Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 3 July 2024, pp: 863-874
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Machine Learning Modeling for Forecasting Repeat Purchases in Online Shopping

Duan Lianzhai[1*], Dian Tri Hariyanto[2]

[1]Department of Master Science and Information Technology, Faculty of Computer Science,
President University, Indonesia
[2]PT Lontar Papyrus Pulp & Paper, Jambi, Indonesia

E-Mail: [1]Lark95@163.com, [2]diantrihariyanto3@gmail.com

Received Mar 28th 2024; Revised Apr 5th 2024; Accepted May 15th 2024
Corresponding Author: Kurniawan Danil

Abstract

Online shopping merchants will conduct a series of marketing activities to increase customers, but in many cases, most of the new customers will not make repeat purchases, which is not conducive to the long-term interests of the merchants. Therefore, it is important for merchants to target users who are more likely to repurchase, as this can reduce marketing costs and increase ROI. Based on the dataset provided by the online shopping website, this paper conducts mining and exploratory analysis of the data, utilizes feature engineering methodology, and modeling analysis using LightGBM, Logistic, Xgboost for machine learning modeling. Meanwhile, parameter optimization and model evaluation verification are performed, Finally, the comparative analysis resulted in Light GBM as the best prediction model, will provide efficient marketing decisions for the operation of online shopping stores.

Keyword: Data Analysis, Data Modeling, Machine Learning, Online Shopping, Repeat Purchase Forecast

1. INTRODUCTION

Merchants sometimes launch large-scale promotions or issue coupons on specific dates to attract consumers. However, many of the buyers attracted are one-time consumers. These promotions may not be helpful to the growth of sales performance in the long term, so for to solve this problem, merchants need to identify which type of consumers can be converted into repeat buyers. By analyzing and positioning these potential loyal customers and conducting precise marketing, merchants can greatly reduce promotional costs and increase return on investment (ROI). As we all know, it is difficult to accurately target customers when advertising online, especially targeting new consumers. With the development of big data technology and the continuous growth of e-commerce platforms, personal information such as users' interests and hobbies, as well as behavioral information such as daily shopping, have been accumulated in the databases of major e-commerce platforms, gradually forming a massive amount of data. It has been found that by mining big data on online shopping behavior, users' repeat purchase behavior can be predicted in advance, and it can even be specifically predicted which merchants' products each user has repeat purchase intentions.

In order to predict customers' decision to purchase, analytical methods such as regression and ML have been used by researchers over the years. The most widely used methods include Stepwise Logistic Regression (SLR) [1], Decision Tree (DT) [2], Random Forest (RF) [3], Support Vector Machines (SVM) [4], and Artificial Neural Networks (ANN) [5]. DT and RF have widespread applications for pre- diction related problems because of their ease of use and the high interpretability of their generated results. Moreover, unlike ANN, DT, and RF are both capable of directly handling categorical variables [2,3]. However, DT is less robust than RF and has been found to be highly sensitive to even small variations in data [6]. Additionally, RF is simpler to tune because it has a smaller number of hyperparameters as compared to neural network-based models [3]. However, ANN has been found to outperform DT and RF in terms of resource utilization and handling of multidimensional complex datasets [6,7]. SLR has been used in extant literature for predictions involving binary dependent variables. However, it suffers from a major limitation that makes it unfit for rigorous empirical analysis. SLR adds or removes variables during analysis in a specific order and studies have found that this order of addition or removal of variables can affect the final outcome [8]. This has prompted scholars to suggest the use of SLR for exploratory research only. Artificial intelligence (AI) has aided in accelerating the digital revolution over the decennium [9]. Machine learning (ML) enables self-learning and

DOI: https://doi.org/10.57152/malcom.v4i3.1388

modeling via data mining and provides ma-chines automatic control and decision making with situation-awareness [10,11]. These ML algorithms help business models evolve through continuous data learning and recommend informed decisions [12,13,14]. While all of these approaches have improved the ability to determine customers' purchases, we believe that recent advances in computing, especially DL techniques, hold much promise, primarily due to their capability to improve predictions through learning. Recently various studies have started embracing this approach for analyzing large and complex datasets. For example, a recent study by Loureiro et al. [15] has adopted DL to forecast sales in fashion retailing. Also, Korpusik et al. [16] have applied a feedback-based DNN model (i.e., Recurrent Neural Network) to a large corpus of tweets of potential customers to predict their choice of products and final purchases.

However, when existing machine learning models are used to predict repeat purchase behavior, their prediction accuracy is low, and it is difficult to produce valuable effects when applied to recommendation systems. Through the analysis of factors affecting repeat purchase behavior, it is found that users' repeat purchase behavior is controlled by a series of complex subjective factors, including perceived value, satisfaction, trust, etc. These influencing factors make users' repeat purchasing behavior diverse and differentiated. When these complex and diverse behavioral patterns are hidden in the data, it will bring great difficulties to the fitting of the machine learning model. The difference in user behavior patterns is an important reason that restricts the prediction accuracy of the machine learning model. This paper takes the prediction of repeat purchase behavior of e-commerce platform users as the research background, and after performing machine learning modeling analysis, as well as parameter optimization and model evaluation. Ultimately, LightGBM was selected as the best predictive model for repeat purchase to reduce the impact of differences in user purchasing behavior patterns on prediction performance, thereby improving the accuracy of repeat purchase behavior prediction.

2. DATASET AND RESEARCH METHOD

2.1. Data Collection and Pre-processing

In order to analyze user repeat purchase rates, I download four data files from the online shopping website, namely training data, test data, user portraits and user history records. The training data provides the dimensions of user, merchant, and whether the user is a repeat purchaser of the merchant (i.e. label). The user portrait data set provides age and gender information corresponding to the user ID; the user history record provides the user's various active statuses and click times in different stores in the past six months; the test data set is a combination of users and merchants to predict the Whether the user is a repeat purchaser of the merchant.

Based on the four data forms, the combination of user ID and merchant ID is given in the test data, and it is necessary to predict the user's repeat purchase probability value at the corresponding merchant. Through mining and analysis of the above data sets, we can predict the user's repurchase behavior and provide efficient marketing decisions for store operations.

2.2. Explore the Various Factors That Influence Repurchase

2.2.1. Analyze the Relationship Between Different Stores and Repurchase

The repurchase situation of different stores is very different, and the repurchase rate of some stores is low.

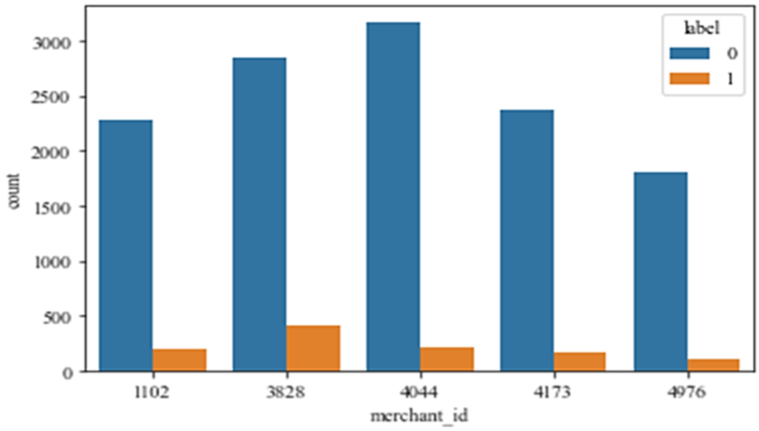


Figure 1. Analyze the Relationship Between Different Stores and Repurchase

2.2.2. Check the Store's Repurchase Distribution

The repurchase rate of user stores is distributed around 0.15. The repurchase rate of users is relatively low, and different stores have different repurchase rates. (The data distribution is right-skewed)

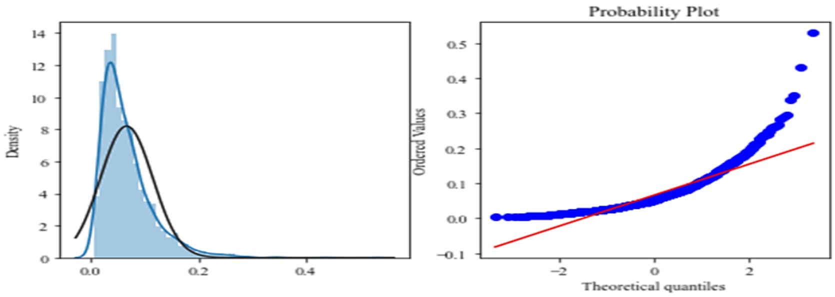


Figure 2. Check the Store's Repurchase Distribution

2.2.3 View the Repurchase Distribution of Users

Users in the past six months have mainly made one purchase, with fewer repeat purchases.

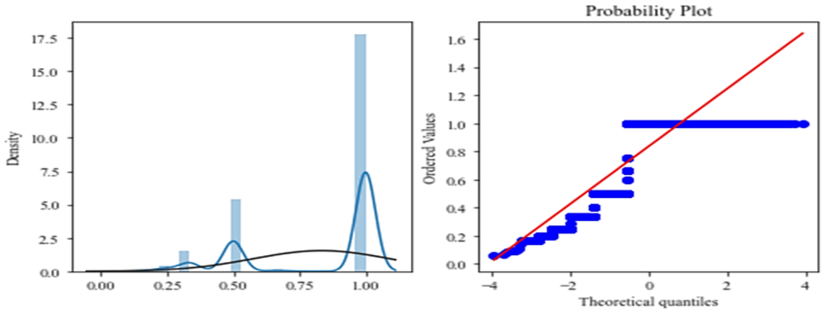


Figure 3. View the Repurchase Distribution of Users

2.2.4 Analysis of User Gender

There are differences in the repurchase situation between men and women, with women's purchase situation being much higher than that of men.

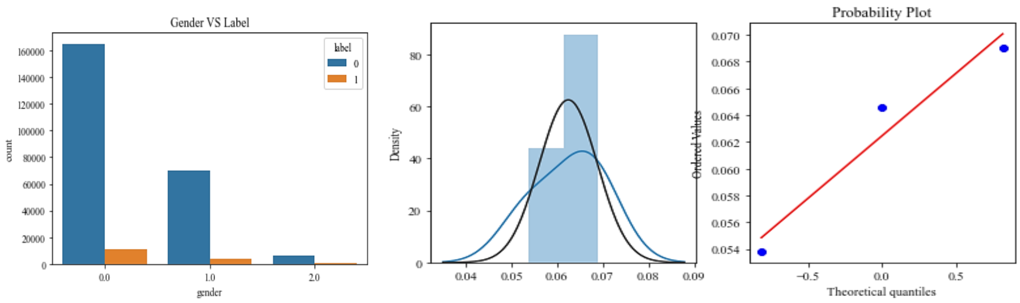


Figure 4. Analysis of User Gender

2.2.5 Analysis of User Age

There are differences in the repurchase situation of users at different age groups, with the focus being on users between 25 and 34 years old.

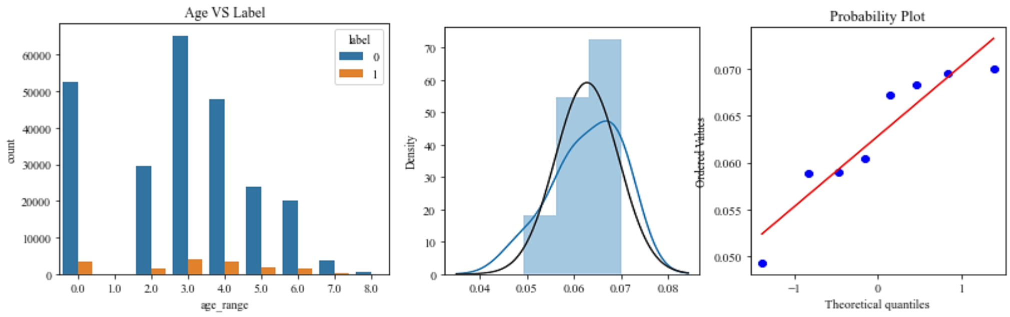


Figure 5. Analysis of User Age

2.3. Feature Engineering

The customer behavior log table contains specific customer, merchant, brand, category and product information. We can extract feature values based on these attributes and the connections between the attributes, and build a model to predict the merchant and the given merchant in the customer purchase behavior table. Whether the merchant's new customers during the promotion period will make repeat purchases in the next six months. Based on the customer behavior logs and customer personal information 6 months before the promotion, a feature table is formed based on basic attributes and basic attribute pairs with associated relationships, and the customer purchasing behavior table expanded in the data processing stage is further expanded to generate the final customer purchasing behavior. table as the data set for subsequent training models.

The following principles are followed when extending the extracted features to the customer purchasing behavior table: basic attribute features are expanded according to the corresponding basic attribute values; associated attribute pairs are expanded according to the associated basic attribute values; gender and age-related attributes are expanded according to the basic attributes and gender, age Expand.

The extracted features are mainly divided into basic data statistical features, integrated features, complex features, age and gender features and recent behavioral features.

We clarified the definition of repeat purchasing customers and the factors that affect customers' repeat purchasing behavior, analyzed the data in the data set, and performed data preprocessing such as data cleaning and data integration based on understanding the data. Finally, the relevant features of customers and merchants were extracted according to manual rules, and a feature project was constructed. The feature project included 5 categories and a total of 82-dimensional feature vectors, and a sample data set was constructed for the training model.

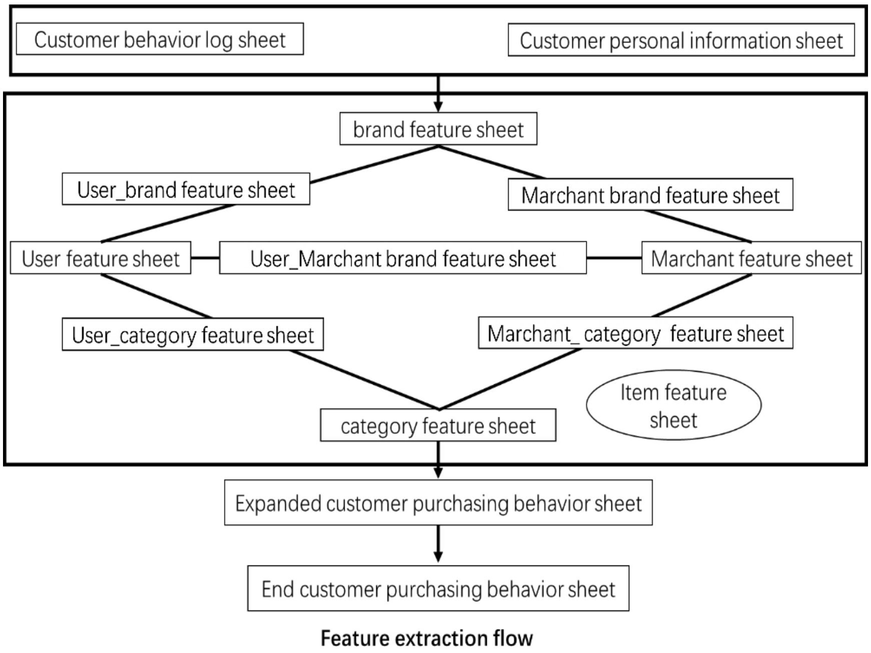

Customer behavior log sheet
Customer personal information sheet
brand feature sheet
User_brand feature sheet
Marchant brand feature sheet
User feature sheet
User_Marchant brand feature sheet
Marchant feature sheet
User_category feature sheet
Marchant_ category feature sheet
Item feature sheet
category feature sheet
Expanded customer purchasing behavior sheet
End customer purchasing behavior sheet
Feature extraction flow

Figure 6. Feature Engineering

3. ALGORITHM TEST AND MODELING

3.1. Algorithm Selection

Model building is the core part of the data mining process. According to the goals of data mining and the characteristics of the data set, one or more methods are selected to model the relationships implicit in the data. This stage requires continuous experimentation, parameter adjustment, and as much as possible. Build the best possible model.

Considering that the data used in this article is large and high-dimensional, and the focus of this article's research is whether the indicator of merchant influence will have an impact on customers' repeat purchasing behavior after promotional activities, this is mainly accomplished by setting up an experimental control group. The final goal is to determine which factors affect customers' repeat purchasing behavior after the big promotion. Therefore, the selected model needs to be suitable for the data of this article. The model must have strong interpretability and high accuracy. Therefore, the constructed models are linear models based on logistic regression algorithm, and Xgboost algorithm, LightGBM algorithm, and Catboost algorithm based on tree models.

Table 1. The Comparison of Advantages and Disadvantages of Algorithms

Algorithm	Advantages	Disadvantages
Logistic	Easy to use and explain	Sensitive to multicollinearity among independent variables in the model
Xgboost	The model loss function uses regularization to control the complexity of the model and prevent overfitting.	Adoption and sorting, when the amount of data is large, the greedy algorithm takes a long time and takes up a lot of memory.
LightGBM	Training is faster and takes up less memory	LightGBM's leaf-wise may produce deeper decision trees

First, according to the degree of feature importance, select features that contribute more to the training model and eliminate features that contribute less to reduce the time complexity of the training model and avoid model overfitting [18], [19]. an initial characterization of the data structure required for modeling is performed, where x denotes the independent variable of the model, the feature dimension of feature engineering, y denotes the model observable dependent variable, i.e., whether repurchase occurs in the next 6 months for users whose first purchase occurred at a merchant during the promotional period, x denotes the customer, w denotes the merchant, xi denotes the customer feature, wi denotes the merchant characteristics (including merchant influence characteristics), zi denotes customer-merchant characteristics, and yij denotes whether the customer finally repurchases from the corresponding merchant, taken as 0 or 1. This is shown in the following table:

Table 2. Data structure

Customers	Merchant	Merchant Influence	Customer-Merchant	Repurchase
$x1, x2, \cdots, xm$	$w1, w2, \cdots, wf$	$wf+1, wf+2, \cdots, wf+q$	$z1, z2, \cdots, zl$	yij

Secondly, the original dataset is divided, where 70% is used as training set and 30% as testing set, the sample number of samples and related details are shown in the following table:

Table 3. Information Table for the Divided Sample Set

Data set	Sample Size	Number of Positive Samples	Percentage of Positive Samples	Positive: Negative
training set	168282	10602	6.3%	1:14.87
test set	72120	4754	6.5%	1:14.17

Finally, in order to measure the impact of the inclusion of the merchant's market position influence on the model's prediction effect, and to compare the prediction accuracy of whether or not to use this dimension of features in the model, this paper considers the use of AUC as a metric for describing the model's prediction accuracy. The feature engineering in this paper contains 82 features, 17 of which belong to the features that measure the influence of the merchant's market position, i.e., through comparison, to explore the influence of the merchant's market position on the prediction of whether the user will repeat purchasing behavior at the merchant.

Although the number of positive samples in this paper is sufficient, the proportion of positive and negative samples is still slightly unbalanced, so we consider to control the proportion of positive and negative samples in the model to conduct experiments and observe whether the four models used in this paper will be disturbed by the proportion of positive and negative samples, so this paper adopts the following sample ratios to conduct the experiments: 1:3, 1:5, 1:10, and 1:14, respectively.

3.2. Logistic Regression Algorithm

3.2.1. Modeling Steps

The modeling steps of the logistic regression algorithm used in this paper are as follows:

1. Set the independent variables and dependent variables from the needs and objectives of this paper. The independent variable is a number of features mined from the four dimensions in the previous chapter, and the dependent variable is a dichotomous variable representing whether new customers repurchase at the corresponding merchant within six months after the promotion.
2. Filtering some important features through the tree model LightGBM algorithm for feature extraction. According to LightGBM, all the mined features are ranked in terms of feature importance, and some important features are extracted into the logistic regression model by setting a threshold value.
3. Evaluate the modeling effect of the training and test sets using the AUC value.

3.2.2. Logistic Regression Model Results

Firstly, a total of 83 features including users, merchants, merchant-users, and merchant influence after feature cleaning are used to train the LightGBM model, the top 50 features are selected, and the features in the training and test sets are normalized into the logistic regression model, and different positive and negative sample ratios are selected to see the impact on the AUC results of the test set.

Secondly, in order to measure the influence of merchant's market position on the model prediction, 17 features that measure the influence of merchant's market position are excluded, and 66 features in three dimensions, namely, customer itself, merchant itself, and customer-merchant interaction, are left to enter into the LightGBM model for training, and the top 50 features are selected, and then entered into the logistic regression model after normalization of the data to see the impact on the AUC results of the test set under different sample ratios. After normalizing the data, we also enter the logistic regression model to see the AUC results of the test set with different sample ratios.

Finally, by comparing the results of the above two steps, we explore whether the influence of merchant's market position has any significant effect on the improvement of prediction accuracy.

Table 4. Logistic regression algorithm results

Sample proportion of Logistic Regression	The AUC of the test set (include influence of merchant status)	The AUC of the test set (exclude influence of merchant status)
1:3	0.6995	0.6772
1:5	0.6990	0.6764
1:10	0.6975	0.6749
1:14	0.6968	0.6743

By observing the above experimental results, it can be found that regardless of whether the latter contains the information of the indicator of the influence of the merchant's market position, the value of AUC gradually decreases with the imbalance of the proportion of positive and negative samples, although the effect of the decrease is not obvious, but it also verifies the logistic regression model is easy to be disturbed by the imbalance of the samples. The AUC value of the test set containing the factor of merchant's market position influence is better than that of the AUC without merchant's position influence in all sample proportions, with an average of 2.2% higher, which indicates that the factor of merchant's market position influence in the logistic regression model has an impact on the prediction of repeat purchase behavior.

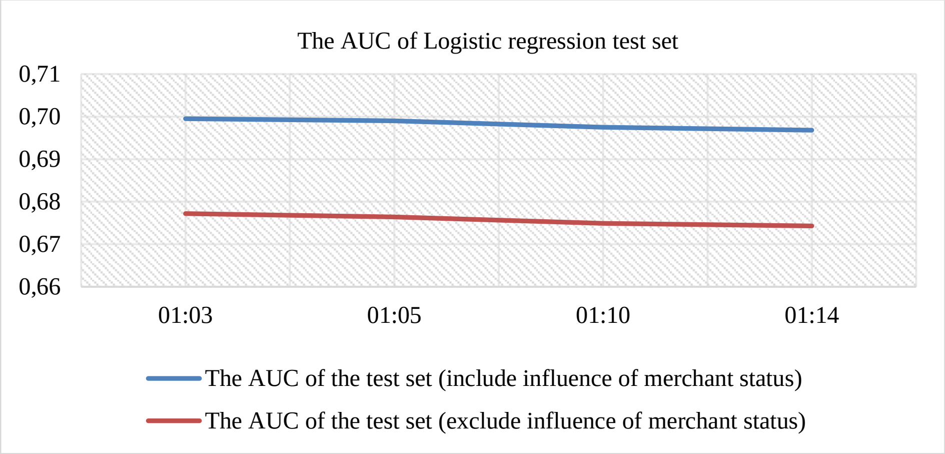


Figure 7. The AUC of Logistic Regression Test Set

3.3. Xgboost Algorithm

3.3.1. Xgboost Algorithm Parameter Settings

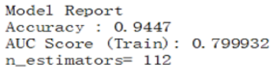
This experiment is based on sklearn library, the following table shows the main parameter settings of Xgboost, using the more common parameter settings:

Table 5. Xgboost Algorithm Parameter Settings

Parameter item	Parameter Description	Selected value in this article
learning rate	Learning rate	0.1
min_child_weight	Minimum leaf node sample weight	1
max_depth	Maximum depth of tree	5
max_leaf_nodes	Maximum number of nodes	default
gamma	Minimum loss function drop value	0

Parameter item	Parameter Description	Selected value in this article
subsample	Random sampling ratio	0.8
colsample_bytree	Proportion of randomly sampled columns	0.8
alpha	L1	1
scale_pos_weight	Handle sample imbalance terms	1

After the common parameter settings, the training set is modeled using the five-fold cross-validation method, in which the number of ideal regression trees is 112, which is determined after determining the learning rate of 0.1, using the "cv" function in Xgboost, and using cross-validation in each iteration, and the AUC of the training set reaches 0.799, as shown in the following figure. The AUC of the training set is 0.799, as shown in the figure below:



```
Model Report
Accuracy : 0.9447
AUC Score (Train): 0.799932
n_estimators= 112
```

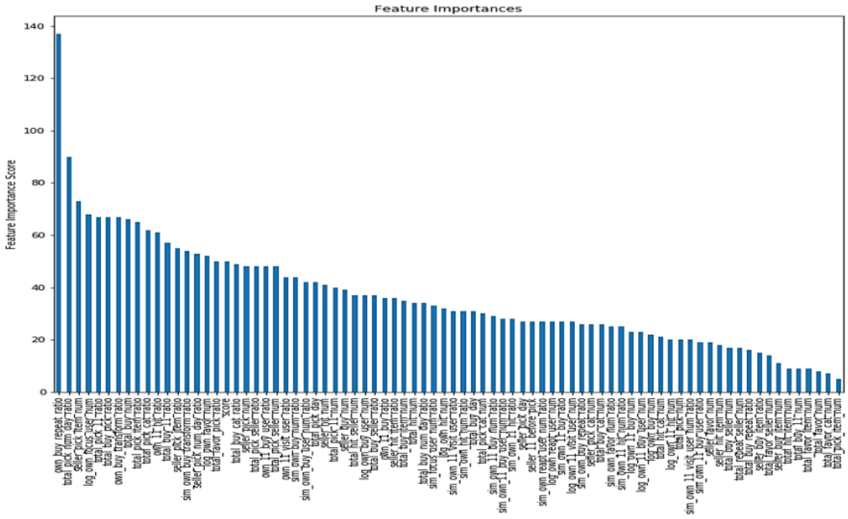


Figure 8. Xgboost Feature Importance Ranking

3.3.2. Xgboost Model Results

The Xgboost algorithm uses the parameters selected in the previous section, the main idea is: divide the training samples into four sample sets with different positive and negative ratios, use all features (83) versus some of the features that do not include the influence of the merchant's market position (66), and finally view their prediction effect on the test set, the results are shown in the following table:

Table 6. Xgboost Algorithm Results

Sample proportion of Xgboost Algorithm	The AUC of the test set (include influence of merchant status)	The AUC of the test set (exclude influence of merchant status)
1:3	0.7246	0.7127
1:5	0.7273	0.7129
1:10	0.7264	0.7147
1:14	0.7281	0.7159

As can be observed in the figure below, the AUC of the test set with merchant status influence is, on average, 1.2% higher than the AUC of the test set without merchant status influence for the four different sample proportion scenarios. The difference between the results of the models in these two test sets is not very obvious, indicating that the Xgboost algorithm is not very sensitive to the sample imbalance problem, but in comparison, the best prediction results are obtained for the model with 1:14 samples containing the influence of the merchant's status, and the best prediction results are obtained for the model with 1:10 samples not containing the influence of the merchant's status.

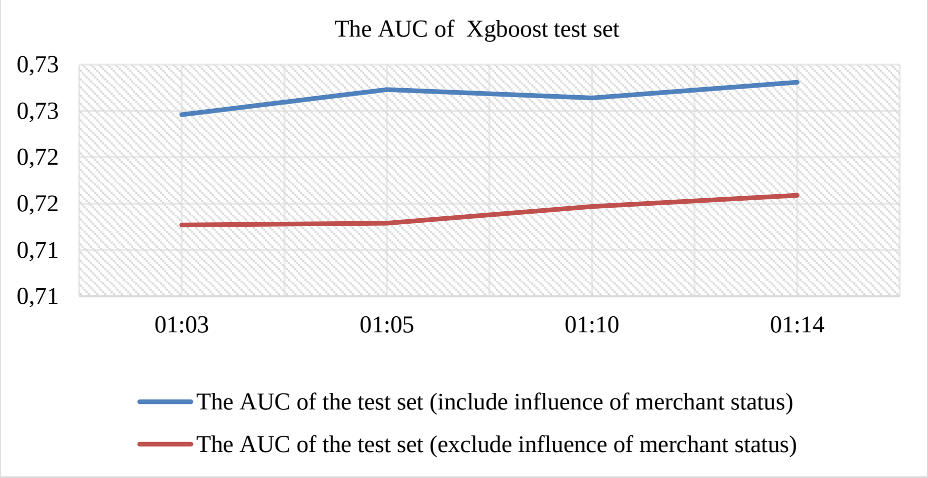


Figure 9. The AUC of Xgboost Test Set

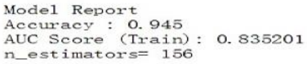
3.4. LightGBM Algorithm

3.4.1. LightGBM Parameter Settings

Table 7. LightGBM Parameter Settings

Parameter item	Parameter Description	Selected value in this article
max_depth	Maximum depth of tree	5
max_bin Feature	The maximum number of bins that will be stored	Default
num_leaves	The value should be <=2^(max_depth)	30
min_data_in_leaf	Minimum amount of data in a leaf	Default
colsample_bytree	Control the proportion of randomly sampled columns for each regression tree	0.8
subsample	Control the proportion of random sampling for each tree	0.8
lambda	Specify regularization	Default
min_gain_to_split	Describes the minimum gain of the split	Default

The LightGBM model is also constructed using 50% discounted cross-validation on the training set, and the optimal ideal decision tree is 156, the AUC on the training set reaches 0.815, which is better than Xgboost, and the feature importance is evaluated, and the results are shown in the following figure:



```
Model Report
Accuracy : 0.945
AUC Score (Train): 0.835201
n_estimators= 156
```

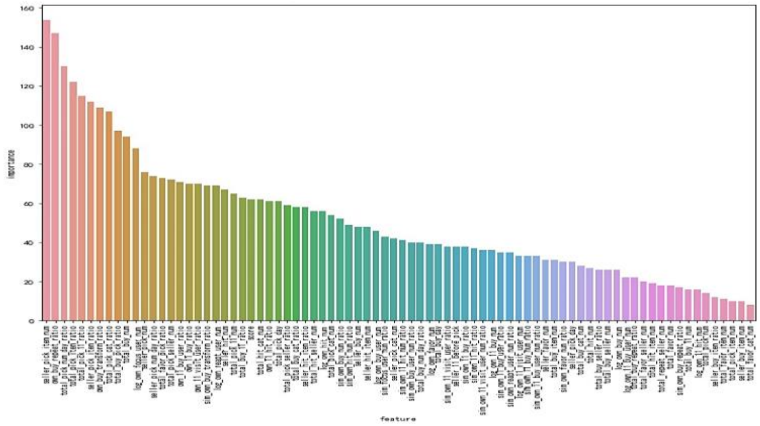


Figure 10. LightGBM Feature Importance Ranking

3.4.2. LightGBM Modeling Results

The LightGBM algorithm adopts the parameters selected in the previous section, the main idea is to use all the features (83) and part of the features (66) that do not include the influence of merchant's market position in four sample sets with different positive and negative ratios, and finally compare the experimental results of the four single models with different sample ratios to see which model is the most effective, and to validate the validity of the innovations in this paper through empirical evidence, i.e., to investigate the influence of merchant's market position on the research. That is, the necessity of studying merchant market position influence to study user repurchase behavior, the detailed results are shown in the following table:

Table 8. LightGBM Algorithm Results

Sample proportion of LightGBM Algorithm	The AUC of the test set (include influence of merchant status)	The AUC of the test set (exclude influence of merchant status)
1:3	0.7397	0.7286
1:5	0.7402	0.7285
1:10	0.7382	0.7276
1:14	0.7411	0.7294

It can be observed from the figure below that, in the case of four different sample proportions, the AUC of the test set containing the influence of the merchant's status is on average 1.1% higher than that of the test set without the influence of the merchant's status, which verifies the validity of the innovation points in this paper. The difference between the AUC values of several models in these two test sets is not very obvious, which shows that the LightGBM algorithm is not sensitive to the sample imbalance problem, but comparatively speaking, no matter whether it contains the indicator of merchant's status influence or not, the prediction effect of the sample ratio of 1:14 is the best.

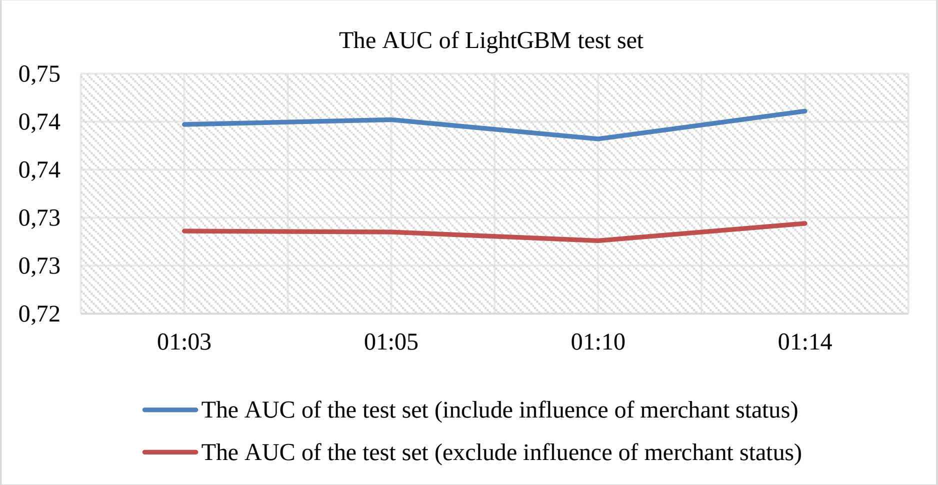


Figure 11. The AUC of the LightGBM Test Set

3.5. Comparison and analysis of single model results

The predictive effects between single models are first compared. Starting from the result analysis according to whether or not the merchant status influence is included, four different models are trained through four data sets with different sample ratios: 1:3, 1:5, 1:10, and 1:14. As shown in the figure below, the three lines in the figure represent the different performances of Logistic Regression, Xgboost, and LightGBM in the two cases respectively. From the figure, it can be seen that the AUC effect of the test set with merchant influence is significantly better than that of the test set without merchant influence, and the average prediction accuracy is 1%-2% higher, indicating that the influence of merchants has an impact on users' repeat purchase behavior. Regardless of whether the feature of merchant influence is included or not, LightGBM's AUC on the test set is not much different and has the best performance, while the least effective is the logistic regression model. The logistic regression model is the most effective with a sample ratio of 1:3, and its effect decreases as the ratio of positive and negative samples gets smaller and smaller; the other three groups are not very sensitive to the imbalance between positive and negative sample ratios, and the positive and negative sample ratios do not have a great impact on them.

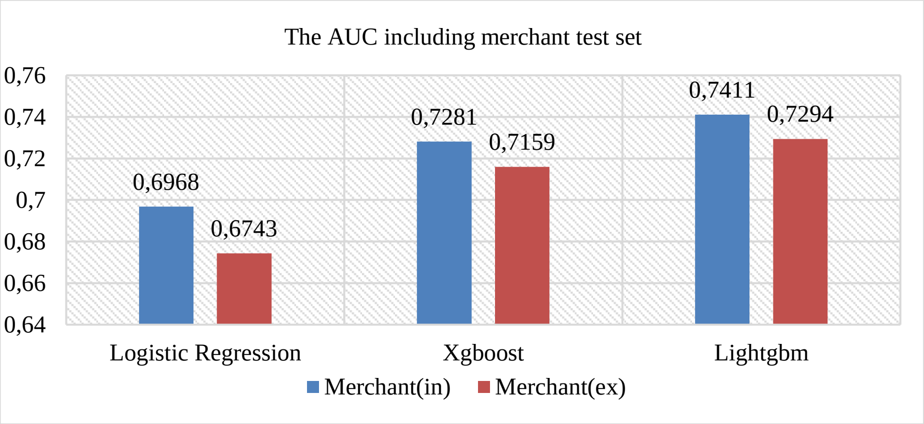


Figure 12. The Comparison of AUC Result on the Test Set

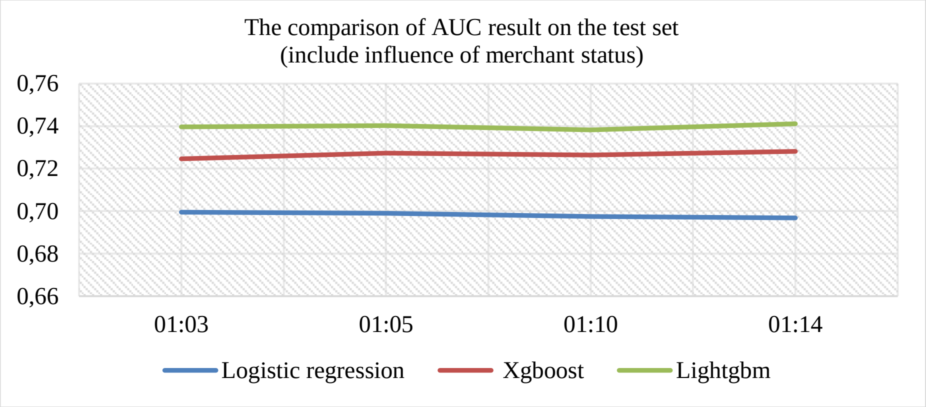


Figure 13. The Comparison of AUC Result on the Test Set (Include Influence of Merchant Status)

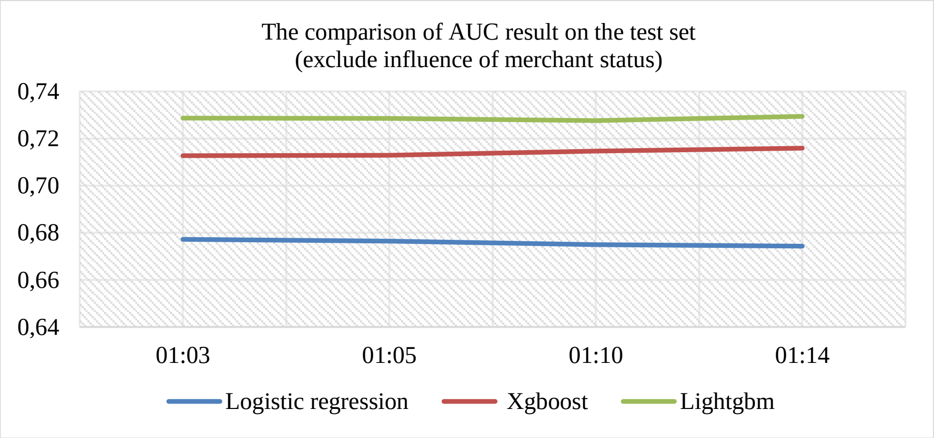


Figure 14. The Comparison of AUC Result on the Test Set (Exclude Influence of Merchant Status)

Through experiments, it is verified that several other models except Logistic Regression model are not sensitive to the problem of imbalance between positive and negative sample ratios, so they are uniformly trained in accordance with the sample ratio of 1:14, and the results of their test set are recorded, and the training effect of each model on the training set is added, and the results are shown in the table below:

Table 9. All Model Results

Model	AUC including merchant status influence indicator		AUC without merchant status influence indicator	
	Training set	Test set	Training set	Test set
Logistic Regression	0.736	0.6968	0.713	0.6743
Xgboost	0.799	0.7281	0.785	0.7159
LightGBM	0.815	0.7411	0.796	0.7294

It can be seen that the LightGBM model performs optimally regardless of whether the test set contains merchant influence indicators or not, and the decision tree algorithm based on the histogram algorithm, LightGBM, is fast, takes up little memory, and is very suitable for the high-dimensional big data used in this paper. Overall, the single model with merchant influence metrics has a higher AUC level than the single model without merchant influence metrics by an average of 1%-2%. Important features were then selected based on the predictive effectiveness of the single model. Since LightGBM is more effective, the top 10 important features are selected based on the feature importance of the LightGBM model to see how important the features are. The top 10 features are:

Table 10. LightGBM feature ranking

LightGBM feature ranking	Corresponding features
1	Repeat purchase rate at the merchant
2	User purchase conversion rate
3	The logarithm of the number of customers followed by the merchant
4	Purchase conversion rate among similar merchants
5	Total number of purchases by users
6	Number of times/total number of times users selected different products
7	Proportion of clicks among similar merchants
8	The number/total number of selections made by users during Double Eleven (promotion sensitivity)
9	The average number of daily selections made by users at the corresponding merchant
10	Merchant’s Influence Score

From a macro point of view, the information of the merchant has an important role to play in whether or not the repurchase occurs at the merchant, i.e.: the merchant's own attributes, as well as the influence of the merchant's market position have an important impact on whether or not the user will choose the merchant to continue to purchase, and from the micro level to the following several factors are very important:

(1) In the dimension of the merchant itself: the flow of customers at the merchant, the loyalty of customers at the merchant. (2) In the customer dimension: the customer's pickiness, sensitivity to promotions, and customer loyalty. (3) In the customer-merchant dimension: the degree of customer interest in the merchant. (4) In the merchant market influence dimension: the influence of the merchant's customer traffic, the influence of user loyalty and the influence of the merchant among all merchants.

4. CONCLUSION

Most consumers have shifted from physical stores to online channels for buying products and services. Repeat purchase behavior prediction is to analyze the user's purchase behavior rules based on the user's past historical behavior data, and then predict the user's future repeat purchase behavior. It is one of the important indicators in the development of online shopping, which directly reflects the user's satisfaction and loyalty to the product or service. A high repeat purchase rate means that users are satisfied with the product or service and are willing to continue purchasing and using it, which has a positive impact on the company's business model and market competitiveness. The repeat purchase behavior prediction can be applied to the recommendation system of e-commerce platforms to help merchants identify users with repeat purchase intentions, allowing merchants to take more targeted marketing measures and improve user experience, thereby allowing users and merchants to establish a lasting and reliable relationship. relationship, prompting users to purchase more products, thus bringing huge profits to the e-commerce platform.

This paper uses real user data disclosed by shopping websites to provide a certain exploration of the repurchase phenomenon, and more of a prediction of whether to repurchase. Combined with current mainstream machine learning algorithms for modeling analysis. In this paper, we propose to use logistic regression, Xgboost, and LightGBM algorithms to construct a single model, and use AUC to evaluate the model's effectiveness, and find that LightGBM is the most effective among the single models, and all three single models show the same feature, which is that the AUC value of the model containing the merchant's influence dimension is 1%-2% higher than that of the one not containing this dimension. Perhaps there is still a lot of room for improvement in the model, and continuous optimization and improvement can be considered from the following aspects:

(1) Considering that studying the problem of customers' repeat purchase behavior should not only focus on the customers themselves, but the merchants' related information will also have an impact on the customers. However, this paper only considers the influence of the merchant's influence on the customer's repeat purchase on the basis of the previous work, and future work can start from the information on the interaction behavior between the customer and the merchant's similar merchants at the merchant, to dig out richer information.

(2) Restricted by the access to data channels, the user behavior time obtained in this paper is measured in days, if there are better access channels, it can be accurate to the hours, minutes and seconds will be more conducive to the study of user behavior analysis.

REFERENCES

[1] S.L. Gortmaker, D.W. Hosmer, S. Lemeshow, Applied logistic regression, Contemp. Sociol. 23 (1994) 159, https://doi.org/10.2307/2074954.

[2] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106, https:// doi.org/10.1023/A:1022643204877.

[3] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/ 10.1017/CBO9781107415324.004.

[4] H. Drucker, C.J.C. Surges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, in: Advances in Neural Information Processing Systems, 1997: pp. 155–161.

[5] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, The Bulletin of Mathematical Biophysics. 5 (1943) 115–133, https://doi. org/10.1007/BF02478259.

[6] S. Lessmann, B. Baesens, H.V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, Eur. J. Oper. Res. 247 (2015) 124–136, https://doi.org/10.1016/j.ejor.2015.05.030.

[7] S. Lessmann, B. Baesens, H.V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research, Eur. J. Oper. Res. 247 (2015) 124–136, https://doi.org/10.1016/j.ejor.2015.05.030.

[8] M. Chau, H. Chen, A machine learning approach to web page filtering using content and structure analysis, Decis. Support. Syst. 44 (2008) 482–494, https:// doi.org/10.1016/j.dss.2007.06.002.

[9] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inform. 35 (2002) 352–359, https://doi.org/10.1016/S1532-0464(03)00034-0.

[10] Srivastava, M., Abhishek, S., Pandey, N., 2023. Electronic word-of-mouth (eWOM) and customer brand engagement (CBE): do they really go hand-in-hand? Electron. Commer. Res. 1-69 https://doi.org/10.1007/s10660-023-09743-z.

[11] Kim, T.S., Sohn, S.Y., 2020. Machine-learning-based deep semantic analysis approach for forecasting new technology convergence. Technol. Forecast. Soc. Chang. 157,120095.

[12] Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260.

[13] Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Wright, R.,2023a. “So what if ChatGPT wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inf. Manag. 71, 102642.

[14] Dwivedi, Y.K., Pandey, N., Currie, W., Micu, A., 2023b. Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: practices, challenges and research agenda. Int. J. Contemp. Hosp. Manag. https://doi.org/10.1108/IJCHM-05-2023-0686.

[15] Grover, P., Kar, A.K., Dwivedi, Y.K., 2022. Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions. Ann. Oper. Res. 308 (1–2), 177–213.

[16] M. Mousavizadeh, D.J. Kim, R. Chen, Effects of assurance mechanisms and consumer concerns on online purchase decisions: an empirical study, Decis. Support. Syst. 92 (2016) 79–90, https://doi.org/10.1016/j.dss.2016.09.011.

[17] A.L.D. Loureiro, V.L. Migu´eis, L.F.M. da Silva, Exploring the use of deep neural networks for sales forecasting in fashion retail, Decis. Support. Syst. 114 (2018) 81–93, https://doi.org/10.1016/j.dss.2018.08.010.

[18] M. Korpusik, S. Sakaki, F. Chen, Y.Y. Chen, Recurrent neural networks for customer purchase prediction on Twitter, in: CEUR Workshop Proceedings, 2016: pp. 47–50.

[19] M Dashand,H Liu . Feature selection for classification [J]. Intelligent Data Analysis , 1997,(03):131-156.

[20] Molina L C, Belanche L, Àngela Nebot. Feature Selection Algorithms: A Survey and Experimental Evaluation[C].IEEE International Conference on Data Mining. DBLP, 2002:306-313.