Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Research on Prediction of E-commerce Users' Purchase Behavior Based on Gradient Boosting Decision Tree Model

Wenrui Xu[1,✉]

[1] *Guangdong Polytechnic of Science and Technology, Dongguan, Guangdong, 523000, China*

ABSTRACT

Driven by big data, e-commerce platforms have accumulated massive user behavior data, which can be transformed into valuable information after cleaning and feature selection. This study analyzes users' historical behavioral data on e-commerce platforms, constructs a gradient boosting decision tree prediction model based on user, product, category, and two-by-two interaction behavioral features, and extracts designed feature data from raw CSV data based on Hiv as the prediction basis of the model. At the same time, clustering analysis is performed based on the user's purchasing behavior (dwell time, browsing frequency) to generate user profiles. The experimental results show that after 7 days, the purchase conversion rate of browsing, collecting, adding to cart and purchasing tends to 0. Therefore, the time window for purchase behavior prediction is chosen to be 7 days. In this paper, the prediction model is only trained to 20 epochs, and the Loss value converges to about 0.14, which shows a good training effect. The model has the best classification performance for user purchase behavior prediction, with precision, recall, and F1 values between 0.91 and 0.97. The clustering algorithm divides the user purchase behavior into four clusters, where cluster class 4 has the best user value. In summary, using the gradient boosting decision tree model, e-commerce platforms can more accurately predict user purchasing behavior, thus improving user experience and platform economic benefits.

Keywords: gradient boosting decision tree; cluster analysis; feature design extraction; purchase behavior prediction; e-commerce platforms

1. Introduction

Gradient boosted decision tree (GBDT) is an integrated learning algorithm based on decision trees. It has strong performance in both classification and regression problems and has achieved good results in many machine learning competitions. The GBDT model uses serialization to construct multiple weak classifiers and combines them to form a single strong classifier [1-4]. The core idea of the GBDT model is to iteratively train a series of decision tree models, where each tree tries to correct the errors of the previous tree. The output of the whole model is a weighted sum of all the weak classifiers. The GBDT model has

✉ Corresponding author.
E-mail address: xuwrr001@163.com (W. Xu).
Received 15 April 2024; Revised 19 August 2024; Accepted 27 October 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-114

been very successful in practical applications and has potential applications in predicting the purchasing behavior of e-commerce users [5-8].

With the development of e-commerce, the shopping habits and behaviors of consumers have changed a lot. In this case, analyzing and predicting consumers' buying behavior is crucial for the management and development of e-commerce platforms. The analysis and prediction of e-commerce users' purchasing behavior is an important part of e-commerce enterprises' decision-making [9-12]. Through the application of data mining, big data analysis, social media analysis and other methods, we can better understand the user's purchase behavior and demand. At the same time, through collaborative filtering-based recommender systems, time series-based prediction methods and other technologies, it is possible to predict users' purchasing behavior and improve sales and customer satisfaction [13-16]. However, e-commerce platforms need to pay attention to data quality and privacy protection when analyzing and predicting users' purchasing behaviors, and analyze and apply them in combination with specific business scenarios [17-18].

Literature [19] proposes an e-commerce data analysis and prediction method based on GBDT deep learning model. Data analysis and prediction is achieved by extracting features of user behavior and creating GBDT models. The results pointed out that GBDT model performs better for data analysis and prediction in e-commerce compared to traditional prediction models. Literature [20] is based on user-product interaction in order to predict user's consumption behavior. Using feature fusion to construct strong features and using Random Forest to filter features, a fusion algorithm based on XGBoost and LightGBM for user purchase behavior prediction is proposed and compared with the prediction effect of Random Forest and GBDT algorithms. The results show that the fusion model has better prediction effect, and merchants can carry out marketing activities based on the prediction results. Literature [21] aims to analyze the strengths and weaknesses of supervised learning algorithms and integrated learning algorithms, as well as their application and performance in predicting user purchasing behavior. The strengths and weaknesses of different algorithms in prediction are emphasized, while integrating models with complementary strengths can improve the efficiency and accuracy of the integrated approach. Literature [22] indicated that purchase prediction helps to improve user experience and marketing promotion effectiveness. An algorithm for user purchase prediction was built based on feature engineering and model construction, and a two-layer fusion model was designed using a superposition-based multi-model fusion approach. The experimental results reveal that the accuracy and stability of the model are better than the benchmark model. An improved GBDT-NN was proposed in the literature [23]. The features and feature combinations are screened by GBDT training, and the screening results are processed using artificial neural networks, and the experimental results emphasize the very high accuracy in the GBDT-NN test set. Literature [24] emphasizes the importance of predicting consumer behavior in e-commerce. Combining artificially engineered features with location information and applying feature sparsification, the performance of models such as Random Forest and GBDT was compared using a large-scale logistic regression model. The results confirm the effectiveness of logistic regression models combining artificial cross-cutting features and location information in consumer behavior prediction this. Literature [25] proposes a SE-stacking model based on information fusion and integrated learning to predict users' purchasing behavior. By using the stacking algorithm to predict users' purchasing behavior, the results validate the effectiveness of SE-stacking, and the model is important for academic research and development in this field. Literature [26] proposes a method for predicting the purchase behavior of O2O goods based on user behavior logs. And based on the data provided by Ali Mobile Recommendation Contest, the purchase behavior of goods was predicted, and the results emphasized that the proposed prediction method has better results. Literature [27] proposed an online store behavior analysis and prediction system based on the shortcomings of traditional prediction methods. Experiments were carried out by using an optimized XGBoost model, and the results specified that the XGBoost model was improved over a single model. And based on comparative experiments, it is proved that the chosen algorithm can filter the features effectively. Literature [28] emphasizes the importance of accurately predicting e-commerce

users' purchasing behavior. The shortcomings of current prediction methods are pointed out to propose an integrated learning prediction model based on the fusion of information from multiple sources. Tests on the Tmall dataset show that the model's exhibit very high accuracy. Literature [29] explored user purchase behavior based on datasets provided by online distribution platforms and Alibaba. And based on the GBDT model, machine learning and training were implemented both online and offline using ODPS and Python, and user purchase behavior was predicted by combining user behavior sequences. Literature [30] is based on O2O scene related data. Through feature engineering, features are extracted from users, merchants, and coupons to generate a dataset and used for XGBoost model training. The model is optimized through parameter adjustment and feature importance evaluation and selection, which achieves effective prediction of coupon usage by users at a specified time.

In this study, we mined the historical data of user purchasing behavior (dwell time, browsing frequency, etc.) in recent years in Ali e-commerce platform, and carried out data preprocessing through data integration and cleaning to improve the data quality. Behavioral features including user, commodity, category, user-commodity-category, user-category in the processed data are designed and extracted to provide strong support for this paper to construct a gradient boosting decision tree model based on the prediction of user purchasing behavior. In order to further construct the user portrait in depth, this paper uses k-means clustering algorithm to divide the class groups according to the user's purchasing behavior and evaluate the purchasing value of the users in different class groups.

2. User purchase data mining processing and feature extraction

2.1. *Data sources*

The raw data studied in this paper is user behavior data randomly selected from the entire mobile terminal users of Ali, involving a total of 20,000 user actions on a specific set of commodities, with 22 million records, with a data period between June 10, 2023 and July 10, 2023, and a number of commodities generating interactions with the sampled users of 3,460,000 during that period, with a total of attributed to 10,365 commodity categories.

2.2. *Data pre-processing*

After getting the data, data preprocessing is usually needed to process the raw data into high-quality data suitable for data mining through operations such as data integration, transformation, missing value processing, and standardization. This section focuses on data integration and cleaning based on the characteristics of the dataset provided by Ali Group.

2.2.1. Data integration. In the field of electronic commerce, given its huge amount of data and the complexity of data, in actual production, data are often categorized according to their characteristics and uses, and then stored in different database tables, in order to reduce the time required for storing and extracting the data. HIVESQL tools are usually used in enterprises to extract the required data from the system database, and at the same time, the data set will be made into a new temporary data table, to facilitate subsequent data conversion and feature extraction. In this experiment, Python is utilized for data integration.

2.2.2. Data cleansing. For data under different scenarios and applications, the way and process of data cleansing are different, and there is no set of usual processing flow. The data cleaning process in general is: remove unique identifiers, handle missing values, identify anomalies, and standardize data.

There are three types of commonly used methods to deal with missing values in general, the first type is to do no processing, directly input into the model training, mainly for the feature under the missing data accounted for a relatively small number of cases. The second category is to delete the feature with missing data, mainly for the case where the missing data account for a large percentage and contain less valid

information, resulting in a weaker contribution of the feature to the model. The third category is to fill the missing data, and the methods of filling are mean/like mean filling, predictive value filling, high-dimensional mapping, multiple interpolation, and manual interpolation.

2.3. Feature design and extraction

2.3.1. Characterization. The object record studied in this paper involves three basic attributes of users, commodities and their categories, and the first thing that comes to mind is to design relevant features based on these three. In addition to the single dimension, the composite features should also be taken into account. From the two-by-two interaction, the composite features of user-commodity, user-commodity category and commodity-commodity category can be derived, but taking into account the inherent subordinate relationship between commodity and commodity category, it is therefore fused with the user-commodity interaction features to obtain the user-commodity-commodity category features. Among them, the interaction between users and commodities and the interaction between users and commodity categories reflect the degree of user preference for a certain commodity or a certain category of commodities, which can be simply understood as the higher value of the interaction statistics of the user's commodity has a higher likelihood of success for the eventual purchase, and get the feature division of the dimensions of users, commodities, and categories as shown in Figure 1.

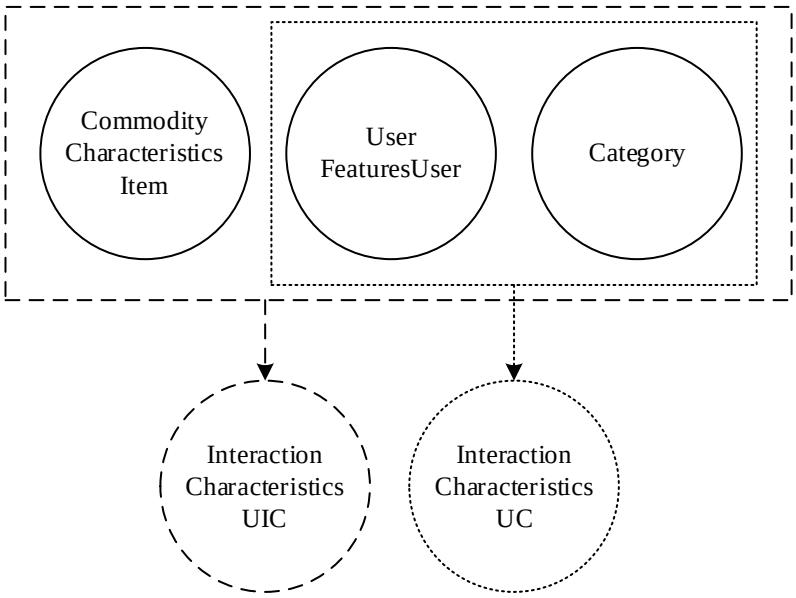

Fig. 1. The characteristics of users, goods and category dimensions are divided

Figure 2 shows all the conversion rate metrics related to purchase behavior. In order to avoid the previous part of data preprocessing found that the user may have purchased but did not generate the behavior of collection or add cart, this paper in the collection and add cart to purchase conversion formula in the denominator for special treatment, to avoid the abnormal situation of the divisor is zero. The approach taken here is to ensure that the denominator is greater than zero by adding a positive number to the denominator. Finally, all behaviors are summed together to form a behavior-to-purchase ratio indicator to reflect the impact of all behaviors on the final purchase decision.

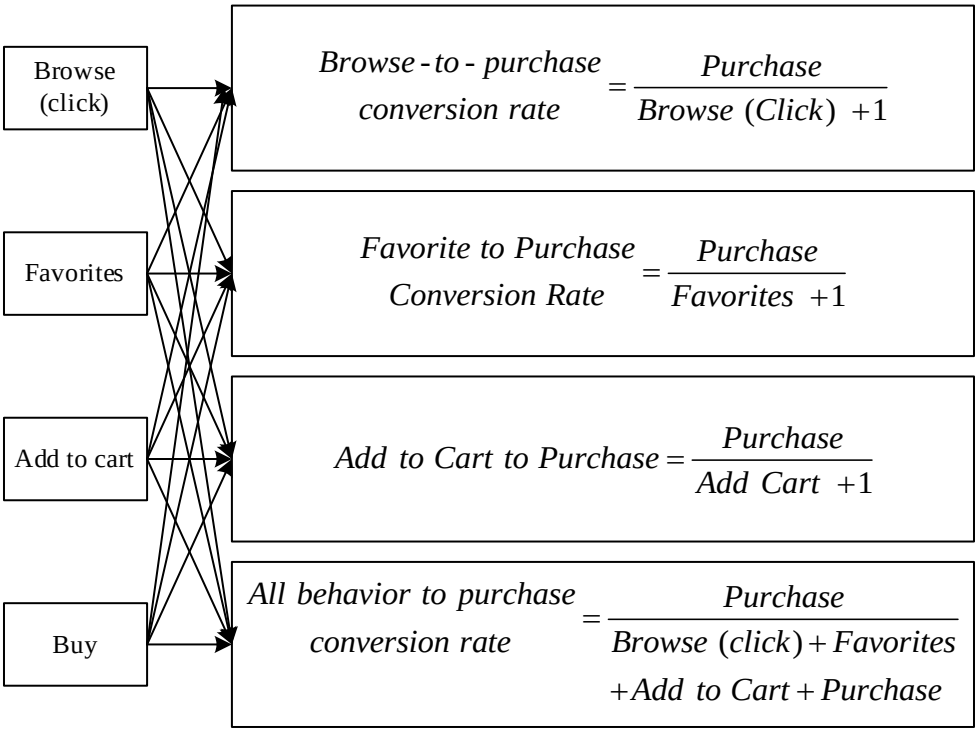

Fig. 2. Based on behavioral design

Through the above division from the three angles of attributes, behavior and time, the feature variables used for purchase prediction are finally obtained as input data, of which the attributes include the user, product, product category and two interaction features, while the behavior includes four ratio data with the goal of purchasing in addition to the four types of behavioral data obtained from the direct statistics, of which three are the conversion rate data involved in the funnel analysis, while the time is Extracted in accordance with 1, 3, 7 days in advance, to maximize the reflection of the nature of the features at the same time also saves unnecessary dimensions, and ultimately get the user, product, product category and two interaction all the characteristics of the variables as shown in Table 1.

Table 1. Users, commodities, commodity categories and two interaction characteristics

1/3/7	u	i	c	uic	uc
click	u_click1	i_click1	c_click1	uic_click1	uc_click1
	u_click3	i_click3	c_click3	uic_click3	uc_click3
	u_click7	i_click7	c_click7	uic_click7	uc_click7
fav	u_fav1	i_fav1	c_fav1	uic_fav1	uc_fav1
	u_fav3	i_fav3	c_fav3	uic_fav3	uc_fav3
	u_fav7	i_fav7	c_fav7	uic_fav7	uc_fav7
add	u_add1	i_add1	c_add1	uic_add1	uc_add1
	u_add3	i_add3	c_add3	uic_add3	uc_add3
	u_add7	i_add7	c_add7	uic_add7	uc_add7
buy	u_buy1	i_buy1	c_buy1	uic_buy1	uc_buy1
	u_buy3	i_buy3	c_buy3	uic_buy3	uc_buy3
	u_buy7	i_buy7	c_buy7	uic_buy7	uc_buy7
bcr	u_bcr1	i_bcr1	c_bcr1	uic_bcr1	uc_bcr1
	u_bcr3	i_bcr3	c_bcr3	uic_bcr3	uc_bcr3
	u_bcr7	i_bcr7	c_bcr7	uic_bcr7	uc_bcr7
bfr	u_bfr1	i_bfr1	c_bfr1	uic_bfr1	uc_bfr1
	u_bfr3	i_bfr3	c_bfr3	uic_bfr3	uc_bfr3
	u_bfr7	i_bfr7	c_bfr7	uic_bfr7	uc_bfr7
bar	u_bar1	i_bar1	c_bar1	uic_bar1	uc_bar1
	u_bar3	i_bar3	c_bar3	uic_bar3	uc_bar3
	u_bar7	i_bar7	c_bar7	uic_bar7	uc_bar7
bsr	u_bsr1	i_bsr1	c_bsr1	uic_bsr1	uc_bsr1
	u_bsr3	i_bsr3	c_bsr3	uic_bsr3	uc_bsr3
	u_bsr7	i_bsr7	c_bsr7	uic_bsr7	uc_bsr7

2.3.2. Feature extraction. After completing the feature design, we proceeded to the extraction of the features, utilizing Hive to manipulate the raw data table. After launching into the Hive program through the SecureCRT client, the original CSV file is uploaded first, the original data table is established based on the data within the CSV file, and the intermediate tables are established on the basis of the original table with the user, goods, category, user-goods-category, user-category, where each table extracts the feature data related to the behavior with a different period of time, and then the longer table is merged with the shorter table to obtain each intermediate table containing the complete behavioral features and temporal features, and then the longer table and the shorter table are merged. The longer table is merged with the shorter time table to get each intermediate table contains complete behavioral and temporal features, and then the five tables are merged as the basis of feature data through user-goods-category, and finally the four intermediate tables of user, goods, category and user-category are merged on the basis of user-goods-category, which is then combined with the purchasing data on the day of the prediction to get the final total table.

2.4. Purchase Conversion Rates for Four Behaviors

This section evaluates the purchase conversion rate for the four behaviors described above, i.e., browsing, bookmarking, adding to cart, and purchasing, with the aim of supporting the accurate prediction of user purchasing behavior below. Figure 3 shows the distribution of purchase conversion rates over time for the four behaviors.

By observing the dredging in the figure, we find that the purchase conversion rate of this behavior is much higher than the other behaviors, and the purchase conversion rate on the first day is higher than the

rest of the three behaviors by 3.58% to 9.80%, and the rate of increase continues to expand with the passage of time before the 4th day. The purchase conversion rate of all four behaviors showed a significant decline inflection point after 4 days, and the purchase conversion rate of the four behaviors gradually decreased over time and converged to 0 after 7 days, which means that the user's interaction with the product is difficult to be converted into a purchase on the day of the inspection after 7 days. Therefore, the time window for dividing the training set and test set of purchase behavior prediction is chosen as 7 days.

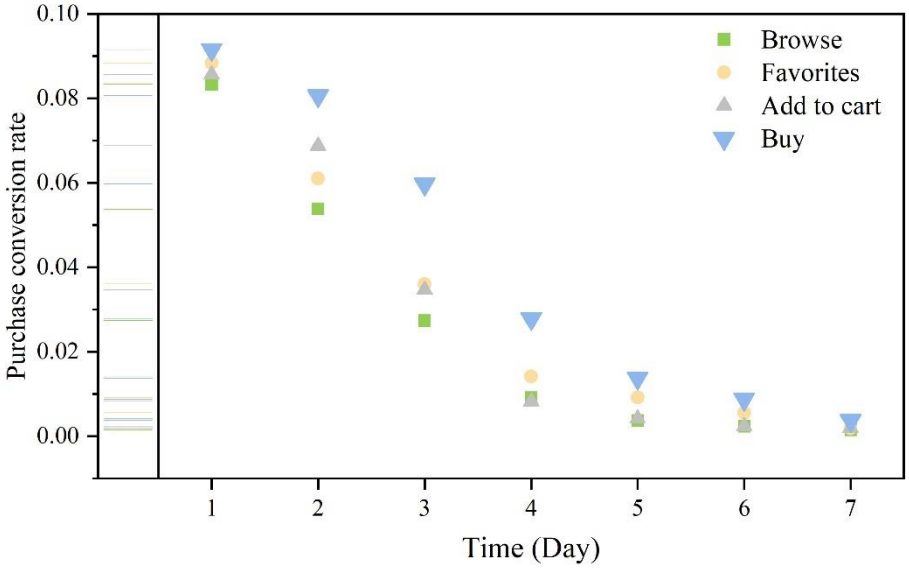

Fig. 3. The four behaviors purchase the conversion rate in time

3. Gradient boosting decision tree algorithm based on e-commerce users' purchasing behavior

The data features in this paper are characterized by large data volume and high dimensionality, while Gradient Boosted Decision Tree (GBDT) is characterized by strong generalization ability, prevention of overfitting situations in decision tree based processing, and high algorithmic efficiency.

3.1. Decision trees

There are three commonly used methods for generating decision trees [31-32], which are ID3, C4.5, and CART. Here, the C4.5 algorithm is used to construct a decision tree, and the decision tree construction process and related knowledge are introduced. The C4.5 algorithm is an improved ID3 algorithm, which employs the information gain ratio to divide the attribute values and realize the branching of the decision tree. Let D denote the division of training data with attribute categories, then the entropy of D is expressed as:

$$\inf o(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

In Equation (1), p_i represents the probability that the i nd attribute category occurs in the whole training data, which can be estimated by the ratio of the number of elements of this attribute category to the total number of elements of the training data. Assuming that the training data D is divided by attribute A, then the expected information of A to D is:

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j) \tag{2}$$

The information gain is the difference between the above entropy and the desired information:

$$gain(A) = info(D) - info_A(D) \tag{3}$$

Because the information gain will favor the selection of attributes with more attribute values when selecting attributes, the C4.5 algorithm uses the information gain ratio to divide the features, and first defines the split information expectation formula as follows:

$$split\_info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{D_j|}{|D|}\right) \tag{4}$$

Then, the information gain ratio is defined as the ratio of the information gain to the splitting expectation with the following formula:

$$gain\_ratio(A) = \frac{gain(A)}{split\_info(A)} \tag{5}$$

The GBDT algorithm is an algorithm that combines a large number of trained decision trees to generate predictions, and it incorporates the idea of boosting on the trained decision trees to combine them into the GBDT algorithm.

3.2. Gradient boosting decision tree

3.2.1. boost thought training process. Gradient boosting method is iterating multiple trees to make a joint decision, it is a method based on boosting idea. boosting boosting idea is actually easy to understand, it is an iterative process. For a training data, it is modeled using multiple weak learners like decision trees, where each tree is modeled to improve the result of the previous tree.The boosting idea training process can be depicted in figure 4.

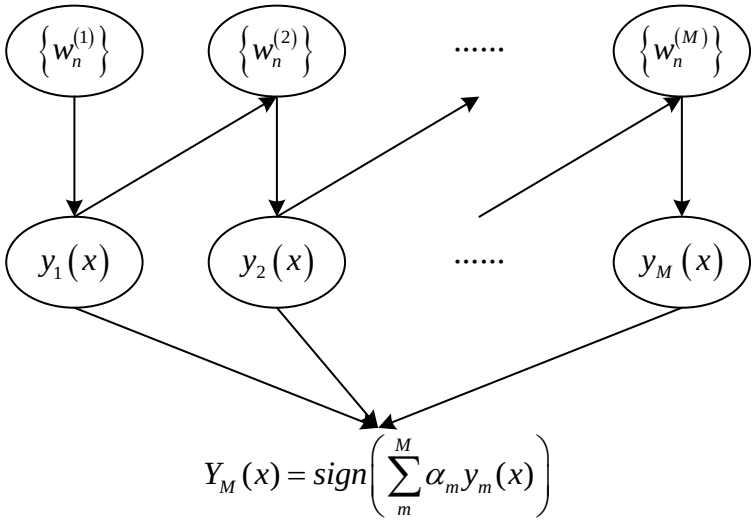

Fig. 4. Boosting training process

As the training continues, the model pays more attention to the points that can be easily misclassified, and keeps on revising the weight values to classify this attribute, improving the results of the last time, and finally combining the weighted combination of these classifiers to get the decision result.The weighted combination of the classifiers of the boosting method can be expressed by the following equation:

$$f(x) = x_0 + \sum_{m=1}^{M} w_m \phi_m(x) \tag{6}$$

where r_m denotes the weight of the m nd classifier and $\phi(x)$ denotes the base classifier.

3.2.2. Loss function. Loss function is used to describe the loss of the system under different parameter values, which can assist us to continuously reduce the variation of the target value through the improvement of the process. In the construction of the algorithm, the loss function is used to describe the credibility of the constructed model, the smaller the value of the loss function, the more reliable the constructed model is, the higher the credibility and the better the robustness.Gradient boosting is not a simple weighted combination of correctly and incorrectly categorized samples, but rather, it is a process of iterative training, and at each time of constructing the model, the gradient of the loss function of the previously constructed model decreases by a factor of 1 to 1 in the direction of constructing a new model. Instead, through iterative training, each time the model is constructed, the gradient of the loss function of the previously constructed model decreases in the direction of constructing a new model, reducing the residuals of the previous model to ensure that the final result is optimal.

The goal of model building is to require that the loss function of the model is constantly decreasing, which indicates that the model has been continuously improved, and the best improvement is to make the model loss function decrease in the direction of its gradient. The loss function in the Gradient boosting algorithm calculates the gradient of the loss function each time the model is trained, and determines the direction in which the training should proceed based on the gradient decrease (i.e., the loss function decreases rapidly).

3.2.3. Forward stepping algorithm. The Gradient boosting algorithm is an additive model based on the underlying learner that utilizes a forward stepping algorithm to implement the training process for learning. The forward stepping algorithm is described as follows:

The inputs to the algorithm include: training dataset $T=\{(x_1,y_1),(x_2,y_3),\ldots,(x_N,y_N)\}$. Loss function $L(y,f(x))$. Base learner set $\{b(x,\gamma)\}$. Output is the algorithm model $f(x)$. The algorithm starts by first initializing the model $f_0(x)=0$, and then for $m=1,2,\ldots,M$. Computes the loss function for the minimization:

$$(\beta_m,\gamma_m)=\arg\min_{\beta\cdot\gamma}\sum_{i=1}^{N}L\left(y_i,f_{m-1}(x_i)+\beta b(x_i;\gamma)\right) \tag{7}$$

The coefficients β_m of the parametric base learner and its parameters γ_m are obtained and the obtained parameters are used to update the next model:

$$f_m(x)=f_{m-1}(x)+\beta_m b(x;\gamma_m) \tag{8}$$

Finally the algorithmic model is obtained by weighting these models of training using an additive model:

$$f(x)=f_M(x)=\sum_{m=1}^{M}\beta_m b(x;\gamma_m) \tag{9}$$

The construction of the above algorithmic model shows that given the training data and loss function $L(y,f(x))$, the learning algorithm model $f(x)$ becomes an empirical risk minimization, i.e., loss function minimization problem, with an optimization objective function Eq:

$$\min_{\beta_m\cdot\gamma_m}\sum_{i=1}^{N}L\left(y_i,\sum_{m=1}^{M}\beta_m b(x_i;\gamma_m)\right) \tag{10}$$

Whereas the forward stepwise algorithm makes the complexity of the algorithm reduced by simplifying the complexity of the optimization at each step, specifically, only the following loss function needs to be optimized at each step:

$$\min_{\beta,\gamma}\sum_{i=1}^{N}L\left(y_i,\beta b(x_i;\gamma)\right) \tag{11}$$

3.2.4. Gradient Boosting Decision Tree Algorithm. Gradient boosting decision tree algorithm [33] (GBDT) is an algorithm that uses a decision tree as a base learner on the Gradient boosting framework, assuming that the algorithm model is represented by the following function:

$$F(x;P)=F\left(x;\{\beta_m,\alpha_m\}_1^M\right)=\sum_{m=1}^{M}\beta_m h\left(x;\alpha_m\right) \tag{12}$$

where P denotes the parameters, there may be multiple parameters that make up $P=\{p_0,p_1,p_2\ldots\}$, i.e., there are multiple different $\{\beta,\alpha\}$, $F(x;P)$ denotes the model function, the whole model is added up from multiple base learner models, β denotes the weights of each model, and α denotes the parameters of each model. For the parameters $\{\beta,\alpha\}$ of the model, it can be expressed as follows:

$$\{\beta_m,\alpha_m\}_1^M=\arg\min\sum_{i=1}^{N}L\left(y_i,\sum_{m=1}^{M}\beta_m h\left(x_i;\alpha_m\right)\right) \tag{13}$$

$\{\beta_m,\alpha_m\}_1^M$ denotes two m-dimensional parameters, and the above formula denotes that for N sample points (x_i,y_i), the loss function under the model $F(x;\{\beta,\alpha\})$ is computed, and the optimal $\{\beta,\alpha\}$ is the one that minimizes this loss function $\{\beta,\alpha\}$. Written as a gradient descent this is the form of the following formula:

$$\beta_m,\alpha_m=\arg\min\sum_{i=1}^{N}L\left(y_i,F_{m-1}\left(x_i\right)+\beta h(x;\alpha)\right) \tag{14}$$

The above formula describes the calculation to obtain the parameter $\{\beta_m,\alpha_m\}$ of model $F_m(x)$ can make the direction of $F_m(x)$ is the fastest descent direction of the loss function of the previously obtained model $F_{m-1}(x)$. For each data point x_i, a $g_m\left(x_i\right)$ can be obtained and eventually we can get a complete direction of gradient descent:

$$\overrightarrow{g_m}=\left\{-g_m\left(x_i\right)\right\}_1^N-g_m\left(x_i\right)=-\left[\frac{\partial L\left(y_i,F\left(x_i\right)\right)}{\partial F\left(x_i\right)}\right]_{F(x)=F_{m-1}(x)} \tag{15}$$

To enable $F_m(x)$ to be in the direction of $g_m(x)$, we can use least squares optimization to obtain the following equation:

$$\alpha_m=\arg\min\sum_{i=1}^{N}\left(-g_m\left(x_i\right)-\beta h(x;\alpha)\right)^2 \tag{16}$$

After calculating the value of α_m, you can calculate to get β_m with the following formula:

$$\beta_m=\arg\min\sum_{i=1}^{N}L\left(y_i,F_{m-1}\left(x_i\right)+\beta h\left(x_i;\alpha_m\right)\right) \tag{17}$$

The final merged algorithm is modeled as:

$$F_m(x)=F_{m-1}(x)+\beta_m h\left(x;\alpha_m\right) \tag{18}$$

3.3. *Implementation of e-commerce user purchase behavior prediction model*

3.3.1. Evaluation of Predictive Model Training Effectiveness. This section of the experiment uses the user purchase behavior dataset provided by Ali Group above.

Training environment: CPU: INTEL-I5-7300HQ GPU: NIVIDIA GTX1050TI Memory: 16G Using the framework python 3.7 and keras-gpu deep learning open source library.

Model training: we first divide the training set, 20% of the data as the validation set, the remaining 80% of the data as the training set, the purpose of doing so is to compare the accuracy of the model on the

training set and the validation set, to prevent the model from overfitting. The results of the purchase behavior prediction model training are shown in Figure 5.

From the figure, we conclude that the training performance of the model on the training set and the test set is basically the same. With the increase in the number of training iterations, the Loss value of the model in this paper shows a sharp downward trend, and when the training is up to 20 epochs, the Loss value converges to about 0.14. The trend of Accuracy and Loss value is opposite, when the training iteration reaches 30 rounds, the Accuracy also stabilizes to 0.95 or so. In order to prevent overfitting we use chekpoint in the keras library to set checkpoints, and save the round with the highest accuracy on the validation set as our final prediction model, i.e., in the 30th round, the model's Loss value and Accuracy both reach the best effect. So we select the model obtained in the 30th round of iteration as the final purchase behavior prediction model.

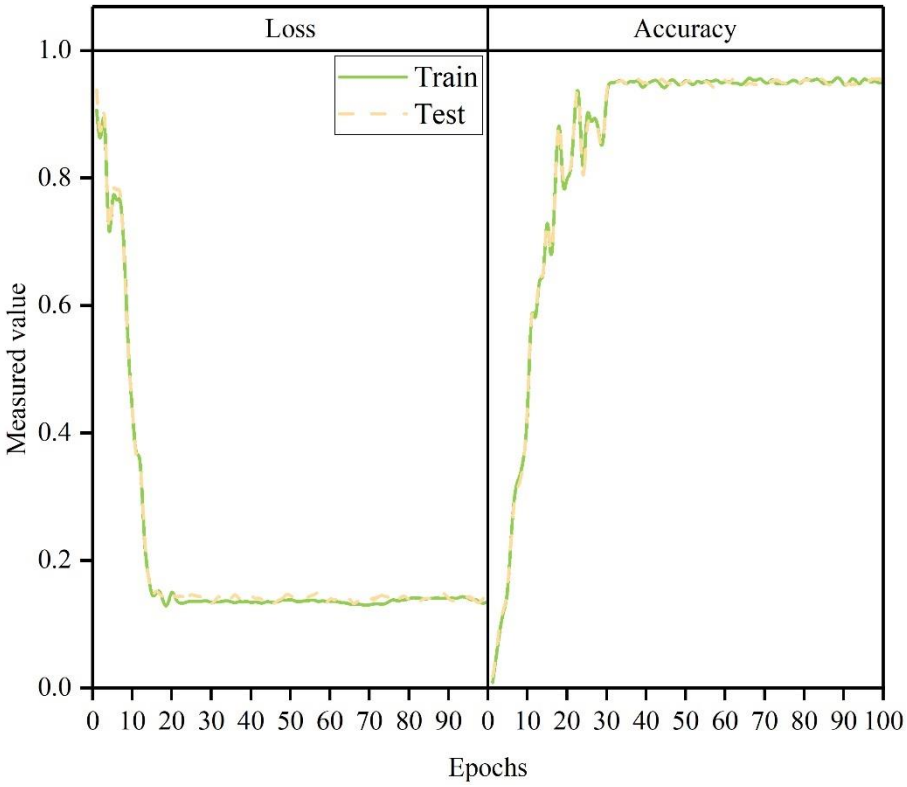

Fig. 5. Purchase behavior prediction model training results

3.3.2. Comparison of Purchase Behavior Prediction Accuracy Models. In this section, the problem of multicollinearity in the constructed feature variables is reduced by doing principal component dimensionality reduction on the processed data, by doing principal component dimensionality reduction. For the dimensionality reduced data composed a new dataset, this dataset is reasonably divided into training dataset and test set, and user purchase behavior prediction analysis is carried out in GBDT (Ours), XGBoost model, CART Tree, LSTM model, CNN network, and BP neural network, and the evaluation metrics are chosen to be conventional precision rate, recall rate, and F1 value. Figure 6 shows the results of each model for user purchase behavior prediction.

The figure shows that comparing the prediction results of the six classification models selected above, it can be found that GBDT, XGBoost model, CART tree, LSTM model, CNN network, and BP neural network in the number of predicted purchasing behaviors are 1833, 1404, 1605, 1270, 1359, and 1153 times, respectively. In the number of predicted non-purchase behavior is 148, 554, 544, 654, 690, 974 times respectively. Among them, the model in this paper predicts the user's purchase behavior more closely to the real situation.

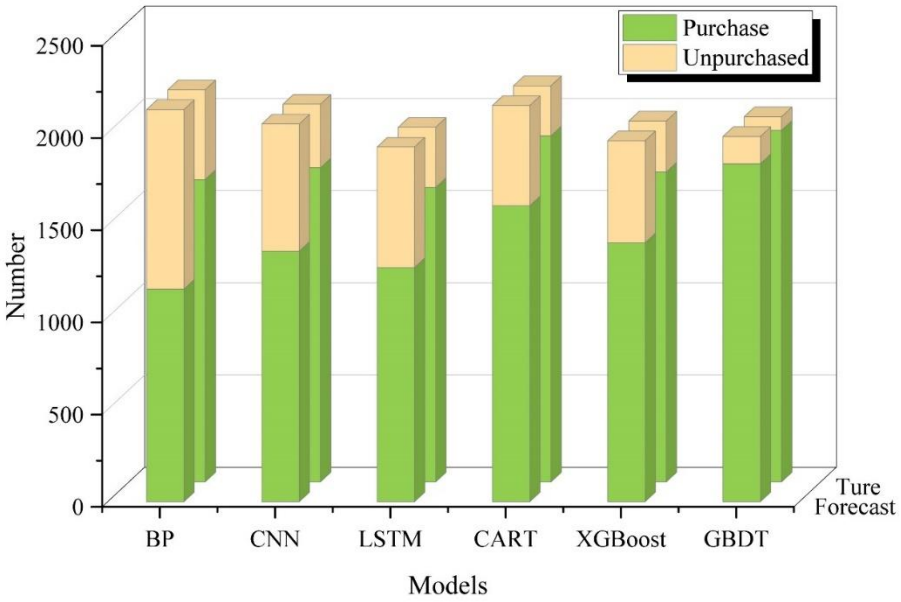

Fig. 6. The model is the result of the user's purchase behavior

The results of the comparison of the evaluation metrics of each model are obtained as shown in Fig. 7. Then the comparison of the evaluation indexes precision rate, recall rate and F1 value found that: six classification models in the prediction, LSTM model, CNN network and BP neural network precision rate, recall rate and F1 value is lower (less than 80%), the rest of the model performance is better, especially this paper's model, the value of the three indexes were 96.12%, 91.27%, 94.46% respectively. Comprehensive consideration leads to the result that this paper's model has the best effect on the prediction of purchasing behavior based on e-commerce users, and is more applicable to e-commerce platforms.

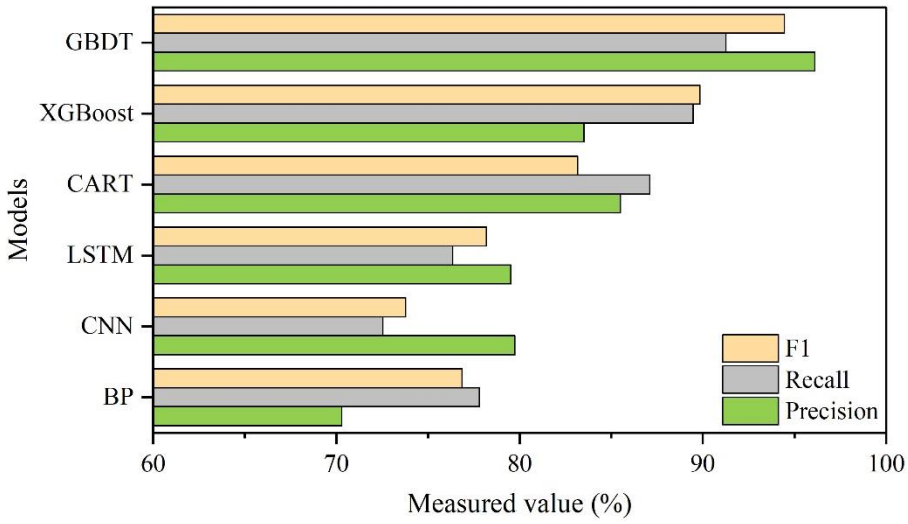

Fig. 7. The evaluation index of each model is compared

4. Cluster analysis algorithms based on user purchase behavior

4.1. Basic model of cluster analysis

Cluster analysis belongs to unsupervised learning in machine learning, which is a multivariate statistical approach widely used in the field of data mining.

Let a dataset $D=\{x_1,x_2\ldots x_n\}$, R be a cluster defined on dataset D, split D into m aggregate classes, $C_1,C_2\ldots C_m$ if the following three conditions are satisfied:

$$C_i \neq \varnothing (i=1,2\ldots,m) \tag{19}$$

$$\bigcup_{i=1}^{m} C_i = D \tag{20}$$

$$C_i \cap C_j = \varnothing, i \neq j (i, j = 1, 2 \ldots m) \tag{21}$$

Only if these three conditions are satisfied can we call it a clustering on D.

4.2. *Process of k-means clustering algorithm*

The most commonly used in cluster analysis is k-means clustering [34-35]. The steps of the algorithm are as follows:

1) Determine k initial points, which can be randomly assigned by the user.
2) Assign the rest of the sample points to the initial point closest to it, roughly forming k classes.
3) Then re-select the center of mass in the already k classes, one by one, and assign the remaining data points again to the new center of mass nearest to it.
4) Repeat step (3) until each data point is stabilized in the same class without further class changes.

4.3. *Determination of the optimal number of clusters*

In this paper, the elbow method is used to achieve the determination of the optimal number of clusters. For example, the central idea of the k-means clustering algorithm is to minimize the total sum of squares (SSE) within a class. And the total SSE measures the tightness of the clusters, so when the total within-class sum of squares is smaller it means the clusters are more tightly packed. Thus the elbow method looks at the total intra-class sum of squares as a function of the number of clusters evaluated:

$$SSE = \sum_{i=1}^{k} \sum_{p=C_i} |p - m_i|^2 \tag{22}$$

where C_i is the ind cluster, p is a sample point in C_i, and m_i is the center of mass of cluster C_i. The specific steps are as follows:

1) For example, for k-means clustering, calculate the clustering results when the value of k is changed, for example, by changing k from 1 to 10 to calculate different clustering results.
2) For different values of k, calculate the total error sum of squares SSE.
3) Plot the corresponding SSE curves for different values of k.
4) Based on the obtained plot, we observe the curve trend, and usually the value of k corresponding to the obvious bend out (elbow) in the plot is the optimal number of clusters.

4.4. *Cluster Analysis of Users Based on Purchasing Behavior*

In this section, based on the indicator data, customers are clustered and divided into groups, combined with the business characteristics of each user group, analyze their user value, and rank each user group.

For the user purchase behavior data, can be based on the display of clustering results and combined with the platform business characteristics of the user value analysis, the following are the key indicators: B (the number of goods purchased), C (the number of goods added to the shopping cart), S (stay time), P (browsing frequency), T (click on the type of goods of the type of species). At present, the user's purchase behavior is mainly the above five behaviors, so these five dimensions are proposed as the object of analysis.

Fig. 8 Trend graph of the relationship between the number of different clusters and SSE. From the figure, it can be seen that when the number of clusters is equal to 4, the SSE value is 27.12, and the decline is more obvious before 4, and the decline is slower after 4, so consider choosing the number of clusters to be 4.

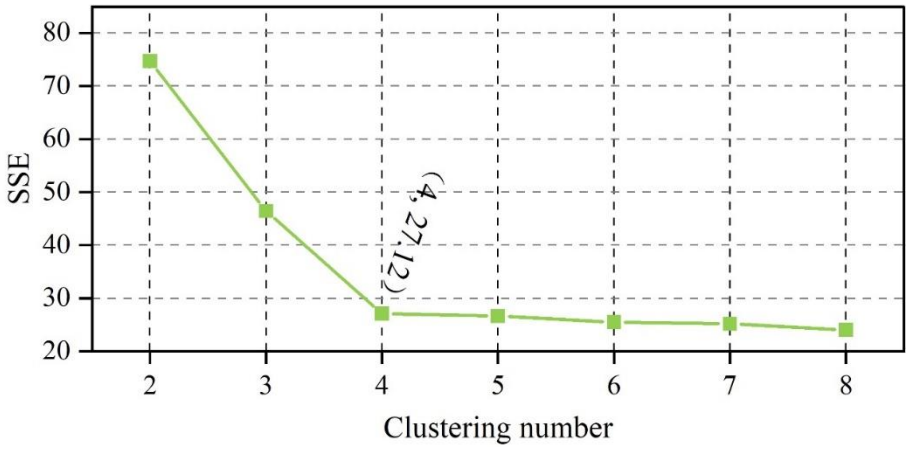

Fig. 8. The trend diagram of different clustering Numbers and SSE relationships

In this section, 100 users of the e-commerce platform are extracted and analyzed by clustering. Figure 9 shows the results of cluster class division based on different purchase behaviors. Table 2 shows the distribution of cluster centers for each dimension.

Combined with the chart, the five dimensions of user behavior are concentrated in four regions, among which, the number of times of collection, the number of types of products in the collection and the click rate are distinguished most obviously. The following is an analysis of the four sets of categories:

Cluster 1: C attribute is the largest, B, S, P attribute is the second largest, defined as important retained users.

Cluster 2: B, C, S, T attributes are the smallest, P attributes are the largest, positioning low-value/churned users.

Cluster 3: S, P attributes are the largest, B, T are the second largest, defined as important development users.

Cluster 4: B, T are the largest, C is larger, S, P are the smallest, defined as important retention customers.

Cluster category 4 (important to keep the user) as an example for detailed analysis: this type of user accounted for 29%, its B (the number of goods purchased), T (the number of types of goods) are the largest, C (the number of goods added to the shopping cart) is larger, S (stay time), P (browsing frequency) is the smallest. It shows that this type of user is willing to spend money on this platform, because the stay time is shorter and the browsing frequency is smaller, which indicates that the customer's goal is clear, and resources should be tilted to this part of the user, and a layered operation strategy should be adopted, focusing on improving the loyalty and satisfaction of this type of user.

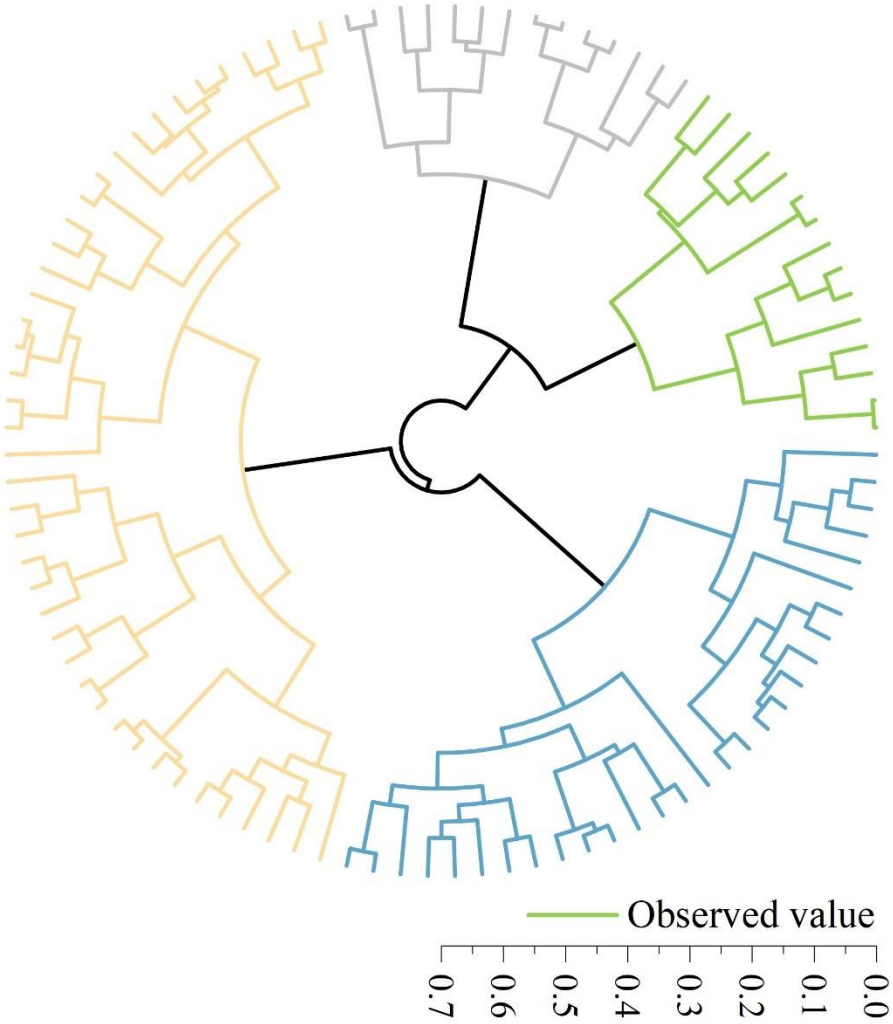

Fig. 9. Clustering of clusters based on different purchase behaviors

Table 2. Distribution of each dimensional cluster center

Clusters	B	C	S	P	T	Proportion
Cluster 1(Green)	5.65	23.11	9.11	126.25	19.81	15%
Cluster 2(Gray)	4.32	7.42	7.72	44.65	11.09	14%
Cluster 3(Yellow)	6.57	9.43	36.74	196.89	25.45	42%
Cluster 4(Blue)	8.61	17.55	5.86	6.41	29.89	29%

5. Conclusion

This paper briefly introduces the data sources of e-commerce users' purchasing behavior, and preprocesses (integrates and cleans) the raw data. According to the data characteristics, user purchase behavior features are designed to provide a basis for subsequent purchase behavior prediction using gradient boosting decision tree algorithm. In addition, a clustering model was used to classify the clusters according to user behavior. The behavior of "purchase" has the highest conversion rate of user purchase, which is more than 3.58%~9.80% higher than the other three behaviors in the near future. In this paper, the training model with 30 rounds of iterations is used as the purchase behavior prediction model, at which the Loss and Accuracy of the model converge to 0.14 and 0.95, respectively, and the precision, recall and F1 value of the model in this paper are 96.12%, 91.27%, and 94.46%, which are applicable to the prediction of e-commerce user's purchase behavior. The clustering algorithm divides the e-commerce users into 4 clusters, in which the cluster class 4 is important to keep the user, accounting for 29%, which has a higher user value.

Funding

1) Discipline Co construction Project of Philosophy and Social Sciences in Guangdong Province's 13th Five Year Plan for 2020 (Project Code: GD20XJY30).
2) Discipline Co construction Project of Philosophy and Social Sciences in Guangdong Province's 14th Five Year Plan for 2022 (Project Code: GD22XGL44).

References

[1] Peng, B., Zhao, J., Sun, Y., & Liu, Y. (2024, April). Research on IWOA-GBDT Regression Prediction Model Based on Gradient Lifting Decision Tree. In 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT) (pp. 636-640). IEEE.

[2] Ma, Z., Zhang, S., Chang, Y., Zhou, Q., Hu, X., & Li, X. (2022, April). An Extraction Method of Network Security Situation Elements Based on Gradient Lifting Decision Tree. In 2021 International Conference on Advanced Computing and Endogenous Security (pp. 1-13). IEEE.

[3] Huang, X., & Li, L. (2023). Location method of garden air pollution source based on gradient lifting regression tree algorithm. International Journal of Environmental Technology and Management, 26(6), 445-456.

[4] Lu, W. J., Huang, Z., Zhang, Q., Wang, Y., & Hong, C. (2023). Squirrel: A Scalable Secure {Two-Party} Computation Framework for Training Gradient Boosting Decision Tree. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 6435-6451).

[5] Ruobing, W. (2023, December). Research on the Application of Gradient Lifting Stochastic Forest Model in Electricity Price Forecasting. In 2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE) (pp. 527-531). IEEE.

[6] Yang, J., Zhao, C., Yu, H., & Chen, H. (2020). Use GBDT to predict the stock market. Procedia Computer Science, 174, 161-171.

[7] Li, D., & Ma, C. (2022). Research on lane change prediction model based on GBDT. Physica A: Statistical Mechanics and Its Applications, 608, 128290.

[8] Li, Q., Gu, M., Zhou, K., & Sun, X. (2015, November). Multi-classes feature engineering with sliding window for purchase prediction in mobile commerce. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 1048-1054). IEEE.

[9] Trudel, R. (2019). Sustainable consumer behavior. Consumer psychology review, 2(1), 85-96.

[10] Ali Hakami, N., & Hosni Mahmoud, H. A. (2022). The prediction of consumer behavior from social media activities. Behavioral sciences, 12(8), 284.

[11] Madanchian, M. (2024). Generative AI for Consumer Behavior Prediction: Techniques and Applications. Sustainability, 16(22), 9963.

[12] Chaudhary, K., Alam, M., Al-Rakhami, M. S., & Gumaei, A. (2021). Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. Journal of Big Data, 8(1), 73.

[13] Kooti, F., Lerman, K., Aiello, L. M., Grbovic, M., Djuric, N., & Radosavljevic, V. (2016, February). Portrait of an online shopper: Understanding and predicting consumer behavior. In Proceedings of the ninth ACM international conference on web search and data mining (pp. 205-214).

[14] Xu, A., Li, Y., & Donta, P. K. (2024). Marketing decision model and consumer behavior prediction with deep learning. Journal of Organizational and End User Computing (JOEUC), 36(1), 1-25.

[15] Hicham, N., & Karim, S. (2022, October). Machine learning applications for consumer behavior prediction. In The Proceedings of the International Conference on Smart City Applications (pp. 666-675). Cham: Springer International Publishing.

[16] Guo, L., Zhang, B., & Zhao, X. (2021). A Consumer Behavior Prediction Model Based on Multivariate Real-Time Sequence Analysis. Mathematical Problems in Engineering, 2021.

[17] Li, Y., Song, L., & Zeng, Y. (2019). Research on information security and privacy protection model based on consumer behavior in big data environment. Concurrency and Computation: Practice and Experience, 31(10), e4881.

[18] Bandara, R., Fernando, M., & Akter, S. (2021). Managing consumer privacy concerns and defensive behaviours in the digital marketplace. European Journal of Marketing, 55(1), 219-246.

[19] Feng, L. (2022). Data Analysis and Prediction Modeling Based on Deep Learning in E - Commerce. Scientific Programming, 2022(1), 1041741.

[20] Zhai, X., Shi, P., Xu, L., Wang, Y., & Chen, X. (2020, October). Prediction Model of User Purchase Behavior Based on Machine Learning. In 2020 IEEE International Conference on Mechatronics and Automation (ICMA) (pp. 1483-1487). IEEE.

[21] Sun, R. (2024). Applications of Machine Learning Algorithms in Predicting User's Purchasing Behavior. Science and Technology of Engineering, Chemistry and Environmental Protection, 1(7).

[22] Zhong, Y., & Li, W. (2022, August). User behavior prediction based on fusion model. In International Symposium on Artificial Intelligence Control and Application Technology (AICAT 2022) (Vol. 12305, pp. 211-219). SPIE.

[23] Wu, H., & Li, B. (2022, January). Customer Purchase Prediction Based on Improved Gradient Boosting Decision Tree Algorithm. In 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 795-798). IEEE.

[24] Jiajie, G. (2024). Research on Mobile E-commerce Recommendation Algorithms Based on Logistic Regression Improved Model Features. Academic Journal of Engineering and Technology Science, 7(5), 111-115.

[25] Xu, J., Wang, J., Tian, Y., Yan, J., Li, X., & Gao, X. (2020). SE-stacking: Improving user purchase behavior prediction by information fusion and ensemble learning. Plos one, 15(11), e0242629.

[26] Li, D., Zhao, G., Wang, Z., Ma, W., & Liu, Y. (2015, November). A method of purchase prediction based on user behavior log. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 1031-1039). IEEE.

[27] Liu, C. J., Huang, T. S., Ho, P. T., Huang, J. C., & Hsieh, C. T. (2020). Machine learning-based e-commerce platform repurchase customer prediction model. Plos one, 15(12), e0243105.

[28] Zhang, H., & Dong, J. (2020). Prediction of Repeat Customers on E - Commerce Platform Based on Blockchain. Wireless communications and mobile computing, 2020(1), 8841437.

[29] Yang, F., Han, X., Lang, J., Lu, W., Liu, L., Zhang, L., & Pan, J. (2018, October). Commodity recommendation for users based on E-commerce data. In Proceedings of the 2nd International Conference on Big Data Research (pp. 146-149).

[30] Mei, H., & Li, X. (2019). Research on e-commerce coupon user behavior prediction technology based on decision tree algorithm. International Core Journal of Engineering, 5(9), 48-58.

[31] Marcin Czajkowski,Krzysztof Jurczuk & Marek Kretowski. (2025). Enhancing multi-omics data classification with relative expression analysis and decision trees. Journal of Computational Science102460-102460.

[32] Caigang Zhuang,Shaoying Li,Haoming Zhuang & Xiaoping Liu. (2025). Mining individual daily commuting patterns of dockless bike-sharing users: A two-layer framework integrating spatiotemporal flow clustering and rule-based decision trees. Sustainable Cities and Society105985-105985.

[33] Kumar Ganti Praful,Naik Hrushikesh & Kanungo Barada Mohanty. (2022). Environmental impact analysis and enhancement of factors affecting the photovoltaic (PV) energy utilization in mining industry by sparrow search optimization based gradient boosting decision tree approach. Energy(PA),

[34] Xiuqi Huang,Hong Tao,Haotian Ni & Chenping Hou. (2025). Debiasing weighted multi-view k-means clustering based on causal regularization. Pattern Recognition111195-111195.

[35] Muhammad Rabiu Abdullahi,Qing Chang Lu,Adil Hussain,Sajib Tripura,Peng Cheng Xu & ShiXin Wang. (2024). Location optimization of EV charging stations: A custom K-means cluster algorithm approach. Energy Reports5367-5382.