ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 80

1st International Conference on Business, Economics, Management Science (BEMS 2019)

P2P Default Risk Prediction based on XGBoost, SVM and RF Fusion Model

Guanlin Li [a], Yuliang Shi [b], Zihao Zhang [c]

Beijing University of Technology, Beijing, China.

[a] 13021093733@163.com, [b] shiyl@bjut.edu.cn, [c] 379882792@qq.com

Abstract. In the P2P platform, the problem of overdue repayment of users often occurs. This phenomenon seriously damages the interests of the platform and creditors. Therefore, how to improve and improve the risk monitoring capability of the P2P online lending platform and reduce the investment risk of investors is the future development of the P2P online lending industry. Very important question. To solve this problem, this paper proposes a P2P default risk prediction model based on XGBoost, SVM and RF fusion model. The model uses the stacking model set framework to model XGBoost, SVM and RF, and combines the advantages of high accuracy, robustness and generalization ability of the three models. The proposed fusion model has better prediction. Effect.

Keywords: risk prediction, XGBoost, SVM, RF, fusion model.

1. Introduction

P2P (peer to peer lending), which means personal-to-individual, peer-to-peer lending, is a business model that gathers small amounts of money and borrows money to people with financial needs. It is an Internet financial service product. It belongs to the private small-scale lending, which can meet the needs of individual funds, develop a personal credit system, and improve the utilization rate of idle funds.

Because the high returns and low thresholds of online lending satisfy the financial needs of investors and financiers, this emerging lending method was once favored and developed rapidly. However, there are more and more platforms that have problems, mainly because the P2P platform credit system is imperfect and users have overdue repayment behavior. Therefore, how to effectively control and prevent users from overdue repayment behavior is a problem that the platform needs to solve while the number of users grows and the scale of business expands. Therefore, constructing an efficient P2P default risk prediction model, identifying users who will overdue payments, and reducing overdue transactions will have important practical significance for the P2P platform.

2. Overview of P2P Overdue Risk Assessment

As an important innovation of Internet finance, P2P network lending plays an important role in solving the problem of personal and SME lending. However, its default risk has been plaguing the development of the P2P platform. It is an important risk control method to prevent P2P users from overdue by analyzing and counting the information related to the credit information of P2P platform users. However, the current P2P platform faces a problem in the analysis and processing of user data: the number of platform users is increasing rapidly, the user data growth rate is increasing rapidly, and the type dimension of user data is also rapidly expanding, so that the lending platform judges according to simple user data information. The traditional method of whether a user defaults cannot meet the requirements of today's large user volume, multiple data types, and high accuracy of risk prediction.

Therefore, based on the actual P2P platform business, the platform user's multi-type behavior data and transaction data are fully utilized, and the user default risk prediction model is constructed according to the processing requirements that meet the actual business requirements, which can solve the existing problems of the P2P risk control business. Improve the accuracy of risk prediction and reduce the bad debt rate of P2P platform.

In the selection of risk prediction models, the main methods include XGBoost, support vector machine, random forest and logistic regression. Yan Sumei used the decision tree and support vector

machine algorithm to construct the borrower default risk assessment model in the platform. It is finally confirmed that the decision tree and the support vector machine model can effectively predict the default probability of the borrower. Xu Tingting proposed that the random forest algorithm can tolerate noise better, is not easy to produce over-fitting, and has high stability. Compared with the traditional single-classifier model, it can better deal with credit risk assessment. Xu Chenxin's research shows that in the three models based on trees, the XGBoost model is the best in the credit field, and the widely used GBDT+LR model is only better than the decision tree.

From the above-mentioned risk prediction research background and results experience, it can be found that most models directly use a single prediction model based on risk assessment indicators, and few use multiple model fusions to improve the model prediction accuracy and generalization ability. Therefore, this paper uses the stacking model fusion framework to fuse the three models XGBoost, support vector machine and random forest which perform well in the example prediction, and construct a new fusion model to improve the prediction ability of the model.

3. P2P Overdue Risk Prediction Fusion Model

3.1 XGBoost, SVM and RF Model Fusion Ideas

XGBoost, SVM and RF are excellent algorithm models in the two-class classification algorithm. Using these three models as the base model can greatly improve the accuracy of the model after fusion. XGBoost has the highest accuracy, but it is easy to overfit. Random forest has good generalization ability under the premise of ensuring accuracy. SVM has the lowest accuracy among the three, but the robustness is the best. In order to combine the advantages of the three, to make up for the shortcomings, this paper based on the prediction results of the model combination.

The method firstly uses three models as the base classifier to train, and obtains the prediction result. The prediction result is taken as the feature. The model is trained by logistics, and the weights of the three models are obtained, and the weighted processing is performed to obtain the final prediction model.

The model fusion ideas is shown in Fig. 1.

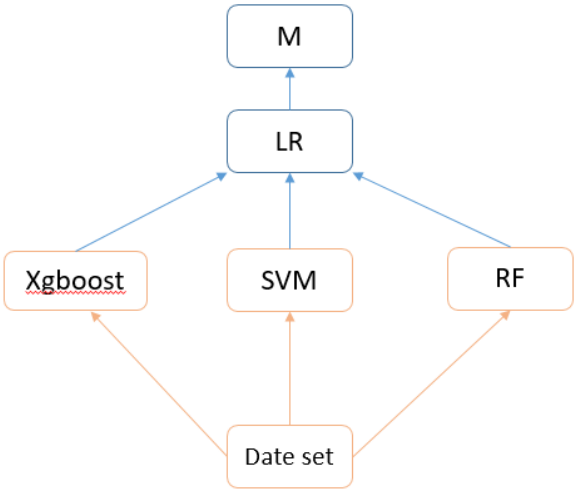

Fig. 1 Model fusion

3.2 Model Fusion Process

According to the optimization idea, this paper designs a specific fusion process, as shown in Fig 2 below.

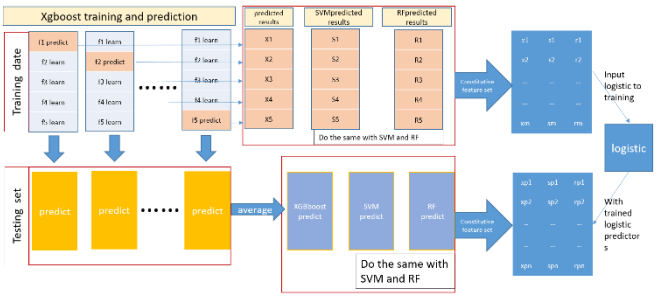

Fig. 2 Fusion process

The First, we use K-Fold cross-validation method to divide the training set. The K value is 5, and the training data set is divided into 5 groups as f1~f5. The first time f1 is used as the verification data set, and the remaining f2 The ~f5 data set is used as a training set. Then use XGBoost to train and predict, and obtain the prediction result X1 and the model XGBoost1. At this time, the obtained XGBoost1 is used to predict the test set to the prediction result P1. Because we are doing 50% cross-validation, we use the same method to carry out the remaining 4 trainings and predictions, and finally get the prediction results of the verification set X1, X2, X3, X4, X5, the prediction results of the test set P1, P2, P3, P4, P5.

After completing the whole process of training and prediction with XGBoost, it can be found that X1, X2, X3, X4, and X5 are the prediction results of the entire training set, and the matrix of 4000 rows and one column after the splicing is recorded as Xn. For the prediction results of the test set P1, P2, P3, P4, P5, we obtain better prediction results by adding and then averaging, which is recorded as XP. The above is the complete execution flow of a model in this model fusion. The SVM model and the RF model are used to repeat the above steps to obtain the prediction results Sn, Rn of the training set, and the prediction results SP, RP of the test set are obtained. Finally, we input Xn, Sn, and Rn as features into the logistic training to obtain the second layer model, and then predict the features of XP, SP, and RP, and obtain the prediction results of the prediction set.

4. Experimental Analysis

4.1 Data Sources and Overview

The data comes from well-known P2P financial companies, and as of August 31, 2018, the cumulative number of borrowing users has been 13.4 million. The public data set using the website has more data dimensions, and the second is that the data is more realistic. Persuasive. This user data contains user data for more than 220 fields. The data includes three data sheets, the borrower's basic information table (third-party data fields by time period, academic status fields, historical behavior fields, borrower transaction time fields, primary key fields for each loan), and the borrower's The information modification record table (modification content, modification time) and the borrower's login record table (operation code, operation category, login time). The user history behavior data of these 220 fields is used to predict whether the user will overdue in the next six months. The specific field information is as shown in Fig 3 below.

Variable name	Variable declaration
idx	The primary key of each loan can match the idx in the other two files.
Target	Default label (C1= loan default, 0= normal repayment)
UserInfo*	Borrower characteristics field (24 variables)
Education Info*	Academic record field (C8 variables)
ThirdParty_Info_ PeriodN_*	Third party data time period N field (a total of seven time periods, 17 variables for each time period)
S ocialNetwork*	Social network fields (17 variables)
LinstingInfo	Loan transaction time
W eblogInfo_*	Info network behavior field (58 variables)
ListingInfol	Loan transaction time
UserupdateInfo1	Modify the content
UserupdateInfo2	Modify the time
LogInfoX	Login operation (three variables)

Fig. 3 Data description

In order to make the user data better training model, speed up the model to solve the optimal solution speed and improve the prediction accuracy, the data is cleaned, data standardized and feature extraction work. Delete or replace the dirty data in the dataset, such as missing data, contradictory data, etc., using data unification, data discretization, and format conversion processing to make the data quality and training effect, delete the abnormal value, and delete Values and outliers are processed for category imbalance. Features that are more relevant to the prediction results are extracted, and then more effective features are constructed to simplify the computational complexity and improve the prediction effect of the model.

4.2 Valuation Indicators

In the classification problem, the model's generalization ability and predictive ability are usually used to evaluate the pros and cons of the model. The generalization ability of the model refers to the applicability of the algorithm to the new sample, that is, the prediction accuracy of the model for similar data. The predictive ability of the model is the accuracy of the model prediction, that is, the difference between the predicted result of the model and the real result, and the predictive ability of the model is generally measured by the following performance indicators.

Precision is also called accuracy. It is defined as the real case divided by the real case plus the false positive case (the data point that is actually a counterexample is classified as a positive example), which reflects that the data found by the model is actually related data. Ability.

The recall rate (recall) is also called the recall rate, which is accurately defined as the real case (the data of the positive case of the correct classification) divided by the real case plus the false counterexample (actually the positive data point is classified as a counterexample). It is the ability of the classification model to identify all relevant data.

Precision - the recall rate trade-off, in some cases, we need to sacrifice one of the indicators of accuracy and recall to maximize another. For example, in the overdue repayment forecast, in the initial loan overdue inspection of the user's follow-up inspection, we want to find all users who are actually overdue, and the recall rate is close to 1. If the cost of follow-up examination is not very high, we can accept lower accuracy.

The ROC curve shows how the relationship between recall and accuracy changes when changing the threshold that is identified as a positive example in the model. The ROC curve plots the true case rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis. TPR is the recall rate, and FPR is the probability that a counterexample is reported as a positive example. The ROC curve can comprehensively and accurately judge the different performances of the model under different thresholds, and it is very good for evaluating the performance of the unbalanced data classifier. The auc value is the numerical value of the roc.

4.3 Experimental Results

The prediction model constructed in this paper is based on the fusion model of XGBoost, SVM and RF model commonly used in default prediction. Firstly, these three models are used as the base classifier for training, and the prediction results are used as features. Finally, the model training is carried out with logistics. The weights of the three models are obtained and weighted to obtain the final prediction model. The resulting model can be used for risk prediction of P2P lending overdue.

In order to verify the prediction effect of the fusion model, three common models and fusion models are used for training and prediction. The predicted results are shown in Table 1.

Table 1. Model prediction result

model	precision	recall rate	F1
XGboost	0.892	0.866	0.878
RF	0.833	0.811	0.821
SVM	0.840	0.823	0.831
Fusion mode	0.901	0.881	0.890

It can be seen from the above table that the fusion model has better prediction effect on user default prediction than XGBoost, SVM and RF single model. The results show that the fusion model has better accuracy and recall rate for the identification of default users, indicating that the risk prediction probability value of the model output is more effective, and the fusion model can improve the risk prediction accuracy.

In order to visually demonstrate the model performance and effectiveness of the fusion model and XGBoost, SVM and RF in risk prediction, we have drawn the ROC curve. The ROC curve comparison method is simple and intuitive. The accuracy of the analysis method can be observed through the illustration, and the naked eye can be used. make judgement. The ROC curve combines sensitivity and specificity in a graphical manner, which accurately reflects the relationship between the specificity and sensitivity of an analytical method and is a comprehensive representation of the accuracy of the test. Fig 4 is a comparison of the fusion model and the single model ROC curve.

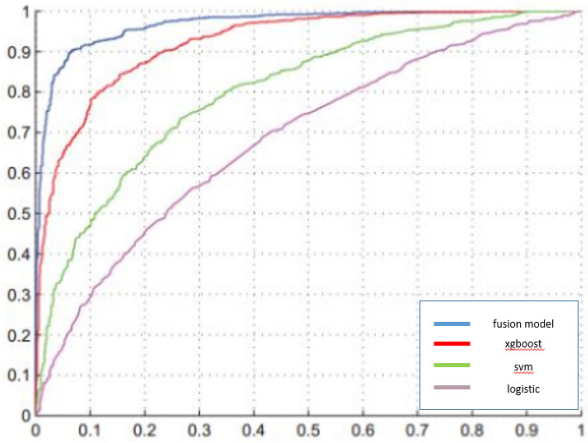

Fig. 4 Model ROC curve

5. Conclusion

Through the empirical analysis of several classification algorithms and the model fusion algorithm, it is found that using XGBoost, SVM and RF fusion algorithm has better adaptability to P2P lending platform user data than other algorithms. The model can more accurately predict the users who borrowed overdue payments, that is, the minority of the sample. It has important reference significance for the operation decision of P2P platform, and provides better solutions and methods for overdue repayment risk control.

However, in the process of predicting the breach of contract, due to the imbalance of the sample data, many violations are judged as non-default, although the model performance is improved by the boundary domain oversampling method. It is optimized based on the premise that the data category imbalance ratio is known. In the process of forecasting in the future, the data imbalance ratio will change, and the model needs to be self-updating according to the specific situation.

References

[1]. Emekter R, Tu Y, Jirasakuldech B , et al. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending [J]. Applied Economics.

[2]. Zhao H, Qi L, Wang G, et al. Portfolio Selections in P2P Lending: A Multi-Objective Perspective [C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2016.

[3]. Tan J, De Silva D G. Better Off Or Worse Off: An Economic Analysis Of Online P2p Lending Market [C]// International Conference on Service Systems & Service Management. 2011.

[4]. Serranocinca C, Gutierreznieto B, Lopezpalacios L. Determinants of Default in P2P Lending [J]. Plos One, 2015, 10 (10): e0139427.

[5]. Huang Zhen. Research on credit risk assessment of Chinese P2P borrowers based on BP neural network model [D]. 2015.

[6]. Huo Jianglin, Liu Surong. Research on Credit Risk Assessment of P2P Network Loan Platform Borrowers [J]. Financial Development Research, 2016 (12).

[7]. Luo C, Xiong H, Zhou W, et al. Enhancing investment decisions in P2P lending:an investor composition perspective [C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. DBLP, 2011.