Сравнение моделей прогнозирования оттока клиентов интернет-провайдеров

А. А. Карякина, А. В. Мельников
suein_i@mail.ru; andmelnikov1956@yandex.ru
Челябинский государственный университет, Россия, г. Челябинск, ул. Братьев Кашириных, 129

На основе данных российского интернет-провайдера прогнозируется отток клиентов. Определены основные подходы к предварительной обработке архивных данных. Для сравнения использованы алгоритмы классификации: деревья решений, случайный лес, наивный байесовский алгоритм, градиентный бустинг, метод k-ближайших соседей. В качестве первой выборки сформирован экспериментальный массив входных данных размера $6 \times$ $\times\, 400\,000$, в который специально подобраны признаки из обращений (id, сервис, признак, причина, результат, уход). В качестве второй выборки сформирован массив входных данных размера $13 \times 400\,000$. Признаками для него были выбраны: id, количество обращений по каждому типу сервиса, по каждому типу результата, общее количество обращений у клиента, уход. Построены модели для прогнозирования с наилучшими параметрами. В таблицах показаны результаты проведенного исследования с разными наборами данных на разных классификаторах.

Ключевые слова: *прогнозирование; отток клиентов; интернет-провайдер; python; обращения клиентов; классификация*

DOI: 10.21469/22233792.3.4.03

1 Введение

Решается задача прогнозирования оттока клиентов. Для сокращения оттока операторы компании организуют обзвон абонентов, находящихся в «группе риска», и мотивируют их не прекращать сотрудничество, что требует больших ресурсов [1].

Анализ признаков поведения клиентов можно совершать на основе сочетаний следующих факторов [2]: социально-демографические (пол, возраст), договорные отношения (дата подключения, тарифный план), типы потребляемых услуг (Интернет, телевидение), тип взаимосвязи (личные обращения, звонки).

В данном исследовании проанализирована взаимосвязь обращений в колл-центр с уходом клиентов.

2 Формулировка задачи

Решается задача прогнозирования вероятности прекращения использования услуг компании клиентом после обращения. Для этого собираются и агрегируются данные по обращениям, выбираются алгоритмы для построения моделей, сравниваются между собой, после чего определяется лучший из алгоритмов.

Планируемые результаты проекта со стороны интернет-провайдера, предоставляющего данные, — модель, определяющая с вероятностью более 50% абонента компании, который закроет договор или прекратит пользоваться услугами после обращений.

3 Анализ существующих систем

В настоящее время многие компании активно заказывают решения по прогнозированию оттока клиентов. Например, Yandex Data Factory решил прооблему ухода клиентов для «Росавтодор», Wargaming и «ВымпелКом» [3], а Айкумен ИБС — «Сбербанку» [4].

Для сегментации и управления оттоком клиентов, формирования финансовой отчетности, анализа отзывов в соцсетях и на форумах «ВТБ24» использует Teradata, SAS Visual Analytics и SAS Marketing Optimizer [5].

Для анализа социальных сетей и поведения пользователей сайта, оценки кредитоспособности, прогнозирования оттока клиентов, персонализации контента и вторичных продаж «Альфа-Банк» использует платформы хранения и обработки Oracle Exadata, Oracle Big data Appliance и фреймворк Hadoop [5].

Компания «Parcus Group» предлагает свои услуги по анализу оттока клиентов телекома. Для предсказания они берут некоторые или все следующие информационные переменные [6]: демографические данные клиентов (почтовый индекс, доход, профессия, адрес, семейное положение), история покупок (покупка места или канала, расходы, количество услуг, дата покупки, дата отмены), данные об использовании сервиса (количество вызовов, вход или выход, SMS-сообщения, минуты, использование данных), платежные данные (общая сумма расходов, SMS, история платежей, задержки платежей, пропущенные платежи), информация о продукте (ассортимент товаров, предоплата или постоплата), маркетинговые данные (существующие виды сегментации, конкуренты, кампании конкурентов и влияние на отторжение, ценообразование конкурентов).

Опыт применения анализа данных в телекоммуникационных компаниях описан в различных исследованиях [6–9] по оттоку клиентов. Ниже рассматриваются те из них, на основе которых выбраны данные и алгоритмы для прогноза.

В [7] выбраны для анализа следующие показатели за 6 мес.: социально-демографические, статистика вызовов (продолжительность звонков), информация о платежах (за что заплатил клиент), жалобы и споры (проблемы удовлетворенности клиентов и предпринятые меры по исправлению положения), кредитная история. Сравнивались два алгоритма: нейронные сети и дерево решений. Точность каждого составила более 90%. Однако в качестве итогового алгоритма было выбрано дерево решений.

В [8] использован набор данных европейской телекоммуникационной компании. Выборка состоит из 30 619 клиентов, для которых собрано 99 переменных из данных за 10 мес. до целевого события (оттока). Список переменных, описанных в данной статье, приведен в табл. 1. Авторы описали не все переменные, также нет объяснения, что имеется в виду под сервисами и каналами продаж. В качестве базовых алгоритмов выбраны логисти-

Таблица 1 Список переменных

Статические переменные	Ежемесячные переменные
Идентификационный номер клиента	Количество исходящих звонков в пиковое время
Стоимость контракта	Продолжительность исходящих вызовов в пиковое время
Способ оплаты	Стоимость исходящих вызовов в пиковое время
Пол	Количество исходящих вызовов в непиковое время
Возраст	Продолжительность исходящих вызовов в непиковое время
Географическая зона активации	Стоимость исходящих вызовов в непиковое время
Канал продаж	Количество входящих вызовов
Наличие сервиса 1 (бинарная)	Продолжительность входящих вызовов
Наличие сервиса 2 (бинарная)	Количество отправленных SMS
	Количество звонков в службу поддержки клиентов

ческая и линейная регрессии, линейный дискриминантный анализ, метод k-ближайших соседей, дерево решений и нейронная сеть с пятью скрытыми слоями с порогом $t = 1/2$.

В [9] показано, что демографические признаки меньше всего влияют на отток клиентов. Для анализа рассмотрены такие алгоритмы, как метод k-ближайших соседей, градиентный бустинг, наивный байес. Лучшие результаты получены на модели случайного леса с точностью 91% и полнотой 87%.

Обзор существующих исследований по оттоку клиентов был сделан в [10]. В большинстве рассмотренных работ в обзоре сравниваются несколько моделей машинного обучения: логистическая регрессия, случайный лес, градиентный бустинг, дерево решений.

Результаты исследований не следует сравнивать между собой, так как итоговые показатели метрик зависят от способа сбора данных и их изначальной структуры, обработки данных и подбора лучших параметров для моделей. Однако можно выделить наиболее перспективные классификаторы, которые демонстрируют лучшие результаты в подобных задачах: деревья решений, случайный лес, метод k-ближайших соседей, наивный байесовский классификатор, градиентный бустинг.

Факторы клиентов, которые используются в [6–9] для анализа: статистика вызовов, статистика оплат, тарифы, жалобы. Однако интернет-провайдеру, предоставляющему данные для исследования, интересно влияние именно обращений на отток клиентов, поэтому в данной работе будет использован только этот фактор.

4 Подходы к агрегации данных

Для агрегации данных в литературе используются такие методы, как суммирование и извлечение среднего значения. Но так как для полей обращений, исследуемых в данной работе, не применим второй метод, будет рассмотрен только первый. Также будет применен свой подход к агрегации. Первый подход к агрегации, используемый в данной работе, заключается в определении наиболее часто встречающегося значения в каждом поле в обращениях и преобразовании его в признак. Под обращением понимается звонок клиента в компанию по какой-либо проблеме, связанной с оказываемыми услугами. В итоге сформирована матрица размером $6 \times 400\,000$. В табл. 2 показаны признаки и часть выборки. Второй подход к решению заключается в анализе типов сервисов и результатов и выделение схожих в категории. В итоге было создано 5 категорий типов сервисов (административные вопросы; эксплуатация сети; жалобы; повреждение оборудования; расторжение) и 5 типов результатов (проблема не решена; есть претензии; проблемы с оплатой; приостановление; расторжение).

В итоге сформирована матрица размером $13 \times 400\,000$, в которую вошли такие признаки как id, количество обращений по каждому типу сервисов и результатов, сколько всего обращений и уход.

Таблица 2 Часть выборки в первом подходе

Id клиента	Тип сервиса	Признак	Причина	Тип результата	Уход
32 646	7864	2473	2423	4527	1
134 577	2558	3475	1254	5653	0
346 457	458	533	456	987	0
78 436	768	2476	561	124	0
...	...	...	...	...	...

5 Подготовка данных

Авторы используют набор данных, содержащий информацию об обращениях клиентов в компанию. Для исследования выбрано около 400 000 клиентов, из них только 20% ушедших, и их данные за полгода из базы российского интернет-провайдера.

Ушедших клиентов в 4 раза меньше, чем активных. Существует несколько методов балансировки классов [11]. В данном исследовании использован метод smote из модуля imblearn.over_sampling.

Перед обучением убрано на время поле «Id» и выборка разделена на показатель, который исследуем («Уход»), и признаки, его определяющие («Тип сервиса», «Признак», «Причина», «Тип результата»).

Набор данных был разделен на две части: обучающая (тренировочная) выборка (75%) и тестовая выборка (25%).

Пропущенные значения можно заполнить несколькими способами [12]. В данной работе выбран вариант заполнения пропущенных значений нулями с помощью DataFrame.fillna(0), так как это неслучайные пропуски и их необходимо учитывать. Перед этапом обучения моделей произведена стандартизация данных с помощью функции scale из модуля sklearn.preprocessing.

Для лучшего результата подобраны лучшие сочетания параметров для каждого алгоритма с помощью GridSearchCV модуля sklearn.model_selection.

В данных было выделено два класса в столбце «Уход»:

0 — абонент продолжит пользоваться услугами компании;

1 — абонент расторгнет договор.

6 Сравнение моделей

Цель прогнозирования — определить клиентов, которые собираются расторгнуть договор, чтобы успеть провести соответствующие мероприятия по их удержанию, поэтому в представленных ниже таблицах приведены только значения критериев по предсказанию оттока.

В первом подходе на собранных данных у моделей получились результаты, представленные в табл. 3; во втором подходе у моделей получились результаты, представленные в табл. 4.

Лучшие результаты показал случайный лес во втором подходе к решению. Этот алгоритм превосходит остальные по точности и полноте в целевом классе. Также стоит отметить, что такие высокие показатели были достигнуты с помощью балансировки классов методом smote.

Код для повторения вышеизложенного эксперимента находится в [13]. Использован язык программирования Python, среды разработки PyCharm и IPython notebook и биб-

Таблица 3 Сравнительная таблица моделей в первом подходе

Алгоритм	Точность	Полнота	F1
Decision Tree	0,96	0,30	0,45
Random Forest	0,12	0,61	0,20
K-nearest neighbors	0,19	0,5	0,28
Naive Bayes	0,13	0,58	0,21
Gradient Boosting	0,13	0,68	0,21

Таблица 4 Сравнительная таблица моделей во втором подходе

Алгоритм	Точность	Полнота	F1
Decision Tree	0,77	0,83	0,83
Random Forest	0,85	0,90	0,87
K-nearest neighbors	0,84	0,78	0,81
Naive Bayes	0,60	0,84	0,70
Gradient Boosting	0,71	0,86	0,78

лиотеки scikit_learn (для построения и настройки классификаторов), pandas (для работы с данными), matplotlib (для визуализации данных), numpy (для работы с массивами).

7 Заключение

В данной статье описан подход к решению проблемы оттока клиентов телекоммуникационных компании. Произведен сравнительный анализ моделей, построенных с помощью собранных данных, на основании критериев точности и полноты. Алгоритм «Случайный лес» показал наилучшие результаты. В будущем планируется учесть сезонность, долги и оплаты абонентов. Также планируется рассмотреть ансамбли моделей.

Литература

[1] Прогнозирование оттока клиентов со scikit_learn. http://datareview.info/article/prognozirovanie-ottoka-klientov-so-scikit-learn/.

[2] *Арустамов А.* Предотвращение оттока клиентов в телекоме. https://vdocuments.site/9-548b9954b479594c5f8b4658.html.

[3] «Яндекс» остановит отток клиентов «Билайна» большими данными. https://roem.ru/19-11-2015/213365/yandex-bees-dactory/.

[4] *Ильин И.* Отток клиентов «Сбербанка» остановит Big Data «Ростелекома», а не «Яндекса»? https://roem.ru/27-10-2016/235566/iqmen-sberdata/.

[5] *Беркана А.* Что такое Big data: собрали все самое важное о больших данных. https://rb.ru/howto/chto-takoe-big-data/.

[6] Telecom Customer Churn Prediction Models. https://parcusgroup.com/Telecom-Customer-Churn-Prediction-Models.

[7] Condamoor Ravi Building Predictive Models for Customer Churn in Telecom. https://www.experfy.com/blog/building-predictive-models-for-customer-churn-in-telecom.

[8] *Canale A., Lunardon N.* Churn prediction in telecommunications industry. A study based on bagging classifiers telecom // Carlo Alberto Notebooks, 2014. Vol. 350. P. 1–11. https://www.carloalberto.org/assets/working-papers/no.350.pdf.

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

[10] *Корыстов М. А.* Применение методов машинного обучения для предсказания поведения абонентов оператора сотовой связи. — СПб.: СПбГУ, 2015. 25 с.

[11] *Гончаров М.* Data Mining: Классификация редких событий. — М.: Microsoft. 23 p. https://rutechdays.blob.core.windows.net/uploads/95127d97-f198-466f-b027-716418372529/acc79412-79de-4fbc-b2d0-2f82c3451b94/24hoursofpassppt_maxgon.pdf.

[12] *Глушко О.* Обработка пропусков в данных — часть 1. https://basegroup.ru/community/articles/missing.

[13] Исходный код. https://github.com/KiraTanaka/Prediction-churn.

Поступила в редакцию 25.06.2017

Comparison of methods for predicting the customer churn in Internet service provider companies

A. A. Karyakina and A. V. Melnikov

suein_i@mail.ru; andmelnikov1956@yandex.ru

Chelyabinsk State University, 129 Bratiev Kashirinykh Str., Chelyabinsk, Russia

The possibility of forecasting the churn of customers based on the data of the Russian Internet service providers (ISP) has been considered. The basic approaches to preprocessing of archived data are defined. For comparison, classification algorithms are used: decision trees, random forest, naive Bayesian algorithm, gradient boosting, and the method of k-nearest neighbors for prediction. As the first sample, an experimental array of input data of size $6 \times 400\,000$ was formed, which contains the fields from the calls (id, type of service, feature, reason, type of result, and leaving). As the second sample, an array of input data of size $13 \times 400\,000$ was formed. For it, there have been selected the following features: id, count of calls for each type of service, for each type of result, total count of calls from the client, and leaving. The models for prediction with the best parameters have been constructed. In the tables, the results of the research with different data sets for various classifiers are shown.

Keywords: *prediction; clients churn; ISP; python; customers calls; classification*

DOI: 10.21469/22233792.3.4.03

References

[1] *Prognozirovanie ottoka klientov so scikit-learn* [Predicting customer churn with scikit-learn]. Available at: http://datareview.info/article/prognozirovanie-ottoka-klientov-so-scikit-learn/ (accessed December 29, 2017).

[2] Arustamov, A. *Predotvrashenie ottoka klientov v telekome* [Prevention churn in telecommunication companies]. Available at: https://vdocuments.site/9-548b9954b479594c5f8b4658.html (accessed December 29, 2017).

[3] *"Yandeks" ostanovit ottok klientov "Bilayna" bol'shimi dannymi* ["Yandex" will stop the outflow of customers "Beeline" big data]. Available at: https://roem.ru/19-11-2015/213365/yandex-bees-dactory/ (accessed December 29, 2017).

[4] Ilin, I. 2016. *Ottok klientov "Sberbanka" ostanovit Big Data "Rostelekoma," a ne "Yandeksa"?* [Customer churn "Sberbank's" going to stop Big Data "Rostelecom" and not "Yandex"?] Available at: https://roem.ru/27-10-2016/235566/iqmen-sberdata/ (accessed December 29, 2017).

[5] Berkana, A. *Chto takoe Big data: Sobrali vse samoe vazhnoe o bol'shikh dannykh* [What is Big data: Collected all the most important information about big data]. Available at: https://rb.ru/howto/chto-takoe-big-data/ (accessed December 29, 2017).

[6] Telecom Customer Churn Prediction Models. Available at: https://parcusgroup.com/Telecom-Customer-Churn-Prediction-Models (accessed December 29, 2017).

[7] Condamoor Ravi Building Predictive Models for Customer Churn in Telecom. Available at: https://www.experfy.com/blog/building-predictive-models-for-customer-churn-in-telecom (accessed June 6, 2017).

[8] Canale, A., and N. Lunardon. 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 (accessed December 29, 2017).

[9] 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 (accessed December 29, 2017).

[10] Korystov, M. A. 2015. *Primenenie metodov mashinnogo obucheniya dlya predskazaniya povedeniya abonentov operatora sotovoy svyazi* [The application of machine learning methods for predicting the behavior of subscribers of the cellular operator]. St. Petersburg: SPbU. 25 p.

[11] Goncharov, M. *Data Mining: Klassifikatsiya redkikh sobytiy* [Data Mining: Classification of rare events]. Moscow: Microsoft. 23 p. Available at: https://rutechdays.blob.core.windows.net/uploads/95127d97-f198-466f-b027-716418372529/acc79412-79de-4fbc-b2d0-2f82c3451b94/24hoursofpassppt_maxgon.pdf (accessed December 29, 2017).

[12] Glushko, O. *Obrabotka propuskov v dannykh — chast' 1* [Handling missing data — part 1]. Available at: https://basegroup.ru/community/articles/missing (accessed December 29, 2017).

[13] Source. Available at: https://github.com/KiraTanaka/Prediction-churn (accessed December 29, 2017).

Received June 25, 2017