каждого атрибута решена не была.

Michael C Mozer, RichardWolniewicz, David B Grimes, Eric Johnson, и Howard Kaushansky в работе «Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry» [8] анализируют отток провайдера услуг беспроводной связи по причине низкого удовлетворения качеством. В работе упор сделан на логистическую регрессию и нейронные сети, а также на нейронные сети и бустинг. Эксперимент базировался на основе данных о 47 тысячах клиентов. Помимо решения задачи об оттоке в работе описана важность корректного представления и хранения данных держателями сырых данных.

Chih-Ping Wei и I-Tang Chiu в работе «Turning telecommunications call details to churn prediction: a data mining approach» [9] анализируют отток у тайваньского оператора связи с помощью дерева решений. Эмпирическая оценка результатов показала, что модель прогноза оттока на основе трафикового поведения клиента имеет бо́льшую точность и эффективность при использовании самых последних голосовых вызовов. Зайдя в исследовании дальше, авторы пришли к выводу, что самые точные результаты получаются в пределах месячного интервала между построением модели и предсказанием оттока.

Yaya Xie, Xiu Li, EWT Ngai, и Weiyun Ying в работе «Customer churn prediction using improved balanced random forests» [10] на основе метода random forest дают рекомендации по предсказанию оттока по базе клиентов китайского государственного банка. Авторы обозначают проблему несбалансированность в распределении данных. Поэтому дополнительно одним из методов кластеризации они рассмотрели метод improved balance random forest. Суть данного метода заключается в том, что лучшие клиентские характеристики исследуются итеративно с помощью вариаций распределений между классами и установкой негативного признака характеристике при неправильной классификации в малых кластерах.

Shin-Yuan Hung, David C Yen, и Hsiu-Yu Wang в работе «Applying data mining to telecom churn management» [11] рассматривают базу клиентов тайваньского оператора, применяя k-means кластеризацию и нейронные сети. Проблема перенасыщенности тайваньского рынка началась в 1997 году, когда регулятор в лице государства перестал контролировать индустрию. Это привело к высокому уровню конкуренции, и проблема оттока старой базы в условиях такой конкуренции стала важна. Результатом этой работы стали выводы о том, что и нейросети и кластеризация k-means показывают одинаковые результаты и эффективность модели.

Что касается западного опыта, тут оказалась интересной работа бельгийских специалистов Kristof Coussement и Dirk Van den Poel. Они анализировали подписчиков печатных изданий бельгийской компании в работе «Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques» [12]. На основе данных, которые собирались в CRM-системах изданий, авторы провели исследование с построением вектор-машин. Техника исследования строилась на основе кросс-оценок, а результаты вектор-машин в дальнейшем сравнивались с кластеризацией методами логистической регрессии и random forest. Исследование показало хорошие показатели обобщения при применении к данным маркетинговых систем. Тем не менее, процедура оптимизации параметров играет важную роль в прогнозировании производительности. Авторы показали, что только при применении оптимальной процедуры выбора параметров векторные машины превосходят традиционную логистическую регрессию, тогда как случайные леса превосходят оба вида опорных векторных машин. В качестве существенного вклада в работе дается обзор наиболее важных драйверов оттока. В отличие от исследований в телекоме, например, стоимость подписки и траты клиентов не играют важной роли в объяснении оттока. Существенным оказалось влияние переменных, описывающих взаимодействие между клиентом и компанией.

В целом в большинстве работ, перечисленных выше, сравниваются несколько моделей машинного обучения на одном наборе данных. Как оказалось,

это — нормальная практика построения хорошего классификатора. В работах, анализировавших данные на основе логистической регрессии, этот базовый алгоритм классификации показал хорошие результаты кластеризации. В нескольких работах перед обучением проводится обработка данных. В работе бельгийских специалистов исходные данные содержат множество характеристик, поэтому перед обучением применяется алгоритм уменьшения размерности PCA (метод главных компонент). Собственно, этот метод мы в дальнейшем применили в работе с предсказанием оттока отечественного Оператора.

В работах тайваньских специалистов лидирующие результаты показывает модель машинного обучения — деревья решений. Нейронные сети в рассмотренных работах показали результаты близкие к лучшим. В целом, итоги 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: 1. Деревья решений; 2. Random forest; 3. Логистическая регрессия; 4. Нейронные сети.

Что касается задач по построению маршрутов клиентов, литературы по данному направлению значительно меньше. Но в последнее время исследование последовательностей активностей абонентов операторов связи становится актуальной темой для исследований. Так в работе Laasonen Kari «Clustering and Prediction of Mobile User Routes from Cellular Data» [13] были исследованы закономерности временных рядов активностей абонента для предсказания направления его движения, а именно для определения следующей базовой станции сотовой связи, которую абонент посетит. В работе авторы также пытались объединять схожие пути в группы, для сравнения путей использовалась мера, сходная с мерой Жаккара, которая для путей p и q определяется по формуле:

$$¿(p,q)=\frac{|p\bigcap q|}{|p\bigcup q|} \quad ,$$

но учитывающая порядок следования элементов (базовых станций). В данной работе исследователи не обладали информацией о пространственном расположении базовых станций, поэтому работали лишь с точным совпадением элементов пути абонента.

Тема сравнения путей и выделения кластеров путей рассмотрена в работе Saravanan Pravinth Samuel и Pavan Holla «Route Detection and Mobility Based Clustering» [14]. В данной работе исследователи обладали информацией о расположении базовых станций. Для сравнения путей абонентов использовался метод поэлементного сравнения базовых станций с определенным временным промежутком. Для каждой активности первого абонента в заданном временном интервале искалась активность второго абонента вблизи этого же места, результаты суммировались. Для кластеризации путей использовался алгоритм *QT*, изначально разработанный для кластеризации геномных последовательностей. Данный алгоритм имеет временную сложность $O(n^3)$, где n — число путей для кластеризации, что не позволяет использовать его для больших объемов данных. Целью же работы являлось определение метода, который позволит исследовать ежедневные маршруты пользователей и анализировать паттерны движения городских масс на основе данных сотовых операторов. Объединение результатов построения маршрутов с данными о клиентах из тех же систем операторов позволит сформировать фреймворк для работы по таргетированным рассылкам в рамках Location-Based-Advertising.

2 Обозреваемые задачи и методы их решения

2.1 Постановка задач об оттоке абонентов и смене аппарата у абонентов

Имеется в наличии абонентская база оператора. По каждому клиенту есть одинаковый набор данных о его денежных начислениях, платежах, тарифе, трафике SMS, передачи данных и голосовых услуг в разрезе различных направлений. Дополнительно есть информация социально-демографического типа: пол, возраст, часто посещаемые места на основе выборки трафика с базовых станций. Понятно, что часть данных о поле и возрасте могут быть некорректными, в случае, когда сим-карты семьи оформлены на одного члена семьи, например. Но, согласно статистике, таких сим-карт в базе менее 2% и на результаты исследований они не влияли. Необходимо определять тех абонентов, которые в течение месяца перестанут быть активными и уйдут в отток, т.е. перестанут пользоваться любым видом мобильного трафика и тратить денежные средства либо абонентов, которые в течение месяца сменят телефонный аппарат.

Важность задачи определения склонных к оттоку абонентов обуславливается тем, что проникновение мобильной связи в стране близится к 200%. Это значит, что на каждого жителя страны в среднем скоро будет приходиться минимум 2 сим-карты любых операторов мобильной связи. Это, соответственно, означает, что с каждым годом операторам всё сложнее становится привлекать новых абонентов. Понятно, что проникновение в 200% не говорит о том, что реально каждый житель имеет по две сим-карты. В России есть населенные пункты, где мобильная связь может быть недоступна, есть социальные группы, которые не являются клиентами какого-либо оператора. Кроме того, в данные 200% попадают сим-карты сегмента M2M, то есть сим-карты, установленные в банкоматы, в автомобили, в устройства категории IoT. Тем не менее, новые подключения операторов в большей степени представляют собой переподключения собственной базы или абонентов конкурентов, склонных к смене оператора из-за выгодных ценовых предложений. Очевидно, что

привлечение нового клиента — это дополнительная расходная нагрузка для оператора, выраженная в комиссионных вознаграждениях, в расходах на маркетинговое привлечение. Таким образом, задача удержания «старого» абонента и предсказания оттока имеет вполне конкретные экономические основания.

Актуальность второй задачи по выявлению абонентов, склонных к смене аппарата обусловлена следующим: согласно имеющимся данным Оператора, клиенты, которые используют смартфоны, имеют больше ARPU (Average Revenue per User), иными словами приносят больше денег Оператору, чем клиенты, пользующиеся так называемыми feature-фонами (обычными телефонами). Основная причина такой разницы кроется в технических возможностях аппаратов. На смартфонах у абонентов есть возможность пользоваться мобильным интернетом, мобильными приложениями, которые потребляют трафик. Таким образом, эта задача должна решать проблему ускорения перехода абонента с простого устройства на более современное (смартфон или планшетофон). Определение таких клиентов позволит Оператору проводить точечную работу силами сегментного маркетинга для смены аппарата клиентами.

Набор анализируемых характеристик и переменных для этих задач, очевидно, должен быть различен. Если в случае с предсказанием клиентов, склонных к оттоку нас будут интересовать данные о потреблении голосового трафика абонента в разрезе направлений, трафика передачи данных и его начислениях и платежах, то в случае задачи по выявлению склонных к смене аппарата клиентов более интересны его потребление трафика передачи данных, тип его устройства, частота выхода в сеть и структура ARPU.

Для решения обеих задач была собрана рабочая группа, целью работы которой была разработка или адаптация метода машинного обучения на основе предоставленных данных оператора связи. Методы обучения должны были решить задачу предсказания оттока клиентов компании и выявление абонентов, склонных к смене типа аппарата.

Как уже было написано выше, клиентские данные, в том числе в телекоммуникационной среде, изучаются за рубежом достаточно давно. Поэтому

мы обратились к европейскому опыту в Ирландии и телекоммуникационным операторам Юго-Восточной Азии и посмотрели, что делали с имеющимися данными исследователи оттока в сингапурских и тайваньских операторах связи. Кроме того, во время поиска примеров машинного обучения попалась достаточно интересная статья по базе данных одного из китайских банков. Все исследования касались движения клиентской базы этих компаний, соответственно, представляли для нас достаточно большой интерес, поскольку задача предсказания оттока - это тоже по сути своей движение клиентской базы. Модели машинного обучения в этих исследованиях включали в себя логистическую регрессию, нейронные сети, деревья решений, random forest и k-means кластеризацию и рассматривались в этих исследованиях в комбинациях друг друга. Подход с комбинацией методов обучения — вполне оправданное решение аналитиков на этапе обучения машины. Да, структура клиентской базы и область деятельности компаний схожи, но на момент старта исследований непонятно, какой из методов сработает. Поскольку в этих работах в итоге был выделен набор перспективных методов, показавший лучшие результаты для типа задач, связанных с оттоком клиентской базы, было решено в нашем исследовании опираться на этот набор методов обучения: «Градиентный бустинг», «Random forest», «Логистическую регрессию» и «Нейронные сети».

2.2 Описание хода решения задачи

Работы по задаче были спланированы в два этапа, для обоих из которых были подготовлены выборки по клиентским данным. Первый этап представлял собой машинное обучение на основе фактических реальных данных клиентов, и на этом этапе машине передавались параметры каждого клиента и интересующий нас в ходе решения задачи результат этого клиента (то есть остается клиент в базе или уходит в отток). На этапе тестирования на второй выборке данных на тех же параметрах новых клиентов мы провели предсказания и оценили результаты по выбранным метрикам.

Для оценки эффективности предсказания данных задач мы определили наиболее важными следующие метрики - precision, recall и AUC. Для определения метрик нам понадобится определить следующие понятия для задач классификации:

- истинно-положительные категории/элементы или true positives — это категории/элементы, которые должны были попасть в выборку и попали в следствие решения задачи классификации
- ложно-положительные категории/элементы или false positives — это категории/элементы, которые не должны были попасть в выборку, но попали в следствие решения задачи классификации
- ложно-отрицательные категории/элементы или false negatives — это категории/элементы, которые должны были попасть в выборку, но не попали в следствие решения задачи классификации
- истинно-отрицательные категории/элементы или true negatives — это категории/элементы, которые не должны были попасть в выборку, и в следствие решения задачи классификации мы их и не взяли

Так вот, precision — это мера точности, которая показывает точность определения положительных ответов. Иначе говоря,

$$\text{Precision} = \text{true positives} / \text{any positives}$$

Чем ближе к единице precision, тем меньше неправильных определений категорий, которые мы посчитали правильными.

Далее, recall — можно определить, как меру полноты. Эта мера показывает, как хорошо мы в ходе классификации угадали положительные ответы из всех положительных в выборке. Иначе говоря,

$$\text{Recall} = \text{true positives} / (\text{true positives} + \text{false negatives})$$

По сути эти два показателя друг друга дополняют и в большей степени зависимы друг от друга в обратной пропорциональности. Третья метрика, которую мы выбрали — AUC. AUC представляет собой количественную интерпретацию ROC-кривой. Area Under Curve — площадь между ROC и осью, на которой расположены ложно-положительные классификации. Чем ближе к единице AUC, тем лучше классификатор.

На первом этапе обучения для каждого метода в качестве исходных данных мы использовали следующие параметры клиента: объем голосового трафика в разрезе направлений звонков. Смотрели входящие и исходящие вызовы внутри сети, звонки на номера сторонних операторов внутри региона подключения, звонки на междугороднее и международное направление. В целом было выбрано 10 направлений вызовов. Дополнительно анализировался SMS-трафик в разрезе входящие и исходящие сообщения и объем трафика передачи данных.

Данные были представлены по выборке чуть менее 400 тысяч абонентов. Было проанализировано потребление этих абонентов в течение последних 15 месяцев. Дополнительно в качестве характеристики абонента был выбран уникальный идентификатор устройства связи — TAC (TypeAllocationCode). Данный идентификатор позволяет установить тип устройства клиента (смартфон, планшет, кнопочный телефон), вендора аппарата, версию операционной системы, год выпуска модели, возможность поддержки Wi-Fi, технические характеристики аппарата (размер дисплея) и характеристики поддерживаемых поколений сети (2G/UMTS/LTE).

Анализ тестовой выборки на первом этапе не показал существенных достижений на этапе обработки. Установить степень влияния каждого параметра на отток не удавалось и не было понимания, насколько правильно проведена кластеризация параметров. До начала работ по исследованию были предположения относительно того, что наличие большого количества параметров в выборках позволит увеличить точность предсказания и оценки. Однако после выполнения первого подхода возникло обратное предположение о том, что параметров слишком много, и они негативно сказываются на обучении и предсказании. Поэтому было решено провести оценку важности выбранных параметров и оценить, какие из них оказывают наибольшее влияние на результат. В ходе оценки выяснилось, что наиболее значимыми характеристиками показали себя дата регистрации (иначе говоря, срок жизни клиента в сети), общее количество голосовых входящих минут и количество потреблённого трафика передачи данных. Среди характеристик, влияющих в меньшей степени на отток

оказались исходящие звонки по направлению межгорода и городских номеров. В целом, данные результаты поддаются логике: в текущих реалиях пользования услугами связи городские телефоны действительно теряют свою популярность, а звонки на межгород являются больше разовой необходимостью, чем каким-то постоянным действием, за исключением целенаправленного использования, например, приезжими студентами, на основании которого можно делать выводы об оттоке.

Учитывая проведенную оценку характеристик, малозначимые параметры были исключены. Выборка была дополнена данными по месту жительства клиента (домашний регион), его полу, возрасту и признаку юридического статуса (т.е. проверяли, физическим лицом является клиент или договор заключался на юридическое лицо). Как мы выяснили, размерность данных может негативно сказываться на результатах кластеризации, поэтому 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 социально-демографические параметры абонентов, были выбраны все, имеющиеся в распоряжении, а из признаков, характеризующих трафиковую и платежную активность были выбраны только те, которые представляли собой отношение долей каждой услуги, связанной с вызовами за первый и третий месяц, а также 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 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 $x^{¿} = \frac{x - \acute{x}}{S^{¿}}$

где $S^{¿}$ — среднее квадратическое отклонение, $\acute{x}$ - 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. И уже с этими параметрами метрики тестовой выборки стали показывать результаты на порядок лучше. Кроме того, перед одним из очередных шагов по тестированию выборки была проведена группировка параметров. Мы объединили несколько параметров вместе и провели тестирование на объединенных группах параметров. В результате перебора вариантов кластеров наиболее успешные результаты были получены для разреза: юридический статус + пол + срок жизни в сети + пользователь телефона/смартфона. Мы увидели, что при разграничении этих показателей мы получаем весомые показатели precision и recall.

2.3 Анализ результатов

Если кратко подвести итоги, то выяснилось, что для абонентов, которые являются пользователями телефонов со сроком жизни в сети Оператора менее 2 лет отток предсказать достаточно сложно. С коммерческой точки зрения это возможно объясняется тем, что в последнее время Операторы проводят достаточно много акций с оборудованием. Например, когда при покупке устройства предоставляется скидка на оборудование и/или бесплатный трафик в течение долгого периода времени. Поэтому вместе с качественными подключениями (абонентами, которые выбрали целенаправленно оператора и планируют активно пользоваться услугами связи) оператор набирает в базу абонентов низкокачественных, которые используют акционный трафик и дальше устройством не пользуются. Что касается пользователей устройств, отличных от модема, вполне логично, что машина более точно и полно предсказывает результаты по клиентам со сроком жизни более 2 лет, поскольку более полные данные позволяют отследить сезонность потребления и наметить точки снижения активности в сети как индикатор того, что клиент склонен к оттоку.

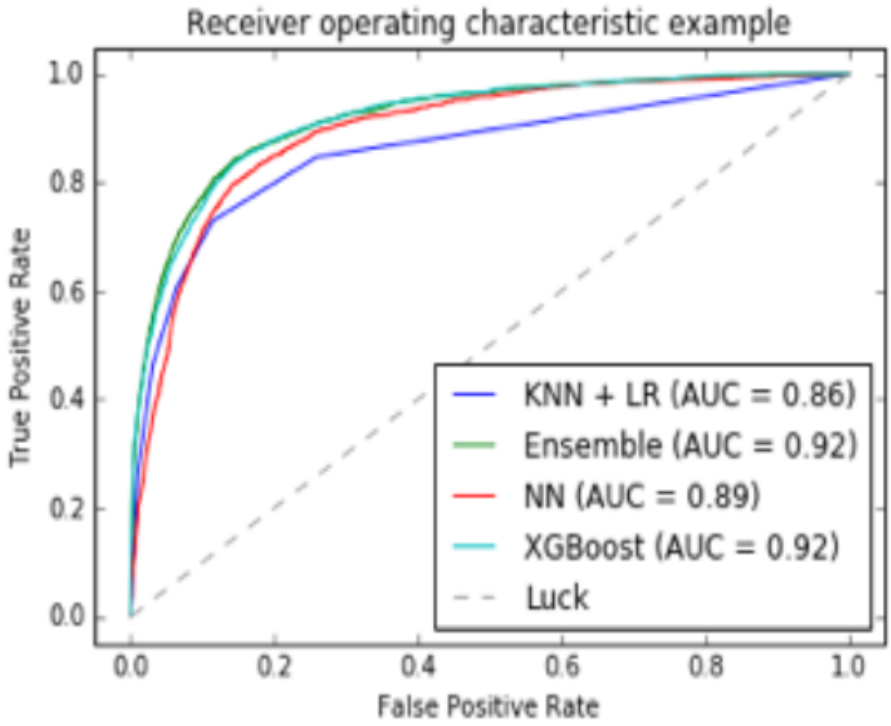
Рассмотрим подробнее результаты классификации на основании графика ROC и сравнении основных метрик в таблице 1. Выбор этих метрик вызван тем, что они в большей степени, чем accuracy учитывают правильность отнесения объекта к положительному классу, в то время как доминирующий в выборке класс - отрицательный. Определение каждой метрики и метод расчета я привел выше.

Таблица 1. Сравнение метрик моделей

Модель	**Precision**	**Recall**	**AUC**	**Accuracy**
Ансамбль	0.75	0.72	0.92	0.88
Градиентный бустинг	0.74	0.69	0.92	0.87
Нейронная сеть	0.70	0.62	0.89	0.86
Ближайшие соседи + случайный лес	0.74	0.60	0.86	0.87

График ROC на рисунке 1 иллюстрирует метрику AUC, видно, что лучшие результаты показывает ансамбль и XGBoost.

Рисунок 1. ROC метрик первой задачи



К минусам исследования вопроса по оттоку клиентов надо отнести то, что, в итоге, всё-таки, не было получено пользовательского интерфейса, с помощью

которого конечные пользователи могли бы загружать данные в необходимой детализации и получать прогнозные данные по клиентскому оттоку. Остался открытым вопрос практического применения на боевой базе, поскольку интерфейс еще предстоит доработать.

Аналогичная работа, проведенная в рамках второй задачи по анализу клиентов, склонных к смене аппарата, набора данных, выгруженных для первой задачи, оказалось недостаточно. В качестве триггера были использован момент (дата) смены TAC. Однако, в имеющейся выборке в 400 тысяч клиентов оказалось только около 25 тысяч человек, которые пользовались телефонами, сменив их в дальнейшем на смартфоны. Группа столкнулась со сложностью разделения выборок на обучение и тестирование ввиду малого количества абонентов и с невозможностью определить значимость параметров ввиду малого количества записей. Решение задачи предполагается повторить и проверить на новых выборках 17-18 годов, когда проникновение смартфонов стало на порядок выше и смена аппарата с кнопочного на обычный уже не такое редкое явление.

2.4 Постановка задач о маршрутах абонентов для размещения рекламных носителей

Еще одна задача, решаемая в рамках работы по сегментации пользовательских данных и построения моделей Big Data в работе сотовых операторов, звучала так: провести анализ маршруты абонентов с целью нахождения наиболее популярных из них. Нахождение популярных маршрутов движения населения должно помочь в планировании размещения рекламных носителей. В маркетинге у рекламных носителей измеряется показатель GRP.Он показывает, сколько раз рекламное сообщение с определенного носителя попадается на глаза целевой аудитории компании за период рекламной кампании. При его расчете учитываются не уникальные, а все контакты аудитории с носителем. Если очень просто, то GRP рейтинг — это соотношение контактов, которые видели рекламное сообщение, по отношению ко всем контактам, которые могли его видеть. Поэтому актуальность задачи построения популярных

маршрутов и размещения рекламных носителей на них связано с тем, чтобы эти носители увидело как можно большее количество существующих и потенциальных клиентов.

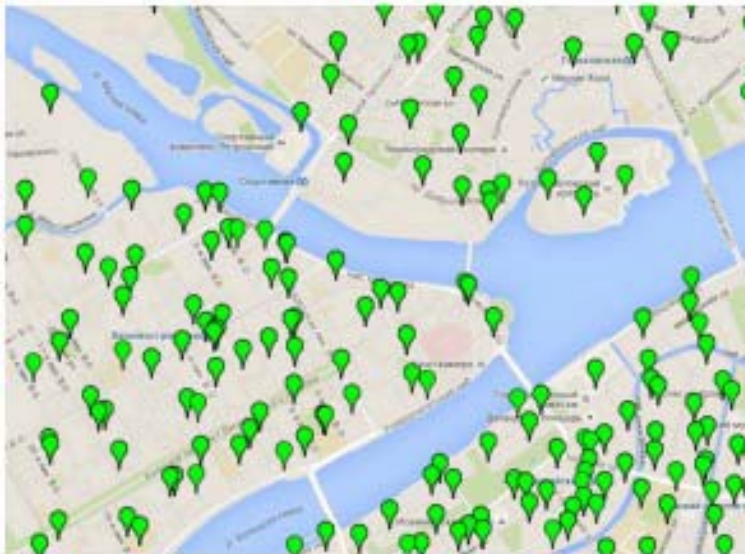
Учитывая, что анализ строился на данных клиентов из Санкт-Петербурга, задача была сформулирована следующим образом: построить наиболее популярные маршруты клиентов на пути их следования на спортивные мероприятия на стадионе «Петровский» (Санкт-Петербург) с целью оптимального размещения рекламных носителей крупных форматов для повышения индексов просмотра рекламной информации. Коммерческая актуальность данной задачи обусловлена следующим: стоимость размещения информации на рекламных носителях в центре города значимо превышает стоимость на его окраинах за счет более объемного охвата аудитории. В условиях ограниченных рекламных и маркетинговых бюджетов требуется наиболее оптимальное использование таких носителей в центре города. При этом задача максимального охвата аудитории у Оператора остается. Построение маршрутов решает задачу экономии расходов на маркетинговые кампании без сокращения эффективности охвата и работы с целевой аудиторией. Определение задачи построения маршрутов в дни спортивных мероприятий на стадионе «Петровский» связано с тем, что на этом стадионе выступает футбольный клуб «Зенит», с которым у исследуемого оператора связи периодически проводятся совместные рекламные акции и кампании, на основе которых можно судить об их постоянных партнерских отношениях. Можно утверждать, что Оператор видит свою целевую аудиторию, в том числе, и среди болельщиков этого клуба, которую можно привлечь в абонентскую базу специальными предложениями.

2.5 Описание решения

Для анализа и решения задачи были взяты обезличенные данные (имелся только User ID абонента из биллинговой системы без персонализации) по всем возможным видам трафика: трафик SMS, голосовой трафик и трафик передачи данных. Дополнительно обрабатывалась информация о местоположении базовых

станций, на которых происходила регистрация вышеупомянутого трафика. Данные по трафику мы взяли за временной промежуток тех дней, когда на стадионе проводилось спортивное мероприятие с участием футбольного клуба. Были взяты данные в день проведения матча, часовой промежуток при этом составлял следующий интервал: за три часа до старта матча и два часа после матча. Мы взяли данный временной интервал на основе экспертного анализа, исходя из того, что данного промежутка времени человеку хватает, чтобы добраться до стадиона и затем уйти со стадиона. Данные по каждой базовой станции были представлены в виде пары (LAC, Cell ID), где LAC — это уникальный код городской зоны, а Cell ID — уникальный номер базовой станции в этой зоне. Таким образом, мы смогли идентифицировать месторасположение базовых станций на карте города как на Рисунке 2

Рисунок 2. Идентификация положения базовых станций в г. Санкт-Петербург



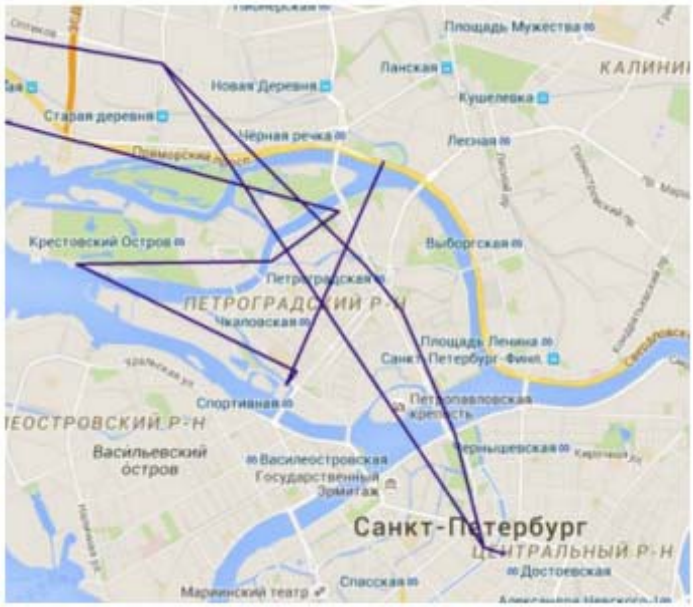
Клиентские данные были взяты в разрезе User ID со временем регистрации любого перечисленного выше трафика на базовой станции. Поскольку пара LAC, Cell ID для каждой базовой станции была определена, фактически по каждому клиенту мы получили Таблицу 2 следующего вида:

Таблица 2. Клиентские данные в разрезе трафика и базовых станций

Время	**Тип передачи информации**	**Cell ID**	**LAC**	**User ID**
10.09.02	Входящие SMS	53119	4708	8875
10.09.04	Входящие SMS	53119	4708	8875
10.24.18	Входящие SMS	33751	4708	8875

Иначе говоря, мы получили возможность изучать расположение клиента в различные моменты времени и его передвижение исходя из временной последовательности регистрации на базовых станциях. Условное передвижение в течение дня теперь можно было наблюдать в следующем виде:

Рисунок 3. Условное передвижение клиента в течение дня



Понятно, что данные маршруты достаточно условны, поскольку клиент не передвигается исключительно по прямой. Если была поездка на метро — это один маршрут, если на автомобиле — другой. Но, тем не менее, общие очертания движения и, самое главное, вершины и ребра графа передвижения абонента были определены, направления движения от одной вершины до другой тоже. Учитывая узкий временной промежуток времени (три часа до матча и два часа после матча),

допущение, что человек выходил за пределы текущего графа, при этом не отправляя сигналы на другие базовые станции, маловероятно.

Далее, учитывая имеющиеся перечисленные данные в разрезе идентификаторов каждого клиента, мы попробовали выполнить кластеризацию вершин и ребер различными методами: кластеризацию EM-алгоритмом, агломеративную кластеризацию с неевклидовыми метриками, GRGPF, алгоритмами кластеризации K-Means и Mini-BatchK-Means, а также алгоритмом кластеризации графов MCL. Задачей каждого алгоритма кластеризации являлось нахождение так называемых важных ребер, т. е. участков передвижения абонентов, которые потенциально представляют наибольший интерес с коммерческой точки зрения — размещения рекламных носителей. Проще говоря, надо было определить ребра и вершины, через которые прошло (зарегистрировалось с трафиком) наибольшее количество абонентов. Напомню, что для каждого абонента (User ID) у нас имеется только соответствие базовой станции в момент трафиковой транзакции. Исходя из таких первоначальных данных в ходе работ необходимо было сделать следующие допущения:

- Положение клиента в определенный момент времени определяется координатами базовой станции — последней, в пределах которой клиент зарегистрировался
- Нам не требуется знать координаты клиента непрерывно в любой момент времени. Перемещение клиента считаем дискретным, и для построения маршрута достаточно знать, где клиент находится через определенный фиксированный интервал времени.
- Не рассматриваем базовые станции, находящиеся вне центральных районов города вдалеке от стадиона. Трафика в исследуемые часы у рассматриваемых клиентов практически не было. Те же абоненты, которые регистрировались на таких базовых станциях, выборку не портили ввиду малого количества.
- Данные с нулевыми позициями, например, там, где по скрытой причине не определились координаты базовой станции, или где не произошла связка User ID с трафиковой транзакцией, которая должна была

зарегистрировать клиента на базовой станции, были приняты за погрешность выгрузки и также не учтены в анализе.

В итоге, если обратиться к результатам кластеризации графов MCL, нам удалось выделить набор важных ребер и итоги кластеризации можно представить следующим образом, как в Таблице 3:

Таблица 3. Распределение загруженности маршрутов по ребрам

День	**1**	**5**	**10**	**20**
03.10.2015	26	13	6	2
20.10.2015	30	15	7	2
24.10.2015	27	14	6	2
31.10.2015	24	11	5	2
21.11.2015	23	10	5	2
24.11.2015	30	16	7	2

Значения в столбцах таблицы 1, 5, 10 и 20 — это количество предложенных важных маршрутов, а значения в ячейках — это процент абонентов, которые прошли хотя бы по 1, 5, 10 и 20 маршрутам в указанный день. Иными словами, интерпретация показателя 6% в первой строке в третьем столбце следующая: по первому маршруту шло всего 26% абонентов, далее на каком-то перекрестке абоненты разделились и пошли часть налево своим вторым маршрутом, часть прямо своим вторым маршрутом и так далее до 10-го маршрута. Таким образом, все вместе по одному набору 10 маршрутов прошло 6% абонентов. Сами маршруты представлены на Рисунке 4.

Рисунок 4. Пример выделения маршрутов в разрезе одного кластера

Кроме того, дополнительно мы решили проверить разброс в популярности маршрутов по дням. В исходных данных видна определенная неравномерность в

объеме прохождения маршрутов. Подобный разброс сначала был предположительно связан с погодными условиями. Проверив данные о погоде в эти дни, стало ясно, что погода не причем. Во все рассмотренные дни она была примерно одинаковой — типично осенней и петербуржской, но без осадков. Тогда было решено проверить значимость дней недели. Предположение было, что в выходные дни большее количество болельщиков предпочитает пройти большие расстояния. Но на самом деле оказалось, что 20.10 и 24.11 — дни, когда маршруты были загружены в большей степени, являлись будними днями. В эти дни проходили игры Лиги Чемпионов, соответственно, приезжали европейские клубы, что вызвало больший интерес у болельщиков и, соответственно, более плотное перемещение у стадиона. Более высокая плотность прохождения людей по одному маршруту свидетельствовала о том, что случайных прохожих, шедших по индивидуальным маршрутам в эти дни было меньше.

Если обратиться к результатам кластеризации методами K-Means и агломеративной кластеризации, результирующие кластеры выглядели следующим образом, соответственно:

Рисунок 5. Представление кластеров после кластеризации методом K-Means

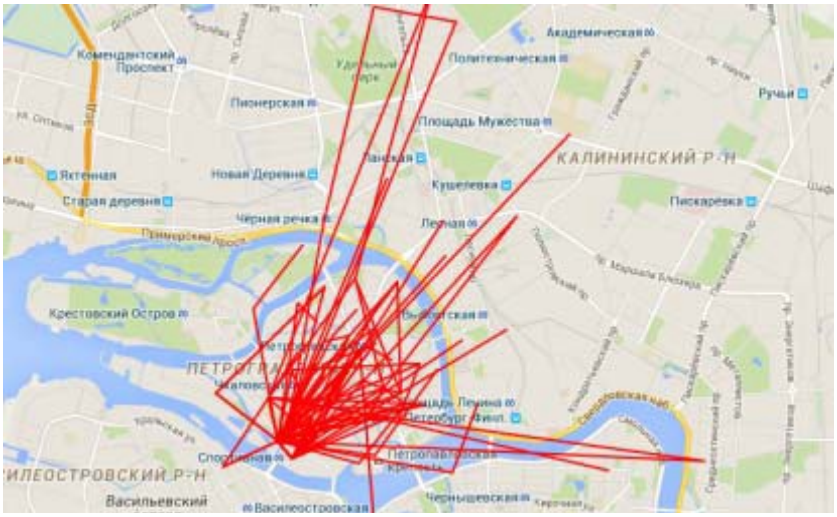
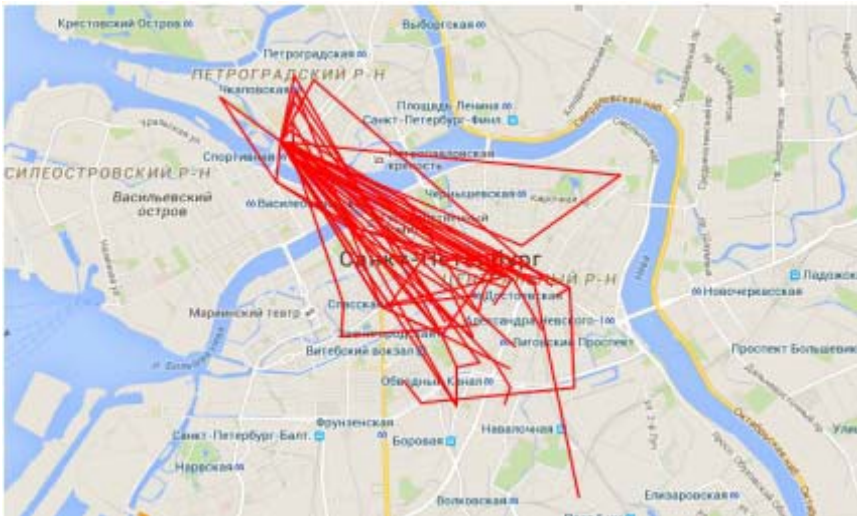


Рисунок 6. Представление кластеров после кластеризации методом агломеративной кластеризации

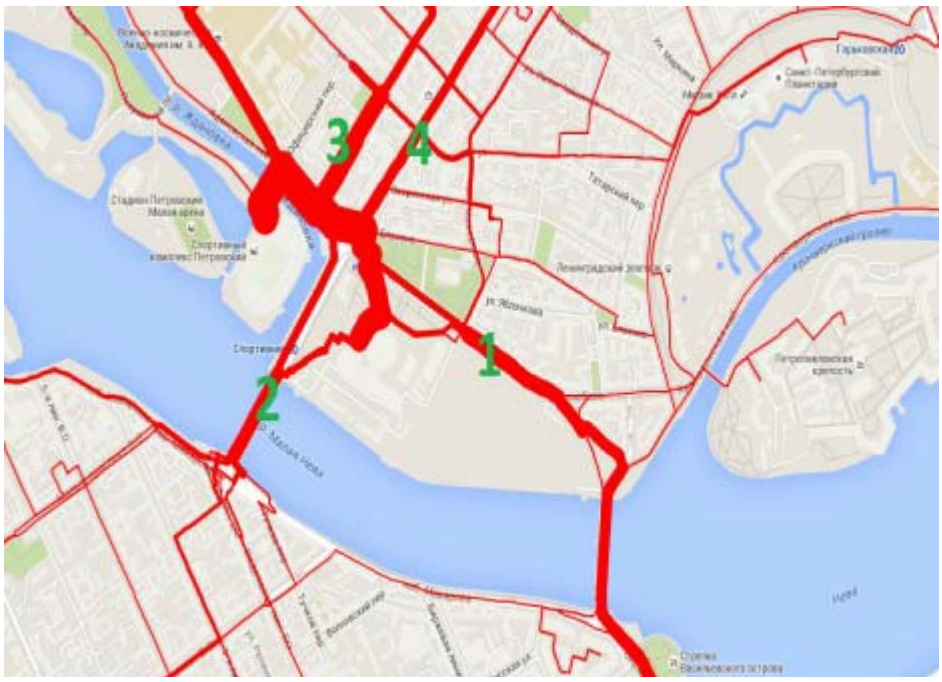


Когда мы попытались выделить самые популярные маршруты данными методами, мы столкнулись с проблемой недостаточно точной информации. Напомню, что мы допустили ряд предположений о дискретности маршрутов, а для большого количества абонентов число активностей было сравнительно малым. Кроме того, расстояния между регистрациями на базовых станциях (когда абонент совершал транзакцию трафика любого вида) оказались большими. Да, мы имели примерный маршрут клиента условно по прямой линии, но расхождение маршрутов по улицам было непонятно. Поэтому для уточнения самых популярных путей в данном случае мы решили прибегнуть к средству прокладывания маршрутов из GoogleMapsDirectionsAPI.

Для каждой пары действий абонента был проложен маршрут и полученные данные для каждого ребра и пары вершин визуализирован с учетом разбиения абонентов по кластерам. Таким образом, нам удалось выделить 4 наиболее популярных маршрута:

1 — пр. Добролюбова и Биржевой мост;
2 — Тучков мост
3 — Малый проспект Петроградской стороны
4 — Большой проспект Петроградской стороны

Рисунок 7. Загруженность маршрутов на пути к стадиону «Петровский»



Схожие результаты показала визуализация маршрутов на основе метода кластеризации GRGPF:

Рисунок 8. Визуализация маршрутов на основе метода кластеризации GRGPF

Качество кластеризации оценивали с помощью квадратов расстояний между объектами каждого кластера (SSQ), а похожие карты популярных маршрутов,

построенных двумя различными методами кластеризации, позволяют говорить о правильной группировке полученных клиентских данных.

2.6 Выводы результатов решения

Решение данной задачи и приведенные примеры построения маршрутов на стадион представляют не только теоретический, но и значимый практический интерес. Как я писал выше, в условиях необходимости сокращения коммерческих маркетинговых затрат и одновременном требовании акционеров о повышении показателей эффективности бизнеса компании вынуждены искать дополнительные меры по оптимизации своих затрат. Телекоммуникационный рынок находится в определенной стагнации в России и проведенный анализ может позволить не только сэкономить на маркетинговых исследованиях, которые, скорее всего, покажут аналогичные результаты и выдадут схожие рекомендации по размещению рекламных носителей, но и использовать эти наработки в решении других задач: размещение офисов продаж и обслуживания абонентов, в которых осуществляется продажа контрактов, оборудования и предоставляется сервис оператора, планирование дополнительных размещений базовых станций в часы пиковых нагрузок и т. д.

Полученные результаты в целом логичны и понятны. Достаточно высокую проходимость демонстрируют все популярные подходы к стадиону. Но результаты и ожидания приоритетов популярности маршрутов внутри группы этих маршрутов разошлись. Ожидания от маршрута по Тучкову мосту совпали с результатами. Данный маршрут оказался одним из двух самых популярных маршрутов. Что касается второго по популярности маршрута, то ожидаемые проспекты Петроградской стороны значимо уступили Биржевому мосту. Мы проверили проводились ли на момент анализа какие-либо ремонтные работы на этих участках, которые мешали пешеходным и автомобильным передвижениям. Подобных работ не велось, что говорит о том, что на результаты не влияли внешние городские факторы. Биржевой мост является хорошей рекламной площадкой, и на подходах к нему и на самом мосту сконцентрировано много

рекламных носителей поставщиков рекламных мест. Соответственно, проведенный анализ говорит о том, что концентрацию на этих маршрутах стоит усилить, снизив ее соответственно на Петроградской стороне, по крайней мере, в дни проведения рекламных акций, связанных с кобрендингом «Зенита» и Оператора. Низкая концентрация потоков на дальних подходах к стадиону также объяснима: достаточно высокий процент зрителей добирается до стадиона на автотранспорте, но в связи с проблемами с парковочными местами оставляет автомобили далеко от стадиона. За рулем трафиковая активность клиентов, разумеется, ниже, плюс перемещение между базовыми станциями происходит достаточно быстро, что не позволяет регистрировать клиента на каждом отрезке его пути. Этот вывод также может служить рекомендацией Оператору, направленной на концентрацию рекламных носителей на пути к стадиону. Коэффициент успешного просмотра рекламных носителей за рулем, если клиент не стоит в пробке, на порядок ниже, поэтому и их эффективность ниже, чем если бы они стояли на пути его пешеходного маршрута. Поэтому размещение на подобных носителях в дни проведения акций и рекламных кампаний стоит производить на маршрутах, по которым клиент идет пешком, уже оставив автомобиль.

ЗАКЛЮЧЕНИЕ

Все выполненные в рамках исследования работы имели не только теоретическую ценность, но и практическую. Основанный на реальных выгрузках и данных анализ по маршрутам применили на практике, и результаты были представлены Оператору. Исследования были признаны успешными, и тестовая рекламная кампания была запущена, основываясь на результатах исследования популярных маршрутов.

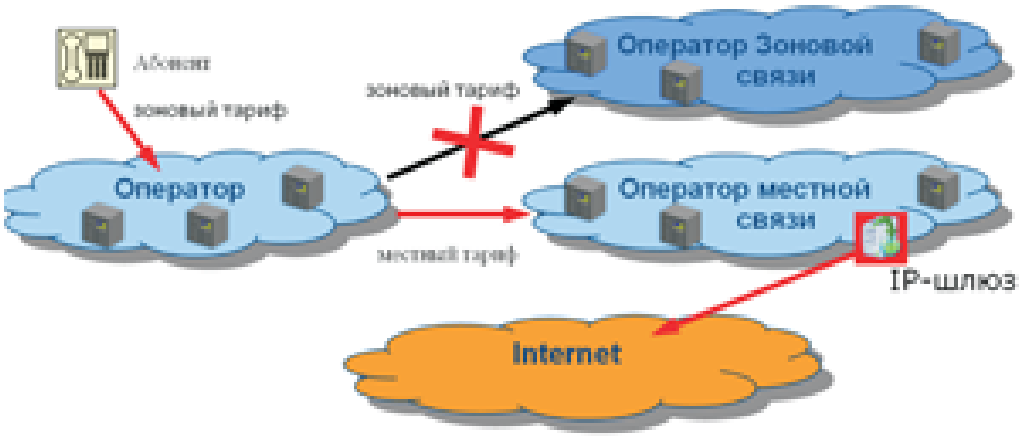
Хотя не всё получилось идеальным образом в части оттока, в частности отсутствие интерфейса для загрузки данных оператору компании, полученные результаты удовлетворили заказчиков исследования и дали нам опыт практического применения нового интересного направления информатики — машинного обучения. Кроме того, алгоритмы применили к фактическим данным существующих клиентов и на основе полученных результатов с частью абонентов была проведена работа по удержанию. К минусам исследования вопроса по оттоку клиентов надо отнести то, что в итоге, всё-таки, не было получено пользовательского интерфейса, с помощью которого конечные пользователи могли бы загружать данные в необходимой детализации и получать прогнозные данные по клиентскому оттоку. Проработка интерфейса лежит в плоскости глобального проекта с Оператором и подразумевает заключение отдельного коммерческого контракта и проведения закупочной процедуры. Что касается маршрутов пользователей, то их построение и кластеризации путей абонентов — задача достаточно объемная и может представлять интерес не только для оператора, но и для государственных органов, для других коммерческих организаций. Результаты кластеризации маршрутов можно связать с задачами роста проникновения продуктов на основе технологий RFID и iBeacon, задачами разделения транспортных и пешеходных потоков, планирование городской инфраструктуры и др.

Следующими этапами планируется продолжение работ в части построения маршрутов и работы с оттоком. Дополнительно к этому прорабатывается задача

по определению фродовых звонков на SIM-боксах. SIM-бокс представляет собой устройство-репозиторий SIM-карт и используется для незаконного подключения к телефонным сетям. Устройство конвертирует международные звонки в локальные звонки домашнего региона. Существует несколько видов мошенничества с использованием данного вида устройств:

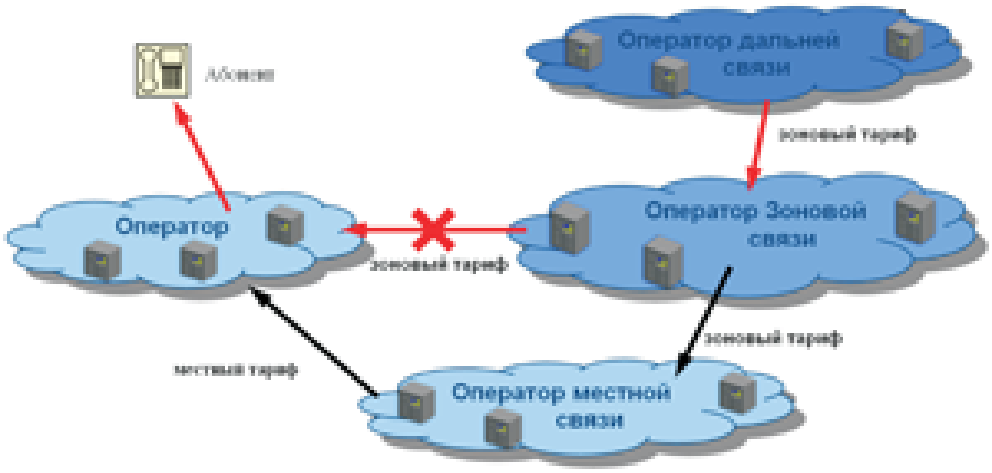
- Мошенничество операторов связи. Оператор прогоняет телефонный трафик с использованием IP-телефонии. Чаще всего речь про присоединенных операторов связи. Присоединенный оператор перенаправляет междугородный и международный (МГ/МН) трафик через IP-шлюзы, минуя операторов зоновой связи, т.е. уводят трафик через IP-шлюзы. Схема, как это выглядит, представлена на Рисунке 9

Рисунок 9. Прогон трафика без зонового приземления



- Мошенничество опять же оператора связи, когда последний организует пропуск трафика через сеть третьей стороны, т.е. оператор направляет вызовы другому присоединённому оператору, вызовы из сети которого тарифицируются (или не тарифицируются вообще) оператором зоновой связи по более низким тарифам. Схема представлена на Рисунке 10. Оба этих способа позволяют операторам обходить стандартные соединения по базовым станциям, присваивая прибыль традиционных операторов.

Рисунок 10. Обход трафика через присоединенного оператора



- Мошенничество без оператора. Группа физических лиц внутри себя договаривается о перенаправлении трафика через шлюз, минуя традиционное соединение через сотовых операторов. Та же схема, что представлена на Рисунке 9, только вместо переадресации трафика присоединенным оператором, абонент напрямую устанавливает соединение со шлюзом, который дальше совершает звонок по цене местного оператора. Таким образом традиционный оператор недополучает деньги на выручке от интерконнекта. Напомню, что интерконнектом называется связь между сетями операторов, а ставка интерконнекта — это стоимость терминации звонка при взаиморасчетах между операторами. Например, если абонент A оператора X звонит абоненту B оператора Y, то X заплатит Y сумму, равную количеству проговоренных минут * ставку интерконнекта
- Еще один вид мошенничества — случай, когда в SIM-бокс вставлена сим-карта оператора, оператор теоретически только выигрывает в выручке за счет выручки от интерконнекта, если звонок совершается с SIM-бокса на другого оператора связи. В этом случае другой оператор связи заплатит согласно установленной ставке интерконнекта. Но в данном случае это

политическое джентельменское соглашение между операторами, и считается фродом. Сим-карта блокируется, чтобы не генерить фродовый трафик

- И, наконец, обратная ситуация. Если в SIM-бокс вставлена сим-карта оператора другого оператора и звонки идут на оператора, чьи интересы мы хотим защитить, то последний теряет в выручке за счет недополучения выручки от интерконнекта, который мог быть получен за счет звонка по стандартным каналам. Кроме того, идет потеря выручки за счет выплат от интерконнекта оператору, чья сим-карта вставлена в SIM-бокс. В этом случае надо определять такую сим-карту стороннего оператора и сообщать ему о проблеме. По соглашениям, упомянутым выше, оператор блокирует такую сим-карту

Задача, которую необходимо решить — научиться за счет машинного обучение определять сим-боксы и находить такие сим-карты, которые описаны выше в двух последних случаях в течение получаса. Сейчас есть два вида фродовых сим-боксов:

- Простые SIM-боксы, которые сразу генерируют много исходящего голосового трафика, у них нет передачи данных и смс. Звонки на разные номера по всей России в течение короткого промежутка. Такие сим-боксы находятся в течение часа. Достаточно данных из биллинга, чтобы такую сим-карту определить. Обычно они подключают большой пакет с небольшим авансовым платежом, который дает возможность в течение одного двух дней использовать весь объем включенного трафика.
- «Умные» сим-боксы. Они генерируют входящие вызовы в том числе, бывает смс трафик и трафик передачи данных. Такие сим-боксы находят в течение 4-5 часов. Данные из биллинга подгружаются в BI-системы, где скриптами без машинного обучения делается гипотеза, что это фродовый клиент. На сим-

карту совершается тестовый звонок из страны, из которой генерируется трафик. Если не отвечает живой абонент в течение нескольких звонков, сим-карта блокируется.

Необходимо научиться за счет машинного обучение ловить сим-боксы первого и второго типов. У задачи вполне конкретное практическое применение с экономическим эффектом. Как было сказано выше, эта задача лежит в сфере интересов проработки и уже сейчас начинаются выгрузки данных для машинного обучения.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity, June 2011 // http://www.slideshare.net/blueeyepathrec/mckinsey-global-institute-big-data-the-next-frontier-for-innovation-competition-and-productivity
2. Пономарёв А.А. Варианты использования больших данных в телекоммуникационном бизнесе // Компьютерные инструменты в образовании. – 2015 – №4: 3-8
3. Slon Magazine – онлайн-журнал об экономике и политике. URL: http://slon.ru/specials/data-economics/articles/target (Дата обращения: 22.05.2015)
4. Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. Churn reduction in the wireless industry. URL: http://www.cs.colorado.edu/~mozer/Research/Selected%20Publications/reprints/churn_nips.pdf
5. Nabgha Hashmi, Naveed Anwer Butt and Dr.Muddesar Iqbal. Customer Churn Prediction in Telecommunication A Decade Review and Classification. URL: http://www.researchgate.net/profile/Nabgha_Hashmi/publication/257920014_Customer_Churn_Prediction_in_Telecommunication_A_Decade_Review_and_Classification/links/00b495261475ba6758000000.pdf
6. Bingquan Huang, Mohand Tahar Kechadi, BrianBuckley. Customer churn prediction in telecommunications. URL: http://www.sciencedirect.com/science/article/pii/S0957417411011353
7. V Umayaparvathi and K Iyakutti. Applications of data mining techniques in telecom churn prediction. International Journal of Computer Applications, 42(20):5–9, 2012 URL: https://pdfs.semanticscholar.org/4f96/e06db144823d16516af787e96d13073b4316.pdf
8. Michael C Mozer, RichardWolniewicz, 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 URL: https://pdfs.semanticscholar.org/ef46/da76583559b112820255104d7b9dfbe25b60.pdf

9. 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 URL: http://isiarticles.com/bundles/Article/pre/pdf/21396.pdf
10. 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 URL: https://pdfs.semanticscholar.org/4359/d76d3b36944553eb1d08befaf219122fbefd.pdf
11. 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 URL: https://pdfs.semanticscholar.org/9e39/a8fa0c10314199d4f99aeaafdb124ea435de.pdf
12. 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 URL: https://pdfs.semanticscholar.org/9223/0fcb6e68a86db7757267813db75f99f4ccc0.pdf
13. Laasonen Kari. Clustering and Prediction of Mobile User Routes from Cellular Data // Knowledge Discovery in Databases: PKDD 2005. –2005 URL: https://www.cs.helsinki.fi/group/context/pubs/pkdd05.pdf
14. Saravanan M Pravinth Samuel V Pavan Holla. Route Detection and Mobility Based Clustering // Internet Multimedia Systems Architecture and Application (IMSAA), 2011 IEEE 5th International Conference on. – 2011 URL: http://ieeexplore.ieee.org/iel5/6151916/6156331/06156372.pdf