ИСПОЛЬЗОВАНИЕ ПРОГНОЗНОЙ АНАЛИТИКИ И АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ ДАННЫХ ДЛЯ ПРЕДОТВРАЩЕНИЯ ЭСКАЛАЦИИ ПРОБЛЕМ И ОТТОКА КЛИЕНТОВ

Д.Л. Северинец
Ведущий бизнес-аналитик компании IBA (международный деловой альянс), Республика Беларусь

А.В .Гавриленко
Ведущий инженер-программист компании IBA (международный деловой альянс), Республика Беларусь

Иностранное предприятие «АйБиЭй АйТи Парк», DzmitrySeviarynets@iba.by, AGavrilenko@iba.by

This report is aimed at describing a practical approach to building a predictive model that helps establishing long-term relations with your customers by increasing their loyalty and reducing churn. The model allows responding to early incidents by proactively leveraging additional resources to solve potential problems before they actually lead to critical situations. The model is based on mathematically grounded forecasts, which are built by analyzing structured and unstructured incident data using IBM analytical software platforms.

О компании-заказчике

Компания заказчик – один из лидеров мировой IT-индустрии, производитель и поставщик аппаратного и программного обеспечения, а также ИТ-сервисов и консалтинговых услуг.

Бизнес проблема

Одной из наиболее актуальных проблем для компании-заказчика является налаживание с клиентами долгосрочных партнерских отношений и повышение их лояльности. Имея широкую линейку продуктов, важно оперативно реагировать на все проблемы, которые могут возникать у клиентов в ходе эксплуатации продуктов (как аппаратного, так и программного обеспечения).

Вначале, когда клиент сталкивается с проблемой при эксплуатации продуктов компании-заказчика, он высылает первоначальную информацию об этой проблеме. Таким образом, оформляется своего рода идентификационный номер проблемы, к которому относится первичная информация о проблеме: какое ПО (либо аппаратное обеспечение) дало сбой, какие проблемы с эксплуатацией есть на данный момент; насколько данная проблема критична для жизнедеятельности бизнеса клиента и т.д. Такая первичная информация содержит как структурированные поля (собранные из анкеты, заполняемой при формировании проблемы), так различные источники текстового описания проблемы. По-

скольку линейка продуктов компании-заказчика достаточно широкая, таких проблем в открытом статусе находится ежедневно несколько десятков тысяч.

Затем, (как правило, в течение одной-двух недель) служба поддержки обрабатывает и решает большинство проблем, и клиенты могут продолжать работать дальше.

Но менее чем 1% таких проблем в течение первых двух недель перерастает в жалобу со стороны клиента. В этом случае, на первый взгляд из «обычной» неполадки (которая была зарегистрирована первоначально), проблема клиента перерастает в ситуацию, когда из-за проблем с оборудованием либо ПО его деятельность может быть частично парализована. Бизнес- клиент в этом случае не может продолжать работать в обычном режиме и несет прямые убытки.

Очевидно, что после появления таких жалоб (критических инцидентов) лояльность клиентов к компании-заказчику будет сильно подорвана и дальнейшее сотрудничество может оказаться под угрозой.

Одной из основных целей проекта стало построение прогнозной (статистической) модели, которая из большого числа открытых на данный момент проблем заблаговременного выявляла бы ситуации, которые могут привести к возникновению жалобы. В этом случае будет возможность превентивно реагировать на них (переопределить дополнительные ресурсы для оперативного решения проблемы) и избежать формирования жалобы (критической ситуации для клиента).

Рассмотрим построение и применение прогнозной модели на одном из продуктов компании-заказчика (далее – Продукт).

По рассматриваемому Продукту в день приходит порядка 500 новых проблем от клиентов со всего мира. На анализ каждой проблемы уходит порядка 20 минут. Таким образом, чтобы ежедневно анализировать все 500 новых проблем по рассматриваемому Продукту, понадобится полная загрузка 20 человек.

Более того, жалобы постоянно дополняются новой информацией, т.е. проанализировав сегодня проблему, сервисный аналитик может сделать вывод, что она не приведет к жалобе. Но не исключено, что на следующий день клиент дополнит проблему новой информацией, которая уже может отнести эту проблему к категории потенциальных критических ситуаций.

Таким образом, обработка проблем не ограничивается лишь теми, которые открылись в определенный день. Для качественной оценки состояния проблем необходимо проводить скоринг всех открытых на данный момент проблем. В среднем их ежедневный объем по рассматриваемому Продукту составляет порядка 5000.

При этом по статистике «критическими» ежедневно становятся 2-3 случая из 5000. То есть основная цель – из 5000 проблем выбрать в первую очередь эти 2-3 потенциально критические ситуации и выделить на их решение основные ресурсы.

Метод решения

Практическую задачу построения прогнозной статистической модели можно разбить на следующие этапы:

1. Формирование исходной выборки
2. Подготовка данных
   а. проверка качества данных
   б. исключение экстремальных (аномальных) значений
   в. трансформация и стандартизация переменных
   г. применение фиктивных переменных (dummy variables)
   д. проверка мультиколлинеарности
3. Анализ переменных, отбор наиболее сильных предикторов
4. Разбивка выборки на 2 части: тренировочную и тестовую
5. Применение модели
6. Оценка качества построенной модели

Для реализации задачи построения модели был использован комплекс программных средств компании IBM: IBM Watson content analytics (для анализа и обработки неструктурированных данных) и IBM SPSS Modeler (для подготовки выборки и построения скоринговой модели).

Формирование исходной выборки

Любые статистически обоснованные прогнозы строятся на обработке исторических данных (выборке). Для построения прогнозной (статистической) модели прогнозирования жалоб от клиентов была собрана выборка с историческими данными (с проблемами за последние три года). В выборку вошли как «легкие» проблемы (которые не привели к появлению жалобы), так и проблемы, которые переросли в жалобу со стороны клиента.

Идея формирования выборки заключалась в следующем. Было выделено аппаратное обеспечение, относящееся к рассматриваемому Продукту, которое было связано с жалобами в прошлом. Затем были собраны «характеристики» данного аппаратного обеспечения и содержание проблемы за двое суток до даты появления жалобы. Двое суток были зарезервированы для того, чтобы на момент скоринга и прогноза появления жалобы у сервисных инженеров было время на принятие соответсвующих мер по предотвращению появления жалобы.

Таким образом была сформирована первая часть выборки, в которой проблемы были связаны с жалобами. Вторая часть выборки содержала «легкие» проблемы. Для ее построения выделялась каждая машина из первой части, которая связывалась с «легкими» проблемами, открытыми примерно в то же время, что и проблемы, приведшие к жалобе. Поскольку у этих «легких» проблем априори нет связи с датой появления жалобы, то характеристика «легких» проблем генерировалась за 2 дня до момента формирования жалобы из проблемы, открытой примерно в то же время и приведшей к жалобе.

В итоговой выборке присутствовало два уровня данных (входных переменных):

- характеристика текущей поломки (ее срочность, сколько дней назад была открыта, сколько времени было потрачено на ее решение и многое другое). Сюда же относятся и текстовые (неструктурированные) данные о проблеме.

- характеристика машины, на которой произошла поломка (как часто на машине происходили проблемы за последние полгода, какой срочности они были, сколько из них привели к появлению жалоб и пр.)

Поскольку данные в выборке являются историческими, по каждой проблеме был известен финальный исход – т.е. какая из проблем привела к жалобе, а какая нет. Последний столбец выборки и содержал данную характеристику – так называемую зависимую переменную, которую нужно предугадать по значениям остальных полей (входных переменных). Зависимая переменная в данном случае принимает 2 значения: единица (1) – соответствует проблеме, которая трансформировалась в жалобу; а ноль (0) – соответствует «легкой» проблеме.

Анализ неструктурированных данных.

Для построения прогнозной (статистической) модели мы провели анализ неструктурированных (текстовых) данных содержащих информацию о проблеме клиента с оборудованием заказчика. Данные были собраны в коллекцию из различных источников и содержали такую информацию как: переписку с сервисом техподдержки, описание проблемы со стороны клиента, различную техническую информацию о состоянии оборудования на момент возникновения проблемы и текстовые документы, полученные в результате распознавания речи.

Для того чтобы понять как мы можем эффективно использовать всю эту информацию, мы обработали ее с помощью продукта IBM Watson content analytics. Данный продукт позволяет работать с большими текстовыми данными (Big Data). В результате мы получили аналитическую коллекцию, в которой все исходные текстовые данные были проанализированы и разделены на пара-

графы, предложения и слова. Для каждого слова была определена часть речи и его начальная форма. Это было сделано с помощью алгоритмов обработки естественного языка (NLP – Natural Language Processing), которые реализованы в продукте IBM Watson content analytics на базе UIMA (Unstructured Information Management Architecture). Кроме этого была получена базовая статистика. К примеру, мы можем узнать частоту использования слов в коллекции (поддержка) и насколько это слово является достоверным для какого-либо подмножества документов (достоверность).

В результате проведенного анализа мы пришли к выводу, что наилучшими входными параметрами для построения прогнозной (статистической) модели будет список, состоящий из всех слов и фраз с высоким уровнем поддержки и достоверности. Перед тем как приступить к поиску данной информации и для улучшения результатов мы провели подготовку данных. Во-первых, сформировали словарь, включающий в себя все уникальные слова в коллекции документов. Для каждого слова мы сохранили: идентификатор документа, порядковый номер предложения в документе, уровень поддержки и уровень достоверности по отношению к документам, связанным с жалобами клиентов Во-вторых, удалили из словаря все шумовые слова (предлоги, суффиксы, причастия, междометия, цифры и частицы), которые встречаются в большинстве документов и обычно не несут смысловой нагрузки. В-третьих, удалили все слова не удовлетворяющие минимальным уровням поддержки и достоверности. Данные уровни были получены экспериментально.

Полученный список мы использовали для поиска фраз с помощью алгоритма ToPMine [1] (Topical Phrase Mining). Так как в поставленной перед нами задачи нет необходимости разбивать документы по темам, то была использована только первая часть алгоритма TopMine, которая описывает алгоритм поиска фраз с высоким уровнем и поддержки. Данный алгоритм основан на двух основных свойствах другого алгоритма – Apriori [2]:

Если фраза P непопулярна (обладает низким уровнем поддержки), то любая другая фраза, включающая в себя фразу P, гарантированно будет непопулярной.

Антимонотонность данных: если в документе нет популярных фраз длиной N, то документ не содержит популярных фраз длиной больше чем N.

Описанные выше свойства алгоритма Apriori применительно к тексту позволяют нам опустить все слова, которые не удовлетворяют уровню поддержки. Это значительно уменьшит число переборов для поиска фраз.

На первом этапе алгоритм формирует список фраз кандидатов длиной k, состоящий из двух последовательных (расположенных рядом) фраз длиной k-1

в границах одного предложения. Данное действие делается для всех предложений во всей коллекции документов. На втором этапе выбирается «фраза-чемпион» для каждого предложения во всех документах. «Чемпионом» становится фраза, имеющая максимальный уровень поддержки в предложении. Повторяем этапы до тех пор, пока не останется «фраз-чемпионов» удовлетворяющим минимальному уровню поддержки или размер фразы станет равен размеру предложения.

Результатом работы данного алгоритма будет список «фраз-чемпионов», полученных на каждом цикле, плюс список слов, обладающих высоким уровнем поддержки и достоверности. Достоверность «фраз-чемпионов» не проверяется на основании правила, применяемого в методах по поиску ассоциативных правил, которое гласит, что при объединении двух фраз с высоким уровнем достоверности, мы получим фразу, которая также обладает высоким уровнем достоверности.

Для автоматизации процесса анализа новых документов используется продукт IBM Watson Content analytics. Для этого был создан пользовательский аннотатор на основе словаря, состоящего из полученного списка фраз. Результатом работы данного продукта является формирование списка для использования его в качестве входного параметра при построении прогнозной модели.

Подготовка данных

На этапе подготовки данных из выборочной совокупности мы исключили нестандартные (аномальные) случаи. Например, из выборки были исключены случаи когда жалоба регистрировалась в течение 1-2 дней после появления проблемы. Данные ситуации не удается решить при помощи описываемого подхода в силу отсутствия необходимого количества времени у сервисных инженеров на реакцию и действия по предотвращению наступления жалобы.

Большинство дата-майнинг методов и моделей, таких как анализ главных компонент и логистическая регрессия, применяются эффективнее, когда переменные распределены нормально (или хотя бы симметрично). В связи с этим, для ряда переменных мы применили наиболее подходящие трансформации числовых полей – натуральный логарифм, извлечение квадратного корня и другие.

Всесте с тем, после применения трансформации полей, в них все еще остаются существенные различия в разбросе значений входных переменных. Для уменьшения этого разброса к числовым полям в выборке мы применили z-стандартизацию. Таким образом, все числовые переменные будут характеризоваться нулевым средним значением и стандартным отклонением равным 1.

Помимо числовых переменных в выборке также присутствовали и качественные показатели. Они были преобразованы в так называемые фиктивные переменные (dummy variables). Фиктивные переменные предполагают, что разница между одной группой качественной переменной и другой одинакова, в этом случае используется кодировка фиктивных переменных значениями 0 или 1. Вместе с тем, использование значения WoE (Weight of Evidence – вес свидетельства) позволяет решить эту проблему так, что отражает точное направление и масштаб зависимости между различными категориями сгруппированных характеристик [3]. WoE – измеряет статистическую значимость каждой категории переменной и рассчитывается как:

$$WoE_{cat} = \ell n(odds) = \ell n(p_{good} / p_{bad}),$$

где p_{good} – отношение количества неперешедших в жалобы проблем в категории к числу всех неперешедших в жалобы проблем, p_{bad} – отношение количества проблем, перешедших в жалобы в категории к числу всех проблем, перешедших в жалобы.

Отдельной интересной задачей была подготовка для моделирования данных, полученных на основании анализа неструктурированных источников. Результат анализа, сделанного в IBM Watson Content analytics был трансформирован в таблицу, в которой каждая строка включала собой идентификационный номер проблемы и флаговые (бинарные) поля, означающие наличие или отсутствие ключевых слов (фраз), содержащихся в данной проблеме. Таблицу можно представить следующим образом:

Таблица 1

ID_проблемы	Фраза_1	Фраза_2	...	Фраза_N
ID_1	1	0	...	0
…	...	...	...	...
ID_N	1	1	...	0

В IBM Watson content analytics было выделено порядка 300 фраз для рассматриваемого продукта, т.е. данная таблица имела порядка 300 полей. При этом таблица была очень разреженной – заполненность большинства полей единицами не превышала 5%.

Для использования этой информации в последующем процессе моделирования такая разреженность данных должна была быть исправлена. Для этих целей был применен кластерный анализ (алгоритм TwoStep [3]). Одно из его преимуществ (к примеру, в отличие от алгоритма K-Means) состоит в том, что оптимальное количество кластеров алгоритм определяет самостоятельно. В ре-

зультате применения алгоритма было выявлено 2 кластера. На Рисунке 1 ниже можно видеть, что один из кластеров содержит в себе порядка 40% проблем, связанных с жалобами (красная заливка):

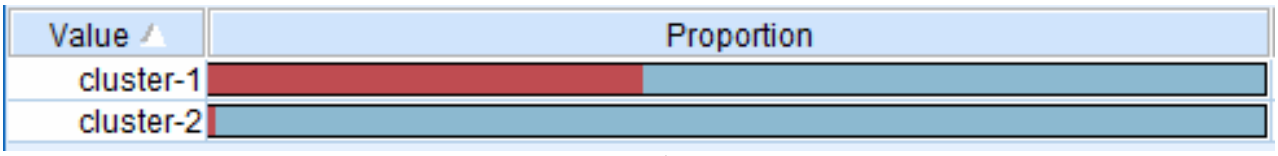

Рис. 1.

Более детальный анализ двух кластеров показал, что концентрация ключевых слов с оттенком срочности, ущерба и общего негатива клиента значительно больше в кластере-1, что логично, когда проблема является сложной и в итоге приводит к появлению жалобы. Как итог, было решено применить результаты кластеризации в последующем моделировании.

Корреляционный анализ также является важным этапом оценки переменных. Все использующиеся в модели переменные необходимо проверить на наличие между ними корреляции. В случае присутствия проблемы мультиколлинеарности необходимо найти оптимальное сочетание между удалением статистически незначимых характеристик и группировкой (к примеру, используя анализ главных компонент) или выбором одной общей переменной из каждого коррелирующего кластера. Решение проблемы мультиколлениарности особенно важно в случае если помимо оценки вероятности наступления жалобы необходима также интерпретация отдельных коэффициентов (профелирование характеристик проблем, приводящих к возникновению жалобы). В нашем случае – профелирование было одним из требований при разработке модели.

Всего были выделены три коррелирующие между собой переменные. Они были удалены из моделирования в силу малой прогнозной силы, оцененной на базе Информационного Значения (Information Value).

Анализ переменных

Следующий этап построения модели – это проверка статистической значимости независимых переменных. Данный анализ позволяет определить, какие переменные являются наиболее точными предсказателями модели.

Информационное Значение (IV) – считается самой распространенной мерой определения статистической значимости переменных. На основе данного показателя можно сделать вывод о значимости той или иной переменной в модели: чем больше значение, тем больше значимость переменной (тем лучше для данной переменной разделяются «легкие» и приведшие к жалобе проблемы).

Информационное Значение определяется по формуле:

$$IV = \sum_{cat} ((p_{good} / p_{bad}) * WoE_{cat})$$

На практике значения данного коэффициента часто трактуются следующим образом:

- менее 0,02 – статистически незначимая переменная;
- 0,02 – 0,1 - статистически мало значимая переменная;
- 0,1 – 0,3 - статистически значимая переменная;
- 0,3 и более – статистически сильная переменная.

Разбивка выборки на две части – тренировочную и тестовую

Подготовленная для моделирования выборка содержала порядка 100000 проблем. Из них около 1000 привели к возникновению жалоб.

Для повышения устойчивости модели, мы разделили выборку на две части: обучающую и тестовую.

Как известно, соотношение числа «плохих» и «хороших» клиентов в выборке может влиять на качество построенной модели. В связи с чем мы сгенерировали несколько вариантов обучающей выборки. Варианты отличались пропорцией «легких» проблем и проблем, приведшим к жалобе. К примеру, первый вариант содержал исходную пропорцию между классами целевой переменной. В других вариантах пропорция жалоб в тренировочной выборке возрастала и последний вариант содержал равные доли (50%/50%) «легких» проблем и проблем, приведших к жалобе.

Тестовая выборка содержала исходную пропорцию данных, без изменений – т.е. из 3000 проблем, 2700 были «легкими», а оставшиеся 300 привели к жалобе.

Построение модели и оценка качества

Поскольку профелирование (создание профелей характеристик проблем, приводящих к возникновению жалобы) было одним из требований при разработке модели, необходимо было использовать аглоритмы, которые предоставляют такую возможность. В результате на этапе построения модели был использован ряд алгоритмов машинного обучения (деревья решений): CHAID, C5.0, Quest, C&R Tree (Classification and Regression Tree). Также была применена логистическая регрессия.

После того, как модели были построены, они были оценены по различным параметрам. Выбор был продиктован путем сравнения статистических показателей качества модели, основными из которых являются статистика Колмогорова-Смирнова, коэффициент Джини (Gini coefficient) и область под ROC кри-

вой (Receiver Operating Curve). Чем выше каждая из этих статистик, тем качественнее считается построенная модель.

В результате анализа моделей лучшей по прогнозным характеристикам оказалась логистическая регрессия при пропорции в тренировочной выборке 50%/50% «легких» проблем и проблем, приведших к жалобе. Для создания профелей характеристик проблем, приводящих к возникновению жалобы, за основу были взяты результаты работы алгоритмов деревеьев решений.

Результаты

Разработанная модель была введена в тестовую эксплуатацию в начале 2015 года.

По итогам 4 месяцев тестирования построенная модель заблаговременно обнаруживает порядка 56% всех проблем, которые привели к возникновению жалобы. Важным моментом производительности модели также является объем ложно положительных прогнозов, поскольку он влияет на загрузку сервисных инженеров. По итогом тестов, для предотвращения 56% всех жалоб необходимо просмотреть лишь 8 проблем в день из 5000 ежедневно открытых проблем (это менее 3 часов работы одного сервисного инженера в день). В результате внедрения модели, в среднем, модель распознает потенциальную жалобу с запасом для реагирования в 7 дней. Таким образом, у службы поддержки есть достаточно времени, чтобы оперативно отреагировать на «указанную» моделью проблему и предотвратить возникновение жалобы.

Такие результаты позволяют существенно сократить риск возникновения жалоб и значительно повысить уровень лояльности клиентов.

Литература

1. Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora, In Proc. VLDB, volume 8, pages 305 – 316, 2015

2. R. Agrawal, R. Sricant, et al. Fast algorithms for mining association rules. In Proc. VLDB, volume 1215, pages 487 – 499, 1994.

3. Ковалев, М. Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц / М. Ковалев, В. Корженевская// Вестник Ассоциации белорусских банков. - 2007. - №46. - С. 16-20.

4. Zhang, T.; Ramakrishnan, R.; Livny, M. (1996). "BIRCH: an efficient data clustering method for very large databases". Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD '96. pp. 103—114.