УДК 004.85

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ СИСТЕМЫ УПРАВЛЕНИЯ ПРОЕКТАМИ

В. С. Попукайло

Приднестровский Государственный Университет им. Т.Г. Шевченко

Аннотация. *На основе данных системы управления проектами ПланФикс прогнозируется отток клиентов после окончания использования пробной версии сервиса. Произведён отбор признаков, значимо влияющих на целевую переменную. Для решения поставленной задачи используются алгоритмы классификации: логистическая регрессия, деревья решений, случайный лес; производится сравнение полученных математических моделей; предобработка и построение моделей производится на языке R.*

Ключевые слова: *моделирование, анализ данных, классификация, логистическая регрессия, деревья решений, случайный лес.*

Введение

В данной статье решается задача прогнозирования оттока клиентов системы управления проектами после окончания использования ими пробной версии сервиса.

Данная задача является актуальной и коммерчески важной для любой компании, представляющей услуги на основе механизма подписки, так как может позволить не только классифицировать клиентов, но и, с некоторой вероятностью, предсказать последующее поведение потребителей. Полученная информация может быть использована для внесения изменений в маркетинговую стратегию, с целью увеличения процента клиентов, перешедших на платные пакеты сервиса.

Прогнозирование оттока клиентов – распространенная задача машинного обучения, решаемая многими компаниями на основе анализа собственных данных, что не позволяет использовать для этих целей готовые программные продукты.

Для решения поставленной задачи применяется язык R с библиотеками, реализующими различные этапы предобработки данных, построения моделей классификации и визуализации полученных результатов.

1. Постановка задачи исследования

Задача исследования состоит в построении предиктивной модели, которая позволит с определённой долей уверенности предсказать отток клиентов на основе их поведения во время бесплатного пробного периода использования сервиса, не накладывающего ограничения на функционал системы управления проектами.

Опыт применения анализа данных для прогнозирования оттока клиентов описан в различных исследованиях [1-3]. На практике для решения поставленной задачи применяются различные инструменты, среди которых язык программирования Python [4], статистический пакет Statsoft STATISTICA [5], облачная платформа Microsoft Azure [6], IBM SPSS Statistics и язык программирования для статистической обработки данных R [7].

Для проведения данного исследования был выбран язык программирования R, как обладающий большим количеством дополнительных библиотек, облегчающих, как предварительный анализ данных, так и построение предиктивных моделей.

При прогнозировании оттока клиентов необходимо решить задачу классификации, которая в данном случае является бинарной.

Предикторами в построенной модели должны быть метрики, описывающие поведение клиента на седьмой, четырнадцатый и двадцать первый день после регистрации, а также на момент окончания бесплатного пробного периода использования сервиса.

Решение поставленной задачи можно разбить на несколько этапов:

1. Преобразование таблиц базы данных, содержащих, как количественные, так и номинативные данные и представляющие собой временные ряды в вид удобный для дальнейшей обработки.

2. Отбор признаков, значимо влияющих на вероятность перехода клиентов на платные пакеты сервиса.

3. Построение классификационных моделей для несбалансированных групп.

2. Предобработка данных

Данные для исследования представляют собой резервную копию базы данных MySQL, содержащей информацию о двух таблицах: «account_event» и «metrics_account». В первой таблице хранится 193639 строк с информацией об идентификаторе пользователя, дате совершения события, типе и подтипе события, а также – опционально дополнительная справочная информация. Во второй таблице хранится информация о метриках активности клиента и каждая строка содержит дату своего создания, уникальный идентификатор пользователя, тип и название метрики, а также числовое поле, содержащее значение метрики. Таким образом, таблица состоит из 5 столбцов и 2785162 строк. Пользу для дальнейшего исследования представляют метрики, относящиеся к пробному периоду использования сервиса и имеющие тип, начинающийся со слова «trial». Таким образом, первым шагом исследования было необходимо отобрать из таблицы «metrics_account» строки, относящиеся к первым 30 дням использования сервиса. На данном этапе было отсеяно 284823 строки. Для удобства дальнейшей обработки данных столбцы, содержащие тип и название метрики, были объединены и заменены на один столбец, содержащий уникальный идентификатор метрики.

Следующим шагом было необходимо трансформировать таблицу таким образом, чтобы полученные названия метрик стали самостоятельными столбцами, а их значения заполнили ячейки на пересечении с соответствующими идентификаторами пользователей.

Однако, при проведении данного преобразования выяснилось, что для некоторых пользователей существуют метрики с одинаковыми названиями, но записанные в разное время и хранящие различные значения. Данное обстоятельство обусловлено некорректным поведением функции, сохраняющей информацию в таблицу базу данных. Для исправления сложившейся ситуации данные были сгруппированы по уникальному идентификатору пользователя и названию метрики, после чего из них были отобраны строки, содержащие наиболее раннюю дату.

После завершения преобразования была получена таблица, содержащая 13895 строк и 133 столбца.

Следующим шагом была получена целевая переменная, для чего из таблицы «account_event» были отобраны идентификаторы пользователей для которых существовали события, показывающие переход с бесплатного пакета на платный: «PlanChangedToPayedFromFree». На основании полученных данных был сформирован бинарный вектор, содержащий признак «1», если пользователь перешел на платный пакет и «0», если этого не произошло.

Последним шагом предобработки данных стала замена пропущенных значений в итоговой таблице на нули, что согласуется с отсутствием информации по данной метрике для данного клиента.

Таким образом, был получен массив исходных данных, пригодный для дальнейшей обработки методами машинного обучения.

3. Отбор признаков

На этапе первичного анализа исследуемых признаков было выявлено, что из рассмотрения необходимо отбросить столбец, содержащий идентификатор пользователя, а также восемь признаков, характеризующих переходы пользователей между бесплатными и платными пакетами. После чего оставшиеся 124 метрики были разделены на бинарные и количественные, исходя из содержащихся в них данных.

Анализ бинарных метрик показал, что 76 показателей, фактически, не используются, то есть содержат только нули. Оставшиеся 48 признаков были проверены на наличие статистической связи с целевой переменной при помощи точного теста Фишера, с поправкой на множественные сравнения по методу Бенджамини-Хохберга, который позволил отклонить нулевую гипотезу об отсутствии связи на уровне значимости в 1% для всех признаков, кроме метрики, фиксирующей блокировку аккаунта на 21 день использования: «trial_21blockedTrialExpired».

Проведение статистического теста при помощи более консервативной поправки Бонферонни дало аналогичные результаты.

Анализ связи количественных метрик с целевой переменной при помощи критерия Уилкоксона-Манна-Уитни с поправкой Бенджамини-Хохберга позволил принять гипотезу об отсутствии различий в группах для метрик, характеризующих количество созданных отчетов на 14 и 21 день после регистрации нового аккаунта пользователем: «trial_14countReportTemplate» и «trial_21countReportTemplate». Применение поправки Бонферонни дополнительно позволило принять гипотезу для метрик, отображающих количество созданных отчетов на 7 день и на момент окончания бесплатного периода.

Таким образом, можно сделать вывод, что применение более консервативной поправки Бонферонни для проверки множественных гипотез, в данном случае, даёт более надежные результаты, так как позволяет определить незначи-

мый фактор, без привязки к времени фиксирования его значения системой.

Следующим шагом для отбора признаков, влияющих на целевую переменную, было построено дерево решений. Для этого таблица данных была разделена на обучающую и тестовую часть, на основе целевого признака в соотношении 80 к 20. Анализ полученного дерева показал, что наиболее значимыми являются факторы, которые характеризуют такие показатели, как: количество новых действий за последний 10 дней, количество новых задач за последние 10 дней, общее количество задач, общее количество контактов, общее количество проектов и прочие.

На рис. 1 представлено изображение построенного дерева.

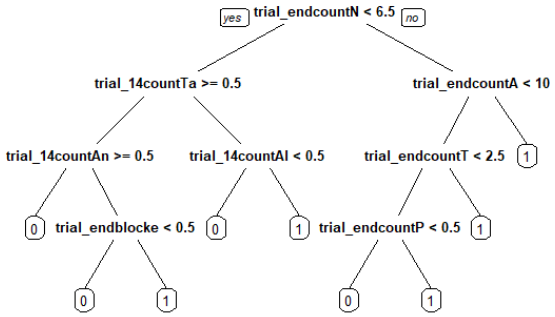

Рис. 1. Дерево решений

Всего дерево, не ограниченное по глубине, позволило отобрать 30 наиболее важных признаков, относящихся к активности пользователей во все временные отрезки и используемых в дальнейшем для построения предиктивных моделей.

4. Построение предиктивных моделей

Первой моделью для предсказания оттока пользователей было решено использовать дерево решений, построенное на предыдущем шаге исследования. Доля правильно классифицированных клиентов на обучающей выборке составила 94,1%, а на тестовой выборке – 90,8%. Однако, учитывая тот факт, что распределение целевой переменной сильно смещено в сторону класса клиентов, не перешедших на платный пакет, следует также рассчитать F-меру, которая на отложенной выборке для класса, отмеченного как «0», будет равна 0, 966, а для класса «1» - 0, 523. Как видно из полученных результатов, в целом, дерево решений хорошо справляется с задачей классификации целевой группы клиентов, но прогнозировать на его основе вероятность перехода клиента на платный пакет – практически не имеет смысла. Следующим шагом для улучшения качество прогнозирования было решено применить процедуры рандомизации и построить ансамбль решающих деревьев. Модель случайного леса, построенная по факторам, отобранным деревом решений и содержащая 5000 деревьев, на тестовой выборке правильно классифицирует 98,8% клиентов, при этом F-мера для каждого и классов равна 0,993 и 0,924, соответственно. Однако, данная точность обусловлена переобучением модели и на отложенной выборке таких результатов достичь не удалось: F-мера для класса «0» с точностью до второго знака после запятой совпадает с аналогичной метрикой для дерева и равна 0,968, а для класса «1» - 0,586. Таким образом, общая доля верно классифицированных клиентов составляет 94,1%. Третьей моделью для прогнозирования оттока было решено использовать логистическую регрессию. На первом шаге модель строилась по параметрам, отобранным деревом решений, после чего применялась итерационная процедура, на каждом шаге которой отбрасывался наименее значимый фактор, до тех пор, пока уменьшался информационный критерий Акаике. Результатом такого подхода стала модель бинарной классификации, состоящая из 20 предикторов с характеристиками качества, не уступающими исходной. Однако, доля правильно классифицированных клиентов при помощи логистической регрессии значительно ниже, чем у моделей, основанных на деревьях решений и составляет 81,2%, при этом F-мера для класса «0» равна 0,888, а для класса «1» - 0,414. Такое качество предсказания может быть обусловлено наличием сильно коррелированных входных переменных, что недопустимо для логистической регрессии [8]. Полученные результаты позволяют сделать вывод о нецелесообразности использовать полученную модель для предсказания оттока клиентов.

Полученные на предыдущих шагах результаты позволяют с приемлемым уровнем точности классифицировать клиентов, не готовых переходить на платные пакеты по окончанию использования пробной версии, что может позволить более достоверно планировать показатели развития проекта на ближайший месяц. Однако, стоит отметить, что в список факторов, наиболее значимо влияющих на целевую переменную, вошли, в основном, показатели, регистрирующие количественные характеристики активности клиента во время использования сервиса. При этом, разработчики системы управления проектами практически не могут повлиять на них, с целью уменьшения оттока клиентов, а значит и увеличения прибыли организации. В связи с этим, было решено дополнительно построить модель по метрикам, описывающим использование различного функционала сервиса на конец бесплатного

пробного периода использования сервиса. Для решения этой задачи из массива данных были отобраны только столбцы, содержащие в своём названии подстроку «_enduse». Так как все одиннадцать полученных характеристик являются бинарными, было решено получить вероятности их влияния на целевую переменную при помощи регрессионной логистической модели. Анализ полученных результатов позволил сделать следующие выводы:

- Коэффициенты перед факторами «Интеграция с социальной сетью Facebook» и «Использование конфигурации для выставления счетов» статистически не значимы, что не позволяет сделать выводы о их влиянии на целевую переменную.

- Все остальные коэффициенты являются положительными и значимыми при вероятности ошибки первого рода в 5%.

Приведём некоторые характеристики, наиболее сильно влияющие на вероятность перехода на платные пакеты сервиса:

- Если клиент использует интеграцию с сервисом для планирования встреч, событий и дел «Google каледарь» или Telegram-ботом, то его шансы оформить подписку на сервис после бесплатного пробного периода возрастают в 5 раз по сравнению с клиентом, не пользующимся данными услугами.

- Более, чем в 2,5 раза возрастают шансы у клиентов, пользующихся одним из следующих сервисов: «Использование протокола передачи почты SMTP», «Интеграция с социальной сетью ВКонтакте», «Использование конфигурации учет рабочего времени».

- Использование сервиса «Автоподпись в почтовых отправлениях» увеличивает шансы перейти на платный пакет в 2,4 раза.

- Остальные исследуемые метрики оказывают влияние на шансы оформить подписку не более, чем в 1,5 раза.

Таким образом, можно сделать вывод о том, что из интеграций, использование которых отражается в данных, полученных для исследования, наибольшее влияние оказывают работа с Google календарём и системой обмена мгновенными сообщениями Telegram.

Выводы

В ходе проведённого исследования удалось построить предиктивные модели, способные с необходимой точностью классифицировать клиентов, которые не перейдут на платные версии продукта после окончания работы с ознакомительной версией. Наилучшее качество в прогнозировании было достигнуто при помощи алгоритма случайного леса. Применение текущей реализации модели затруднительно в силу высокой вычислительной сложности. Следующим этапом исследования должно стать построение наиболее эффективной модели для внедрения в программный продукт. Для этого предлагается минимизировать количество решающих деревьев, настолько, чтобы это не отразилось на качестве классификации. При помощи модели логистической регрессии среди параметров, характеризующих использование функционала пользователем на момент времени окончания использования пробной версии, были отобраны наиболее значимо влияющие на шансы перехода на платные версии сервиса. Полученные результаты могут повлиять на направления маркетинговой стратегии компании с целью уменьшения оттока клиентов системы управления проектами.

Список использованной литературы

1. Canale, A. Churn prediction in telecommunications industry. A study based on bagging classifiers telecom. [Electronic Resource] / A. Canale., N. Lunardon //Moncalieri, Italy: Collegio Carlo Alberto. – 2014. Available at: https: //www.carloalberto.org/assets/working-papers/no.350.pdf

2. Khan, A. A. Applying data mining to customer churn prediction in an Internet service provider. [Electronic Resource] / A. A. Khan, J. Sanjay, M. M. Sepehri // Int. J. Comput. Appl. – 2010. – №9(7). – P.8–14. Available at: http://www.ijcaonline.org/volume9/number7/pxc3871889.pdf

3. Mitkees, I. M., Customer churn prediction model using data mining techniques. [Text] / I. M. Mitkees, S. M. Badr, A. I. B. ElSeddawy //In Computer Engineering Conference (ICENCO), 13th International. IEEE. "Universal decimal Classification. Summary". – 2017. pp. 262–268. doi:10.1109/ICENCO.2017.8289798

4. Карякина, А. А. Сравнение моделей прогнозирования оттока клиентов интернет-провайдеров. [Текст] / Карякина А. А., Мельников А. В.// Машинное обучение и анализ данных. – 2017. - Том 3, № 4. – С. 250–256.

5. Пальмов, С. В. Анализ и прогнозирование оттока клиентов в телекоммуникационных компаниях на основе технологии Data Mining. [Текст] Автореферат диссертации ... кандидата техн. наук : 05.13.13 / С. В. Пальмов. – Поволжская государственная академия телекоммуникаций и информатики. – Самара, 2005г. – 16с.

6. Analyzing Customer Churn by using Azure Machine Learning. [Electronic Resource]. – Access Mode: https://docs.microsoft.com/azure/machine-learning/studio/azure-ml-customer-churn-scenario

7. Груздев, А. В. Прогнозное моделирование в IBM SPSS Statistics и R. Метод деревьев решений. [Текст] /Груздев А. В. – ДМК-Пресс, 2016г. – 278с.

8. Алексеева, В. А. Использование методов интеллектуального анализа в задачах бинарной классификации [Текст] / В. А. Алексеева //Известия Самарского научного центра Российской академии наук. – 2014. – Т. 16. – №. 6–2. – С.354–356.

References

1. Canale, A., Lunardon, N. (2014) Churn prediction in telecommunications industry. A study based on bagging classifiers telecom. Carlo Alberto Notebooks 350:1–11. Available at: https: //www. carloalberto.org/assets/working-papers/no.350.pdf

2. Khan, A. A., J. Sanjay, and M. M. Sepehri. (2010) Applying data mining to customer churn prediction in an Internet service provider. Int. J. Comput. Appl. 9(7):8–14. Available at: http://www. ijcaonline.org/volume9/number7/pxc3871889.pdf

3. Mitkees, I. M., Badr, S. M. and ElSeddawy, A. I. B., (2017) Customer churn prediction model using data mining techniques. In Computer Engineering Conference (ICENCO), 2017 13th International (pp. 262–268). IEEE. "Universal decimal Classification. Summary",
doi: 10.1109/ICENCO.2017.8289798

4. Karyakina, A. A. Mel'nikov, A. V. (2017) Comparison of models predicting the outflow of Internet service providers. [Sravnenie modelei prognozirovaniya ottoka klientov internet-provaiderov] Mashinnoe obuchenie i analiz dannykh (Machine learning and data analysis), Vol. 3, No. 4, pp. 250–256.

5. Pal'mov, S. V. (2005) Analysis and forecasting of the outflow of customers in telecommunication companies based on Data Mining technology. [Analiz i prognozirovanie ottoka klientov v telekommunikatsionnykh kompaniyakh na osnove tekhnologii Data Mining], extended abstract of dissertation… Ph.D. in Engineering Science. Povolzhskaya gosudarstvennaya akademiya telekommunikatsii i informatiki, Samara, 16p.

6. Analyzing Customer Churn by using Azure Machine Learning. avaliable at: https:// docs.microsoft.com/azure/machine-learning/studio/ azure-ml-customer-churn-scenario

7. Gruzdev, A. V. (2016) Predictive modeling in IBM SPSS Statistics and R. Method of decision trees. [Prognoznoe modelirovanie v IBM SPSS Statistics i R. Metod derev'ev reshenii.] DMK-Press, 2016г., 278p.

8. Alekseeva, V. A. (2014) The use of methods of intellectual analysis in problems of binary classification [Ispol'zovanie metodov intellektu-al'nogo analiza v zadachakh binarnoi klassifikatsii] Izvestiya Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk, T. 16, №. 6–2, pp.354–356

ANALYSIS AND CHURN PREDICTION OF THE PROJECT MANAGEMENT SYSTEM

V. S. Popukaylo

Pridnestrovian State University

Abstract. *Based on the data of the project management system planfix.com, we predict customers' churn on the end of the free trial period. Predictive lead scoring (the process of predicting which trial users are going to convert into paying customers) is a common task of machine learning, which is solved by many companies based on the analysis of their own data, which prevents the usage of general purpose software for this task. We describe the algorithm for preprocessing data from the MySQL database to obtain a dataframe, which is a "tidy data", which is convenient for processing by machine learning methods, and does not contain unnecessary information and missing values. We made the selection of quantitative and binary features, that significantly influence the target variable using statistical criteria for testing multiple hypotheses. To solve this problem, we use classification algorithms, such as logistic regression, decision trees, random forest. We show that each of these algorithms copes well with the task. The most significant features were selected via the decision tree; these features were later used as parameters for more complex models. The random forest model was the most accurate on the task of classifying customers by the target attribute. The logistic regression usage made possible to calculate the probability of converting into paying customer based on customers' usage of various additional services of the system. We make a comparison of the obtained models. We show the characteristics of the customer's account that most affect the chances of the customer switching to a paid version after the end of free trial period. We give recommendations on the continuation of research, including the selection of the most effective form of a random forest model to facilitate the usage*

of predictive analysis of customers in the software product. Preprocessing and model building is done using the R programming language.

Keywords: *mathematical modeling, data mining, classification, logistic regression, decision trees, random forest.*

АНАЛІЗ І ПРОГНОЗУВАННЯ ВІДТОКУ КЛІЄНТІВ СИСТЕМИ УПРАВЛІННЯ ПРОЕКТАМИ

В. С. Попукайло

Піднестровскій Державний Університет ім. Т.Г. Шевченко

Анотація. *На основі даних системи управління проектами ПланФікс прогнозується відтік клієнтів після закінчення використання пробної версії сервісу. Прогнозування відтоку клієнтів - поширена задача машинного навчання, яке вирішується багатьма компаніями на основі аналізу власних даних, що не дозволяє використовувати для цих цілей готові програмні продукти. Описується алгоритм попередньої обробки даних, отриманих з бази даних MySQL, для отримання таблиці, що представляє собою «охайні дані», зручною для обробки методами машинного навчання, яка не містить зайвої інформації, а також пропущених значень. Зроблено відбір кількісних і бінарних ознак, значимо впливають на цільову змінну за допомогою статистичних критеріїв з поправками, призначеними для перевірки множинних гіпотез. Для вирішення поставленого завдання використовуються алгоритми класифікації, такі як логістична регресія, дерева рішень, випадковий ліс; показано, що кожен з цих алгоритмів добре справляється з поставленим завданням: за допомогою дерева рішень були відібрані найбільш значимі ознаки, які потім були параметрами для більш складних моделей; модель випадкового лісу найбільш точно дозволяє класифікувати клієнтів за цільовим ознакою, а логістична регресія дозволяє розрахувати ймовірності оформлення передплати для клієнтів, які використовують різні додаткові сервіси; проводиться порівняння отриманих моделей; вказані бінарні характеристики записів клієнтів, найбільш сильно впливають на шанси переходу клієнта на платний пакет після закінчення безкоштовного пробного періоду. Дано рекомендації щодо продовження досліджень, серед яких підбір найбільш ефективної форми моделі випадкового лісу для полегшення впровадження інтелектуального аналізу клієнтів в програмний продукт; попередня обробка і побудова моделей проводиться на мові програмування R.*

Ключові слова: *моделювання, аналіз даних, класифікація, логістична регресія, дерева рішень, випадковий ліс.*

Получено 15.03.2018

Попукайло Владимир Сергеевич, старший преподаватель кафедры Информационных технологий и Автоматизированного Управления Производственными Процессами Инженерно-технического Института Приднестровского Государственного Университета, ассоциированный сотрудник Института Математики и Информатики Академии Наук Молдовы. ул. Восстания, 2-а, 3200-MD, Тирасполь, Молдова, E-mail: vsp.science@gmail.com, тел. +373-533-73762

Vladimir Popukaylo, Senior Lecturer of the Department of Information Technologies and Automated Management of Production Processes of the Engineering and Technical Institute of the Pridnestrovian State University, associate researcher of the Institute of Mathematics and Informatics of the Academy of Science of Moldova, Vosstania str., 2-a, 3200-MD, Tiraspol, Moldova, E-mail: vsp.science@gmail.com, tel. +373-533-73762

ORCID ID: 0000-0001-7742-7959