International Journal of Smart Business and Technology
Vol.7, No.1 (2019), pp.45-56
http://dx.doi.org/10.21742/ijsbt.2019.7.1.05

Predictive Analysis of User Purchase Behavior Based on Machine Learning

Zhenyu Liu[1] and Xinyi Ma[2]

[1,2]*Harbin University of Science and Technology, China*
[1]*liuzhenyu@hrbust.edu.cn*

Abstract

In corporate customer management, companies must evaluate the costs and benefits of investment expenditures and determine the optimal resource allocation for marketing and sales activities within a period. Understanding customers' buying behavior in the future is a key driving force for the sales and marketing departments to allocate resources effectively. This paper proposes a combined prediction model using the Stacking method to integrate multiple decision tree models to predict whether users will buy in the future and their purchase time. The model uses the idea of stacking model fusion to fuse the prediction results of three different integrated decision tree models of Light GBM, XG Boost, and Random Forest, and then uses a simple logistic regression classification model and a linear regression model to predict separately based on the fused prediction results Whether the user will buy in the future and the specific time of purchase. In addition, in this study, we used real retail sales data to evaluate the predictive performance of the proposed method.

Keywords: *Machine learning, Buying behavior prediction, Light GBM algorithm, Stacking integration*

1. Introduction

The prediction of user purchase behavior has attracted the attention of scholars from a very early time [1]. Still, the lack of historical data has led to long-term stagnation of research in this area. The most challenging problem in predicting user purchase behavior is the prediction of user purchase behavior when the user's current state cannot be directly observed, and there are very few available historical records [2]. In the past few years, information technology's rapid development has greatly increased user transaction data availability [3]. The initial analysis of these user transaction data is usually carried out in summary statistics, such as the average or average order quantity and information characteristics related to user behavior.

With the greatly increased data availability, machine learning, and data mining techniques are often used in user-based predictions, and user churn prediction is one of the important issues in this field. In recent years, the concept of user churn and related predictive analysis has been well studied [4][5][6]. Accurately predicting user purchase behavior can provide a basis for companies to formulate inventory and sales plans, thereby reducing sales losses and unnecessary inventory costs. Therefore, many recent studies have focused on predicting users' future purchase behaviors. Martínez et al. proposed a dynamic, data-driven framework for

Article history:
Received (January 25, 2019), Review Result (March 20, 2019), Accepted (May 7, 2019)

Print: 2288-8969, eISSN: 2207-516X, IJSBT

predicting whether users intend to purchase within the company shortly in a non-contractual environment.

The above-mentioned research on predicting future user purchase behavior only focuses on whether the user will purchase within a period in the future and does not predict the specific purchase time of the user. Therefore, this article will further predict the specific purchase time of users based on predicting whether users will buy in the future.

Many methods can be used to predict purchase behavior, including time series analysis, panel data models, machine learning-based models, and random models (such as BG/NBD models). Time series analysis includes many different methods, such as the exponential smoothing method [7], the moving average method [8], and the autoregressive integrated moving average (ARIMA) model [9], etc. The research on these methods is more mature, but these methods do not include enough factors or take into account personal influence. Panel data [10] and random data model [11] contain more factors than time series analysis and have been successfully applied in various business scenarios related to forecasting. On the other hand, models based on machine learning can consider more factors and variables [12].

With the improvement of data availability, more research on predictive models based on machine learning is used to predict users' future purchase behaviors. These prediction methods based on machine learning mainly include logistic regression, support vector machines [13][14], artificial neural networks [15], gradient boosting decision trees (GBDT), and so on. In addition, Martínez et al. proposed a dynamic and data-driven framework to compare the prediction performance of three machine learning algorithms: logistic regression, neural network, and Gradient Boosting Decision Tree (GBDT). The results show that the gradient-boosting decision tree has the best prediction. Most of these studies use a single forecasting model, and by comparing multiple different single models, one or more forecasting methods with better performance are found.

Although the research method of adopting a single prediction model is relatively mature, it is easily affected by other random factors, resulting in a low prediction accuracy rate. The model is generally only suitable for specific environments, and the generalization ability is not strong enough. Therefore, to effectively reduce or offset the influence of random factors in a single model and improve the prediction accuracy and credibility of the prediction model, some scholars use different combination models to solve the prediction problem. The results show that the forecast accuracy of the integrated forecasting model is higher than that of the single-stage model; the cost of the integrated model of demand forecasting and inventory decision-making is much lower than that of the non-integrated model. Some scholars use historical sales data and online comment data from the automobile industry to predict automobile demand by combining the Bass model and sentiment analysis. The results show that the combined model has higher forecasting accuracy than the standard Bass model and other sales forecasting models.

From the above research, the combined forecasting model performs better in most cases than the single forecasting model. Therefore, based on the above research, this paper effectively combines the cutting-edge machine learning technology XG Boost algorithm, Light GBM algorithm, random forest algorithm, and stacking integrated learning method and proposes a combined prediction model based on multiple differentiated models to predict users' Future buying behavior.

2. Research methods

This paper proposes a prediction method for user purchase behavior based on the stacking fusion model. [Figure 1] describes the overall framework of the entire method.

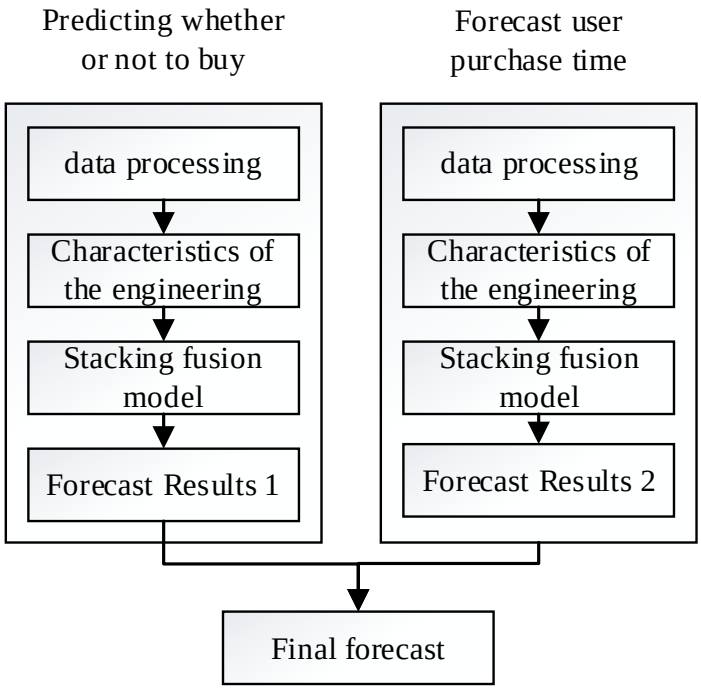

Figure 1. Overall framework

First, the original data set of all users is used to predict whether the user will buy or not and when the user will buy. The use of complete data to make predictions in two aspects can ensure the consistency of the data and, at the same time, enable the model to make the most of the data, thereby improving the overall prediction performance of the model.

Among them, the goal of predicting whether the user purchases is to use a classification model to predict whether the user will purchase the target product in the next period, that is, to obtain the set of users who will purchase the target product. The details are as follows: first, collect the required data, analyze and process simultaneously, and obtain a less noisy and more structured data set. Secondly, feature selection and construction are performed based on user purchase behavior analysis, and a higher-dimensional feature data set is obtained. Then, the feature data set will be inputted into the stacking classification fusion model for effective training. Finally, the trained classification fusion model performs out-of-sample prediction and obtains prediction result 1, that is, predicts the set of users who will purchase the target product within a period in the future.

The goal of user purchase time prediction is to use a regression model to predict the specific purchase time of each user in the next period. This step is similar to whether the user buys or not, that is, first data collection and processing, then feature selection and feature construction, then use of the Stacking regression fusion model for training and finally performing out-of-sample prediction on the trained regression fusion model. Obtain the prediction result 2, that is, predict the purchase time of all users in the future.

The ultimate goal of this article is to predict the specific purchase time of users who will purchase the target product in the next period. Therefore, prediction results one and two are integrated to obtain the final prediction result; that is, the users who will purchase the target product and the specific purchase time are predicted in the future.

2.1. Data collection and processing

In this study, we mainly collected historical sales data of a certain retailer. We extracted the following attributes for each piece of data: membership code, order code, purchase time, purchase quantity, purchase price, cost, product capacity, and product category. These attributes are shown in Table 1.

The collected historical sales data includes the purchase records of members and non-members. Members have a unique identification Vipcode, and the non-member Vipcode attribute value is null. Therefore, we deleted all records with the Vipcode attribute value being null. At the same time, the Number attribute value is generally a positive integer. Still, there is a phenomenon where the number attribute value is equal to or less than 0 in the collected sales data. After analysis, it was found that data less than 0 is the return data, and data equal to 0 is the gift data during the promotion. Therefore, to avoid these noisy data from affecting the prediction results, we offset the data with the Number attribute value less than 0 and the customer's corresponding purchase data and delete all the data with the Number attribute value of 0 to ensure that all the number attribute values are greater than 0. Finally, we processed null and duplicate values.

Table 1. Sales data attribute table

Attributes	Description
Vipcode	Member code
Order	Order code
Date	Purchase time (_ year _ month _ day)
Number	Purchase quantity (pcs)
Price	Purchase price (yuan)
Cost	Spend (Not necessarily equal to quantity * unit price)
Weight	Product capacity (g)
Cate	Product category

2.2. Feature engineering

In the research, we used the time-sliding window method for feature extraction to effectively expand the amount of sample data and make the training of the prediction model cover all historical data as much as possible. Each data set is 5 months, of which the last month is the label month, the first four months are used to extract features, and the length of each sliding is one month. The feature time window is divided into five feature extraction windows, which are 7 days, 14 days, 1 month, 2 months, and 4 months from the first day of the label month. Then, statistically analyze the features in these five windows.

In the study of user purchase behavior prediction, the main factors affecting prediction accuracy are user and product factors. After fully analyzing the sales data, we extracted user and product features from a small feature extraction window based on factors such as user buying behavior habits and product attributes, as shown in Table 2.

Table 2. Characteristic variable table (one)

Feature	Description
User characteristics	Total cost of the target product
	Total cost of all goods
	Total target product capacity
	Characteristics of purchase days (target product, all products, ratio of the two)
	Characteristics of the number of purchase orders (target product, all products, ratio of the two)
	Purchase quantity characteristics (target product, all products, ratio of the two)
	Average number of items per order
	Average number of items purchased per day
	Average number of purchases per target product
Commodity characteristics	Price characteristics (maximum, minimum, mean, median)
	Unit price characteristics (maximum, minimum, mean, median)
	Capacity characteristics (maximum, minimum, mean, median)

In addition, considering the global nature of the data and the relatively fixed consumption rate of the product, we also extracted the following user characteristics in the largest time window, as shown in Table 3.

Table 3 Characteristic variable table (two)

Feature	Description
User characteristics	Date of the first order (target product, all products, number of days from the label date)
	The date of the last order (target product, all products, number of days from the label date)
	The time difference between the last target product order and the last orders
	The time difference between the first and the last order of the target product
	The target product purchase time interval (maximum, minimum, mean, standard deviation)
	The average consumption rate of goods
	The total purchase capacity of the last target product
	Inferred feature 1 (how many days are available for the last purchase of capacity)
	Inferred feature 2 (inferred date of purchase)

2.3. Application model

This article uses various machine learning classification and regression algorithms, including the Light GBM algorithm [19], the XG Boost algorithm [20], the logistic regression algorithm, and the Lasso regression algorithm. The first three algorithms are integrated

models, and the latter two are simple models. The first three algorithms are used to compare the prediction performance of the fusion model and are also the base learners of the fusion model. In comparison, the latter two algorithms are only used as the meta-learners of the fusion model.

Light GBM (Light Gradient Boosting Machine is a distributed gradient boosting framework based on a decision tree algorithm. Its advantage lies in reducing data on memory, ensuring that a single machine uses as much as possible without sacrificing speed. At the same time, it reduces the cost of communication, improves the efficiency of multiple machines in parallel, and achieves linear acceleration in the calculation.

XG Boot (eXtreme Gradient Boosting) and Light GBM are algorithms based on decision trees. Its advantage lies in using many strategies to prevent overfitting while supporting parallelization, adding processing of sparse data, fast training speed, and high accuracy of training results.

Random Forest is an algorithm using multiple decision trees to train and predict samples. The random forest algorithm is an algorithm that contains multiple decision trees, and the multiple trees of individual decision tree output categories determine the output category. Its advantage is that its classification effect is better for most of the data. It can handle high-dimensional features, is not prone to overfitting, and the model training speed is relatively fast, especially for big data. When determining the category, it can assess the importance of the variable. It has strong adaptability to data sets, which can handle discrete and continuous data, and data sets do not need to be standardized.

Logistic regression is an algorithm very similar to linear regression. In essence, the types of problems handled by linear regression are inconsistent with logistic regression. Linear regression deals with numerical problems, while logistic regression is a classification algorithm. In other words, the logistic regression prediction result is a discrete classification. For example, logistic regression is a classic two-classification algorithm to determine whether an email is spam. Logistic regression adds a Sigmoid function to the calculation result of linear regression, converts the numerical result into a probability of 0 to 1, and then makes predictions based on this probability. For example, the email is spam if the probability is more significant than 0.5.

Lasso regression and Ridge regression are both types of generalized linear regression models. Both Lasso regression and ridge regression belong to the posterior probability model.

2.4. Stacking model fusion

The stacking model fusion method divides the original feature data set into several sub-datasets, input into each base learner of the first-layer prediction model, and each base learner outputs its prediction result. Then, the output of the first layer is used as the input of the second layer to train the meta-learner of the prediction model of the second layer. Then, the model in the second layer outputs the final prediction result. The stacking model fusion method can improve the overall prediction accuracy by generalizing the output results of multiple models.

In this study, we use three different integrated model algorithms of Light GBM, XG Boost, and Random Forest as the base learner to obtain three sets of prediction results and then apply the three sets of prediction results to the second layer using a meta-learner, including logic Regression or Lasso regression is trained to get the final prediction results, as shown in Figure 2 and Figure 3. Among them, the steps of the stacking model fusion method are used to predict whether the user purchases and the user purchase time are roughly the same. The

difference is that the base learner for whether the user purchases use the classification algorithm among the three integrated algorithms, and the meta-learner uses logistic regression. The classification algorithm and the base learner for user purchase time prediction use the regression algorithm among the three ensemble algorithms, and the meta-learner uses the Lasso regression algorithm.

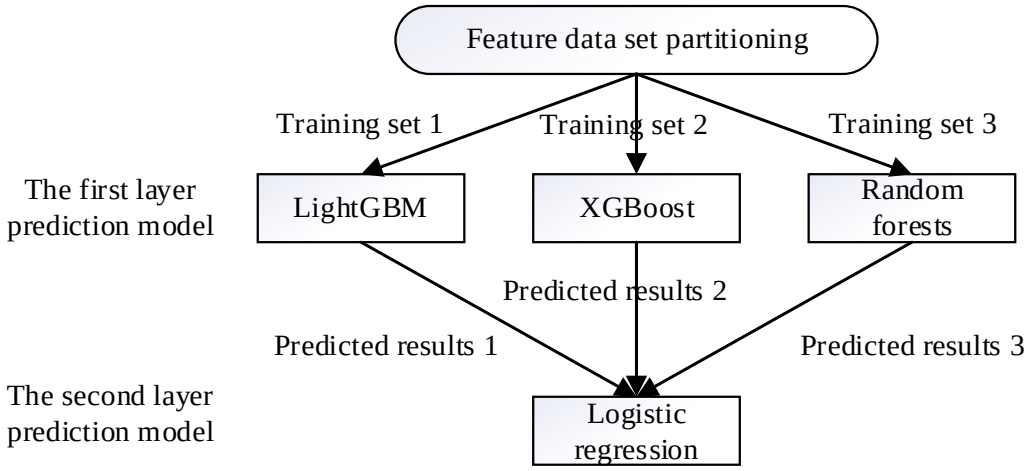

Figure 2. Whether the user will buy the prediction Stacking model fusion method

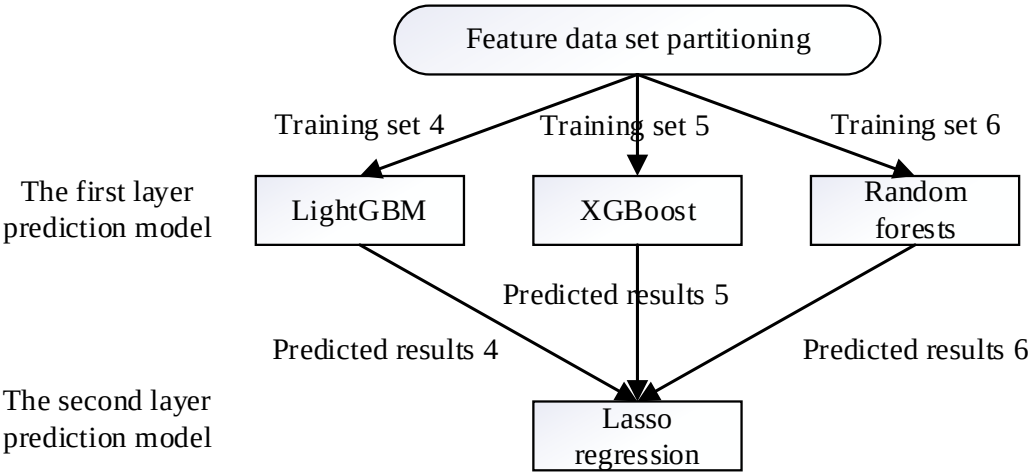

Figure 3 User purchase time prediction Stacking model fusion method

3. Empirical analyses

3.1 Data

This article uses the sales data of a chain retail company as an empirical sample, in which coffee is the target product, to predict whether users will buy in the future and the specific purchase time.

The data set in this article includes the daily sales volume, price, cost, and production capacity of 8 types of coffee from April 1, 2019, to October. Considering the relatively small amount of data in the data set, to effectively expand the amount of sample data and make the

training of the prediction model cover all historical data as much as possible, we used the time-sliding window method for feature extraction. After processing the raw data, we set the maximum time window to 120 days and the feature extraction window to 7 days, 14 days, 30 days, 60 days, and 120 days. Since the time interval for the user to purchase each product is 15 to 30 days, a sliding window is set every 15 days. After determining the time window, we extracted feature vectors, including 159 dimensions.

3.2. Model performance evaluation

This paper researches whether users will buy in the future and the prediction of purchase time. This section uses classification problems and regression problem evaluation indicators commonly used in machine learning to evaluate the prediction model's performance. To make the model interpretable, we also analyzed the feature importance results of several models with better prediction performance.

Predicting whether a user will buy in the future is a typical two-category problem, so we use the commonly used evaluation indicators in two-category issues, including accuracy, precision, recall, AUC value, and Roc curve evaluate model performance. At the same time, we compared the stacking fusion model with the results of a single base learner and meta-learner model. The evaluation results of the test set are shown in Table 4.

Table 4. Comparison results of classification models

Model	Accuracy	Accuracy rate	Recall rate	Auc value
Logistic regression	0.5561	0.5882	0.0018	0.586
Light GBM	**0.8256**	**0.7439**	**0.9264**	**0.91**
XG Boot	0.7741	0.6898	**0.8929**	0.872
Random forest	0.7619	0.6959	0.8241	0.847
Fusion model	**0.8504**	**0.8083**	0.8693	**0.928**

Note: The bolded ones are the optimal results of each indicator, and the underlined ones are the sub-optimal results of each indicator.

As shown in Table 4, the performance of the logistic regression algorithm is the worst. The reason is that a linear classification model such as logistic regression is unsuitable for this study's nonlinear problem, and it cannot handle complex nonlinear classification problems. The random forest result performed the worst among the three single ensemble models, and all the evaluation index results were lower than the other two models. There are three sub-optimal results in Light GBM's evaluation indicators, indicating that it has better predictive performance. Comparing the results of the other four models except for logistic regression, it can be found that the accuracy and precision of the fusion model and the AUC value are all optimal, which shows that the method of using the Stacking model fusion has good predictive performance for predicting user purchase behaviour.

The Roc curve of the fusion model is shown in [Figure 4]. The curve is very close to the y-axis and y=1, and the area under the curve, the AUC value, is above 0.928. This also shows that the fusion model has good predictive performance for user purchase behaviour.

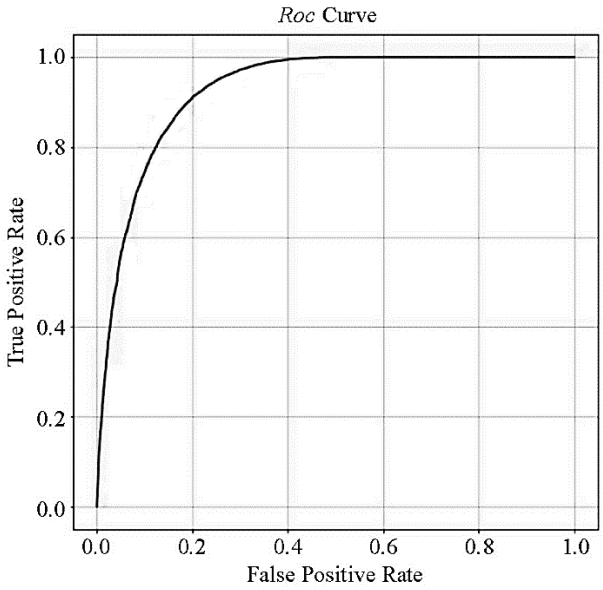

Figure 4. Roc curve of the fusion model

Predicting the purchase time of users is a regression problem. We use the three commonly used indicators in regression problems: Mean Square Error (MSE), Mean Absolute Error (MAE), and Explained Variance (EVS) to evaluate model performance. At the same time, we compared the stacking fusion model with the results of a single base learner and meta-learner model. The evaluation results of the test set are shown in Table 5.

As shown in Table 5, the linear regression model Lasso regression performed the worst, which also shows that linear regression models such as Lasso regression are unsuitable for this study's nonlinear problems, and they cannot effectively deal with complex nonlinear regression problems. Among the other four nonlinear models, Light GBM has the best prediction performance, with MSE and EVS values being the best and MAE being the second best. Two of the three evaluation index results of the fusion model are sub-optimal and are not far from the optimal result. This shows that the Light GBM model is the best in solving the problem of buying time prediction, followed by the fusion model.

Table 5. Regression model comparison results

Model	MSE	MSE	EVS
Lasso return	79.761	7.698	0.0198
Light GBM	**40.335**	**4.041**	**0.5087**
XG Boot	47.344	5.008	0.4181
Random forest	44.449	**3.924**	0.4629
Fusion model	**42.003**	4.342	**0.4808**

Note: ①The MSE indicator calculates the mean value of the sum of squares of errors. The smaller the value, the better the fitting effect. ② MAE is used to evaluate the closeness of the predicted result to the actual data set. The smaller the value, the better the fitting effect. ③ The value range of EVS is [0,1]. The closer to 1, the more the independent variable can explain the variance of the dependent variable. The smaller the value, the worse the effect

Feature importance analysis can be used to evaluate the predictive ability of the constructed feature or its importance to the predictive model. Through feature importance analysis, the predictive ability of the constructed feature can be directly observed to explain the model to a certain extent or further adjust the model structure. Here, we mainly consider the importance of the features of those models that perform well, whether users buy or not consider the fusion model. The user purchase time prediction considers the Light GBM model.

We constructed a comprehensive evaluation function, as shown in formula (1), to evaluate the comprehensive prediction performance of various model combinations.

$$S_{score} = \frac{\sum_{k \in K_p} f(k)}{|K_p|} \#(1)$$

$$f(k) = \begin{cases} 0, k \notin K_r \\ \dfrac{15}{15 + d_k^2}, k \in K_p \end{cases} \#(2)$$

Among them, Kr is the set of users who purchase the target product, Kp is the set of users who are predicted to purchase the target product, and dk represents the distance between the actual and expected purchase time.

As mentioned earlier, both the Light GBM and the fusion models have good prediction performance, so we use four combinations of these two models to evaluate the comprehensive prediction performance. The results are shown in Table 6.

Table 6. Table of comprehensive evaluation results

Classification	Return	S-score
Light GBM	Light GBM	**0.5214**
Light GBM	Fusion model	0.4837
Fusion model	Light GBM	**0.5431**
Fusion model	Fusion model	0.4964

As shown in [Table 6], the fusion model is used to predict whether the user purchases or not, and the comprehensive prediction performance of the Light GBM model is the best for predicting user purchase time. This model combination method improves the prediction performance by 9.4% compared with the two prediction problems using the fusion model.

4. Conclusion

Predicting users' future purchase behavior and purchase time can support the company's inventory decision-making and user marketing management. Although existing studies have researched from different perspectives, most of the studies only focus on whether users will buy in a period in the future, and there are few studies on the specific purchase time of users. This paper proposes a combined forecasting model that uses the Stacking method to integrate multiple decision tree models to predict users' purchase behavior and their specific purchase time. To this end, we fuse the prediction results of three different integrated decision tree models of Light GBM, XG Boost, and Random Forest, and then use a simple logistic regression classification model and a linear regression model to predict the purchase behavior of users based on the fused prediction results. And the specific time of purchase. Finally, we

used accurate retail sales data to verify and evaluate the model in this article. The results show that the fusion model has the highest accuracy and AUC value when predicting whether a user will buy, with an accuracy rate of 85% and an AUC value of 0.928. In addition, when predicting the user's purchase time, we found that the Light GBM algorithm has the best predictive performance compared to the fusion model. At the same time, if the fusion model and the Light GBM algorithm are used in different problem stages, combining the Light GBM algorithm improves the prediction performance by 9.4% compared to the fusion model for both prediction problems. A complete user purchase record includes attribute values such as purchase time, purchase quantity, purchase cost, etc. Therefore, the possible future research direction is to predict future purchases' actual quantity or value. Training predictive models based on the same feature processing methods and constructing predictive models to predict the number of purchases and purchase value of users is an important research topic in the future, which can provide more favorable support for the operation and strategic decision-making of enterprises.

References

[1] J. D. Herniter, "An entropy model of brand purchase behavior," Journal of Marketing Research, vol.10, no.4, pp.361-375, **(1973)**

[2] M. Platzer and T. Reutterer, "Ticking away the moments: Timing regularity helps to better predict customer activity," Marketing Science, vol.35, no.5, pp.779-799, **(2016)**

[3] X. Wu, X. Zhu, and G. Q. Wu, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol.26, no.1, pp.97-107, **(2013)**

[4] Y. Richer, E. Tom-Tov, and N. Slomin, "Predicting customer churn in mobile networks through analysis of social groups," Proceedings of the 2010 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, pp.732-741, **(2010)**

[5] C. S. Lin, G. H. Tzeng, Y. C. Chin, "Combined rough set theory and flow network graph to predict customer churn in credit card accounts," Expert Systems with Applications, vol.38, no.1, pp.8-15, **(2011)**

[6] A. Amin, S. Anwar, and A. Adnan, "Customer churn prediction in the telecommunication sector using a rough set approach," Neurocomputing, no.237, pp.242-254, **(2017)**

[7] M. Hussan, A. Shome, and D. M. Lee, "Impact of forecasting methods on variance ratio in order-up-to-level policy," The International Journal of Advanced Manufacturing Technology, vol.59, no.1,2,3,4, pp.413-420, **(2012)**

[8] S. Lee, and D. B. Fambro, "Application of subset autoregressive integrated moving average model for short term freeway traffic volume forecasting," Transportation Research Record, vol.1678, no.1, pp.179-188, **(1999)**

[9] P. Ramos, N. Santos, and R. Rebello, "Performance of state space and ARIMA models for consumer retail sales forecasting," Robotics and Computer-integrated Manufacturing, no.34, pp.151-163, **(2015)**

[10] S. Ren, T. M. Choi, and N. Liu, "Fashion sales forecasting with a panel data-based particle-filter model," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol.45, no.3, pp.411-421, **(2014)**

[11] P. S. Fader and B. G. S. Hardie 'A note on an integrated model of customer buying behavior," European Journal of Operational Research, vol.139, no.3, pp.682-687, **(2002)**

[12] T. M. Choi, C. L. Hui, N. Liu, "Fast fashion sales forecasting with limited data and time," Decision Support Systems, no.59, pp.84-92, **(2014)**

[13] C. J. Lu, "Sales forecasting of computer products based on variable selection scheme and support vector regression," Neurocomputing, no.128, pp.491-499, **(2014)**

[14] A. Candelieri, "Clustering and support vector regression for water demand forecasting and anomaly detection," Water, vol.9, no.3, pp.224, **(2017)**

[15] M. E. Gnay, “Forecasting annual gross electricity demand by artificial neural networks using predicted values of socio-economic indicators and climatic conditions: the case of Turkey,” Energy Policy, no.90, pp.92-101, **(2016)**