

УДК 007.00

УПРАВЛЕНИЕ ЛОЯЛЬНОСТЬЮ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ-РЕСУРСОВ С ПОМОЩЬЮ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Романова Е. Р (ФГБОУ ВО «Государственный университет управления»)

Научный руководитель – к.э.н., доцент Писарева О.М.
(ФГБОУ ВО «Государственный университет управления»)

Представлена оригинальная авторская модель клиентского поведения, позволяющая оценивать уровень лояльности пользователей к компании, ведущей деятельность в on-line режиме. В качестве данных для построения модели используются «цифровые» следы пользователей на Интернет-ресурсах. Предложена и протестирована процедура анализа действий пользователей на основе компьютерных технологий обработки естественного языка.

Введение. Любая коммерческая компания стремится увеличить свою прибыль. При этом удержание и привлечение пользователей - важнейший приоритет развития бизнес-моделей. Для процветания организации важно, чтобы лояльных пользователей становилось как можно больше. В этой связи ключевой задачей управления интернет-компанией является поиск и удержание клиентов, возможно изначально настроенных нейтрально или негативно, для изучения их отношения к компании и привлечения к более активному взаимодействию с её продуктами/услугами. Важный индикатор лояльности - фиксация систематического использования продукта/услуги, т.е. чем чаще пользователь взаимодействует с ними, тем он представляется более лояльным и к компании. Для поиска таких клиентов активно применяются алгоритмы машинного обучения. При стандартных подходах здесь применяют агрегированные данные о пользователях, например, количество входов в приложение в единицу времени; время, потраченное клиентом на изучение приложений до совершения целевого действия и пр. Автор предлагает использовать «цифровые» следы пользователей, т.е. данные о пользовательском поведении в приложении или на сайте компании. Для решения подобных задач в работах [1-3], например, используется подсчет количества действий пользователя, алгоритмы снижения размерности, анализ последовательностей в рекуррентных нейронных сетях. Однако в этом случае представление признаков пространства является чрезвычайно большим, и становится критическим, если исследуемый ресурс представляет собой многостраничный сайт или приложение с большим выбором блоков и «кликабельных» ссылок. Снижение размерности позволяет снизить остроту данной проблемы. Однако при этом теряется часть информации о действиях пользователя, а также невозможно установить интерпретируемую связь между его последовательными действиями.

Основная часть. В настоящее время в подходах к решению задачи оценки лояльности есть ряд серьёзных недостатков: 1) при представлении действий пользователя в пригодном для моделирования виде признаков пространство имеет очень большую размерность; 2) методы снижения размерности ведут к потере части информации о действиях пользователя; 3) при представлении действий пользователя через подсчет количества их действий не учитывается взаимосвязь между действиями. Для преодоления обозначенных проблем автор предлагает использовать при моделировании пользовательского поведения алгоритмы обработки естественного языка, в предположении, что с сессиями пользователя можно работать также, как и с текстами. Это дает возможность представить действия клиента в сжатом векторном пространстве, учитывая как их взаимосвязь, так и их последовательность. В этой связи целью работы являлось построение модели классификации, отвечающей на вопрос о вероятности возврата клиента к использованию продукта/услуги в течение определенного промежутка времени, для чего решались задачи: 1) подготовки данных для моделирования; 2) построения оригинальной модели классификации; 3) идентификации и верификация модели-классификатора; 4) экспериментального моделирования и анализа его результатов.

Предложенный автором подход к моделированию лояльности пользователя будет сравнен с традиционным (базовый), который состоит в подсчете количества действий и

снижением размерности полученного векторного пространства с помощью метода главных компонент [4]. Авторская модель реализована на языке программирования Python 3.8 [5] с использованием библиотек Sklearn [6], Xgboost [7], Gensim [8]. В качестве информационной базы проведенного исследования использовались данные крупного сетевого онлайн-кинотеатра. Хранение и обработка данных производится с помощью инструментов ClickHouse [9]. Метрикой оценки качества модели классификации выбран Auc-Roc критерий [4, 9, 11]. Оценка качества модели классификации осуществлялась перекрестной проверкой (cross-validation) по десяти подвыборкам. Специфика разбиения данных на обучающую и тестовую выборки состояла в том, что все сессии одного пользователя находились только в одной подвыборке, что позволяет избегать «утечки» данных.

Первым шагом для достижения поставленной цели стало представление сессии пользователя в векторном пространстве. Это сделано на основе алгоритмов «мешок слов» (bag of words), Tf-Idf (term frequency - inverse document frequency), Fasttext, группы моделей Word2vec [12 - 14]. Для агрегирования действий пользователя до проведения сессии предложено использовать усреднение и взятие минимума по каждому измерению вектора действия. В качестве моделей классификации использованы логистическая регрессия и бустинг [4, 11, 15]. Автором были построены модели классификации с использованием традиционного и предложенного подхода. Auc-Roc базовой модели классификации составил 0.8884. Лучший результат показало использование для представления действий в векторном пространстве модели Fasttext, способ агрегации до сессии – взятие минимума, итоговая модель классификации Xgboost. Качество этой модели по Auc-Roc-критерию составило 0.9597.

Выводы. Таким образом, предложенный автором подход к конструированию модели-классификатора позволил существенно улучшить качество классификации пользовательского поведения по сравнению с ранее используемыми методами моделирования. По итогам построения модели и ее тестирования на реальном статистическом материале можно сделать вывод о том, что она обладает достаточно высокой разделяющей способностью и предсказывает возвращение пользователя в on-line-приложение с высокой степенью надежности. Данная модель полезна для оптимизации коммуникационной политики компании (например, отправляя меньше рекламных писем активным и лояльным пользователям; обращая внимание на клиентов, не вернувшихся на портал в заданный временной интервал, что возможно сигнализирует о несовершенстве предлагаемого продукта или услуги, если например они стали неинтересны или невыгодны), а также для обоснования эффективной мотивационной политики компании в отношении целевых клиентов (например, для стимулирования продаж путем предложения бонусов и/или скидок).

Использованные информационные источники:

1. Predicting online user behaviour using deep learning algorithms // arxiv URL: <https://arxiv.org/abs/1511.06247> (дата обращения: 07.01.2021).
2. The Odd One: Applying Machine Learning to User Behavior Anomaly Analysis // HITBSecConf URL: <https://conference.hitb.org/hitbsecconf2018ams/sessions/the-odd-one-applying-machine-learning-to-user-behavior-anomaly-analysis/> (дата обращения: 07.01.2021).
3. What is User Behavior Analytics? // Varonis URL: <https://www.varonis.com/blog/what-is-user-behavior-analytics/> (дата обращения: 07.01.2021).
4. Писарева О.М. Математические основы бизнес-аналитики: Учебное пособие. - М.: ГУУ, 2011. – 207 с.
5. Официальный сайт python URL: <https://www.python.org/> (дата обращения: 25.02.2021).
6. Документация библиотеки sklearn URL: <https://scikit-learn.org/stable/> (дата обращения: 25.02.2021).
7. Документация библиотеки xgboost URL: <https://xgboost.readthedocs.io/en/latest/> (дата обращения: 25.02.2021).

8. Документация библиотеки gensim URL: <https://radimrehurek.com/gensim/> (дата обращения: 25.02.2021).
9. Документация базы данных ClickHouse URL: <https://clickhouse.tech/docs/ru/> (дата обращения: 25.02.2021).
10. AUC ROC // Анализ малых данных. Блог Александра Дьяконова URL: <https://dyakonov.org/2017/07/28/auc-roc-%D0%BF%D0%BB%D0%BE%D1%89%D0%B0%D0%B4%D1%8C-%D0%BF%D0%BE%D0%B4-%D0%BA%D1%80%D0%B8%D0%B2%D0%BE%D0%B9-%D0%BE%D1%88%D0%B8%D0%B1%D0%BE%D0%BA/> (дата обращения: 19.02.2021).
11. Писарева О.М., Денисова А.И. Математическое и компьютерное моделирование рисков лицензируемых товарных рынков Российской Федерации // «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками»: материалы VI Междунар. Молодёжной науч.-практ. Конф. – Саратов: ООО Изд-во «Научная книга», 2017. – 264 с. С.70-76.
12. Прикладные задачи анализа данных // Coursera. URL: <https://www.coursera.org/learn/data-analysis-applications> (дата обращения: 08.01.2021).
8. Efficient Estimation of Word Representations in Vector Space // arXiv URL: <https://arxiv.org/abs/1301.3781> (дата обращения: 08.01.2021).
13. Distributed Representations of Words and Phrases and their Compositionality // arXiv URL: <https://arxiv.org/abs/1310.4546> (дата обращения: 08.01.2021).
14. Enriching Word Vectors with Subword Information // arXiv URL: <https://arxiv.org/abs/1607.04606> (дата обращения: 08.01.2021).
15. Писарева О.М., Суязова С.А. Методы и модели машинного обучения: начальный курс: Учебное пособие. – М.: ГУУ, 2018. – 87 с.

Романова Е.Р. (автор)

Писарева О.М. (научный руководитель)