Research on Early Warning of Customer Churn Based on Mutual Information and Integrated Learning—Taking Ctrip as an Example

Wei Yang

School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei, 066099, China

Abstract: *With the increasing competition pressure among tourism e-commerce platforms, how to reduce customer churn to the greatest extent is of great significance to tourism e-commerce platforms. Based on this, this paper takes Ctrip's hotel customer-related data as an example, and first uses a supervised feature selection method based on mutual information to select features that have an important impact on customer churn. Then, by using the cross-validation method, combined with evaluation indicators such as accuracy rate, F1-sorce, AUC, etc., select the best model set from Logistic regression, Support Vector Machine, Decision Tree, Random Forest, GBDT, XGBoost, and LightGBM. A subset of the optimal models, and then the optimal model fusion is performed. The empirical results show that the multi-model fusion has higher accuracy and stability. In addition, based on the model fusion results, this paper obtains the importance ranking of customer personal characteristics. Finally, this paper puts forward relevant suggestions on how to accurately manage Ctrip and reduce the customer churn rate.*

Keywords: *customer churn, feature selection, machine learning*

1. Introduction

In recent years, big data and machine learning have become more widely used and mature, and their application in the field of e-commerce has become more and more popular. One of the biggest characteristics of e-commerce shopping is that customers are unstable, the competition in the online market is intensified, and the amount of customer churn is high. Therefore, how to retain customers and minimize the rate of customer churn is a problem that all merchants must solve. Mining customers' behavior preferences and their important influencing factors from massive access data, and using the data to predict customer churn trends will help the platform to formulate retention measures and improve merchants' profits.

As for the loss of e-commerce customers, the current classification research of e-commerce customers at home and abroad includes: Dule (2014) is based on customer value, and customers are divided into four categories: gold customers, silver customers, copper customers and iron customers. At present, the prediction methods for the loss of e-commerce users are as follows: The EBURM model proposed by Yanfang and L. Chen (2017) can predict user churn behavior with high confidence; Li Jin (2018) pointed out that the model constructed by XGBoost algorithm has higher performance than the model constructed by random forest and C4.5 algorithm; Chen Peng (2021) proposed that the stacking fusion model based on 4 groups of models of Logistic, SVM, RF and XGBoost algorithms is suitable for the study of customer churn in the online travel booking market.

In the context of predicting customer churn, the existing personal characteristics of customers are usually a high-dimensional set, and most of the current research on predicting customer churn early warning pays more attention to prediction models, and there are few researches on feature selection. Based on this, this paper first uses the mutual information criterion to perform supervised feature selection on the original set of personal characteristics of customers, and obtains a subset of personal characteristics of customers that has an important impact on whether customers churn.Then construct a candidate model class with the full set of machine learning methods such as logistic regression, support vector machine, decision tree, gradient boosting tree, random forest, GBDT, XGBoost, LightGBM, etc. By using the cross-validation method, combined with the correct rate, F1- Sorce, AUC and other evaluation indicators to select the optimal model subset. In the data background of this paper, the models in the optimal subset are random forest, XGBoost and LightGBM respectively. Then, based on the weighted fusion model, the optimal hyperparameters are selected by the cross-validation method to establish a customer churn risk early warning model. The model results show that the prediction ability of the fusion model is better than that of the single model. In addition, based on the model fusion results, the importance ranking of the individual customer characteristics is obtained. Finally, this paper puts forward relevant suggestions for Ctrip on how to carry out accurate management and reduce customer churn rate.

2. Methodology

2.1. Data sources and preprocessing

The data in this article comes from the official website of Hewhale Community.

In this paper, 48 features are preliminarily selected according to the method of selecting features of Ling Fei (2019) Credit Assessment. The dependent variable is whether the customer has churn. When its value is 1, it is considered that the customer has lost, and when the value is 0, it is considered that the customer has not lost. This paper selects a total of 3000 labeled samples and 1200 unlabeled data. Model fit is trained and evaluated using labeled samples, and unlabeled data is used to predict the distribution of customer churn in future potential markets. Regarding the preprocessing of the data, specifically, since the dimensions of the variables are different, the variables are standardized.

2.2. Model principle

In this paper, random forest, XGBoost and LightGBM are used as the base classifiers, and the machine learning model is constructed as a weighted model fusion ensemble learning method. The specific process is as follows: select Ctrip customer access data from the Hewhale Community public data set and perform basic preprocessing, and perform feature screening on the data features based on mutual information to obtain the optimal feature subset. Part of the subset is used as a training set using cross-validation method, combined with evaluation indicators such as accuracy, F1-sorce, AUC, etc., to screen the optimal model subset, and train the optimal fusion model. The final model can rank the importance of the customer's personal characteristics, and another part of the optimal feature subset is input to the trained model as the test set, and the customer classification probability can be output.

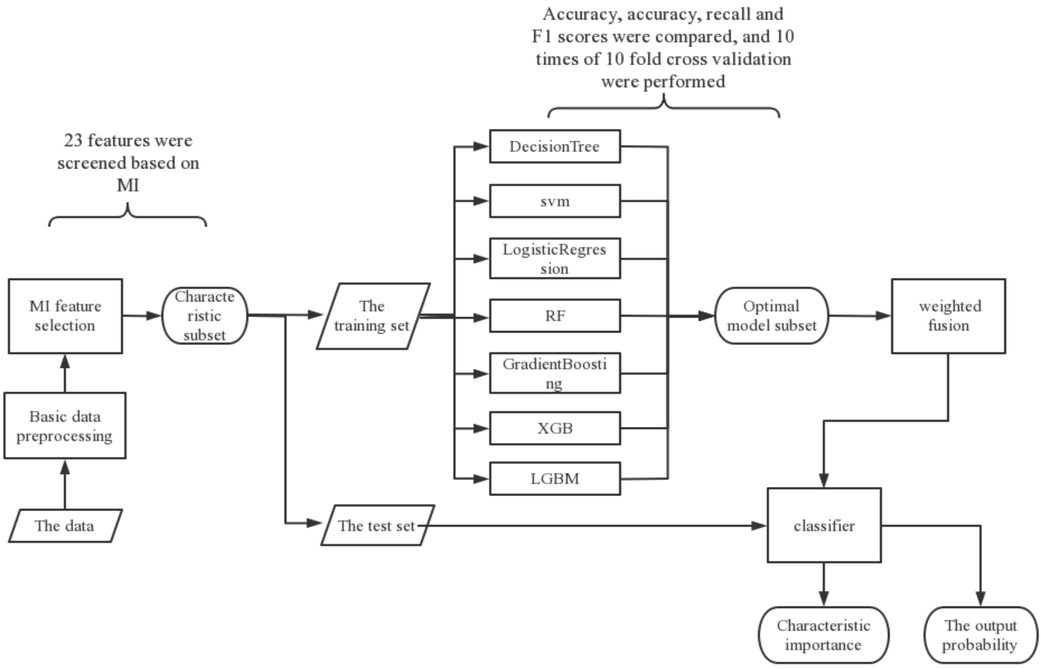

Figure 1: Model flow chart

3. Research results

In this paper, the method based on mutual information is used for feature selection, which is solved by using python programming, and uses the cross-validation method to select 23 of the 48 features as the feature subset for the input of the model. The selected important features are as follows:

Table 1: Important customer characteristics based on mutual information

Index	Indicator meaning	Index	Indicator meaning
x1	User cancellation rate within one year	x13	value of customer
x2	Number of orders cancelled by users in one year	x14	24h access to the hotel can book the lowest average price
x3	star preference	x15	24-hour history of the most viewed number of unique visitors in the history of the hotel
x4	Average daily number of hotel visits by users in the past 3 months	x16	The time since the last order within one year
x5	Number of hotel detail pages visited within 7 days	x17	The number of app uvs that visited the current city yesterday with the same check-in date
x6	User Preferred Rates - Most Viewed Hotel Rates in 24 Hours	x18	The number of app orders submitted yesterday in the current city with the same check-in date
x7	User annual orders	x19	Time since last visit within one year
x8	average price	x20	user conversion rate
x9	The most viewed hotel reviews in 24 hours	x21	Session id, sid=1 can be considered as a new visit
x10	The value of customers in the past year	x22	Annual visits
x11	Historical cancellation rate for the most visited hotel in 24 hours	x23	interview time
x12	The 24-hour historical average of the number of people who viewed the hotel's rating	-	-

In this paper, taking the previously described indicators to compare the selection. The specific results are shown in the figure. According to the judgment of each evaluation index, Random Forest, XGBoost and LightGBM are finally selected as the optimal model sub-set. This paper uses the training set for cross-validation to determine the weight of the weighting method, and uses the area under the ROC curve AUC value to evaluate the performance of the single model and the fusion model with ten-fold ten-fold cross-validation method. The specific results are that the AUC value of random forest is 0.81266, the AUC value of XGBoost is 0.81656, the AUC value of LightGBM is 0.80202, and the AUC value of weighted fusion model is 0.83761.Compared with the random forest model, the prediction accuracy of the weighted fusion model is improved by about 2.5%;Compared with the XGBoost model, the prediction accuracy of the weighted fusion model is improved by about 2.1%;Compared with the LightGBM model, the prediction accuracy of the weighted fusion model is improved by about 3.6%.In summary, the fusion model effectively improves the prediction accuracy of a single model.

The 23 input features are ranked according to the feature importance measure features_importance_ score of the above algorithm, as shown in Figures 2, 3, 4, and 5.

According to the common important characteristics of the model, it is found that the user's personal behavior characteristics are more correlated with the prediction of customer churn. From the user's point of view, the customer churn factors can be summarized into two aspects: The first is the user's personal needs and the degree of personal attention to information; the second is whether the price of the hotel and the details of the hotel page are attractive to the user.

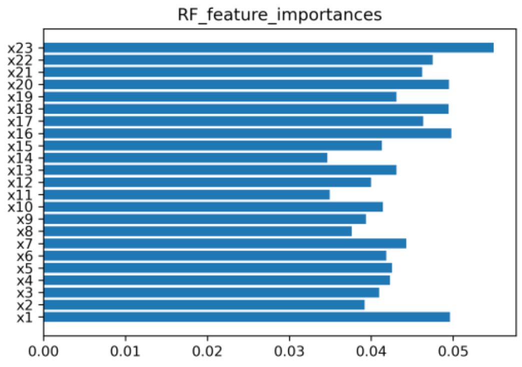

Figure 2: Random Forest feature importance

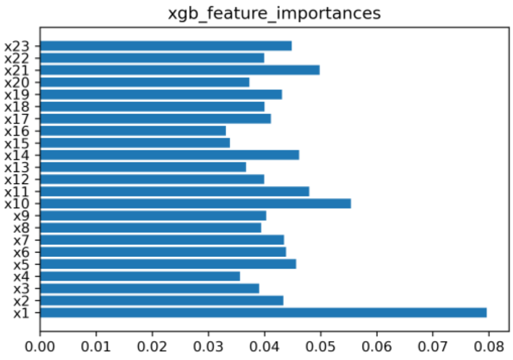

Figure 3: XGBoost feature importance

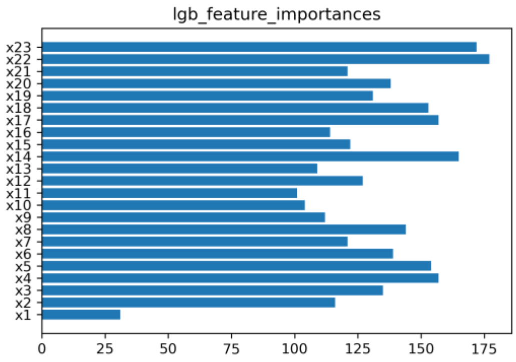

Figure 4: LightGBM feature importance

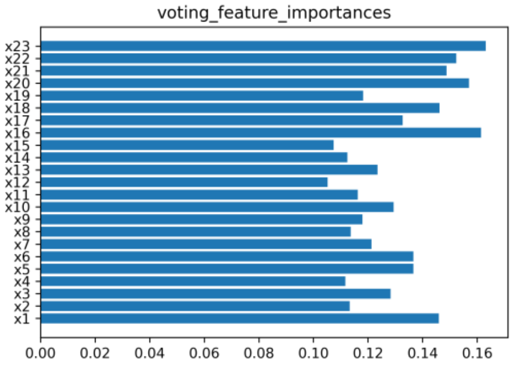

Figure 5: Fusion model feature importance

Finally, this paper uses the test set data and uses the kernel density function method to study the customer distribution. Predict it according to the sample and directly calculate its probability distribution, as shown in Figure 6. It can be seen that the predicted customer churn probability is concentrated between 0.6 and 0.9, indicating that most of the customers in this sample are predicted to be classified as 1, that is, they are judged to be lost customers. Companies should take retention measures to improve corporate profitability. Starting from the two factors that affect the loss of customers, on the one hand, it is to track customer dynamics in real time according to the data obtained from the web page, and to grasp the preferences of consumers to make accurate recommendations; The user's praise has increased, and the hotel room price has been reasonably arranged.

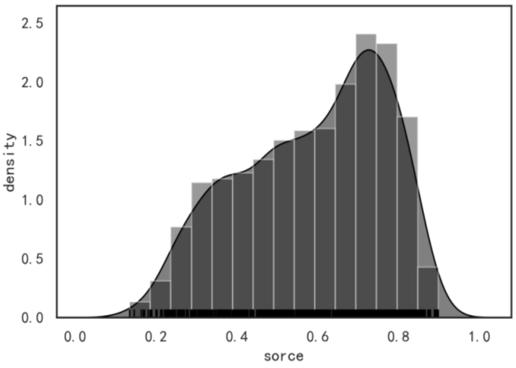

Figure 6: Fit map of the customer loss distribution for the kernel density estimation

4. Conclusions and discussion

In this paper, the weighted fusion model based on mutual information feature selection is efficient. This model can mine the customer loss situation for Ctrip, and if the company wants to effectively stabilize customers, it should collect customer behavior characteristic data, and it can also do: 1. One of the advantages of e-commerce is that it is easy to obtain visitor records, and the platform can retain huge amounts of browsing data. Provide customers with accurate recommendations based on their browsing history on the web, which makes it easier to attract visitors to the app for long periods of time; 2. According to Wirtz B W, Lihotzky N (2003) study of customer retention problems, customers are more likely to be driven by the psychology of intuitively receiving more enjoyment for the same effort and are more willing to choose to cooperate with the platform. Organizing promotions to recommend preferential options to customers, or provide appropriate free services;3. Allowing customers to have a good user experience will increase the probability of customers' secondary consumption, so it will increase the stickiness of customers to cooperate with e-commerce companies. Strengthen hotel management, give customers a good consumption experience, and do a good job in communication with customers.

References

[1] Aronoff S. Voting system [J]. Advances in Computers, 2021, 121: 495-500.
[2] Chen Peng. Study on customer loss prediction and influencing factors of Ctrip hotel based on Stacking [D]. Central University for Nationalities, 2021.

[3] Du Le. B2C E-commerce Enterprise Customer Classification Research [D]. Northern University of Technology, 2014.

[4] Fernandez-Peralta R, Massanet S, Mir A. A New Edge Detector Based on SMOTE and Logistic Regression [J]. 2017.

[5] Lee Jin. Crisis analysis of Ctrip network customer loss [D]. Guizhou University of Finance and Economics.

[6] Qiu Y, Li C. Research on E-commerce User Churn Prediction Based on Logistic Regression [C]// 0.

[7] Wirtz B W, Lihotzky N. Customer Retention Management in the B2C Electronic Business[J]. Long Range Planning, 2003, 36(6): 517-532.

[8] Yu Xiaobing, Cao Jie, Gong in Wu. Customer Loss Research Review [J]. Computer Integrated Manufacturing System, 2012, 18 (10): 11.