Service Optimization of p2p Online Loan Platform Based on Big Data Analysis

Tiantong Yang

School of Finance, Shanghai Jiaotong University, Shanghai, China

ytt0919@sjtu.edu.cn

Abstract. In view of the credit risk loss brought by incomplete loan transactions to the online P2P lending platform, based on the data set of Prosper Company, this paper, on the one hand, establishes machine learning models such as logistic regression, decision tree, random forest, etc. to predict whether the loan application can be completed, so as to optimize the ranking recommendation logic of the platform, and put forward suggestions according to the borrower's situation to reduce the ultimate credit risk; on the other hand, formulates the OLS linear regression model, so that through exploratory analysis of loan data and coefficient analysis of the regression model, important characteristics highly related to loan default are obtained, including total income, occupation type, working life, debt-to-income ratio, loan amount, loan term, etc., which helps the platform to better identify valuable potential customers.

Keywords: P2P lending platform; loan completion improvement; machine learning algorithm; linear regression model

1 Introduction

With the rapid development of internet finance, the demand for loans from credit platforms is increasing and the situation facing the platforms is more complicated. Since artificial intelligence technology offers a means to effectively and comprehensively mine and analyze big data, addressing the challenges posed by financial risks and financial technology, this paper would like to utilize classification algorithms commonly employed in machine learning [1], such as logistic regression, decision tree [2], and random forest [3], to enhance the predictive power of these platforms.

P2P peer-to-peer lending is a business model that collects small amounts of funds and borrows them to fund demanders. Take Prosper, a relatively large peer-to-peer lending platform in the United States, as an example, such a platform only provides information to help borrowers and investors connect, while itself does not directly participate in transactions. In addition, the platform does not provide guarantees and is not responsible for recovering overdue loans. To achieve risk control, Prosper mainly uses credit rating as an indicator, that means the better the credit, the lower the service fee will be charged. Moreover, a customer group model is set up for high-

F. Balli et al. (eds.), *Proceedings of the 2023 2nd International Conference on Economics, Smart Finance and Contemporary Trade (ESFCT 2023)*, Advances in Economics, Business and Management Research 261,
https://doi.org/10.2991/978-94-6463-268-2_15

rated borrowers to enjoy preferential interest rates while also monitoring each other to reduce default.

Due to the weak post-loan management of the platform, lenders tend to adopt more conservative investment strategies, including a preference for multiple small-scale loans and increased focus on credit rating. However, it remains unclear whether the credit ratings given by the platform are reliable indicators of the borrower's likelihood of receiving the loan. Meanwhile, in addition to the situation where the lender is unwilling to lend money, another reason why the transaction cannot be completed is that the applicant voluntarily cancels the loan after its approval. The occurrence of both conditions will result in a reduction in the service fee charged by the platform and negatively impact the operating efficiency and long-term reputation of the platform.

Previous data analysis on Prosper mostly focused on the prediction of the borrower's default. However, this paper notes that, there exists another reason for transactional failure, which is the cancellation at the onset of the process, regardless of whether the borrower initiates and then cancels, or the lender is unwilling to release the investment. Incomplete applications are not accounted for, thus, rendering the platform's primary reference metric, "credit rating", more likely to affect whether the borrower defaults in the repayment process [4]. Therefore, this indicator may not entirely align with whether the initial application is completed, which also leads to the platform still having room for optimization in this regard. On the basis of providing credit ratings, machine learning model may enable advanced big data mining for enhanced analysis [5], so as to optimize the recommendation ranking logic for borrowers, or put forward some suggestions, such as loan amount controls, etc., to reduce incomplete transaction occurrences. This could better align transactional parties and raise platform efficiency and accuracy [6].

In addition, conducting a linear regression analysis of the borrower's characteristics, including but not limited to occupation, income, and time, can assist the platform in identifying valuable customers during the marketing promotion phase. These customers typically demonstrate a greater propensity for loan needs and are less inclined to cancel their loan application.

2 Data preprocessing and analysis

A public dataset is obtained from Prosper.com, which contains 38 values and 114150 data. After deleting data columns with redundancy exceeding 30% and removing some insignificant features, 29 values are finally obtained. In addition, sequential coding and unique hot coding have been applied to transform some features which have multiple values, so they can be better understood by the algorithm. Extremum normalization is also adopted to transform the continuous features to make sure that each feature could be treated equally by the algorithm.

The data are roughly divided into three categories: loan information, basic information of borrowers and credit information of borrowers. Specific information is shown in Table 1.

Table 1. Categories of features

Loan information	total loan amount
	loan annuity
	whether to accept partial funding
	interest rate
	application time
	loan term
Basic information of borrowers	occupation
	months employed
	income
	state
	debt-to-income ratio
Credit information of borrowers	credit rating
	prior loans
	expected loss rate
	active loans

Firstly, loan information. It comprises several key features, including the total loan amount, loan annuity, whether to accept partial funding, interest rate, application time and loan term. For example, the acceptance of a partial loan means that Prosper allows the borrower to indicate when initiating the application, whether to accept a loan at a lower amount (usually the total amount of the application multiplied by 0.7). By comparing the proportion of applicants who accept a partial loan, the transaction is completed (53.4%) significantly less than the transaction cannot be completed (74.3%).

Secondly, the basic information of the borrower. The key characteristics are occupation, months employed, income, state and debt-to-income ratio. From the perspective of the proportion of transactions that cannot be completed, the proportion of occupations with low skills and instability is high, among which 51.51% are students, 23.4% are truck drivers, waiters, salesmen and laborers. Judging from the working life, most borrowers are concentrated within 5 years; as the length of their months employed increases, so does the proportion of transactions fulfilled.

Thirdly, the borrower's credit information is also a crucial factor in loan evaluation, comprising expected loss rate, credit rating, prior loans and active loans. The total number of prior loans is a reminder of credit, and the more repayments, the higher the percentage of transactions completed. Prosper's credit ratings are divided into seven grades: AA, A, B, C, D, E, and HR (High Risk). The corresponding default rates are AA 13.86%, A 14.47%, B 20.68%, C 23.32%, D 27.54%, E 30.37%, and HR 27%, respectively. The difference between AA and A grades is only 1%, while the difference between A and B grades is a significant 6%. The gap between the highest credit rating and the rest of the credit ratings is obvious. Such speculation may exist, and there is room for optimization in the recommendation logic of the Prosper platform.

3 Establish a forecast model

3.1 Construction of Random Forest Model and Parameter Adjustment

The data set is divided into 7:3 training set and test set. Firstly, the random forest model is used for prediction. This paper mainly adjusts several important parameters, such as the number of learners, the maximum number of selected features, the maximum depth of the tree, etc. Through grid search, the best parameters are obtained according to the scoring results.

The parameter adjustment process is as follows: Firstly, Increase the number of learners to improve the fitting ability of the model. The setting range is 100 ~ 200 and the step size is 10. The results show that the best number of learners is 130. Secondly, the maximum number of selected features is adjusted, the setting range is 1 to 20, and the optimal value is 1. Thirdly, the maximum depth of the tree is adjusted to improve the fitting ability of each sub-model, the setting range is 3 to 20, and the optimal value is 14. Finally, the model is re-trained and evaluated with the adjusted parameters and characteristic variables. Table 2 shows the results.

Table 2. adjustment of parameter

Index	Before adjustment	After adjustment
accuracy score	0.6956	0.7171
recall score	0.6205	0.6312
F1	0.8412	0.6559

The importance ranking of the features obtained from the random forest model is basically consistent with the previous data analysis, and some of the results are shown in Table 3:

Table 3. importance ranking of the features

1	Prior_prosper_loans	0.242279
2	Funding_threshold	0.146236
3	Prior_prosper_loans_active	0.077204
4	Estimated_loss_rate	0.070468
5	Estimated_return	0.063240
6	Monthly_payment	0.054610
7	Listing_amount	0.040878
8	Prosper_rating_A	0.037513

The importance ranking of the features obtained from the random forest model is basically consistent with the conclusions drawn from the exploratory data analysis. The total number of existing liabilities, whether there is a partial loan, total repayment of historical loans, loss rate, rate of return, loan amount, credit rating, working hours, etc. rank in the top 10, among which the contribution of the top 8 features accounts for about 70%, which is consistent with the previous forecast. It shows that the above variables play an important role in customer default forecast.

3.2 Comparison and Summary of Multiple Models

The evaluation results of each model are obtained by using the commonly used classification algorithm, and then the evaluation results of each model are compared. Table 4 shows that the decision tree model is more effective.

Table 4. Comparison of models

Model	Accuracy rate	Recall rate	F1
Logistic	0.5809	0.6031	0.5918
Decision Tree	0.7683	0.7867	0.7774
Random Forest	0.7171	0.6312	0.6714

4 OLS regression analysis

The above independent variables such as loan amount, credit rating, occupation, etc. are put into OLS linear regression model for verification to see which factors significantly affect the ultimate completion of the transaction, and to explicate the positive and negative coefficients. While the relationship between the independent and dependent variables is not wholly linear, we believe that in some cases the linear relationship is predominant. Henceforth, the error term s introduced to represent the nonlinear relationship among the variables. Thus, a multiple regression model is available from the above assumptions.

The autocorrelation between some variables is considered, for example, the loan annuity is actually determined by the total loan amount, the loan term and the interest rate, so that it is not included in the model; Table 5 shows some significant regression terms and their coefficients using the 5% confidence level T-value test.

Table 5. some coefficients

Funding amount	Partial funding	Interest rate	Loss rate	Active loans	Prior loans	Months employed
-0.080	-0.142	0.131	-0.418	-0.215	0.328	0.056

Firstly, considering the loan information, the larger the total loan amount, the more unfavorable it is for the transaction to be completed. From the perspective of screening applicants by investors, the risk of diversifying small investments is lower than that of a single large investment; those who choose to accept part of the loan (with a value of 0, and those who choose to accept only the full amount of the loan with a value of 1) will not have easy access to the loan, possibly because this is a signal that implies whether the borrower's need for access to the loan is strong, which makes it difficult to cancel the loan. In addition, the loan interest rate also plays a positive role. It is worth noting that there is a seasonal difference in whether the transaction can be completed. In the independent variable coefficients represented by the three months of 1, 2 and 3, a significant positive impact of about 0.1 is found, therefore, the transaction is easier to be completed during the winter period.

Secondly, in terms of the basic information of the borrower, the length of their employment tenure exerts a certain positive influence. However, the impact of different occupations varies significantly, as shown in Table 6. Low-skilled or unstable occupations, such as students, drivers, waiters, salespersons, and manual laborers, exhibit negative influence. On the contrary, certain high-income and high-skill jobs, including doctors, dentists, analysts, investors, etc. also exert negative impact. This may be due to the fact that individuals engaged in such professions are unlikely to suffer from financial constraints, and investors may not perceive the necessity for their investments hence, the transaction is not completed in the end. Affected by this, the negative impact of the income range is comparatively feeble. More likely, the negative impact of low income and high income dilutes the positive impact of medium income. The influence of state is relatively weak, among which Hawaii (HI) has obvious negative influence.

Table 6. occupation coefficients

student	driver	Waiter	salesman	labor	doctor	investor
-0.138	-0.912	-0.081	-0.045	-0.053	-0.108	-0.158

Finally, the credit information of the borrower shows a negative impact on the loss rate. In addition, the total number of loans that borrowers are repaying on this platform (mostly 0 or 1) has a significant negative impact, which implies a high debt ratio, making investors doubt their solvency. The historical total loan repayments, on the other hand, send a positive signal of good credit.

5 Conclusions

Based on the public data set provided by Prosper peer-to-peer lending platform, this paper predicts through machine learning algorithm and analyzes through linear regression model respectively. On the one hand, the platform can optimize the ordering of customer recommendations through the prediction results, reducing the situation that the transaction cannot be completed; on the other hand, by focusing on a number of salient features, the platform can more accurately find potential customers and carry out targeted promotion.

Similar to the peer-to-peer lending model of "light platform intervention" adopted by Prosper Company, the risk control is a challenge to both the platform and the lenders. Through analysis, it is concluded that the lenders tend to be more inclined to the characteristics of small loans and high credit ratings [7], while ignoring the impact of other factors. The primary approach for the platform to implement risk control is also through credit rating, but it is mainly pertains to the borrower, while whether the transaction can be completed depends on both of them [8]. Therefore, this indicator does not necessarily match the likelihood of completing a transaction. The platform can consider using machine learning algorithm to make higher-precision prediction, and after integrating all the features, put forward a new feature, namely, the probability that the loan application will be completed, and prompt investors for the possibility that the borrower will cancel the application, or the possibility that the borrower

will not be able to raise the specified amount eventually [9]. On the basis of this reference index, it is also possible to optimize the ranking of recommendations for borrowers and the grouping logic. It is no longer necessary to regard credit rating as the only grouping threshold, allowing lenders with good comprehensive performance to join the high credit group. Additionally, the platform used to charge rates based on credit ratings. Now the platform can also refer to the indicator of transaction completion probability. For borrowers with high probability, the current credit rating fee will be appropriately reduced to provide incentives to borrowers, thus improving the probability of transaction completion. Facing the borrower, the platform can make some suggestions to help them propose the appropriate borrowing amount based on their credit rating or the probability of completion of the transaction, and decide whether to accept part of the loan, thus improving the possibility of obtaining the loan. In terms of promotion, the platform can pay more attention to indicators such as time, occupation, region, etc. in addition to the more intuitive income and debt situation with borrowing, so as to more accurately screen and find target customers. Finally, for other similar third-party financing platforms [10], this idea can also be followed to optimize the service level, enabling more borrowers to secure loans and investors to invest. This, in turn, can bolster the platform's credibility and augment its service fees.

References

1. Kotsiantis, S.B., Zaharakis, I.D., Pintelas,P.E. (2007) Machine learning: a review of classification and combining techniques. Artificial Intelligence Review., 26:159-190. https://doi.org/10.1007/s10462-007-9052-3.
2. Quinlan, J.R. (1986) Induction of decision trees. Machine Learning., 1:81-106. https://doi.org/10.1007/BF00116251.
3. Breiman, L. (2001) Random Forests. Machine Learning., 45:5-32. https://doi.org/10.1023/A:1010933404324.
4. Pope, D.G., Sydnor, J.R. (2012) What's in a Picture?: Evidence of Discrimination from Prosper.com. Journal of Human Resources., 46:53-92. https://doi.org/10.1353/jhr.2011.0025.
5. Aslam, M., Kumar, S., Sorooshian, S. (2019) Predicting Likelihood for Loan Default Among Bank Borrowers. International Journal of Financial Research., 11:318-318. https://doi.org/10.5430/ijfr.v11n1p318.
6. Berkovich, E. (2011) Search and herding effects in peer-to-peer lending: evidence from prosper.com. Annals of Finance., 7:389-405. https://doi.org/10.1007/s10436-011-0178-6.
7. Li, H., Zhang, Y., Zhang, N., Jia, H. (2016) Detecting the Abnormal Lenders from P2P Lending Data. Procedia Computer Science., 91:357-361. https://doi.org/10.1016/j.procs.2016.07.095.
8. Loureiro, Y.K., Gonzalez, L. (2015) Competition against common sense. International Journal of Bank Marketing.,33:605-623. https://doi.org/10.1108/IJBM-06-2014-0065
9. Emekter, Tu, Jirasakuldech, Lu (2015) Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics., 47:54-70. https://doi.org/10.1080/00036846.2014.962222.
10. Guo, Y., Zhou, W., Luo, C., Liu, C., Xiong, H. (2016) Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research., 249:417-426, https://doi.org/10.1016/j.ejor.2015.05.050.