Hindawi
Scientific Programming
Volume 2022, Article ID 1041741, 12 pages
https://doi.org/10.1155/2022/1041741

Research Article

Data Analysis and Prediction Modeling Based on Deep Learning in E-Commerce

Lei Feng

School of Digital Commerce, Beijing Information Technology College, Beijing 100018, China

Correspondence should be addressed to Lei Feng; fengl@bitc.edu.cn

Received 15 December 2021; Revised 27 December 2021; Accepted 25 January 2022; Published 24 March 2022

Academic Editor: Hangjun Che

Due to the low efficiency of traditional data analysis methods for massive e-commerce data analysis, an e-commerce data analysis and prediction method based on the GBDT deep learning model was proposed. Purchase behavior is divided into another category, which transforms the problem of e-commerce data analysis and prediction into a binary classification problem. At the same time, we extract 107 features that can reflect the user behavior and construct the GBDT model. The characteristics include counting class, sorting class, time difference class, conversion rate class, and so on. It follows from the above that the analysis and prediction of e-commerce data are realized. In addition, the results show that when the learning rate of GBDT model parameters is 0.05, the number of basic learners is 200, the tree depth is 20, the threshold is 0.5, the model prediction effect is best, and the F1 value can reach 0.12. Compared with the traditional prediction model based on logistic regression and neural network, the proposed GBDT model is more suitable for e-commerce data analysis and prediction.

1. Related Work

The development of computer technology and Internet technology has accelerated the construction and popularization of e-commerce platform. In the era of economic globalization, e-commerce has played a positive role in national economic and social development. Therefore, the use of e-commerce can improve business efficiency and promote sustained and healthy economic development, which is the focus and difficulty of current economic research. The premise of using e-commerce to improve business efficiency is to fully understand the e-commerce platform, which requires the analysis and prediction of e-commerce data. At present, the analysis and prediction methods for data mainly include two categories based on logistic regression and neural network, and the analysis and prediction of data can be realized by constructing the corresponding optimal prediction model using training sets. For example, Ozgur and Franklin analyzed multiple linearity in independent variables of the logistic regression model and applied it to actual data analysis cases [1]. The results show that the logistic regression model is easy to operate in data analysis and can obtain relatively comprehensive prediction results. Cioci et al. and Rekha et al. believed that multiple logistic regression (MLR) used to analyze categorical variables or continuous variables has a positive impact on the single dichotomy by reviewing the statistical methods of adjusting baseline differences in nonrandom studies [2, 3]. It can be seen that the data analysis can be realized according to the data format. [4] realized the prediction of Indian stocks by analyzing the data of Indian stocks based on the advantages of machine learning, especially the recurrent neural network, which can better extract the features of text and data [4, 5]. On the basis of deep learning, Son et al. and Jin et al. adopted the LSTM model to achieve spatiotemporal data prediction and conducted visual analysis [6, 7]. Guo et al. and Agafonov realized the prediction of microinternal leakage by analyzing the data of hydraulic cylinder based on the neural network model [8, 9]. The above method has made some research results in data analysis and prediction. However, due to the huge amount of e-commerce data, the data structure is complex and the data characteristics are rich; if the above method is used to analyze e-commerce data, there is usually a problem of the low efficiency of data analysis, prediction, or the accuracy of prediction. In order to solve the above problems, this paper, with the help of the gradient boosting decision tree (GBDT) model, which has the advantages of high prediction accuracy, few parameters, and stable training process, an e-commerce data analysis and prediction method based on GBDT deep learning model is proposed.

2. Introduction to GBDT Model

The GBDT model is an iterative decision tree algorithm, consisting of multiple decision trees as the base learner. The accuracy of the whole model is improved by trying to reduce the deviation of each decision tree. For the regression and classification problems, the decision trees adopted by the GBDT model are CART regression trees [9]. The CART regression tree is generated by traversing all the data features and dividing the data set into nodes in turn. Firstly, the features are selected according to the least square error, then each region is divided into two regions, and finally, the mean value of the current region is output to establish a regression tree [10]. The steps are as follows:

(1) Suppose the training dataset is D, the feature with the least square error is j, and the corresponding partition node is s. Solve (1) to obtain the optimal partition feature.

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]. \tag{1}$$

(2) Select the best (j, s) to divide regions and output corresponding region values:

$$\begin{gathered}R_1(j,s)=\left\{x|x^{(j)}\le s\right\}R_2(j,s)=\left\{x|x^{(j)}>s\right\}\\ \widehat{c}_m=\frac{1}{N_m}\sum_{x_1\in R_m(j,s)}y_i,\quad x\in R_m, m=1,2.\end{gathered} \tag{2}$$

(3) Repeat the above operations for the two divided regions until the termination conditions are met.

(4) Divide the input space into $R_1, R_2, \ldots, R_M$ subregions, where M is the number of subregions, and the final decision tree is generated.

$$f(x)=\sum_{i=1}^{M}\widehat{c}_m I(x\in R_m). \tag{3}$$

2.1. Gradient Boosting Tree. The GBDT model usually adopts gradient boosting tree to optimize the model learning process. Using the negative gradient of loss function as the descent mode, the regression tree can be constructed rapidly. The generation method of gradient boosting tree is as follows:

(1) Set input training dataset as $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$ and loss function as $L(y,f(x))$. Then, initialize $f_o(x)=0$.

(2) Calculate the pseudo-residual of sample $i=1,2,\ldots,N$ according to the following formula [11]:

$$r_{mi}=-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x)-f_o(x)}. \tag{4}$$

(3) Conduct fitting learning for the residuals and then obtain a regression tree $h_m(x)$, $m=1,2,\ldots M$.

(4) Update:

$$f_m(x)=f_{m-1}+h_m(x). \tag{5}$$

(5) Finally, get the gradient boosting tree:

$$f_M(x)=\sum_{m-1}^{M}h_m(x). \tag{6}$$

2.2. Selection of Loss Function. Common loss functions include squared error, hinge loss, and logistics regression loss. The mathematical expressions are shown in formulas (7)–(9) [12]. Among them, squared error loss function is mainly used for regression model, and hinge loss function is mainly used for SVM classifier. Therefore, this paper adopts logistics regression loss function as GBDT model loss function.

$$\begin{aligned}L(y,f(x))&=\sum_{i=1}^{n}(f(x_i)-y_i)^2\\ L(x,f(x))&=\sum_{i=1}^{n}\max(0,1-y_i*f(x_i))\\ L(y,f(x))&=\sum_{i=1}^{n}1og(1+\exp(-y_i*f(x_i)).\end{aligned} \tag{7}$$

2.3. Classification Method of GBDT Model. Above all, the classification calculation process of GBDT model using logistics regression loss function can be summarized as follows:

(1) Let the training dataset be $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$, and the loss function be $L(y,f(x))=1\text{n}(1+\exp(-2yf(x))), y\in\{0,1\}$; initialize:

$$f_0(x)=\frac{1}{2}1\text{n}\frac{P(y=1|x)}{P(y=0|x)}. \tag{8}$$

(2) Calculate the pseudo-residual of sample $i=1,2,\ldots,N$:

$$r_{mi} = \frac{2y_i}{1 + \exp\left(2y_i f_{m-1}\left(x_i\right)\right)}. \tag{9}$$

(3) Adopt regression tree to fit (12) and then obtain the leaf node region R_{mj}, $j = 1, 2, \ldots, J$ of m trees.

$$\left\{\left(x_1, r_{m1}\right), \left(x_2, r_{m2}\right), \ldots, \left(x_N, r_{mN}\right)\right\}. \tag{10}$$

(4) Calculate $j = 1, 2, \ldots, J, i = 1, 2, \ldots, N$:

$$c_{mj} = \frac{\sum_{x_i \in R_{mj}} r_{mj}}{\sum_{x_i \in R_{mj}} \left|r_{mj}\right| \left(2 - \left|r_{mj}\right|\right)}, \tag{11}$$

where $m = 1, 2, \ldots, M$.

(5) Update:

$$f_m(x) = f_{m-1}(x) + \sum_{j-1}^{J} c_{mj} I\left(x \in R_{mj}\right). \tag{12}$$

(6) Finally, get the classification tree:

$$f(x) = \sum_{m-1}^{M} \sum_{j-1}^{J} c_{mj} I\left(x \in R_{mj}\right). \tag{13}$$

(7) Use the difference between the predicted category probability value and the real probability value to fit the loss, then obtain the probability of different categories, and select the prediction category with high probability [13].

$$\begin{aligned} P(y = 1|x) &= \frac{1}{1 + \exp(-2f(x))}, \\ P(y = 0|x) &= \frac{1}{1 + \exp(2f(x))}. \end{aligned} \tag{14}$$

3. E-Commerce Data Analysis and Prediction Model Based on GBDT

Based on the above GBDT model analysis, the design of the e-commerce data analysis and prediction method is as follows:

(1) Firstly, delete missing values and desensitize all selected e-commerce data. Then, for better data analysis and prediction, the overall distribution of the data is described to obtain the distribution of user behavior.

(2) User browsing, collection, and additional purchase behavior are divided into one category. Purchase behavior is divided into another category. In addition, the problem is converted into binary classification problem.

(3) Select the features that can reflect the data to build the GBDT model and initialize the parameters of GBDT model, including learning rate, the number of base learners, thresholds, and others.

(4) Use the training set to train the GBDT model and tune the model parameters by random search [14]. When the model reaches the maximum number of iterations or optimal parameters, the training is stopped, and the optimal model is output.

(5) Use the optimal model obtained by training to predict the data to predict and then output the prediction result. Thus, the analysis and prediction of e-commerce data are realized.

4. Simulation Experiment

4.1. Construction of Experimental Environment. This experiment is carried out on 64 bit Windows 7 operating system, and the GBDT model is constructed on Python and TensorFlow framework. The CPU is Intel(R)Core(TM) i7-7770hq 2.8 GHz with 8 GB memory.

4.2. Data Sources and Processing

4.2.1. Data Sources. Select tianchi offline competition data as the experimental dataset to predict user purchase data on December 19, 2014. This dataset includes the historical e-commerce behavior data of 20,000 users on the complete collection of goods from November 18 to December 18, 2014 [15]. Its source data include user behavior dataset D and product subset P on the complete collection of goods. Dataset D contains 4758484 kinds of commodities and 4 behavior types where a total of 9557 commodity categories are missing commodity spatial identification [16]. The field description is shown in Table 1, and some data are shown in Table 2. Dataset P contains 422858 kinds of commodities. Here, the spatial identification of 1,054 commodity categories is missing. The field description is shown in Table 3, and some data are shown in Table 4.

4.2.2. Data Description. For better data analysis and prediction, it is necessary to understand the overall distribution of data. First, the operational purchase conversion rate of the data is calculated, that is, the proportion of the user's purchase behavior to its total behavior [17]. Through calculation, the complete behavior distribution of commodities is obtained as shown in Figure 1. It can be seen from the figure that users' browsing behavior on the complete collection of goods accounts for the largest proportion among all behaviors. Except for the abnormal behavior on December 12, the user's behavior on other days is relatively stable. The analysis of the reason for the abnormal behavior on December 12 is related to the promotion of "Double 12" on e-commerce platform.

Figure 2 shows the behavior distribution of users on a subset of goods. It can be seen from the figure that the user's behavior on the subset of goods is mainly browsing. The number of behaviors on December 12 is higher than that on other dates, which is related to the promotion of "Double 12" e-commerce platform. Compared with the user's behavior on the complete collection of goods, the user's behavior on the subset of goods varies greatly.

Table 1: Field description of dataset D.

Field	Meaning	Note
User_id	User ID	Be unique
Item_id	Commodity identification	Be unique
Behavior_type	There is a many-to-one relationship with commodities	Browse Collection Add shopping cart 4-purchase order
User_ geohash	The product category ID of the longitude and latitude area encoded by geohash where the user is located	Encoded by the GeoHash algorithm
Item_category		There is a many-to-one relationship with commodities
Time	Time	The format is year/month/day/hour

Table 2: Partial data examples of dataset D.

Item_id	Behavior_type	Item_category	Time
275254735	1	4076	2019-12-08
436**9947	1	5503	2019-12-12
436**8907	1	5503	2019-12-12
536**6768	1	9762	2019-12-02
151**6952	1	5232	2019-12-12
536****76	4	9762	2019-12-02
2900**061	1	5503	2019-12-12
2983**524	1	10894	2019-12-12
3210**425	1	6513	2019-12-12

Table 3: Field description of dataset P.

Field	Meaning	Note
Item_id	Commodity identification	Be unique
Item_geohash	The product category ID of the longitude and latitude area encoded by geohash where the user is located	Encoded by the GeoHash algorithm
Item_category		There is a many-to-one relationship with commodities

Table 4: Partial data examples of dataset P.

Item_id	Item_category
100**2303	3368
100**3592	7995
100 **6838	12530
1000**089	7791
100**2750	9614
10****072	1032
100****63	9023
10****387	3064
10**23812	6700
1003****7	5827

Since the purpose of this paper is to predict the user's purchase behavior on e-commerce platforms to achieve accurate product recommendation, this paper focuses on the distribution of users' purchase behavior. Figure 3 shows the user's purchase behavior on the complete collection of goods, and Figure 4 shows the user's purchase behavior on the subset of goods. It can be seen from Figure 5 that the distribution of users' purchasing behaviors on the complete collection of goods is relatively stable, stable at about 7000. The number of purchases jumped to 30,000 on December 12th. It can be seen from Figure 5 that the fluctuation range of users' purchasing behavior distribution on the subset of goods is larger than that on the complete collection of goods. The purchasing behavior of users on November 22, November 28, and December 22 is quite different from the usual purchasing behavior, reaching more than 4000.

To further analyze the user behavior, this paper studies the user behavior distribution in 24 hours from the vertical time dimension. The user's behavior distribution in the complete collection of goods is shown in Figure 4, and the behavior distribution in the subset of goods is shown in Figure 6. It can be seen from the figure that the number of user behaviors is related to the user's daily life rules [18]. The number of user behaviors is less in rest time (00:00–08:00), more balanced in working time (09:00–18:00), and reaches the peak in the evening (19:00–23:00) during leisure time.

Figure 7 shows the distribution of user's purchasing behavior on the complete collection and subset of goods, respectively. It can be seen from Figure 7 that the users' purchase behaviors are relatively low during 16:00–18:00, and the number of purchase behaviors is generally flat between day and night. This shows that in the complete collection of goods, users have clear intention to purchase goods during working hours, and the purchase conversion rate is high. The purchase conversion rate is lower due to the longer time in the evening. As can be seen from Figure 8, in the subset of goods, the purchasing power of users is lower in

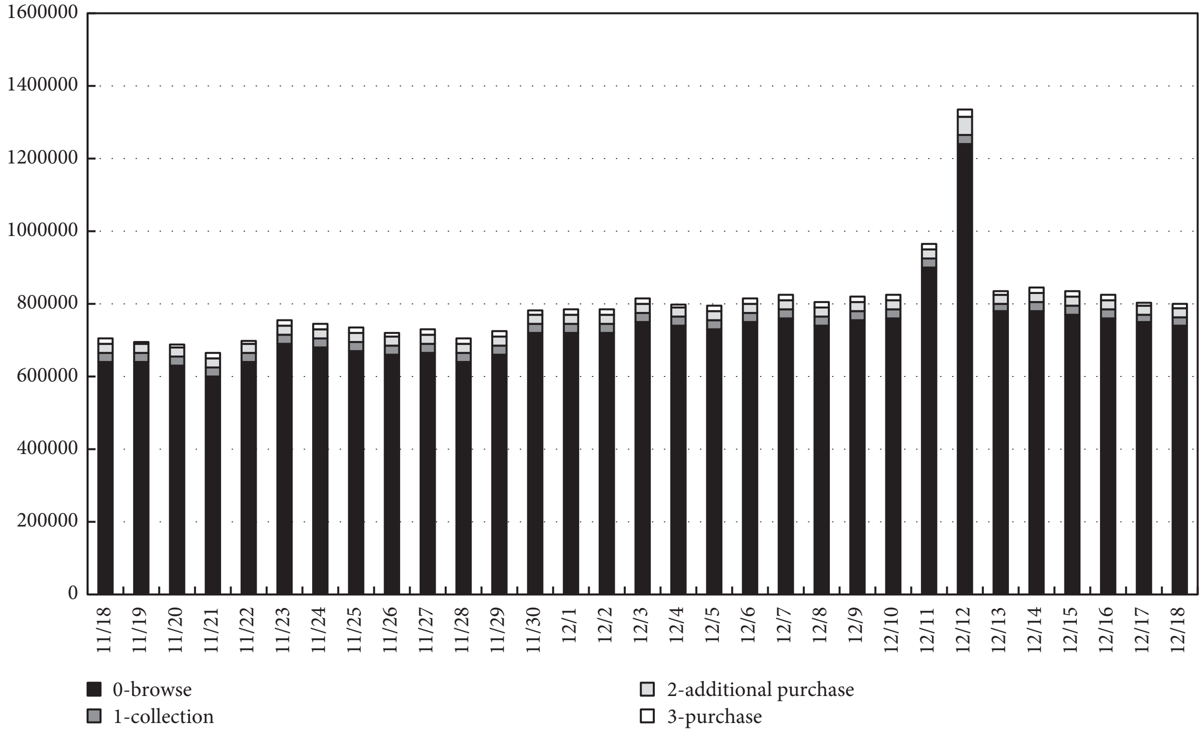

Figure 1: Distribution of user behavior in the complete collection of goods.

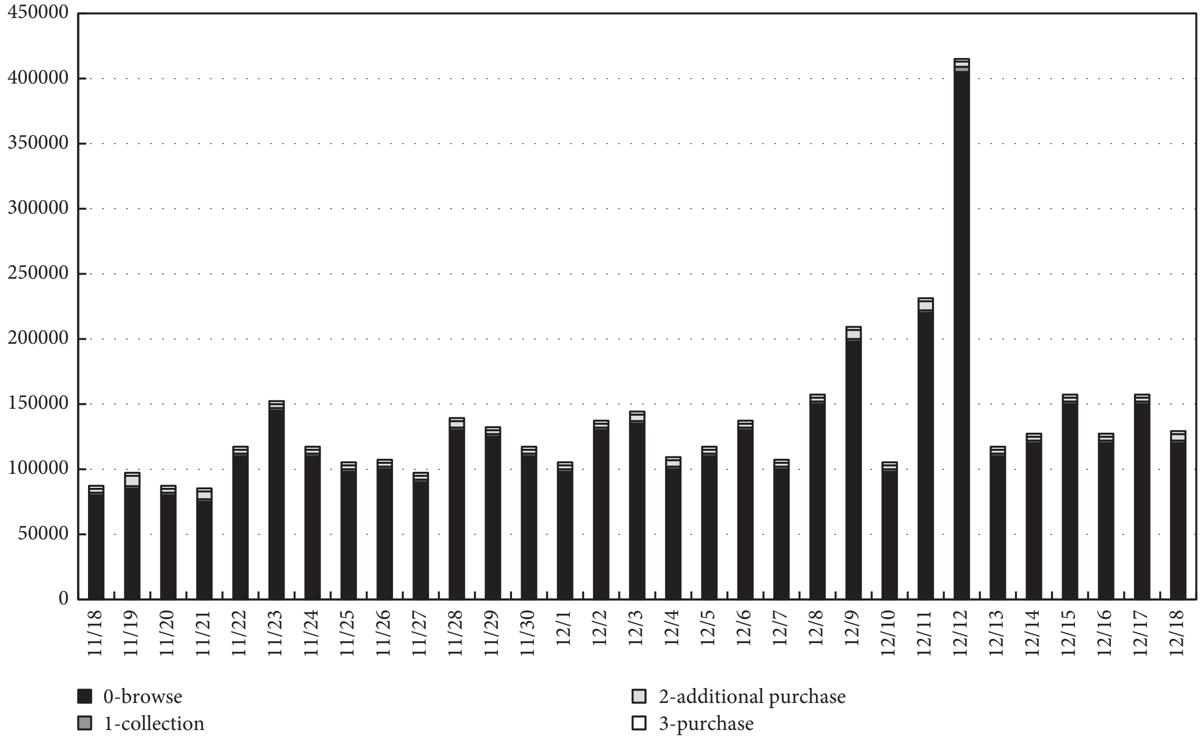

Figure 2: Distribution of user behavior in the subset of goods.

the daytime than in the evening, but it will reach a maximum point in the daytime. Compared with the complete collection of goods, the distribution of users' purchasing behavior on the subset of goods is not stable.

4.2.3. Data Processing. To better predict the user purchase data on December 19, this paper selects the purchase data on December 18, which is close to the date, as the basis and constructs the data characteristic cycle by analyzing the

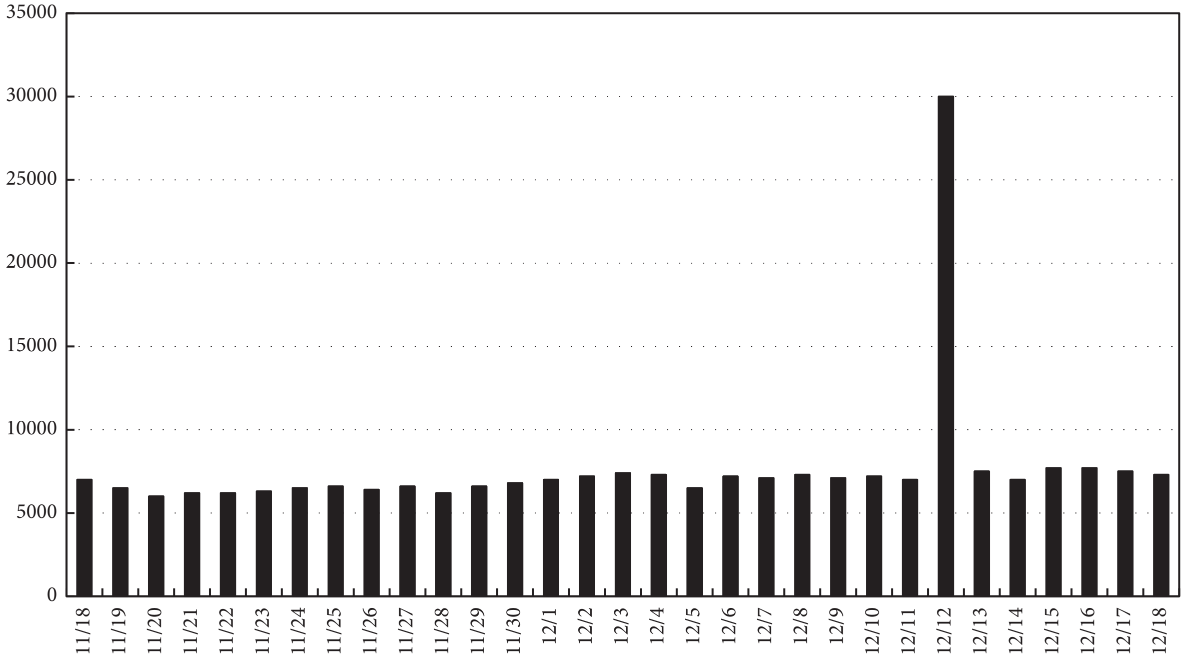

Figure 3: Distribution of user's purchasing behavior on the complete collection of goods.

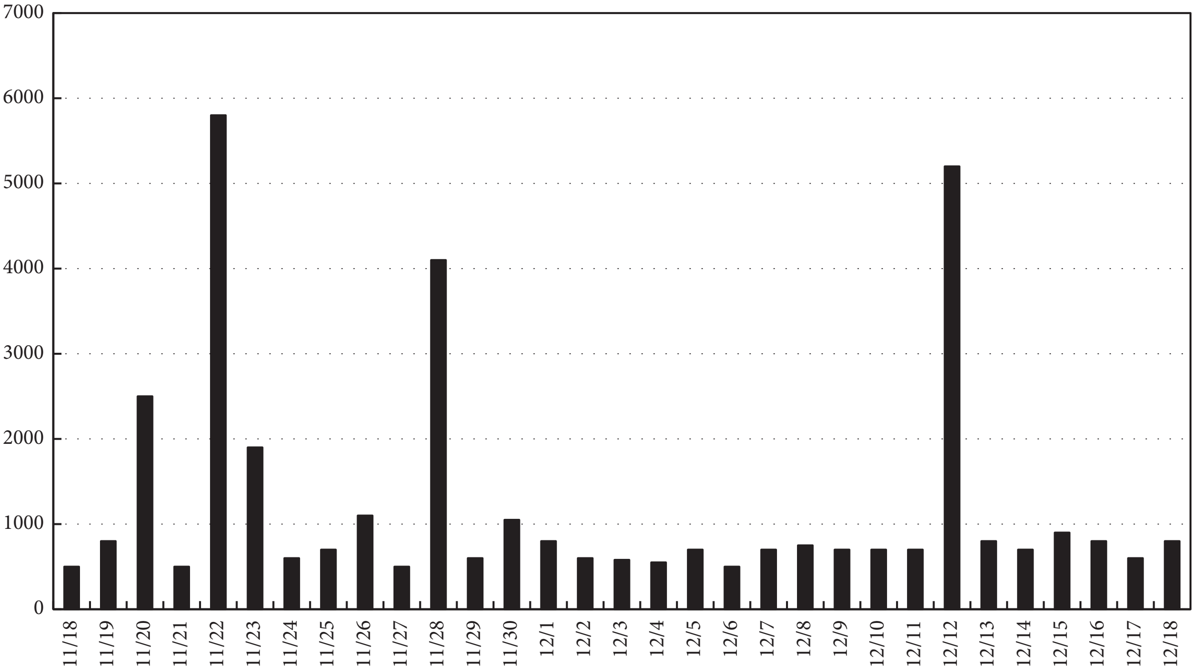

Figure 4: User's behavior distribution in the complete collection of goods (h).

interaction between the purchase behavior on December 18 and other dates. There are 6,925 items of purchase data on December 18, among which 4,662 items cannot be matched with the historical data of one month, and the remaining data can be matched with the historical data of one month. Figure 9 shows the historical data distribution that interacts with the behavior that existed on December 18. As can be seen from the figure, the number of interaction data increases significantly in the week before December 18 and reaches the maximum value on December 17. Therefore, this paper adopts the data of one week before the forecast date to predict it.

Considering that part of the user's purchase behavior on December 18 comes from the direct purchase on that day, there is no interaction with the previous period. There is no positive effect on the user's purchase behavior prediction. The previous browsing, collecting, additional purchasing, and purchasing behavior have a positive impact on the

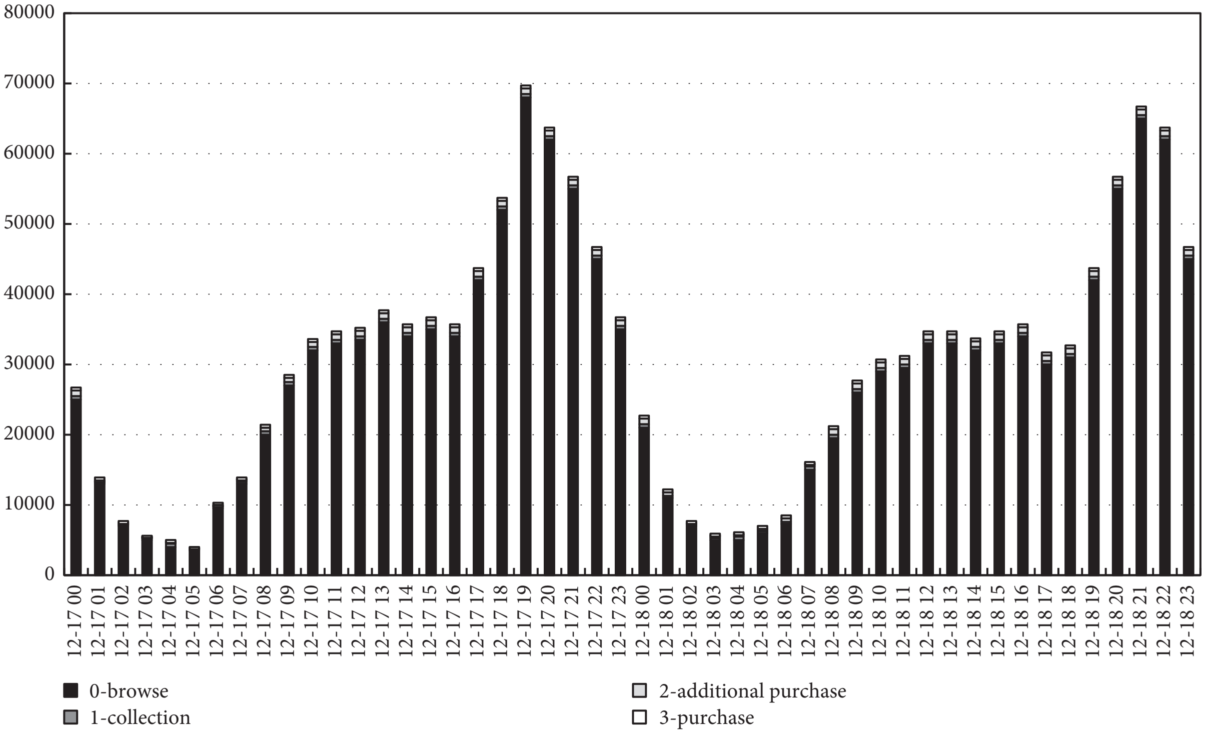

Figure 5: Distribution of user's purchasing behavior on the subset of goods.

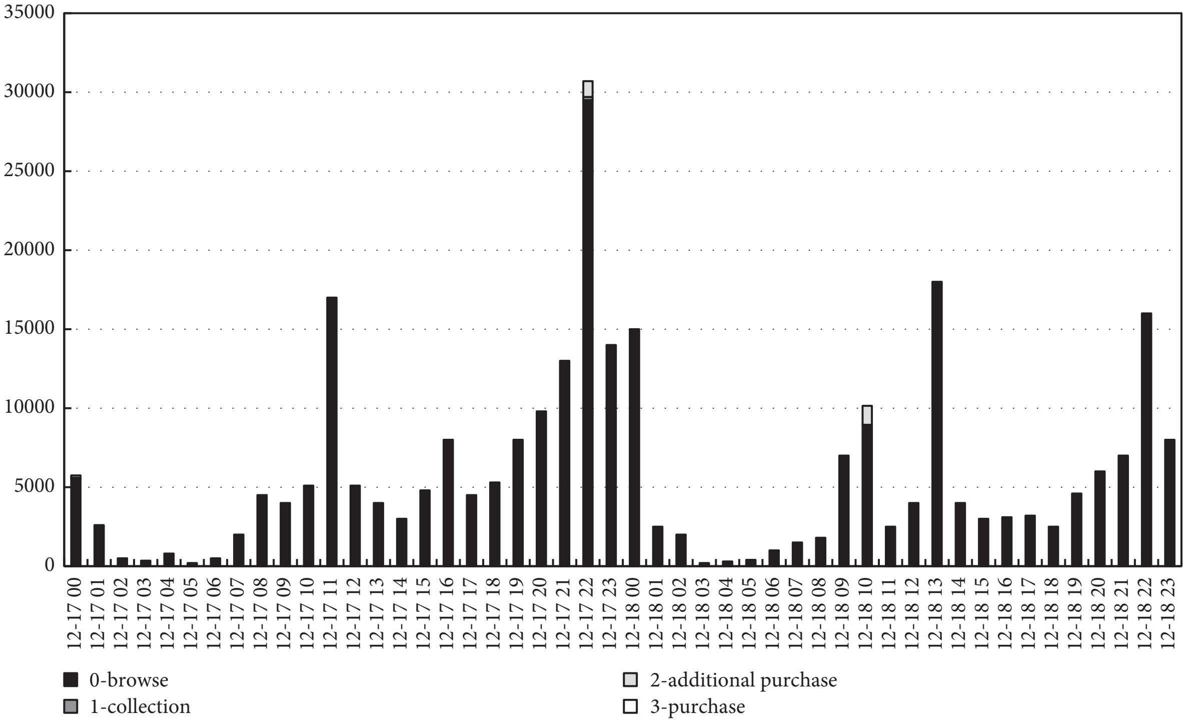

Figure 6: User's behavior distribution in the subset of goods (h).

prediction of the day. Therefore, this part is selected as the main target of this paper for prediction.

To sum up, this experiment divides the longitudinal dimension according to the data weeks, and the horizontal dimension takes Friday data as the goal to build the model. November 22nd to November 28th, November 29th to December 5th, and December 6th to December 12th are split into training sets. Meanwhile, December 13th to December 18th are split into test sets. Then, the problem is transformed into a binary classification problem by taking the user's

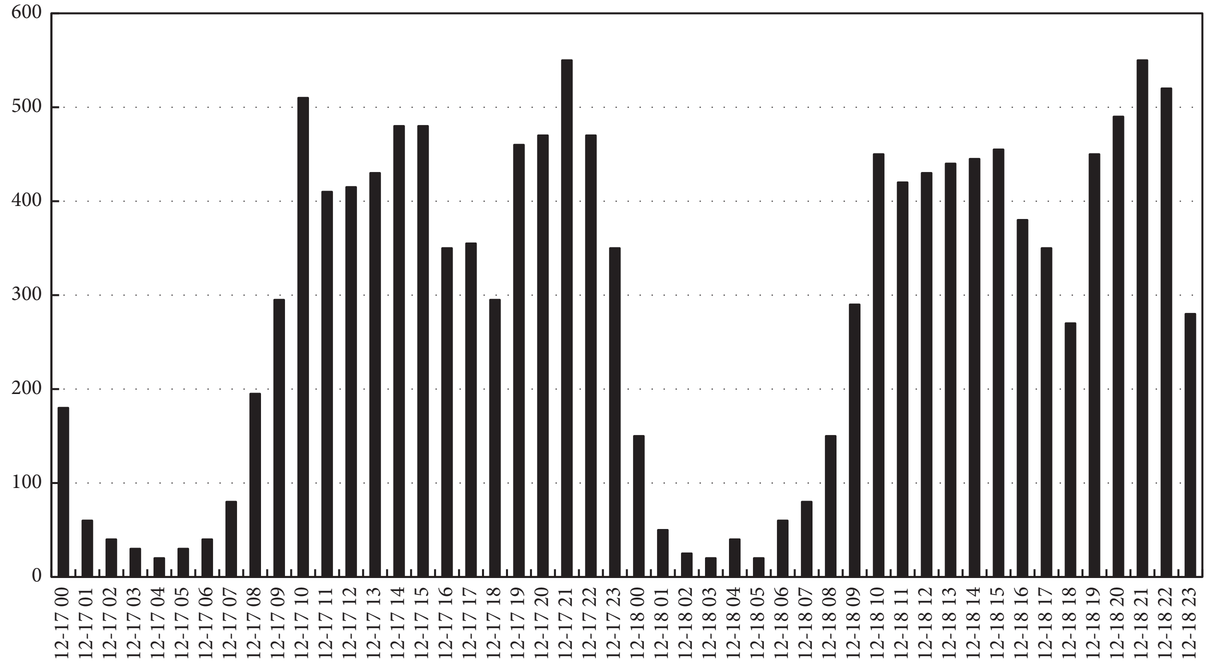

Figure 7: Distribution of user's purchasing behavior on the complete collection of goods (h).

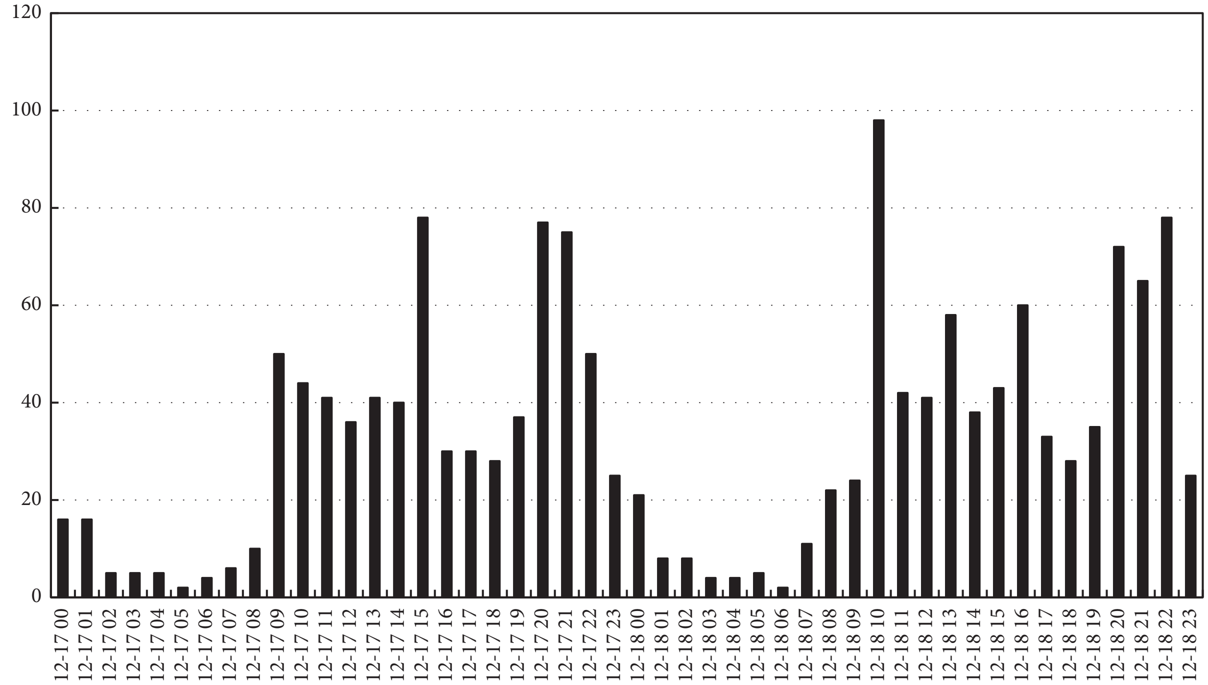

Figure 8: Distribution of user's purchasing behavior on a subset of goods (h).

browsing, collection, and additional purchase behavior as one category and the purchase behavior as another category. In addition, because the data on December 12 are obviously abnormal, they are deleted in order to avoid the influence of the data on the predicted results [19].

4.2.4. Feature Extraction. It is difficult to mine information from the existing feature dimensions because the dataset includes users, commodities, commodity categories, user behavior types, operation time, and other data [20]. Therefore, in order to better mine useful information from data, the 107 features of counting class, sorting class, time difference class, and conversion rate class are selected from the aspects, such as commodities, commodity categories, user-commodity interaction, user-commodity category interaction, and commodity-commodity category interaction, to construct the model [21, 22]. The characteristics of each category and their meanings are shown in Table 5. At the same time, according to the purchase situation of the

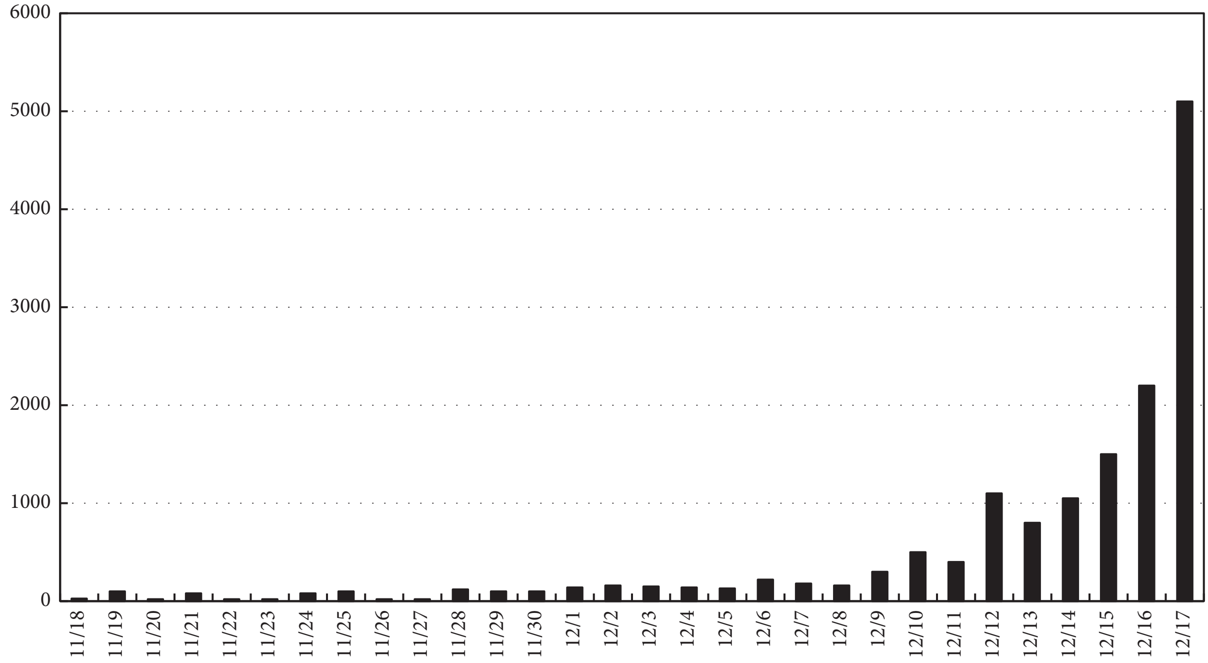

Figure 9: Number of interactions.

Table 5: Feature description.

Characteristic category	Characteristic meaning	Characteristic number
Counting	Number of behaviors of user/product/product category and interaction on 1/3/6 days prior to the inspection date	81
Sorting	Sorting of commodities and user-commodity interactions	12
Time difference	Average time difference between user/product/product category click purchase and interactive product behavior, and check the time difference of date	11
Conversion rate	User/product/product category click-to-buy conversion rate	3

corresponding group on the last day, the data are labeled as follows: 0 means no purchase and 1 means purchase.

4.3. Evaluation Indicators. The F1 value is adopted as an indicator to evaluate the e-commerce data analysis and prediction model, and its calculation method is as follows [23]:

$$\text{precision} = \frac{|\cap\,(\text{prediction set}, \text{reference set})|}{|\text{prediction set}|},$$
$$\text{recall} = \frac{|\cap\,(\text{prediction set}, \text{reference set})|}{|\text{reference set}|}, \tag{15}$$
$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$

where prediction set represents the predicted purchase data and reference set represents real purchase data.

4.4. Experimental Result

4.4.1. Parameter Setting. There are many parameters involved in the training process of GBDT model, and different parameters have some influence on model training and prediction. In order to determine the best model parameters, the control variable method is used to carry out experiments on the positive and negative sample ratio, learning rate, number of base learners, tree depth, and threshold that affect the model fitting results. However, because there is an extreme imbalance between positive and negative samples of e-commerce data and this factor has a great influence on the model fitting results, the positive and negative sample ratio needs to be determined first. Then, the learning rate of the model itself and the number of base learners are determined. Finally, the depth and threshold of the tree are determined.

Figure 10 shows the F1 value changes of the model under different ratios of negative samples and positive samples. As can be seen from the figure, when the ratio of negative samples to positive samples is less than 50, the F1 value increases gradually. When the ratio of negative samples to positive samples is between 50 and 100, the F1 value begins to fluctuate. When the ratio of negative samples to positive samples is greater than 100, the F1 value decreases gradually. The reason is that when the negative and positive samples are less than 50, the F1 value of the model is gradually increased as the number of iterations increases and the model underfitting is reduced. When the negative and positive

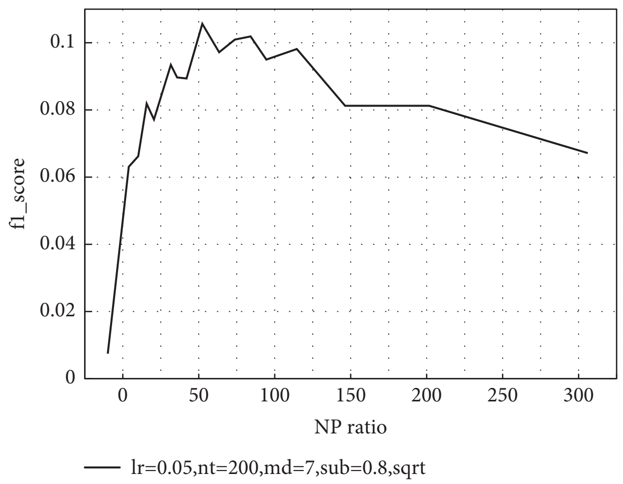

Figure 10: Influence of different negative sample/positive sample ratios on the model.

samples are larger than 100, the model is subject to overfitting, which leads to the reduction of its generalization ability. After comprehensive comparison, the negative and positive sample ratio of the input model selected in this paper is 60.

Figure 11 shows the influence of different learning rates on the predicted results of the model. As can be seen from the figure, different learning rates have different influences on the F1 value of the model. When the learning rate is 0.05, the F1 value of the model is the maximum, and with the increase of the learning rate, the F1 value of the model gradually decreases. Therefore, the learning rate of the proposed model is set at 0.05 in this experiment.

After determining the learning rate of 0.05, the number of base learners of the model is determined, and the results are shown in Figure 12. As can be seen from the figure, with the increase of number of base learners, the F1 value of the model begins to fluctuate and decrease after it starts to rise. When the number of base learners is 30, the F1 value of the model begins to fluctuate. When the number of base learners is 400, the F1 value of the model begins to decline. Therefore, it can be determined that the number of base learners of the model is between 30 and 400, and the median 200 is taken as the number of model base learners in this paper.

Figure 13 shows the influence of different tree depths on the model. It can be seen from the figure that with the increase of tree depth, the F1 value of the model rises first and then decreases. When the tree depth reaches a certain value, the F1 value of the model finally shows an upward trend. The final selected tree depth in this paper is 20.

Figure 14 shows the influence of different thresholds on the model. As can be seen from the figure, when the threshold is less than 0.4, the F1 value of the model rises gradually with the increase of the threshold. When the threshold value is between 0.4 and 09, the F1 value of the model fluctuates between 0.130 and 0.135. When the threshold value is greater than 0.9, the F1 value of the model decreases rapidly. Finally, the threshold of the model is determined to be 0.5.

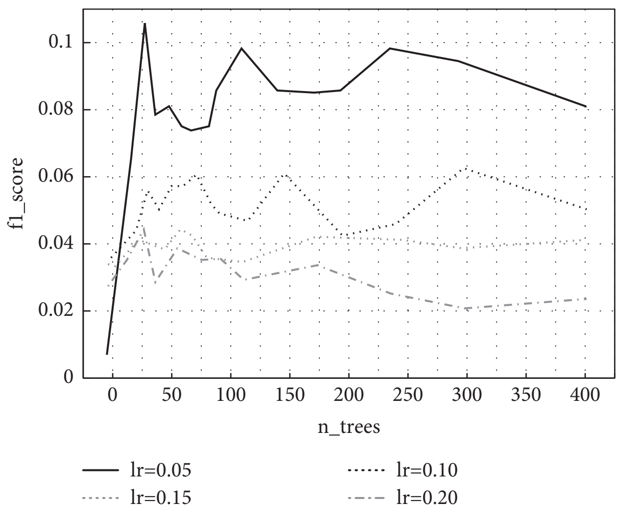

Figure 11: Influence of different learning rates on the model.

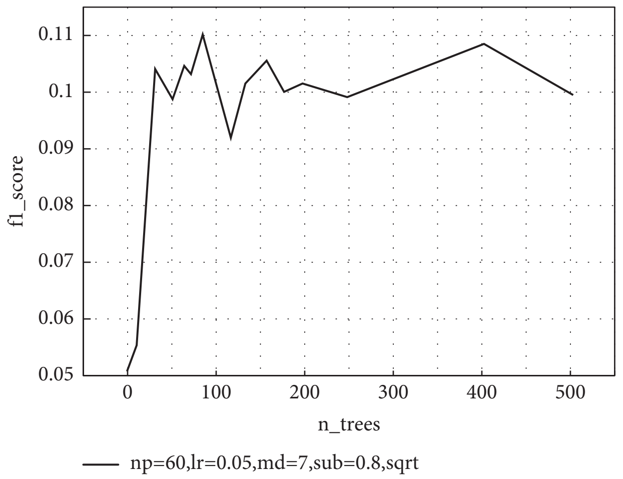

Figure 12: Influence of different number of base learners on the model.

In summary, the parameters of the proposed GBDT model are set as follows: learning rate = 0.05, base learner number = 200, tree depth = 20, and threshold = 0.5.

4.4.2. Comparison of Prediction Results. To further verify the effectiveness of the proposed model, the experiment compares the prediction effect of the proposed model with the traditional logic regression-based prediction model and the neural network-based prediction model. The GBDT model parameters are set according to the parameter setting. Meanwhile, the prediction model parameters based on logical regression of the comparison model are set to the threshold of 0.6. The prediction model parameters based on neural networks are set to the maximum number of iterations of 300. The single training samples are 64, and there are two layers of hidden layers. In addition, the number of nerve nodes per layer is 65.

The GBDT model and comparison model are trained on the training set with a ratio of negative samples to positive

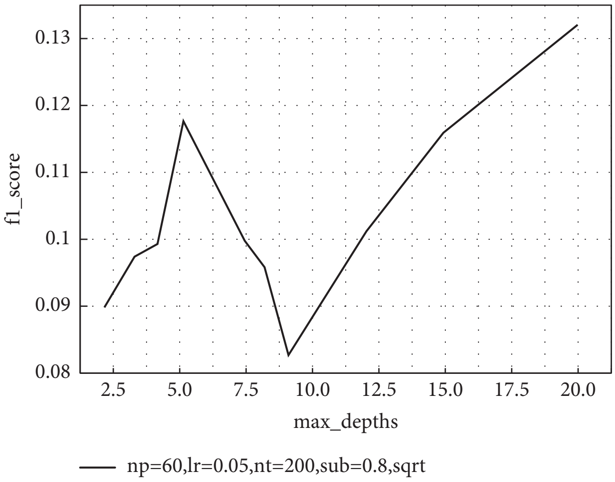

Figure 13: Influence of different tree depths on the model.

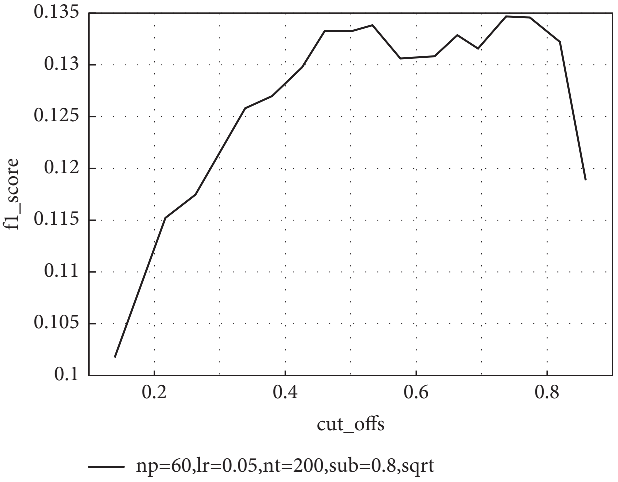

Figure 14: Influence of different model thresholds on the model.

Table 6: Comparison of prediction results.

Model	F1 score
Logistic regression	0.06
Neural network	0.06
GBDT	0.12

samples of 55, and the test set with a ratio of negative samples to positive samples of 60 is used for testing. The results are shown in Table 6. As can be seen from the table, compared with the comparison model, the proposed GBDT model has the highest F1 score of 0.12, which increases about 50%. Therefore, the proposed GBDT model is more suitable for data analysis and prediction in e-commerce.

5. Conclusion

To sum up, the e-commerce data analysis and prediction method based on GBDT model is proposed in this paper. It can be seen that the e-commerce data are preprocessed with missing values, desensitization, etc. At the same time, according to the user behavior, the browsing, collecting, and additional purchase behavior are divided into one category, and the purchase behavior is divided into another category. The problem of e-commerce data analysis and prediction is transformed into a binary classification problem. Then, a total of 107 characteristics of counting class, sorting class, time difference class, and conversion rate class that can reflect the users' behavior characteristics are extracted to build a GBDT model. Finally, the efficient analysis and prediction of e-commerce data are realized. Compared with the traditional prediction model based on logical regression and based on neural network, the proposed GBDT model is more suitable for e-commerce data analysis and prediction. Also, when the learning rate of the GBDT model is 0.05, the number of basic learners is 200, the tree depth is 20, and the threshold is 0.5, and the prediction effect of the proposed model is best. Meanwhile, the F1 value can reach 0.12. Although this paper has obtained some research results, there are still some deficiencies, which are that the random search method is adopted to tune the GBDT model parameter, whose time cost is higher, and the efficiency generally needs to be improved. Therefore, the model automatic tuning parameter can be adopted in the future to improve the efficiency of model training, so as to realize higher precision analysis and prediction of e-commerce data [24, 25].

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] B. E. Ozgur and D. Franklin, "Multicollinearity in logistic regression models[J]," *Anesthesia & Analgesia*, vol. 133, no. 2, pp. 362–365, 2021.

[2] A. C. Cioci, A. L. Cioci, A. M. A. Mantero, J. P. Parreco, D. D. Yeh, and R. Rattan, "Advanced statistics: multiple logistic regression, cox proportional hazards, and propensity scores," *Surgical Infections*, vol. 22, no. 6, pp. 604–610, 2021.

[3] D. Rekha, J. Sangeetha, and V. Ramaswamy, "Digital document analytics using logistic regressive and deep transition-based dependency parsing," *The Journal of Supercomputing*, vol. 78, pp. 1–17, 2021.

[4] J. Singh and A. Chhabra, "Indian stock markets data analysis and prediction using macroeconomics indictors in machine learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 10, pp. 484–486, 2020.

[5] F. Halawa, S. Al-Hihi, W. Shen, and D. Won, "A model-based approach of data analysis and prediction in chronic kidney diseases (CKD)," in *Proceedings of the 2017 Industrial and Systems Engineering Conference*, Pittsburgh, PA, USA, 2017.

[6] H. Son, S. Kim, H. Yeon, Y. Kim, Y. Jang, and S.-E. Kim, "Visual analysis of spatiotemporal data predictions with deep learning models," *Applied Sciences*, vol. 11, no. 13, 5853 pages, 2021.

[7] X.-B. Jin, J.-H. Zhang, T.-L. Su, Y.-T. Bai, J.-L. Kong, and X.-Y. Wang, "Modeling and analysis of data-driven systems through computational neuroscience wavelet-deep optimized model for nonlinear multicomponent data forecasting," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[8] Y. Guo, G. Xiong, L. Zeng, and Q. Li, "Modeling and predictive analysis of small internal leakage of hydraulic cylinder based on neural network," *Energies*, vol. 14, no. 9, 2456 pages, 2021.

[9] A. A. Agafonov, "Short-term traffic data forecasting: a deep learning approach," *Optical Memory & Neural Networks*, vol. 30, no. 1, pp. 1–10, 2021.

[10] S. Omer and R. Lior, "Approximating XGBoost with an interpretable decision tree," *Information Sciences*, vol. 572, pp. 522–542, 2021.

[11] Y. Zou, Y. Chen, and H. Deng, "Gradient boosting decision tree for lithology identification with well logs: a case study of zhaoxian gold deposit, shandong peninsula, China," *Natural Resources Research*, vol. 30, no. 5, pp. 1–21, 2021.

[12] T. Syed AsSadeq and Y. Chen, "Analysis of severe injuries in crashes involving large trucks using K-prototypes clustering-based GBDT model," *Safety Now*, vol. 7, no. 2, p. 32, 2021.

[13] W. Zhang, J. Yu, A. Zhao, and X. Zhou, "Predictive model of cooling load for ice storage air-conditioning system by using GBDT," *Energy Reports*, vol. 7, pp. 1588–1597, 2021.

[14] W. Qiu, Z. Lv, Y. Hong, J. Jia, and X. Xiao, "BOW-GBDT: a GBDT classifier combining with artificial neural network for identifying GPCR–drug interaction based on wordbook learning from sequences," *Frontiers in Cell and Developmental Biology*, vol. 8, 2021.

[15] B. Zhang, J. Ren, Y. Cheng, B. Wang, and Z. Wei, "Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm," *IEEE ACCESS*, vol. 7, pp. 32423–32433, 2019.

[16] X. Ye, J. Wang, T. Wang, X. Yan, Q. Ye, and J. Chen, "Short-term prediction of available parking space based on machine learning approaches," *IEEE ACCESS*, vol. 8, pp. 174530–174541, 2020.

[17] J. Yang, Y. Sheng, and J. Wang, "A GBDT-paralleled quadratic ensemble learning for intrusion detection system," *IEEE ACCESS*, vol. 8, pp. 175467–175482, 2020.

[18] G. Rong, S. Alu, K. Li et al., "Rainfall induced landslide susceptibility mapping based on bayesian optimized random forest and gradient boosting decision tree models—a case study of shuicheng county, China," *Water*, vol. 12, no. 11, 2020.

[19] J. Zhang, Q. Feng, X. Zhang, Q. Hu, J. Yang, and N. Wang, "A novel data-driven method to estimate methane adsorption isotherm on coals using the gradient boosting decision tree: a case study in the qinshui basin, China," *Energies*, vol. 13, no. 20, pp. 5369–5379, 2020.

[20] G. He, "Enterprise E-commerce marketing system based on big data methods of maintaining social relations in the process of E-commerce environmental commodity," *Journal of Organizational and End User Computing*, vol. 33, no. 6, pp. 1–16, 2021.

[21] Y. Kim, H. J. Lee, and J. Shim, "Developing data-conscious deep learning models for product classification[J]," *Applied Sciences*, vol. 11, no. 12, 2021.

[22] S. P. Goldman, H. van Herk, T. Verhagen, and J. W. Weltevreden, "Strategic orientations and digital marketing tactics in cross-border e-commerce: comparing developed and emerging markets," *International Small Business Journal: Researching Entrepreneurship*, vol. 39, no. 4, pp. 350–371, 2021.

[23] M. Zhu, "Implementation of support-vector machine algorithm to develop a model for electronic commerce energy regulatory system," *Energy Reports*, vol. 7, pp. 2703–2710, 2021.

[24] N. Sharma, A. Raj, V. Kesireddy, and P. Akunuri, "Machine learning implementation in electronic commerce for churn prediction of end user," *International Journal of Soft Computing and Engineering*, vol. 10, no. 5, pp. 20–25, 2021.

[25] J. Hou, Q. Li, Y. Liu, and S. Zhang, "An enhanced cascading model for E-commerce consumer credit default prediction," *Journal of Organizational and End User Computing*, vol. 33, no. 6, pp. 1–18, 2021.