ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20200808113807
Original scientific paper

An XGBoost Algorithm for Predicting Purchasing Behaviour on E-Commerce Platforms

Peiyi SONG, Yutong LIU*

Abstract: To improve and enhance the predictive ability of consumer purchasing behaviours on e-commerce platforms, a new method of predicting purchasing behaviour on e-commerce platforms is created in this paper. This study introduced the basic principles of the XGBoost algorithm, analysed the historical data of an e-commerce platform, pre-processed the original data and constructed an e-commerce platform consumer purchase prediction model based on the XGBoost algorithm. By using the traditional random forest algorithm for comparative analysis, the *K*-fold cross-validation method was further used, combined with model performance indicators such as accuracy rate, precision rate, recall rate and F_1-score to evaluate the classification accuracy of the model. The characteristics of the importance of the results were found through visual analysis. The results indicated that using the XGBoost algorithm to predict the purchasing behaviours of e-commerce platform consumers can improve the performance of the method and obtain a better prediction effect. This study provides a reference for improving the accuracy of e-commerce platform consumers' purchasing behaviours prediction, and has important practical significance for the efficient operation of e-commerce platforms.

Keywords: e-commerce platform; purchasing behaviour prediction; XGBoost algorithm

1 INTRODUCTION

In recent years, the e-commerce market has developed increasingly rapidly. In a complex and chaotic market environment, the e-commerce market is facing opportunities and challenges. The rapid growth of e-commerce platforms can provide consumers with more convenient methods of consumption and further expand the development space of e-commerce. Users leave considerable user data on e-commerce platforms, but only a small quantity of data is converted into purchasing behaviours. Data mining of information related to user purchase intention can determine the value and meaning of the data hidden within the big data and effectively improve user purchase intention. Consumer buying behaviours are affected by many factors, and different consumer demands lead to great differences in consumer buying behaviours. To predict consumer purchasing behaviours is to determine the hidden data characteristics within the massive data information left by users on the e-commerce platform and then determine the consumption willingness of future users on the e-commerce platform. Based on the predictive analysis of e-commerce platform consumer purchase behaviours, this paper provides an effective method for predicting users' purchasing behaviours, which can increase the transaction volume of e-commerce platforms and further promote the development of e-commerce platforms. The development of e-commerce platforms plays an extremely important role in the development of the entire e-commerce industry. Therefore, it is of great guiding significance for the development of e-commerce platforms and e-commerce to use scientific forecasting methods to predict and analyse the purchasing behaviours of e-commerce platform consumers.

Many different methods have been proposed in the literature for studying the prediction problem. For example, regression models with comprehensive factors have been used to provide PM 2,5 prediction methods at different time periods [1]. In another study, a group of researchers provided a predictive model of equipment failures in urban transportation systems based on the Internet of Things and big data [2]. Currently, instead of using traditional methods, scholars use more complex algorithms or models to solve problems in fields such as the internet, transportation, agriculture and industry [3, 4]. Some scholars have used genetic algorithms, Markov models and other methods in their research. Other scholars combined Bayesian network methods to study the selection of green suppliers in agricultural production [5]. In one study, an integrated metring strategy was adopted to estimate instant vehicle emissions by measuring the speed and acceleration of vehicles on slopes, thus improving the study on the impact of ramp metring on the environment [6]. In another study, scholars applied an improved whale optimization algorithm and proposed an algorithm framework for the location model of electric vehicle charging stations based on the calculation of a large number of examples [7]. To solve the problem of airport ground vehicle support in the aviation business, some scholars used the standard particle swarm optimization method in their research [8]. Although the research field of these documents is different from this study, the ideas and methods of these studies are still worthy of reference for this study. There are many types of research on network user behaviours. To study the usability of e-commerce platforms, scholars have proposed a new method for evaluating the usability of e-commerce platforms to help e-commerce retailers evaluate the usability of their websites [9]. To classify user behaviours, scholars proposed an improved K-means clustering method, analysed and studied network user behaviours, and improved the scalability of the algorithm [10]. Other scholars explained the important role of trust factors in e-commerce on user usage and user satisfaction [11]. Another type of research was based on user behaviours to study product recommendation systems. By investigating users' online shopping habits, scholars improved the assisted filtering algorithm and built a product recommendation system using the innovator's concept [12]. The improved game model of social e-commerce users' purchasing behaviours from the perspective of commodity information dissemination has also been studied [13].

It is of great value to use machine learning algorithms to predict consumer purchasing behaviours on e-commerce platforms [14]. In view of the correlation between the behaviour data information left by consumers on e-commerce platforms and their purchase intentions, it is an

important research subject to be improved to build a prediction model with high prediction accuracy to predict users' consumption behaviours. This study introduces the eXtreme gradient boosting (XGBoost) algorithm into the prediction of consumers' purchasing behaviours on e-commerce platforms. With the help of Python tools, this paper studies and analyses the characteristics of the purchase intention data of e-commerce platform consumers, which can provide enterprises with a new method for analysing and predicting the purchase behaviours of e-commerce platform consumers to improve the marketing level of enterprises. The obtained results are conducive to targeted planning and strategic decisions of e-commerce platforms and timely adjustment of marketing decisions to attract more consumers, achieve good economic benefits and promote sustainable development of the industry.

The remainder of this paper is organized as follows. In Section 2, the basic principle of the XGBoost algorithm and the dataset information are proposed. The calculation results are presented in Section 3. Finally, in Section 4, some concluding remarks are given.

2 METHODS

Because the XGBoost algorithm can effectively capture the dependencies of complex data, it can also use extensible learning systems to learn from large data sets and get models. Therefore, in view of the advantages in data processing and analysis, to solve the problem of predicting purchasing behaviour on e-commerce platforms, this study uses Python tools to build a prediction model based on the XGBoost algorithm and uses the XGBoost algorithm to realize dataset analysis, training and prediction. XGBoost is an algorithm that realizes efficient classification under the gradient boosting framework. It improves the gradient boosting machine (GBM), which has the features of high efficiency, flexibility and portability, and can provide a gradient boosted decision tree. The basic idea of GBM is the idea of gradient descent, in which each generated tree is based on the previous result to minimize the objective function. When the data are more complex, the XGBoost algorithm can utilize a multicore CPU to perform parallel computation and improve the accuracy of the algorithm [15].

Assuming that the given dataset is D, the number of samples is n, and the number of eigenvalues is m, $D = \{(X_i, y_i)\}$, ($|D| = n$, $X_i \in R^m$, $y_i \in R$). The main model structure of XGBoost is generated by the addition model of K tree models, each tree fitting the residuals of the previous tree. We can express the integrated model of the tree as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

$$F = \left\{ f(x) = w_{q(x)} \right\} \left(q: R^m \to T, w \in R \right) \tag{2}$$

where F represents the regression tree model space, x_i represents the eigenvectors of data, T represents the number of leaf nodes in the tree, and f_k represents the independent tree structure q and leaf weight w.

The objective function contains the difference between the predicted value $\hat{y}_i$ and the real value y_i of training and the complexity of the model. $\sum_i z(y_i, \hat{y}_i)$ is the loss on training data, which is used to measure how well model fits on training data. Because the objective function can not be used in the Euclidean space of the traditional optimization method, each iteration generates a tree based on the previous results, adding a new function to the model. XGboost has the characteristics of the decision tree and also considers the over-fitting of the decision tree. The model contains regularization, which is often used to measure the complexity of the model. XGboost mainly uses the L2 regularization method, $\Omega(f)$ includes the total number of leaf nodes and the regularization obtained from leaf nodes. The result of each iteration is denoted as $\hat{y}_i^{(t)}$. f_t is used to continuously optimize the objective function, and Taylor expansion is carried out at $\hat{y}_i^{(t)}$ to accelerate the optimization of the objective function. g_i is the first derivative of the loss function, and h_i is the second derivative of the loss function. Finally, the training objective function is obtained. The relevant functions are defined as follows.

$$Z(\phi) = \sum_i z(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|W\|^2 \tag{4}$$

$$f(x + \Delta x) \simeq f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \tag{5}$$

$$g_i = \partial_{\hat{y}(t-1)} z\left(y_i, \hat{y}^{(t-1)}\right) \tag{6}$$

$$h_i = \partial^2_{\hat{y}(t-1)} z\left(y_i, \hat{y}^{(t-1)}\right) \tag{7}$$

$$\tilde{Z}^{(t)} \cong \sum_i \left[g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{8}$$

The data used in this paper come from an open dataset of consumers' purchase intentions of an e-commerce platform [16, 17]. The dataset was collected from a real e-commerce environment and recorded information from 12 330 users reflecting online purchase intention in one year, including 10 numerical attributes and 4 classification attributes. In the original dataset, "Administrative" indicates the number of management pages users visited in the session, "Administrative_Duration" denotes the total time users spent accessing the management pages during that session and "Informational_Duration" represents the total time users spent on informational pages. "ProductRelated" indicates the number of product-related pages that users visited in this session, "ProductRelated_Duration" denotes the total time users spent on product-related pages, the values of which were derived from the URL information of the pages the user visited. The "BounceRates" of a web page refers to the percentage of users who entered the site from that page and then left, without triggering any other requests for Google analysis during that session. The "ExitRates" of a particular page is the percentage of visits to all pages according to the last page in the session. The "PageValues" denotes the average value of the page visited before a user completed an online shopping transaction or simply exited the site

without completing the transaction. "SpecialDay" indicates how close the time of the visit to the site was to a particular day, such as Mother's Day. The higher the proximity, the more likely an online transaction was to occur for that session. Another type of attribute is the category attribute. "Month" means the month of the user's visit, "VisitorType" represents the type of users and "Weekend" denotes whether the user access date was a weekend. "Revenue" indicates that the user completed an online shopping transaction.

Among the online browsing information of 12 330 users, 10 422 samples that did not end with shopping are negative examples, and the remaining 1908 samples that end with shopping are positive examples. In the original dataset, the classification result of "Revenue" is FALSE, which represents a negative example, while a TRUE classification result represents a positive example. According to the basic information of the output dataset, it can be seen that the data used in this paper are normalized and cleaned adequately. The null value number of each data point is 0, and the missing rate is 0%.

Since the default data of XGBoost are numeric, this paper adopts the method of one-hot encoding to convert other forms of data to numeric data and uses get_dummies to one-hot coding. In the original dataset, the values of the classification feature of "Month" are Feb, Mar, May, June, Jul, Aug, Sep, Oct, Nov, and Dec, and the values of the classification feature of "VisitorType" are "Returning_Visitor", "New_Visitor" and "Other". The values of the "Weekend" classification feature are FALSE and TRUE. None of these values can be substituted into the model calculation. Therefore, the data of classification features such as "Month", "VisitorType" and "Weekend" are encoded separately. In this paper, the characteristics are selected based on the Pearson correlation coefficient of the "Revenue" feature. The results reach statistical significance. The selected features are combined into a feature dataset. Nine features are selected as the features of the model. The descriptive statistical results of the data set are shown in Tab. 1.

Table 1 Descriptive statistical results of dataset samples

Features	Min	Max	Mean	Standard deviation
Administrative	0,00	27,00	2,32	3,32
ProductRelated	0,00	705,00	31,73	44,48
ProductRelated_Duration	0,00	63973,52	1194,75	1913,67
BounceRates	0,00	0,20	0,02	0,05
ExitRates	0,00	0,20	0,04	0,05
PageValues	0,00	361,76	5,89	18,57
Month_Category_Nov	0,00	1,00	0,24	0,43
VisitorType_Category_ Returning_Visitor	0,00	1,00	0,86	0,35
VisitorType_Category_ New_Visitor	0,00	1,00	0,14	0,34

3 RESULTS AND DISCUSSION

3.1 Test Scheme Design

The dataset is divided into a training set and a test set. The train_test_split method in the sklearn.model_selectiond package is used to set the ratio between the training set and test set to 4:1. Based on the XGBoost algorithm, DictVectorizeris used to vectorize the feature, and the classification and prediction of model data are studied by applying XGBClassifier, which is imported from the XGBoost package. To better reflect the applicability of the XGBoost algorithm to this dataset, this paper also uses the random forest algorithm to compare and analyse the model results [18]. RandomForestClassifier can be imported from sklearn.ensemble to implement the random forest algorithm. For e-commerce platform consumers purchasing behaviour prediction issues, a better method can be found by comparing the two methods.

3.2 Evaluation Indicators

The commonly used evaluation indicators include accuracy rate, precision rate, recall rate, F_1-score and area under the receiver operating characteristic curve (AUC). To better assess the accuracy of model classification, cross_val_score is imported from sklearn.model_selection, and the K-fold cross-validation method is used to obtain the average accuracy of classification by calculating the average recognition rate of K test sets, which can well-evaluate the model effect. In this paper, K is 5, and the performance of the default configuration of RandomForestClassifier and XGBClassifier is evaluated on the training set using the five-fold cross-validation method. The average classification accuracy score of the output five-fold cross validation is high, both of which are greater than 0,89, indicating that a better model effect can be obtained. Therefore, the operation of the model is completed under the default configuration initialization RandomForestClassifier and XGBClassifier. The accuracy of prediction models can be effectively evaluated by the evaluation indicators.

3.3 Prediction Results

It is of great significance for e-commerce platforms to accurately predict the purchasing behaviours of e-commerce platform consumers. According to the prediction results, e-commerce platforms can further understand the user's purchasing intention and increase operating income. The prediction of the test set is carried out, and then classification_report is imported from sklearn.metrics to further analyse the model using the evaluation indicators to assess the prediction effect obtained by using the random forest algorithm and XGBoost algorithm.

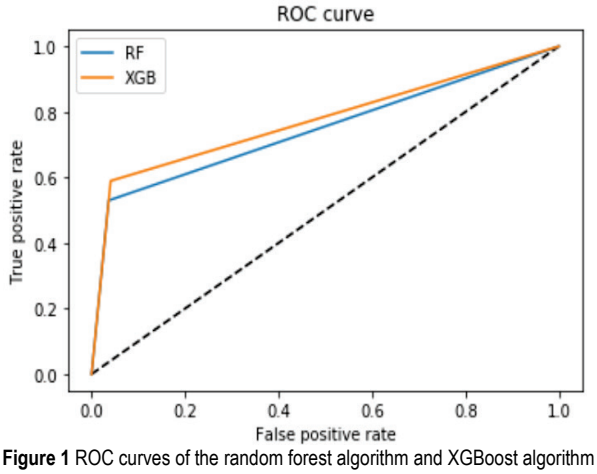

Figure 1 ROC curves of the random forest algorithm and XGBoost algorithm

Table 2 Comparison of the random forest algorithm and XGBoost algorithm indicators

Indicators	Random forest algorithm	XGBoost algorithm
Accuracy rate	0,8958	0,9015
Positive precision rate	0,53	0,59
Negative precision rate	0,96	0,96
Positive recall rate	0,73	0,73
Negative recall rate	0,92	0,93
Positive F_1-score	0,61	0,65
Negative F_1-score	0,94	0,94

From the output calculation results in Tab. 2, the accuracy of the XGBoost algorithm is 0,9015, and the accuracy of random forests is 0,8958, indicating that the XGBoost algorithm has a better classification accuracy than the random forest algorithm. The effect of the XGBoost and stochastic forest algorithms can be intuitively compared through the receiver operating characteristic (ROC) curve. When the ROC curve tends to point (0, 1), the model's effect is better. As shown in Fig. 1, the orange curve represents the ROC curve for the XGBoost model and the blue curve represents the ROC curve for the random forest model. The XGBoost model has a better ROC curve, which is higher than that of the random forest algorithm, indicating that XGBoost can better integrate the characteristics of different dimensions. Therefore, XGBoost is more suitable for the dataset in this study. The results indicate that using the XGBoost algorithm can be an appropriate and good solution to the problem of e-commerce platform consumers purchasing behaviour prediction issues. The AUC of XGBoost is 0,7744, which is higher than that of random forest (0,7467), indicating that XGBoost has higher credibility than random forest.

3.4 Significance Analysis of Features

After predicting the model results, the importance of features can be further analysed, which is of great significance to better predict consumers' purchasing behaviours. The basic idea of the feature importance evaluation of the random forest algorithm is to analyse the contribution value of each feature in each tree in the random forest, then take the mean value, and evaluate the feature importance by comparing the contribution value between features. The feature importance of the output random forest algorithm and the XGBoost algorithm are shown in Fig. 2 and Fig. 3. Obviously the fifth feature ranks first in the order of importance in both models, and its corresponding feature is "exit rate". Since the classification accuracy of the random forest algorithm is lower than that of the XGBoost algorithm, the results of the latter should be used as the main basis for the feature importance analysis.

The XGBoost algorithm's feature importance bar is shown in Fig. 3 to determine which feature variables have a more significant impact on user purchasing behaviour, so as to better understand the user's purchasing habits and purchasing needs. It can be seen in Fig. 3 that the score of feature importance of "ExitRates", "BounceRates" and "VisitorType_Category_New_Visitor" is greater than 0,1, indicating that "ExitRates" is crucial and that it is necessary for e-commerce platforms to conduct an in-depth analysis of exit rates. High exit rates means that web content is less attractive to users, and e-commerce platforms need to improve web content to reduce exit rates. Although "BounceRates" is not critically important, the bounce rate of web pages is an essential feature measuring the performance of web pages. It will directly affect the user experience, and improving user stickiness is the key to reducing the bounce rate. The purchase intention of new users is very strong, indicating that appropriate marketing strategies can maximize the attraction of new users and retain new users, thus completing the transformation of new users. By studying these variables, e-commerce platforms can explain and analyse consumer behaviours in a deeper way and then determine the factors that have an important influence on users' purchasing behaviours and further optimize and improve platform operations.

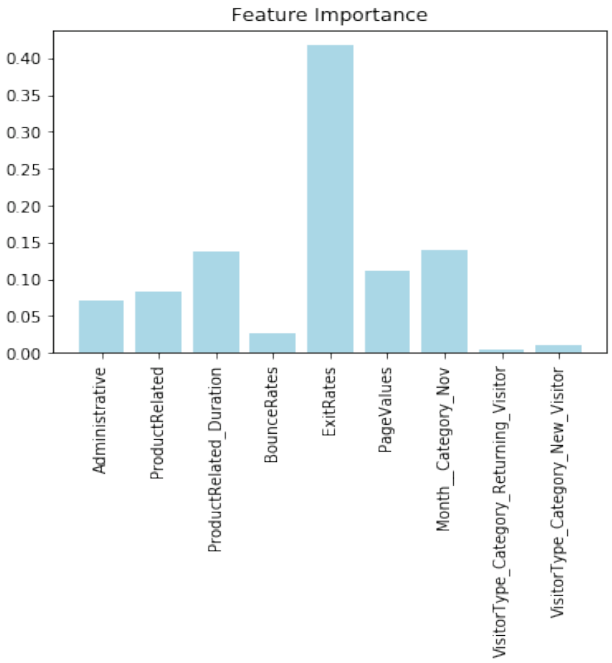

Figure 2 Feature importance of the random forest algorithm

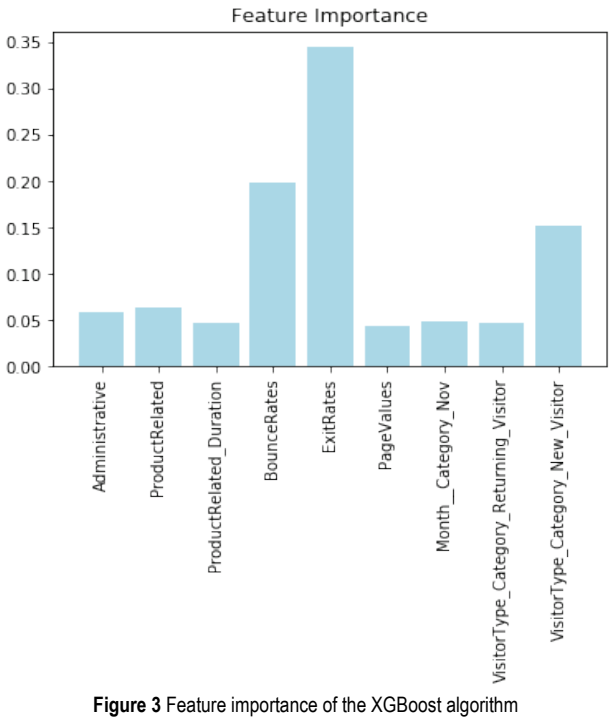

Figure 3 Feature importance of the XGBoost algorithm

4 CONCLUSIONS

Due to the rapid development of e-commerce, major e-commerce platforms pay increasing attention to valuable user data on platforms. It is not negligible to mine the data information and formulate marketing decisions and development strategies for e-commerce platforms more effectively. This paper first expounded the basic principle of the XGBoost algorithm and used the XGBoost classification algorithm to train and predict a real e-commerce platform user dataset with the help of Python tools. By mining the rich information behind the data, this paper studied the prediction of consumers' purchasing behaviours on e-commerce platforms. The random forest algorithm was introduced for comparison with the XGBoost algorithm by extracting and classifying data features. Based on the XGBoost algorithm and random forest algorithm, the test set was trained and predicted, and the importance of features and model performance were further evaluated. The study found that the average classification accuracy precision ratio and recall ratio and the AUC value and other model indicators of XGBoost were better than those of the random forest model, indicating that using the XGBoost algorithm can obtain better prediction results. The application of theXGBoost algorithm can effectively improve the accuracy of e-commerce platform consumers' purchasing behaviour prediction issues, and it also provides new ideas for the efficient operation of e-commerce platforms.

5 REFERENCES

[1] Du, J., Qiao, F., & Yu, L. (2019). Temporal characteristics and forecasting of PM2.5 concentration based on historical data in Houston, USA. *Resources, Conservation & Recycling, 147*, 145-156. https://doi.org/10.1016/j.resconrec.2019.04.024

[2] Li, H., Xu, W., Cui, Y., Wang, Z., Xiao, M., & Sun, Z. (2020). Preventive Maintenance Decision Model of Urban Transportation System Equipment Based on Multi-Control Units. *IEEE Access, 8*(1), 15851-15869. https://doi.org/10.1109/ACCESS.2019.2961433

[3] Zhang, D., Sui, J., & Gong, Y. (2017). Large scale software test data generation based on collective constraint and weighted combination method. *Tehnički vjesnik, 24*(4), 1041-1049. https://doi.org/10.17559/TV-20170319045945

[4] Awad, M. A. & Khalil, I. (2012). Prediction of User's Web-Browsing Behavior: Application of Markov Model. *IEEE Transactions on Systems, Man & Cybernetics: Part B, 42*(4), 1131-1142. https://doi.org/10.1109/TSMCB.2012.2187441

[5] Zhang, H. & Cui, Y. (2019). A model combining a Bayesian network with a modified genetic algorithm for green supplier selection. *Simulation, 95*(12), 1165-1183. https://doi.org/10.1177/0037549719826306

[6] Jianbang, D., Qing, L., Fengxiang, Q., & Lei, Y. (2018). Estimation of Vehicle Emission on Mainline Freeway under Isolated and Integrated Ramp Metering Strategies. *Environmental Engineering & Management Journal (EEMJ), 17*(5), 1237-1248. https://doi.org/10.30638/eemj.2018.123

[7] Zhang, H., Tang, L., Yang, C., & Lan, S. (2019). Locating electric vehicle charging stations with service capacity using the improved whale optimization algorithm. *Advanced Engineering Informatics, 41*, 100901. https://doi.org/10.1016/j.aei.2019.02.006

[8] Zhao, P. X., Gao, W. Q., Han, X., & Luo, W. H. (2019). Bi-Objective Collaborative Scheduling Optimization of Airport Ferry Vehicle and Tractor. *International Journal of Simulation Modelling (IJSIMM), 18*(2), 355-365. https://doi.org/10.2507/IJSIMM18(2)CO9

[9] Hasan, L., Morris, A., & Probets, S. (2013). E-commerce websites for developing countries - a usability evaluation framework. *Online Information Review, 37*(2), 231-251. https://doi.org/10.1108/OIR-10-2011-0166

[10] Gang, Z. (2016). The Application of User Behavior Analysis by Improved K-means Algorithm Based on Hadoop. *Revista de La Facultad de Ingenieria, 31*(7), 111-120. https://doi.org/10.21311/002.31.7.11

[11] Tam, C., Loureiro, A., & Oliveira, T. (2020). The individual performance outcome behind e-commerce: Integrating information systems success and overall trust. *Internet Research, 30*(2), 439-462. https://doi.org/10.1108/INTR-06-2018-0262

[12] Wang, C.-D., Deng, Z.-H., Lai, J.-H., & Yu, P. S. (2019). Serendipitous Recommendation in E-Commerce Using Innovator-Based Collaborative Filtering. *IEEE Transactions on Cybernetics, 49*(7), 2678-2692. https://doi.org/10.1109/TCYB.2018.2841924

[13] Lv, J., Wang, T., Wang, H., Yu, J., & Wang, Y. (2020). A SECPG model for purchase behavior analysis in social e-commerce environment. *International Journal of Communication Systems, 33*(6), 1-12. https://doi.org/10.1002/dac.4149

[14] Athanasiou, V. & Maragoudakis, M. (2017). A Novel, Gradient Boosting Framework for Sentiment Analysis in Languages where NLP Resources Are Not Plentiful: A Case Study for Modern Greek. *Algorithms*, *10*(1), 34. https://doi.org/10.3390/a10010034

[15] Johnson, R. & Zhang, T. (2014). Learning Nonlinear Functions Using Regularized Greedy Forest. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 36*(5), 942-954. https://doi.org/10.1109/TPAMI.2013.159

[16] Dua, D. & Graff, C. (2019). UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[17] 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 & Applications, 31*(10), 6893-6908. https://doi.org/10.1007/s00521-018-3523-0

[18] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Contact information:

Peiyi SONG, Prof., PhD supervisor
School of Economics and Management,
Communication University of China,
Beijing 100024, China
E-mail: pysong@cuc.edu.cn

Yutong LIU
(Corresponding author)
School of Economics and Management,
Communication University of China,
Beijing 100024, China
E-mail: yutong8819@126.com