Моделирование и анализ данных
2020. Том 10. № 4. С. 31–40
DOI: https://doi.org/10.17759/mda.2020100403
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2020. Vol 10, no. 4, pp. 31–40
DOI: https://doi.org/10.17759/mda.2020100403
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.85

Использование методов машинного обучения для решения задач прогнозирования суммы и вероятности покупки на основе данных электронной коммерции

Мамиев О.А.*
Московский авиационный институт
(национальный исследовательский университет),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-1137-4019
e-mail: olegios@mail.ru

Финогенов Н.А.**
Московский авиационный институт
(национальный исследовательский университет),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-7680-9496
e-mail: finogenov.nik@gmail.com

Сологуб Г.Б.**
Московский авиационный институт
(национальный исследовательский университет),
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-5657-4826
e-mail: glebsologub@ya.ru

Для цитаты:
Мамиев О.А., Финогенов Н.А., Сологуб Г.Б. Использование методов машинного обучения для решения задач прогнозирования суммы и вероятности покупки на основе данных электронной коммерции // Моделирование и анализ данных. 2020. Том 10. № 4. С. 31–40. DOI: https://doi.org/10.17759/mda.2020100403

***Мамиев Олег Аланович,** Московский авиационный институт (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-1137-4019, e-mail: olegios@mail.ru

****Финогенов Никита Андреевич,** Московский авиационный институт (национальный исследовательский университет), г.Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-7680-9496, e-mail: finogenov.nik@gmail.com

*****Сологуб Глеб Борисович,** кандидат физико-математических наук, доцент кафедры, Московский авиационный институт (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-5657-4826, e-mail: glebsologub@ya.ru

Работа направлена на исследование возможности применения методов машинного обучения для построения моделей прогнозирования вероятности покупки и суммы покупки клиентов интернет магазинов. Предлагаются к рассмотрению ранее не используемые в рамках конкретной задачи подходы к решению. В качестве выборки использованы данные о транзакциях пользователей сайта ponpare.jp в период с 01.07.2011 по 23.06.2012. Приводится описание и сравнительный анализ наиболее распространенных методов решения аналогичных задач. Описываются метрики, использованные для оценки результатов в случае прогнозирования факта и суммы покупки. Полученные результаты дают понять, что в рамках задачи предсказания вероятности покупки градиентный бустинг, а именно его реализация LGBMClassifier, показывает наиболее точную оценку. Для задачи прогнозирования суммы покупки клиента использование градиентного бустинга также дало наилучшие результаты.

Ключевые слова: прогноз вероятности и суммы покупки, классификация, регрессия, анализ данных, обработка данных, машинное обучение

1. ВВЕДЕНИЕ

В наше время проблема «Больших данных» является основной как с точки зрения научных исследований, так и с точки зрения бизнес-аналитики, для понимания и оптимизации бизнес процессов. Данные широко используются для решения различных задач, возникающих в повседневной жизни – это и рекомендация товаров в интернет – магазинах, и выдача результатов в поисковых движках, и обработка данных с камер наблюдения, и создание беспилотных автомобилей и т.д.

Данная работа посвящена решению задач прогнозирования вероятности и суммы покупки клиентов сайта ponpare.jp. Решение данных проблем является крайне актуальной задачей.

Во-первых, цифровые платформы нацелены на оптимизацию видимости и целевого маркетинга своих продуктов. Такая стратегия не является чем-то новым и уже много лет реализуется в физических магазинах. Исследования поведения клиентов и попытки предсказать их намерения предпринимались еще в конце прошлого века [1]. Улучшенное размещение продуктов в супермаркетах приводит к увеличению их видимости, что в свою очередь обуславливает увеличение продаж. Аналогичные концепции могут быть применены и для интернет-магазинов. Умение заранее распознать намерение клиента приобрести товар (и как следствие улучшенный целенаправленный персональный маркетинг) может снизить затраты и повысить эффективность работы компании.

Во-вторых, почти вся часть современной экономики основана на информации и по этому различные виды электронной коммерции (ecommerce) становятся самыми востребованными и уверенно вытесняют физическую коммерцию, а торговля непосредственно между клиентами (в виде торговых интернет площадок) (B2B ecommerce), становиться наиболее распространенной. Так в 2015 оборот B2B ecommerce составил 21.5 миллиона евро и Россия занимала первое место по количеству электронных покупателей, и их количество неуклонно растет [2]. Подобное развитие интернет-

торговли в том числе обусловлено желанием компаний понимать потребности покупателей и их финансовые возможности. Знание о средней суммы покупки пользователя позволяет оптимизировать цену предоставляемой услуги, скидки, акции и т.д.

Существуют различные методы машинного обучения, применяемые для аналогичных задач. В данной статье проводится сравнительный анализ основных и представляется подход к решению задач прогнозирования вероятности и суммы покупки.

2. ОПИСАНИЕ ДАТАСЕТА

В качестве исходного датасета были использованы данные о транзакциях 22873 клиентов сайта ponpare.jp в период с 01.07.2011 по 23.06.2012. Они структурированы в шести таблицах.

1. Данные о пользователях (22873) содержат следующие атрибуты: идентификатор пользователя, возраст, пол, дата регистрации, дата удаления аккаунта (если происходило), область проживания;
2. Данные о купонах (19413) содержат такие атрибуты как: идентификатор купона, название категории и подкатегории, размер скидки, цена без скидки, цена со скидкой, начало действия купона, окончания действия купона, дата начала и конца скидки, возможность использования купона в конкретный день недели и праздничный день, название области, в которой купон можно использовать;
3. Журнал с просмотрами купонов (2833180) содержит: статус покупки (состоялась она или нет), идентификатор покупки, дата просмотра, идентификатор купона, идентификатор пользователя, идентификатор сессии;
4. Журнал с покупками пользователей (168996) включает информацию о: идентификаторах пользователя, купона и покупки, количестве приобретенных товаров, дате произведения покупки, короткое название области;
5. Данные об областях действия купонов (138185) содержат: идентификатор купона, короткое название области и название префектуры;
6. Данные о местоположении областей (47) содержат: название области, широту, долготу.

3. ПРОГНОЗИРОВАНИЕ ФАКТА ПОКУПКИ КЛИЕНТА

Решение задачи прогнозирования факта покупки сведем к решению задачи бинарной классификации.

На вход получаем набор признаков $(x_1, ..., x_n)$, а на выход выдаем вектор y прогноза факта покупки, то есть бинарное значение, соответствующее одному из двух классов. Первый включает в себя потенциальных покупателей, а второй – пользователей, не планирующих приобретать товар.

Проанализируем наиболее распространенные методы решения подобных задач.

Логистическая регрессия (Logistic Regression) – алгоритм классификации, используемый для определения вероятности успеха и неудачи события. Он поддерживает категоризацию данных по дискретным классам путем изучения взаимосвязи

из заданного набора помеченных данных. Изучается линейная зависимость из заданного набора данных, а затем вводится нелинейность в форме сигмоидной функции.

Достоинства данного алгоритма заключаются в легкости реализации, интерпретации и эффективности при достаточно быстрой скорости обучения. Однако, если количество наблюдений меньше, чем количество признаков, логистическую регрессию использовать не следует, иначе это может привести к переобучению. К тому же, она способна строить только линейные границы.

Деревья решений (Decision trees) – алгоритм, решающий проблему машинного обучения путем преобразования данных в представление дерева. Каждый внутренний узел древовидного представления обозначает атрибут, а каждый листовой узел обозначает метку класса. По сравнению со многими другими алгоритмами деревья решений требуют меньше усилий для подготовки данных во время предварительной обработки. Для обучения нет необходимости нормализовывать или масштабировать данные. К сожалению, даже небольшое изменение данных может вызвать существенное изменение структуры дерева решений, что приводит к нестабильности. Сами вычисления могут быть намного более сложными по сравнению с другими алгоритмами и зачастую занимать много времени.

Алгоритм KNN (K-Nearest Neighbors) предполагает, что похожие вещи существуют в непосредственной близости. Другими словами, похожие вещи находятся рядом друг с другом. KNN достаточно прост, эффективен и интуитивен, но несмотря на это все же имеет несколько ограничений. При достаточно большом тренировочном наборе алгоритм может иметь длительное время выполнения. Он очень чувствителен к несущественным или избыточным признакам, однако при тщательном применении отбора признаков (feature selection) или взвешивания признаков (feature weighting) этого можно избежать. Кроме того, при обучении, основанном на дистанциях, не всегда понятно в чем эту дистанцию измерять с целью добиться наилучших результатов.

Градиентный бустинг (Gradient Boosting) – крайне популярный алгоритм, способный решать задачи классификации, регрессии и ранжирования. В качестве модели градиентного бустинга был использован LightGBM. LightGBM – это фреймворк, в котором используется алгоритм обучения основанный на деревьях решения [3]. Это распределенный и очень эффективный метод для решения задач классификации и регресии . К его достоинствам можно отнести следующие:

1. Быстрая скорость обучения и высокая эффективность, LightGBM использует подход на основе гистограммы (histogram-based) [4], преобразуя непрерывные значения признаков в дискретные.
2. Малый объем памяти.
3. Более высокая точность (в сравнение с другими моделями бустинга): LightGBM создает более сложное дерево, чем метод поэтапного разделения, вследствие чего достигается более высокая точность. Однако это может привести к переобучению, и поэтому необходимо грамотно подходить к глубине дерева.
4. Возможность обработки данных большого размера: в сравнении с другой популярной моделью XGBoost [5], LightGBM позволяет построить более точную модель из-за сокращения времени обучения [6].

4. ПРОГНОЗИРОВАНИЕ СУММЫ ПОКУПКИ КЛИЕНТА

Сформулируем задачу прогнозирования суммы покупки клиента как построение модели машинного обучения, которая позволит построить вектор ответов y (сумма, которую клиент тратит на покупку купонов) в зависимости от набора признаков $(x_1, ..., x_n)$ (данным о клиенте и транзакциях):

$$y = f(x_1, ..., x_n) + \varepsilon,$$

где ε – вектор отклонений модельных данных от исходных.

Цель – используя обучающие данные построить функцию $\hat{f}(x_1, ..., x_n)$, которая могла бы служить аппроксимацией для функции $f(x_1, ..., x_n)$. Существует множество способов для успешного решения поставленной задачи: полиномиальная линейная регрессия (Ordinary Linear Regression) [7], частичная полиномиальная линейная регрессия (Partial Least Squares Regression) [8], метод опорных векторов (support vector regression) [9], нейронные сети [10], модели бустинга [11].

В нашем случае решено было использовать следующие подходы:

1. Полиномиальная линейная регрессия.
2. Градиентный бустинг (Gradient Boosting).

Использование этих методов имеет свои преимущества и недостатки. К достоинствам многомерные регрессионные модели (multivariate regression models) можно отнести простоту реализации и быстроту обучения модели, что позволяет проводить большое количество экспериментов. Вследствие чего использование такой простой, хоть и не очень точной модели, предоставляет возможность проверять различные гипотезы о структуре данных, генерировать новые признаки, производить отбор признаков. В данной работе программной реализации была использована модель LinearRegression из пакета sklearn. Для получения итогового результата был использован LightGBM.

5. МЕТРИКИ ОЦЕНКИ КАЧЕСТВА

Для задачи прогнозирования вероятности покупки клиента в основу были выбраны такие метрики, как: precision, recall и f-score.

Точность (precision) в пределах класса – это доля объектов действительно принадлежащих данному классу относительно всех объектов, которые модель определила в этот класс:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

где TP – истинно – положительное решение, TN – истинно – отрицательное решение, FP -ложно – положительное решение, FN – ложно – отрицательное решение.

Полнота (recall) – это доля найденных объектов, принадлежащих к данному классу, относительно всех объектов из этого класса в тестовой выборке:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

К сожалению, в реальной жизни максимальная точность и полнота одновременно бывают достижимы крайне редко. Поэтому, хотелось бы опираться на метрику, превносившую некий баланс, объединяя в себе информацию о точности и полноте.

F1 -мера (f1-score) определяется как взвешенное гармоническое среднее значение точности и отзыва теста:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

Для задачи прогнозирования суммы покупки клиента были выбраны RMSE, R^2_{score}.

RMSE (Среднеквадратичная ошибка) – это мера того, насколько близка оценка к фактическим данным:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2},$$

где $\hat{y}$ – значения полученные моделью, y – реальные значения. Чем меньше RMSE, тем лучше предсказание модели.

R^2_{score} (коэффициент детерминации) – показывает близость предсказанной модели к реальной модели. Для расчета необходимо вычислить TSS (общая сумма квадратов отклонений – число равное сумме квадратов разности элементов выборки и среднего) и SSE (сумма квадратов невязок – число равное сумме квадратов отклонений модельных данных от исходных):

$$\mathrm{TSS} = \sum_{i=1}^{n}(\bar{y} - y_i)^2,$$

$$\mathrm{SSE} = \sum_{i=1}^{n}(\widehat{y_i} - y_i)^2,$$

где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Тогда R^2_{score}:

$$R^2_{score} = 1 - \frac{\mathrm{SSE}}{\mathrm{TSS}}.$$

R^2_{score} принимает значения от 0 до 1 и чем выше значение, тем точнее полученная модель.

6. АНАЛИЗ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА

В качестве моделей для решения задачи прогнозирования факта покупки товара были выбраны логистическая регрессия, деревья решения, метод k-ближайших соседей и градиентный бустинг.

Model	Precision	Recall	F1-score
Logistic Regression	0.602	0.607	0.604
Decision trees	0.566	0.554	0.56
KNN	0.604	0.619	0.611
LGBM	0.652	0.648	0.65

Рис. 1. Результаты прогнозирования факта покупки

```
Index(['SEX_ID', 'AGE', 'PREF_NAME', 'CAPSULE_TEXT', 'GENRE_NAME',
       'PRICE_RATE', 'CATALOG_PRICE', 'DISCOUNT_PRICE', 'DISPPERIOD',
       'VALIDPERIOD', 'USABLE_DATE_MON', 'USABLE_DATE_TUE', 'USABLE_DATE_WED',
       'USABLE_DATE_THU', 'USABLE_DATE_FRI', 'USABLE_DATE_SAT',
       'USABLE_DATE_SUN', 'USABLE_DATE_HOLIDAY', 'USABLE_DATE_BEFORE_HOLIDAY',
       'large_area_name', 'ken_name', 'small_area_name', 'IS_ACTIVE_USER',
       'target', 'I_DATE_year', 'I_DATE_month', 'I_DATE_dayofweek',
       'REG_DATE_year', 'REG_DATE_month', 'REG_DATE_dayofweek',
       'DISPFROM_year', 'DISPFROM_month', 'DISPFROM_dayofweek', 'DISPEND_year',
       'DISPEND_month', 'DISPEND_dayofweek', 'VALIDFROM_year',
       'VALIDFROM_month', 'VALIDFROM_dayofweek', 'VALIDEND_year',
       'VALIDEND_month', 'VALIDEND_dayofweek'],
```

Рис. 2. Набор признаков для решения задачи прогнозирования суммы покупки

Результаты работы, представленные на рис. 1, дают понять, что реализация градиентного бустинга LGBM выдает наиболее точный ответ одновременно и относительно precision, и recall. Деревья решений же справляются с данным датасетом хуже всего, при этом занимая внушительный промежуток времени в сравнении с прочими моделями.

Для решения задачи прогнозирования суммы покупки были обучены модели multivariate regression и LightGBM с использованием следующих признаков:

В результате имеем следующие значения метрик оценки качества:

Model	RMSE	R^2_{score}
Multivariate regression	4285.964	0.216
LGBM	3947.887	0.334

Рис. 3. Результаты прогнозирования суммы покупки

Рис. 3 показывает, что использование градиентного бустинга для данной задачи, позволяет получить наиболее точный ответ.

7. ВЫВОД

В данной работе были рассмотрены возможности применения методов машинного обучения для прогнозирования вероятности факта и суммы покупки клиентов интернет магазинов. В качестве исходных данных использовались данные о клиентах и их транзакциях с сайта ponpare.jp. Были построены и обучены следующие модели машинного обучения Logistic Regression, Decision trees, KNN, LightGBM Classification для задачи прогнозирования вероятности покупки и multivariate regression, LightGBM Regression для задачи прогнозирования суммы покупки клиента. В рамках обеих задач наилучшие результаты показал градиентный бустинг в реализации LGBM.

Литература

1. *Day, D., Gan, B., Gendall, P. and Esslemont, D.* Predicting purchase behaviour // Marketing Bulletin. 1991. P.18–30.
2. *Starostin, V.S. and CHERNOVA, V.Y.* E-commerce development in Russia: trends and prospects // The Journal of Internet Banking and Commerce. 2016.

3. *Kuhn M, Johnson K.* Applied predictive modeling // New York: Springer. 2013.
4. *Glasbey, C.A*. An analysis of histogram-based thresholding algorithms // CVGIP: Graphical models and image processing. 1993. P. 532–537.
5. https://github.com/dmlc/xgboost
6. *Yang S, Zhang H.* Comparison of several data mining methods in credit card default prediction // Intelligent Information Management. 2018. P. 115.
7. *Wu, H., Jiao, H., Yu, Y., Li, Z., Peng, Z., Liu, L. and Zeng, Z.* Influence factors and regression model of urban housing prices based on internet open access data // Sustainability. 2018. P. 1676.
8. *Liu, L., Ji, M. and Buchroithner, M.* Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra // Remote Sensing. 2017. P. 1299.
9. *Wu, J.Y.* Housing Price prediction Using Support Vector Regression. 2017.
10. *Limsombunchai, V.* House price prediction: hedonic price model vs. artificial neural network // In New Zealand agricultural and resource economics society conference. 2004. P. 25–26.
11. *Li, J.Z.* Monthly Housing Rent Forecast Based on LightGBM (Light Gradient Boosting) Model // International Journal of Intelligent Information and Management Science, 2018.

Using Machine Learning Methods to Solve Problems of Forecasting the Amount and Probability of Purchase Based on E-Commerce Data

Oleg A. Mamiev*
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000-0003-1137-4019
e-mail: olegios@mail.ru

Nikita A. Finogenov**
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000-0001-7680-9496
e-mail: finogenov.nik@gmail.com

Gleb B. Sologub****
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000-0002-5657-4826
e-mail: glebsologub@ya.ru

The study is aimed at investigating the possibility of using machine learning methods to build models for predicting the probability of purchase and the amount of purchase by online store customers. As a sample, we used data of users transactions of the site ponpare.jp in the period from 01.07.2011 to 23.06.2012. The description and comparative analysis of the most common methods for solving similar problems are given. The metrics used to measure the results in the case of forecasting the fact and amount of the purchase are being described. The results obtained make it clear that within the framework of the problem of predicting the probability of a purchase, gradient boosting, namely its implementation of LGBMClassifier, shows the most accurate estimate. For the problem of predicting the amount of a customer's purchase, using gradient boosting also gave the best results.

Keywords: probability and purchase amount forecast, classification, regression, data analysis, data processing, machine learning.

For citation:
Mamiev O.A., Finogenov N.A., Sologub G.B. Using Machine Learning Methods to Solve Problems of Forecasting the Amount and Probability of Purchase Based on E-Commerce Data. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2020. Vol. 10, no. 4, pp. 31–40. DOI: https://doi.org/10.17759/mda.2020100403 (In Russ., abstr. in Engl.).

****Oleg A. Mamiev***, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0003-1137-4019, e-mail: olegios@mail.ru

*****Nikita A. Finogenov***, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0001-7680-9496, e-mail: finogenov.nik@gmail.com

******Gleb B. Sologub***, PhD (Physics and Mathematics), Associate Professor, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0002-5657-4826, e-mail: glebsologub@ya.ru

References

1. Day, D., Gan, B., Gendall, P. and Esslemont, D. Predicting purchase behaviour // Marketing Bulletin. 1991. P.18–30.
2. Starostin, V.S. and CHERNOVA, V.Y. E-commerce development in Russia: trends and prospects // The Journal of Internet Banking and Commerce. 2016.
3. Kuhn M, Johnson K. Applied predictive modeling // New York: Springer. 2013.
4. Glasbey, C.A. An analysis of histogram-based thresholding algorithms // CVGIP: Graphical models and image processing. 1993. P. 532–537.
5. https://github.com/dmlc/xgboost
6. Yang S, Zhang H. Comparison of several data mining methods in credit card default prediction // Intelligent Information Management. 2018. P. 115.
7. Wu, H., Jiao, H., Yu, Y., Li, Z., Peng, Z., Liu, L. and Zeng, Z. Influence factors and regression model of urban housing prices based on internet open access data // Sustainability. 2018. P. 1676.
8. Liu, L., Ji, M. and Buchroithner, M. Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra // Remote Sensing. 2017. P. 1299.
9. Wu, J.Y. Housing Price prediction Using Support Vector Regression. 2017.
10. Limsombunchai, V. House price prediction: hedonic price model vs. artificial neural network // In New Zealand agricultural and resource economics society conference. 2004. P. 25–26.
11. Li, J.Z. Monthly Housing Rent Forecast Based on LightGBM (Light Gradient Boosting) Model // International Journal of Intelligent Information and Management Science, 2018.