Predicting Default Situations in the P2P Lending
Based on Machine Learning

Chenzhou Mo(✉)

Beijing Normal University, Zhuhai 510630, Guangdong, China
1349696871@qq.com

Abstract. As a flexible and efficient new financial format, P2P lending suffers from breach of contract and lack of trust due to the uneven credit, income, and region of borrowers. Therefore, we plan to use machine learning algorithms to predict the default situation in the P2P market in the future, and compare the prediction accuracy of various models to find the optimal default prediction model. The research data in this article includes P2P lending data from 33,105 users in 50 states in the United States. It includes variables such as investment income loss percentage, borrower income, and loan term. To simplify subsequent analysis, missing values were cleaned and data on borrower state and loan date were classified and simplified. T-test and chi-square test were used to preliminarily analyze data-type data and categorical-type data, and the results showed that all relevant variables are statistically significant and need to be considered in subsequent research. To further determine the significance of each variable in the default situation, a logistic regression model was introduced, which has practical significance for lending platforms in user selection. Finally, four types of models were used for constructing default prediction models, which are logistic regression, decision trees, random forests, and GBDT. The ACC and AUC values of different models on the training and testing sets were compared. The conclusion is that the GBDT model has the highest prediction accuracy and a high AUC value, which can serve as a prediction model for future lending platforms to predict user default situations.

Keywords: P2P · Logistic regression model · default · Machine learning · Gradient Boosting Decision Tree Model · T-test · Chi-square test

1 Introduction

Peer-to-peer lending has become a popular means of financing in recent years, used to meet the funding needs of users. P2P refers to the lending model of aggregating small amounts of funds and providing them to borrowers, and is a specific operating model of internet finance. With the development of the internet, the integration of P2P and internet finance has become increasingly high [1]. As a new type of financial business, P2P has the advantages of flexibility, transparency, and efficiency, but also has many problems, the most prominent of which is the risk of default. On P2P lending platforms,

S. Tehseen et al. (Eds.): ICEDBC 2023, AEBMR 258, pp. 607–613, 2024.
https://doi.org/10.2991/978-94-6463-246-0_73

due to differences in credit, income, and regional indicators of borrowers, the quality of users varies, making it impossible for some borrowers to repay their loans, resulting in investors losing their principal and returns. The lack of effective risk warning makes it difficult for investors to make accurate judgments about credit risk. Relying on the internet, P2P lending reduces the authenticity of transactions. Therefore, obtaining early warning of debt default and minimizing the risk of default is of great significance for the safe development of the lending market.

Beaver and Altman first proposed using discriminant analysis to predict corporate financial distress in 1968 [2, 3]. With the development of information technology, machine learning methods have been introduced into default risk prediction. In China, machine learning was first applied to the prediction of individual credit risk on P2P lending platforms. Li et al. used the Classification and Regression Tree (CART) algorithm to predict financial risk for listed companies and confirmed its effectiveness, performing better than Support Vector Machine (SVM) and other methods in terms of predictive performance and significance testing [4]. Cai [5] used the random forest algorithm to analyze default on P2P lending platforms and used the SMOTE algorithm to balance the lending set. Ma Chunwen [6] et al. studied P2P credit risk factors based on a random forest classification model. Akanmu [7] et al. proposed a P2P lending prediction method that improved decision tree models, which had good fitting effects on small business loan data in the United States. Liu [8] et al. proposed the use of rough sets for credit evaluation, which can more accurately identify credit poor users. Although the above methods have good predictive accuracy, they are complex in structure and difficult to analyze specific indicators, making them difficult to apply in practical situations.

This article intends to use relevant machine learning methods to analyze the lending situation of users in 50 states on P2P lending platforms in the US, and predict their default situations. We analyzed nearly 30 indicators of these users, established a model based on these indicators, and used the Gradient Boosting Decision Tree (GBDT) method to predict their default situations. We analyzed nearly 30 indicators of these users, built models based on these indicators, and employed various machine learning techniques to predict their default situations. By comparing the prediction performance of multiple models, we provide a comprehensive evaluation, and ultimately verified whether the GBDT model is the best predictive model for this experiment based on accuracy, with the aim of building a default assessment model with better warning functions.

2 Data Preprocessing and Analysis

2.1 Dataset

The data used in this study was collected from P2P lending platforms and includes loan data from 33,105 users across 50 states in the United States. The loan period ranges from 2001 to 2023. Variables that were deemed irrelevant or did not show any changes were excluded from the analysis, while variables with analytical significance are presented in Table 1.

Table 1. Summary statistics of each variable (variable type, name, and description).

Variable	Variable type	Meaning
estimated_return	numerical variable	Percentage of the expected return on investment that an investor expects to receive in a given loan, relative to the total investment amount
estimated_loss_rate	numerical variable	Percentage of expected investment loss that an investor expects to incur in a given loan, relative to the total investment amount
borrower_rate	numerical variable	Investment interest rate of the given loan
listing_amount	numerical variable	Loan amount
listing_monthly_payment	numerical variable	Monthly loan repayment amount that the borrower needs to pay
dti_wprosper_loan	numerical variable	Ratio of debt to income
stated_monthly_income	numerical variable	Borrower's income
loan_origination_date	numerical variable	Date of obtaining the loan
scorex	numerical variable	Borrower's credit score, used to measure the borrower's creditworthiness
partial_funding_indicator	categorical variable	Indicator of whether the borrower is willing to accept partial payment of the loan
income_verifiable	categorical variable	Verifiability of the borrower's income
prosper_rating	categorical variable	Simplified evaluation score of borrower's integrity based on Scorex
borrower_state	categorical variable	State of borrower
employment_status_description	categorical variable	Employment status of the borrower
prior_prosper_loans_active	categorical variable	Repayment status of the borrower's loan.

2.2 Data Preprocessing

Firstly, we examined the missing value situation of each column of data and found that three features had missing values, as shown in the Table 2. We handled the missing values by not counting them in the statistics for categorical data and replacing them with NAN. For numerical data, missing values were replaced with the mean value.

Due to the wide distribution of borrowers across as many as 5889 cities, which is not conducive to subsequent regression and prediction analysis, the borrower's state can reflect their location information to some extent. Therefore, we removed the borrower's city indicator. For the convenience of subsequent analysis, the 50 states of the United

Table 2. Missing value statistics.

Variable	Number of missing values
scorex	10
borrower_state	928
borrower_city	928

Table 3. The division of the United States into 50 states.

Types	States
A	AK,MD,NH,MA,ND,VA,UT,WA,CT,MN
B	AL,AZ,CA,CO,DE,GA,HI,IL,IN,IA,KS,MI,MO,MT,NE,NV,NJ,NC,OH, OK,OR,PA,RI,SC,SD,TN,TX,VT,WI,WY
C	MS,WV,AR,NM,NY,LA,FL,KY,ME,ID

States were ranked according to their per capita GDP, and divided into three levels: A, B, and C. A and C levels respectively included the top and bottom 10 states in terms of per capita GDP, while the B level included 40 states with moderate per capita GDP. The division results are shown in Table 3 (state names are abbreviated).

For the loan_origination_date, which represents the date when the loan was obtained, in order to simplify subsequent classification, the date was converted to Monday-Sunday. To facilitate subsequent regression and prediction, all categorical data were converted to numerical data through one-hot encoding.

2.3 Variable Selection

To obtain the initial relationship between default and various indicators, the significance of each indicator in studying defaults was preliminarily judged. T-tests and chi-square tests were conducted on both numerical and categorical variables, and the obtained p-values are as follows. If $p < 0.05$, H0 is rejected at the level of $\alpha = 0.05$, indicating that the difference is statistically significant.

From Table 4, it can be seen that all variables have a p-value less than 0.05, indicating that the differences are statistically significant. H0 is rejected at the level of $\alpha = 0.05$, and therefore all variables need to be considered in the analysis of default.

3 Establish a Forecast Model

3.1 Logistic Regression

First, construct a logistic regression default prediction model based on the above variables to observe the coefficients and corresponding P-values of different variables in logistic regression, and determine the importance of different variables in logistic regression., and the results are shown in Table 5.

Table 4. T-test and chi-square test

Numerical variable	**Categorical variable**
listing_amount**	income_verifiable**
estimated_return**	partial_funding_indicator**
estimated_loss_rate**	prosper_rating**
borrower_rate**	borrower_state**
listing_monthly_payment**	employment_status_description**
dti_wprosper_loan**	prior_prosper_loans_active**
stated_monthly_income**	

Table 5. Results of logistic regression analysis

Variable	coef	Variable	coef
amount_funded	3.465e−05**	prosper_rating_C	0.3576**
amount_remaining	−0.0003*	prosper_rating_D	0.4082**
percent_funded	−3.1972**	prosper_rating_E	0.4863**
estimated_return	−6.2560**	prosper_rating_HR	0.5932**
estimated_loss_rate	−8.3862**	employment_status_description_Full-time	−0.1044**
borrower_rate	11.9950**	employment_status_description_Not available	−0.7396*
dti_wprosper_loan	−3.721e−07*	employment_status_description_Not employed	0.2120*
stated_monthly_income	−3.473e−05**	employment_status_description_Other	0.5637**
partial_funding_indicator	0.4683**	employment_status_description_Part-time	−0.4902**
income_verifiable	−0.3455*	employment_status_description_Retired	0.2370**
prosper_rating_AA	−0.7497**	employment_status_description_Self-employed	0.2865**
prosper_rating_B	0.1486*	Intercept	0.5925
prior_prosper_loans_active_1	0.3452**		

Note: * and ** denote statistical significance at the 10% and 5% levels, respectively.

If $P > 0.05$, the impact of the indicator on default is less significant. The p-value less than 0.05 indicates that these variables have a significant impact on whether the user defaults in the logistic regression model. From the results table, it can be seen that most of the variable p-values are significant, with only a few variables being non-significant. Since the logistic regression coefficient can be used to measure the degree of influence of a variable on the target variable, as well as its relative importance, it represents the strength of each variable's effect. If the coefficient value is high, it indicates that the variable has a greater impact on the accuracy of the model. The results of this model have practical significance for screening and controlling user default situations. Controlling loan interest rates and selecting lower rates can help reduce default risk, while increasing the percentage of expected investment loss amount and return amount to investment amount and screening such users can also help reduce default risk.

Table 6. Accuracy and AUC values of each model on the test set

model	ACC	AUC
Logit	0.7234	0.6808
Decision Tree	0.7292	0.7002
Random Forest	0.7375	0.7202
GBDT	0.7614	0.7774

3.2 Machine Learning

After preprocessing the data of 33,105 users, the users were divided into training and testing sets in an 8:2 ratio. The constructed model mainly uses Logit, Decision Tree, Random Forest, and GBDT models. To improve the prediction accuracy, random grid search was used to optimize hyperparameters for all four algorithms. After adjusting the parameters, the optimal model parameter combination for P2P default prediction was established. After adjusting the model parameters, the four models were used to predict and calculate the evaluation indicators for the testing set, as Table 6.

The higher the ACC and AUC scores, the better the model's ability to predict defaults. Among the four models in the comparison table, GBDT had the highest AUC value, indicating a relatively more accurate prediction. Comparing the prediction accuracy of the testing set, it can be seen that the GBDT method had the highest accuracy, with a value of 0.7614. These results indicate that in this dataset, the GBDT model has the most advantages and the best generalization ability, and can be used as a model for predicting default situations in P2P and other online lending platforms.

4 Conclusion

By analyzing the significance of each indicator through the logistic regression model, it provides a reference for the analysis of user indicators by lending platforms: focus on and select users with low loan interest rates and high investment returns and loss ratios. After comparing the prediction results of four machine learning algorithm models on this default dataset, it is found that the performance of GBDT is higher than that of other models. The prediction performance of logistic regression, decision tree, and random forest models are quite similar. The GBDT method has fast calculation speed during the prediction stage and can perform parallel calculations between trees. In dense data sets (such as P2P data), it has good generalization and expression ability, good interpretability and robustness, and can automatically discover high-order relationships between features. It also does not require special preprocessing of data such as normalization. In the predictive experiments of machine learning, we concluded that GBDT has higher accuracy and AUC values in the testing set than the other three groups. Therefore, we can use the GBDT algorithm for default prediction in various loan websites.

For some online lending platforms, risk control is a challenge. The platform can consider using machine learning algorithms for more accurate predictions, based on the

machine learning prediction results to obtain the probability of customers defaulting and reveal the likelihood of transaction completion, optimize the recommendation ranking for borrowers, and assist lenders in decision-making.

References

1. Xiuting, Q., Liyun, Q., (2023) On the Improvement of Risk Regulation System for P2P Online Lending Model in China. China Price.
2. BEAVER, W. H., (1966) Financial ratios as predictors of failure[J]. Journal of accounting research, 4:71–111.
3. ALTMAN E., I., (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. Journal of finance,23:589–609.
4. Li, H., Sun, J., Wu, J., (2010) Predicting business failure using classification and regression tree: an empirical comparison with popular classical statistical methods and top classification mining methods. Expert systems with applications,37:5895–5904.
5. Cai, H. J, (2019) Analysis of P2P online lending default based on random forest[J]. Journal of Physics: Conference Series 1237(2):1-6.
6. Chunwen, M, Hui, Z, Qi, L, (2019) Study on Credit Risk Factors of P2P Online Lending Projects Based on Random Forest Classification Model. Journal of Jilin University (Social Sciences Edition), 59: 39-48.
7. AKANMU, S.A, GILAL, A.R, (2019) A boosted decision tree model for predicting loan default in P2P lending communities. International Journal of Engineering and Advanced Technology, 9: 1257–1261
8. Liu, G. J, Zhu, Y (2006) Credit assessment of contractors: a rough set method. Tsinghua Science&Technology,11:357–362.