DOI: 10.38061/idunas.848233

IDUNAS | NATURAL & APPLIED SCIENCES JOURNAL | 2021 Vol. 4 No. 2 (32-37)

Predicting Purchase Interest of Online Shoppers Using Boosting Algorithms

Research Article

Başak Esin Köktürk Güzel[1*], Devrim Ünay[1*]

[1]Department of Electrical and Electronics Engineering, Izmir Democracy University, İzmir, Turkey.

Author E-mails
basakesin@zoidata.com
devrim.unay@idu.edu.tr

*Correspondance to: Başak Esin Köktürk Güzel, Department of Electrical and Electronics Engineering, Izmir Democracy University, İzmir, Turkey.
DOI: 10.38061/idunas.848233

Received: 28.12.2020; Accepted: 30.06.2021

Abstract

Data driven marketing is becoming more and more vital for businesses day-by-day. Understanding customer behavior has the potential to decrease marketing costs as well as increase sales both in conventional marketing and online marketing. Since online users can access information faster, prices have become more competitive and customer behavior analysis has become more important. The purpose of this study is to predict the purchase interest of the users in an e-commerce web page by using the user session data such as pageview, duration etc. To this aim we used clickstream data for an e-commerce web page which is publicly available. Since only 16.5 percent of the sessions are completed with purchase in the dataset, increasing true positive rates rather than accuracy is more important. To this aim, we have explored the performance of boosting algorithms on the dataset and compared to those of state-of-the-art methods that were previously applied on the same dataset. Results show that boosting algorithms have better performance for identification of the sessions that end with a purchase.
Keywords: Online shopping intention prediction, boosting algorithms, adaboost, gradient boosting, extreme gradient boosting.

1. INTRODUCTION

Online shopping has experienced a rapid growth in recent years, as it offers solutions to 24/7 service needs, is less costly and has a wide range of products.

The global Coronavirus outbreak has had a major impact on societies around the world as well as brands and retailers including the way of shopping. The new conditions, caused by pandemic, have given rise to many brick-and-mortar stores temporarily closed, and thus people are going online to buy goods and services. In the first and second quarter of 2020, revenue of e-retailers have grown dramatically from 11.8% up to 16.1% of total retail sales in the US (Statistica, 2020). This situation has increased the competition

among electronic commerce (e-commerce) sites. During shopping, online shoppers can not only compare the price of the product they are interested in, but also decide based on the reliability and customer satisfaction of each site. Therefore, it is important to measure the purchasing tendencies of online buyers, to make improvements or to reveal the causes of operations that enable and/or inhibit purchasing.

Traditional retail stores can target their potential customers through conventional ways. But it is not possible to identify potential customers by e-commerce sites. However they are able to predict potential buyers by the help of collected data. The decision-making with these predictions can increase both the experience of the users and the recycling rates of sales. Several studies have been conducted to evaluate parameters that determine the online shopping trends of users.

Predicting the online shoppers purchase interest by observing their behavior in the considered online platform (usually by analyzing their behavior in the shopping website) is an interesting topic because if you can predict whether a user will make a purchase or not you got a massive economical information. Some of the related research focused on minimizing shopping cart abandonment by predicting user behavior in real-time (Awad & Khalil (2012), Budnikas (2015), Fernandes (2015)) while others aimed at segmenting customers according to their navigational patterns (Carmona et.al. (2012), Kau et.al. (2003), Moe (2003)). Sakar et al. (2019) compared the performances of different models such as Random Forests (RF), Support Vector Machines (SVM) and Multilayer Perceptron (MLP), and found that the accuracy and f-1 score of the MLP model were significantly higher than those of the RF and SVM models.

This study was designed to estimate whether a user section will end with a sale or not. By predicting purchase interest of a user, marketing campaigns can design according to the user and this provides to increase the conversion rate of the campaign. To this aim, we have applied boosting classifiers on the "Online Shoppers Purchasing Intention Dataset" which is published by Sakar et.al. (2019) and compared the performance of three boosting classifiers.

This paper is structured as follows. Section 2 describes the boosting methods, while Section 3 introduces the dataset used in this study. Section 4 presents the results of the boosting methods for prediction of users' purchase interest in e-commerce web pages. Section 5 concludes the paper by detailed discussions on the results and the potential future works.

2. METHODS

"Boosting" refers to creating a strong learner from the weak learners in machine learning applications. The boosting algorithms learn from the mistakes of weak learners and construct a potentially more robust and accurate classifier. They have an iterative approach where each learner focuses on correcting the mistakes of the previous learners. Adaboost and Gradient Boosting are the most popular boosting classifiers (Köktürk Güzel & Önder ,2018). In recent years, Extreme Boosting got attention in almost all data mining competitions because of its high classification accuracy and low computational cost. In this study we have used three types of boosting algorithms to predict buyers intention in online shopping web pages. Following sections introduce background information on these algorithms.

Adaboost

Adaboost assigns same weights to all samples in the beginning, but then updates the sample weights sequentially by giving higher weights to misclassified samples in the later iterations (Figure 1). To create a strong classifier, the algorithm takes the weighted sum of each learner output (Freund & Schapire (1995)).

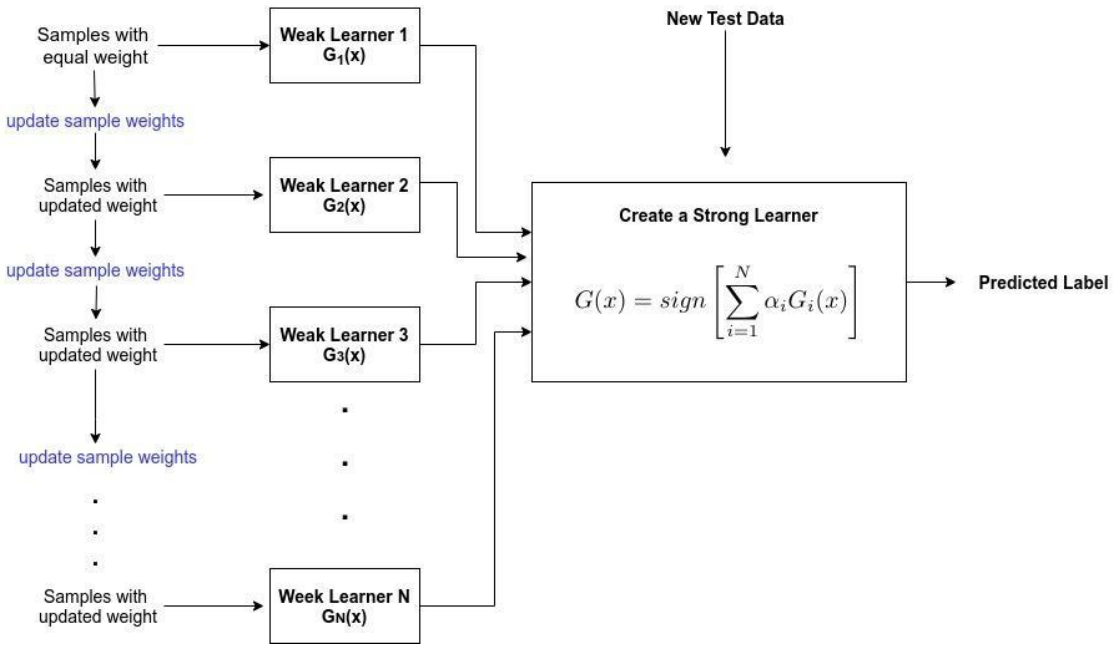

Figure 1. The schematic of the Adaboost algorithm.

Gradient Boosting

Like Adaboost, the Gradient Boosting algorithm also combines weak learners into a single strong learner in an iterative fashion (Figure 2). However, the gradient boosting algorithm trains learners with the error of the previous learner. Therefore it fits a function on the residual of the previous learner and attempts to correct this error (Friedman, 2001).

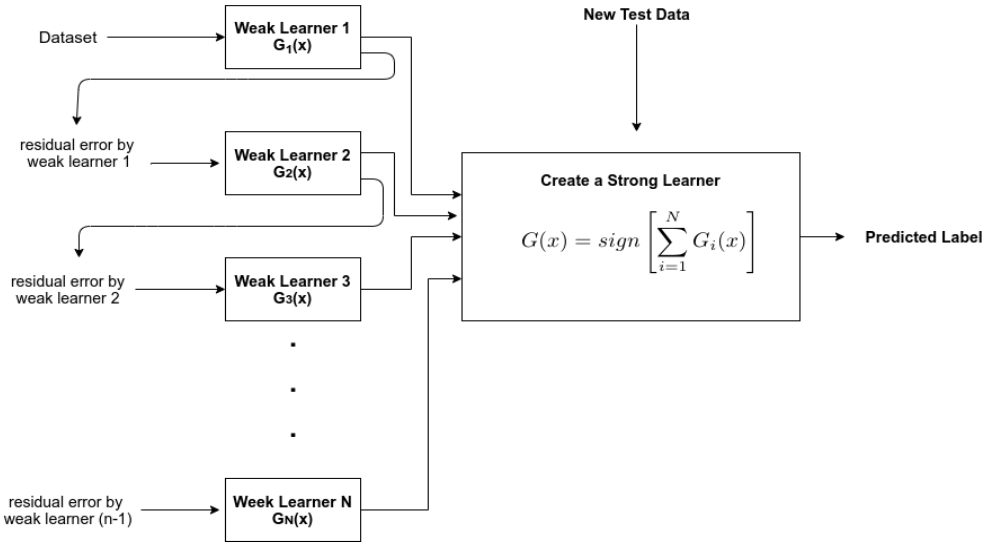

Figure 2. The schematic of the Gradient Boosting algorithm.

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) basically has a gradient boosting framework with more powerful features such as parallel computing, handling missing values, and regularization to avoid overfitting or bias. In other words, XGBoost is an implementation of the gradient boosting algorithm where it gets its power from system optimization and algorithmic enhancements. The XGBoost algorithm solves

the optimization problem by determining the step size and step direction in the same time (Chen and Guestrin 2016).

In the next section, we briefly describe our dataset and the features which we used in this study.

3. DATASET

The dataset has 12330 sessions (each belonging to different unique users) recorded with attributes related to web page analytics in a 1-year period. This 1-year duration has been selected according to certain criteria to avoid any tendency to a specific campaign, special day, user profile, or period. Only 16.5 percent of the samples have positive labels which mean the sessions ended with purchases (Sakar et al. 2019). The dataset, publicly available in the UCI Machine Learning Repository platform, has attributes listed with their descriptions in Table1.

Table 1. Names and Descriptions of the Attributes present in the Dataset (Sakar et al. 2019).

Attribute Name	**Attribute Description**
Administrative	Number of pages visited by the visitor about account management
Administrative_Duration	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Number of pages visited by the visitor about Web site, communication, and address information of the shopping site
Informational_Duration	Total amount of time (in seconds) spent by the visitor on informational pages
ProductRelated	Number of pages visited by visitor about product related pages
ProductRelated_Duration	Total amount of time (in seconds) spent by the visitor on product related pages
BounceRates	Average bounce rate value of the pages visited by the visitor
ExitRates	Average exit rate value of the pages visited by the visitor
PageValues	Average page value of the pages visited by the visitor
SpecialDay	Closeness of the site visiting time to a special day
Month	Month value of the visit date
OperatingSystems	Operating system of the visitor
Browser	Browser of the visitor
Region	Geographic region from which the session has been started by the visitor
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)
VisitorType	Visitor type as “New Visitor,” “Returning Visitor,” and “Other”
Weekend	Boolean value indicating whether the date of the visit is weekend
Revenue	Class label indicating whether the visit has been finalized with a transaction

In the next section we presented the results of classifiers on this dataset.

4. RESULTS

We have used the boosting classifiers to predict the result of a session (whether ended with a purchase or not) in an online shopping website. Note that, in this problem the number of positive samples is 16% of all the samples. So, estimating true positives are more challenging. If we predict all samples as negative our accuracy will be ~84% so to understand the performance of a classifier we also listed f1 measures which is a harmonic mean of precision and recall. Table 2 reports the performances of the algorithms on the cross-validated training data via accuracy and f1 measure scores that are computed using the corresponding confusion matrices.

Table 2. Cross-validation scores of the boosting algorithms on training data reported as mean (standard deviation).

	Adaboost	Gradient Boosting	XGBoost
Accuracy	0.892873 (0.014213)	0.904614 (0.013199)	**0.906672 (0.014580)**
f1-measure	0.622838 (0.050687)	0.659376 (0.042769)	**0.664804 (0.051171)**

Table 3. Test scores of the boosting algorithms.

	Adaboost	Gradient Boosting	XGBoost
True Positive Rate	0.541864	**0.586098**	0.578199
True Negative Rate	0.950815	**0.957509**	0.956636
Accuracy	0.887196	**0.899730**	0.897764
f1-measure	0.599127	**0.645217**	0.637631

Since the results show that performances of the algorithms are very close, we have applied t-test on the prediction results in a pairwise manner. The p-value between the predictions of XGBoost and AdaBoost classifiers is 0.043, which shows that XGBoost classifier has significantly better performance than Adaboost classifier for predicting the shoppers' purchase intention. However, when we calculate the p-value between predictions of XGboost classifier and Gradient Boosting classifier, we obtain a p-value of 0.19 which means that the performances of the classifiers are statistically indifferent.

Python programming language was used in our implementation and all source codes of this study are publicly available at Github page13 in order to allow for reproducibility of our results.

5. CONCLUSION

In this paper, we have presented a comparative study on the classification performances of boosting algorithms for predicting the shopping intention. Previously, Sakar et. al. (2019) published the classification results of multilayer perceptron and support vector machines on the same dataset. Here, we demonstrated that performance of boosting algorithms to identify positive samples is better than the state-of-the-art. We have achieved an average f1 measure of 0.664 using the XGBoost algorithm. We have obtained our results

on an open database and made our implementation publicly available to help further the research in this domain.

Our results support the argument that shopping intention from a clickstream data can be predicted using boosting methods. The technology discussed here is working as an offline classifier, however it has potential in real-time prediction problems. Designing a system that predicts the purchase interest in an online session has significant economic gain for e-retailers. Towards that aim, further research can be performed with the help of marketing experts to exploit domain knowledge.

6. REFERENCES

1. Awad, M. A., & Khalil, I. (2012). Prediction of user's web-browsing behavior: Application of markov model. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(4), 1131-1142.

2. Budnikas, G. (2015). Computerised recommendations on e-transaction finalisation by means of machine learning. Statistics in Transition. New Series, 16(2), 309-322.

3. Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur. com. Expert Systems with Applications, 39(12), 11243-11249.

4. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

5. Fernandes RF, Teixeira CM (2015) Using clickstream data to analyze online purchase intentions. Master's thesis, University of Porto.

6. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

7. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

8. Kau, A. K., Tang, Y. E., & Ghose, S. (2003). Typology of online shoppers. Journal of consumer marketing, 20(2), 139-156.

9. Köktürk Güzel, B. E., & Önder, D. (2018, May). Performance comparision of boosting classifiers on breast termography images. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

10. Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. Journal of consumer psychology, 13(1-2), 29-39.

11. Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing and Applications, 31(10), 6893-6908.

12. Statistica 2020 – www.statista.com/statistics/187439/share-of-e-commerce-sales-in-total-us-retail-sales-in-2010/

13. https://github.com/basakesin