Компьютерные инструменты в образовании, 2015
№ 4: 3–8
УДК: 004.622
http://ipo.spb.ru/journal

инженерия программного обеспечения

ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ В ТЕЛЕКОММУНИКАЦИОННОМ БИЗНЕСЕ

Пономарёв Артемий Александрович

Аннотация

В данной статье рассмотрены общие вопросы использования больших данных в практической плоскости коммерческой деятельности компаний. Автор делает акцент на задачи, которые могут быть решены с использованием больших данных в телекоммуникационной сфере. Проанализированы характеристики клиентов компании, которые позволяют судить об их потенциальном оттоке. Выявлены кластеры, для которых получены наиболее успешные результаты предсказания.

Ключевые слова: *большие данные, телекоммуникации, отток, svm, нейронные сети, вектор-машины.*

1. ВВЕДЕНИЕ

В последние годы всё большую популярность набирает словосочетание «большие данные». Его используют по поводу и без повода в статьях о будущих возможностях маркетинга, промышленности, телекоммуникационных услугах, сервисных и обслуживающих организациях, каким-либо образом собирающих информацию о своих клиентах. Но, используя этот термин, авторы не всегда представляют, что на самом деле кроется за этими словами, обычно имея в виду неструктурированные массивы каких-то данных, которыми обладают компании. Как такие массивы обрабатывать и, тем более, как потом результаты этой обработки интерпретировать, представляют немногие. За рубежом анализ и использование корпорациями данных о своих клиентах — дело далеко не новое, но развитие данное направление получило лишь в последние десятилетия, когда у компаний появились технические возможности для накопления, обмена и обработки больших объемов информации. Широкое применение «bigdata» получила в государственных структурах Соединенных Штатов, в медицине, финансовой сфере и телекоме [1]. Анализ, структурирование и последующее 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 телекоммуникаций на данный момент есть существенный разрыв между большим объёмом накопленной за многие годы информации об абоненте и тем, как эта информация используется внутри компании. Лишь малый процент данных о своих абонентах учитывается при решении проблем, возникающих, например, во время планирования сети или при развитии абонента.

Однако, имея данные о потреблении и платежах абонента, его геосоциальных признаках, о том, в каких местах чаще бывает клиент, спектр задач, которые может решать BigData, можно рассматривать достаточно широко.

Как конкретно мобильные операторы могут использовать данные о клиентах, подсказывает иностранный опыт. Первые опыты по работе с большими данными использовались в Соединенных Штатах для планирования радиосети: на основе текущей загруженности базовых станций операторы делали выводы о том, как быстро растёт население того или иного района, где ведётся активная застройка и заселение и, исходя из этих данных, развивали радиочастотную сеть. В дальнейшем интерес сдвинулся в сторону развития абонента: индивидуальность абонента была поставлена во главу угла работы с большими данными. С появлением смартфонов, вовлечением населения в социальные сети персонализация и индивидуализация только усилились, позволяя компаниям, в том числе и сотовым операторам, работать с каждым клиентом практически индивидуально, направляя ему таргетированные предложения. Известен случай [2], когда одна из торговых сетей в США направила SMS своему клиенту — молодой девушке с рекламой товаров для беременных. 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) таргетинга. Возможно, Вы и сами обращали внимание, что стоит Вам посетить сайт определённой тематики, и на следующий день Вы видите соответствующую рекламу в браузерах и, возможно, получаете SMS с тематикой сайта, который посещали еще вчера.

2. ПОСТАНОВКА ЗАДАЧИ

На первом этапе нашего исследования мы поставили перед собой несколько актуальных коммерческих задач:

1. Определение из выборки номеров абонентов, склонных к оттоку.
2. Определение из выборки номеров абонентов, склонных к смене аппарата.

Актуальность первой задачи обуславливается следующим: проникновение сотовой связи в России близится к 200%. Это говорит о том, что на каждого жителя страны скоро будет приходиться минимум 2 сим-карты любых операторов мобильной связи. Это, в свою очередь, означает, что с каждым годом операторам всё сложнее становится привлекать новых абонентов. Новые подключения больше представляют собой «перемалывание» либо собственной базы, либо абонентов конкурентов, более чувствительных к новым ценовым предложениям. Понятно, что привлечение нового абонента — это дополнительные затраты для оператора, выраженные в комиссионных вознаграждениях. Таким образом, задача удержания «старого» или существующего абонента имеет определённые экономические основания.

Актуальность второй задачи обусловлена следующим: согласно имеющимся данным операторов, абоненты, пользующиеся смартфонами, имеют большую доходность, чем абоненты, пользующиеся так называемыми feature-фонами (обычными телефонами). В основном, это происходит за счет того, что на смартфонах абоненты имеют больше возможностей пользоваться мобильным интернетом. Соответственно, вторая задача долж-

на решать проблему ускорения перехода абонента с более простого устройства на более современное (смартфон или планшетофон).

Набор анализируемых переменных для двух задач, разумеется, должен быть различен. Если в первом случае нас скорее будут интересовать данные о голосовом потреблении абонента в разрезе направлений и его начислениях и платежах, то во втором случае более интересны показатели потребления data-трафика абонента, тип его устройства, частота выхода в сеть и структура ARPU (средняя доходность клиента в месяц).

Для решения первой задачи (предсказания оттока) 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 клиента, склонного к смене типа аппарата связи. Результатом деятельности группы стали дипломные работы кафедры системного программирования СПбГУ, представленные на ресурсе: http://se.math.spbu.ru/SE/diploma/2015/.

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 со своими данными сингапурские и тайваньские операторы связи [3–5]. Кроме того, во время поиска примеров машинного обучения попалась достаточно интересная статья [6] по базе данных одного из китайских банков. Все исследования касались движения клиентской базы этих компаний и 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, randomforest и k-means кластеризацию и рассматривались в этих исследованиях в комбинациях друг с другом. 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», «Randomforest», «Логистическую регрессию» и «Нейронные сети».

3. ОПИСАНИЕ ХОДА РЕШЕНИЯ ЗАДАЧИ

Мы спланировали работу по задаче в два этапа. Для обоих этапов были подготовлены выборки: на этапе обучения мы сообщали машине параметры каждого клиента и результат этого клиента (то есть остается клиент в базе или попадает в отток), а на этапе тестирования на второй выборке мы провели предсказания и оценили результаты по выбранным метрикам.

В качестве метрик в исследовании использовались параметры точности предсказания (precision) и полноты предсказания (recall). Проще говоря, сколько из реально ушедших в отток клиентов модель смогла предсказать и какой процент точности был в выборке клиентов, которую модель посчитала склонной к оттоку. Еще одним показателем качества, который мы использовали для сравнения классификаторов, была выбрана метрика AUC.

На первом этапе обучения для каждого метода в качестве исходных данных была выгружена информация о голосовом трафике, SMS-трафике и трафике передачи данных в разрезе десяти направлений вызовов. Данные были представлены по выборке почти 400 тысяч абонентов. Мы взяли потребление этих абонентов в разрезе 15 месяцев, начиная с середины 2013 года. Кроме того, дополнительным параметром служил номер устройства TAC (TypeAllocationCode), по которому возможно было установить тип устройства клиента (смартфон, планшет, простой телефон), производителя аппарата, операционную систему, год выпуска модели, возможность поддержки Wi-Fi, технические характеристики аппарата (размер дисплея) и характеристики поддерживаемых поколений сети (2G/UMTS/LTE).

Тестовая выборка не показала значимых достижений на первом этапе обработки. Фактически, не было понимания, какие параметры де факто влияют на отток. До начала исследования у нас были мысли относительно того, что, чем больше параметров в выборках мы сможем подготовить и загрузить, тем выше будет 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. В связи с этим мы решили провести оценку важности характеристик и посмотреть, 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 наименее влиятельных характеристик оказались показатели типа аппарата клиента и исходящие звонки по направлению межгорода и городских номеров. Это, в общем-то, поддается логике, поскольку в текущих реалиях городские телефоны действительно теряют свою популярность, а звонки на межгород являются больше разовой необходимостью, чем каким-то постоянным действием, на основании которого можно делать выводы об оттоке.

Учитывая вышесказанное, мы исключили малозначимые параметры и дополнили выборку данными по месту жительства клиента, его полу, возрасту и признаку юридического статуса (то есть проверяли, является клиент физическим лицом или договор заключался на юридическое лицо).

И уже с этими параметрами метрики тестовой выборки стали показывать результаты на порядок лучше. Кроме того, перед одним из очередных шагов по тестированию выборки группа предложила провести группировку параметров. Мы объединили несколько параметров вместе и провели тестирование на группе 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, а AUC составил значение 0,88.

4. АНАЛИЗ РЕЗУЛЬТАТОВ

Оказалось, что для пользователей телефонов со сроком жизни в сети менее 2 лет отток предсказать достаточно сложно. С коммерческой точки зрения это можно объяснить тем, что с оборудованием в последнее время проводится достаточно много акций, когда при покупке устройства предоставляется скидка на оборудование и/или бесплатный трафик. Поэтому зачастую вместе с качественными подключениями 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.

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, интерфейс еще предстоит дорабатывать.

Для второй задачи по исследованию клиентов, склонных к смене аппарата, набора данных, выгруженных для первой задачи, оказалось недостаточно. Мы установили в качестве триггера момент смены TAC. Но в выборке из 400 тысяч клиентов оказалось лишь порядка 25 тысяч человек, которые пользовались телефонами, сменив их в дальнейшем на смартфоны. Здесь мы столкнулись со сложностью разделения выборок на обучение и тестирование и невозможностью определить значимость параметров ввиду малого количества записей. Скорее всего, это связано с тем, что данные по клиентам брались с начала 2013-го года, когда проникновение смартфонов отличалось от проникновения на текущий момент на порядок. Поэтому мы планируем подготовить выборку по 2014–15 годам, когда рост количества смартфонов в розничных сетях и, соответственно, у клиентов, стал происходить большими темпами.

5. ЗАКЛЮЧЕНИЕ

Хотя не всё получилось идеально, уже первые результаты удовлетворили заказчиков исследования и дали нам опыт практического применения нового интересного направления информатики — машинного обучения. Как обычно бывает в исследовательских проектах, в процессе работы появилось много новых интересных задач и предложений по их решению. Без всякого сомнения, исследования в этой области будут продолжены.

Список 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 (дата обращения 27.08.2015).
2. Slon Magazine — онлайн-журнал об экономике и политике. URL: http://slon.ru/specials/data-economics/articles/target/ (дата обращения 06.08.2015)
3. Customer churn prediction in telecommunications. URL: http://www.sciencedirect.com/science/article/pii/S0957417411011353 (дата обращения: 22.05.2015).
4. *V Umayaparvathi and K Iyakutti.* Applications of data mining techniques in telecom churn prediction.// International Journal of Computer Applications, 2012. № 42(20). С. 5–9.
5. *Chih-Ping Wei and I-Tang Chiu.* Turning telecommunications call details to churn prediction: a data mining approach // Expert systems with applications, 2002. № 23(2). P. 103–112,
6. *Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying.* Customer churn prediction using improved balanced random forests // Expert Systems with Applications, 2009. № 36(3). P. 5445–5449.

BIG DATA USAGE IN TELECOMMUNICATIONS

Ponomarev A. A.

Abstract

The article observes general questions of big data usage in practice in commercial activities of companies. The author makes an accent on the problems that can be solved in the telecommunications with the help of big data. He analyzes clients' characteristics that help to judge about their potential churn rate. The author identifies clusters, for which the best prediction results were received.

Keywords: *big data, telecommnications, churn, support vector machines, neural networks.*

Пономарёв Артемий Александрович,
аспирант кафедры системного программирования СПбГУ,
artem.ponomarev@gmail.com