Risk Management in P2P Lending Markets

Jialun Lyu

Haide College, Ocean University of China, Qingdao, China

jialunlv@163.com

Abstract. P2P lending has garnered considerable attention and utilization owing to its minimal entry barriers, reduced expenses, and enhanced efficiency compared to conventional financial institutions. However, alongside its popularity, P2P lending markets also encounter risks, with personal credit risk emerging as the most salient concern. This study aims to explore the determinants that influence borrower risk preferences through the examination of loan data obtained from the Prosper platform.

Using a fixed-effects regression model, this study examines the relationship between key variables and risk preferences. Temporal factors, especially weekdays and weekends, are found to be important factors influencing investor risk preferences. Additionally, this study utilizes machine learning algorithms to develop a default risk prediction model in P2P lending. Through rigorous comparative analysis and experiments, the random forest model demonstrates robust predictive capabilities. Furthermore, a combined learning model utilizing voting and bagging techniques is constructed by integrating random forest, linear regression, and Xgboost models. This ensemble model provides auxiliary support for P2P lending platforms in recommending investable orders to investors.

The findings of this study provide valuable insights into risk management within P2P lending markets, particularly in terms of borrower risk preferences and the utilization of machine learning algorithms for risk prediction. The knowledge acquired from examining loan data from the Prosper platform carries practical implications for P2P lending platforms and risk management practitioners seeking to enhance risk assessment and control strategies.

Keywords: P2P lending; Prosper platform; risk preferences

1 Introduction

P2P lending entails the establishment of a peer-to-peer lending relationship and the completion of associated transaction procedures through online lending platforms. Compared to traditional financial institutions, P2P lending offers advantages such as low barriers to entry, low costs, and high efficiency, thereby garnering significant attention and application in recent years[1]. However, the P2P lending market has witnessed an increase in risks, with personal credit risk serving as a noteworthy concern. To gain a deeper understanding of the risks prevalent in the P2P lending market,

A. Bhunia et al. (eds.), *Proceedings of the 2023 International Conference on Finance, Trade and Business Management (FTBM 2023)*, Advances in Economics, Business and Management Research 264,
https://doi.org/10.2991/978-94-6463-298-9_44

this study analyzes the borrowing data from the Prosper platform to explore the factors that influence borrowers' risk preferences.

2 Empirical Analysis

2.1 Data and Sample

The data for this study is sourced from the Prosper official website, spanning from November 9, 2005, to October 14, 2008. After excluding samples with missing values and samples with significant outliers, the final sample size consists of 582,970 observations.

2.2 Descriptive Statistical Analysis

Table 1. The descriptive statistical analysis of the main variables

	N	Min	Max	Mean	Std
Monthly income	582970	0	4833333	540.90	15210.25
Monthly debt	582970	0	101500	885.39	944.68
Daily listings number	582970	3	5295	3112.73	707.21
Borrower maximum rate	582970	0	48	20.26	7.00
Funded percent	582970	0	43.590910	0.71	0.87
Is homeowner	582970	0	1	0.51	0.50

According to Table 1, which presents the descriptive statistical analysis of the main variables, several key factors have been examined. In this study, the borrower maximum rate is defined as the upper bound of the interest rate that borrowers are willing to accept, which can be considered as a risk assessment indicator[6]. By observing Fig. 1, there is a distinct group of outliers that far exceeds the range of other data. These outliers severely distort the distribution and statistical characteristics of the monthly income data. Consequently, it becomes imperative to exclude these outliers in order to ensure the precision and dependability of the subsequent analysis outcomes. Additionally, given the substantial variation in income across different groups, it is deemed necessary to logarithmically transform the income data. By doing so, the detrimental influence of inter-group differences on the analysis results can be mitigated, thereby facilitating a more accurate depiction of the underlying characteristics inherent in the data[4].

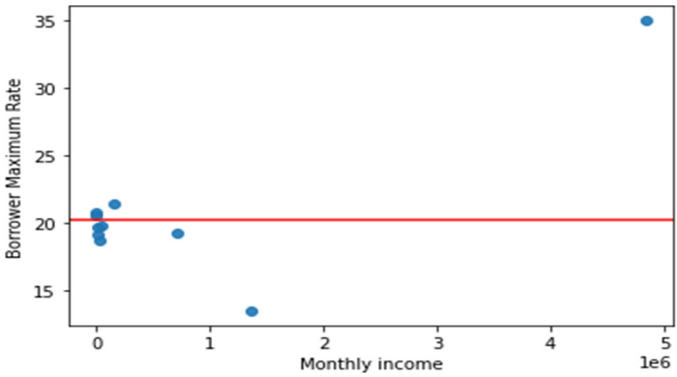

Fig. 1. Relationship between Monthly income and Borrower Maximum Rate

Studying the impact of time factors on investors' risk preferences is a captivating avenue of exploration, wherein weekdays and weekends hold significance as crucial temporal indicators. To investigate the effect of weekdays and weekends on borrowers' maximum acceptable interest rates, pandas library in Python can be adopted to process our data. Employing this tool will facilitate data aggregation and computation based on the categorization of weekdays and weekends.

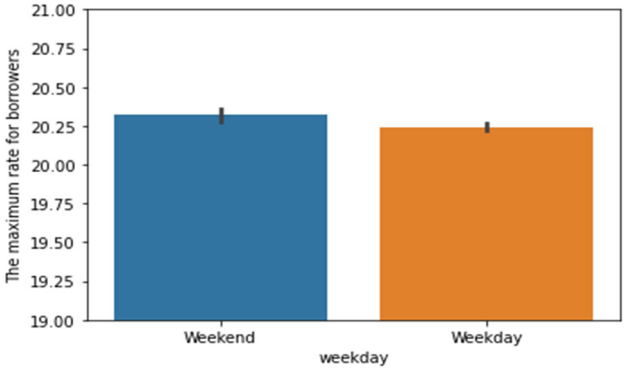

Fig. 2. Impact of weekends versus weekdays on the maximum rate for borrowers

From Fig.2, it is evident that there exists a notable disparity in the borrowers' maximum acceptable interest rates between weekends and weekdays in the sample. The mean borrowing rate during weekends is higher than that during weekdays, with a t-value of -3.40 and a p-value of 6.83e-04. This result indicates that weekends have a significant impact on borrowers' interest rates in this sample, suggesting that risk preferences vary during different time periods. This difference may be attributed to the heightened financial needs of individuals during weekends, coupled with relatively lower investor supply, consequently leading to an increase in the maximum rate borrowers are willing to accept.

2.3 Model Construction

This study selected daily listings number, funded percent, daily bid number, is homeowner, monthly debt, monthly income is weekend, risk preference, and month as independent variables, and borrower maximum rate as the dependent variable. The PanelOLS.from_formula() function was employed to fit the fixed effects model[9]. MemberKey was used as the entity fixed effect variable to eliminate the influence of individual characteristics that may interfere with the model, thus improving the reliability of the model.

To investigate the impact of borrower risk preference on loan interest rates, the research introduced a variable named "Risk Preference" as an independent variable. The dataset was hence grouped by MemberKey, and the average maximum annual interest rate for each borrower's history was calculated as their "Risk Preference" value.

Considering the competitive nature of the P2P market, the paper generated two variables: daily_bid_number and daily_listings_number. Daily_bid_number represents the number of investors providing investment to borrowers each day, and daily_listings_number represents the number of borrowers posting loan demands daily. These two variables could reflect the competitive relationship in the P2P market.

Table 2. Results of the fixed effects model

Variable	coef	Variable	coef
daily_listings_num	-0.0192****	month_4	0.0210***
Funded_per	0.0411****	month_5	0.0816****
daily_bid_number	-0.0198****	month_6	0.0556****
is_homeowner	-0.2484****	month_7	0.0626****
monthly_debt	0.0755****	month_8	0.0493****
income_log	-0.0315****	month_9	0.0431****
is_weekend	0.0017	month_10	-0.0237****
RiskPreference	0.5371****	month_11	-0.0477****
month_2	0.0529****	month_12	0.0162**
month_3	-0.0377****		

According to Table 2, which presents the results of the fixed effects model, several variables and their coefficients have been analyzed.

1. The coefficient of Daily Listings Number is -0.0192. One possible reason for this is that when there is an increase in the number of loan listings on the platform, competition among borrowers becomes more intense, leading to a decrease in the borrower's maximum interest rate.
2. The coefficient of Funded Percent is 0.0411. One possible reason for this is that when a certain proportion of the borrower's funding needs are met, they may be willing to accept higher interest rates in order to secure more funds.
3. The coefficient of Daily Bid Number is -0.0198. This may be attributed to the enhanced borrower-investor matching mechanism facilitated by an increased number of bids. Consequently, this improved matching dynamics potentially lead to a decrease in the borrower's maximum interest rate.

4. The coefficient of Is Homeowner is -0.2484. This observation indicates that homeowners potentially obtain lower maximum loan interest rates in comparison to non-homeowners. This may be because borrowers who possess real estate typically exhibit elevated creditworthiness and decreased risk, thereby enabling them to access loans at diminished interest rates.
5. The coefficient of Monthly Debt is 0.0755. One potential explanation for this is the positive correlation between the borrower's debt level and their credit risk. Consequently, to secure financing from investors, borrowers may be compelled to offer higher interest rates.
6. The coefficient of Income Log is -0.0315. One plausible reason for this is that a higher income level may be indicative of greater borrower capacity to fulfill debt obligations, thereby leading to a diminished credit risk profile.
7. The coefficient of Is Weekend is 0.0017. One possible reason for this could be the presence of elevated personal financial requirements during weekends or a reduced availability of investors, thereby necessitating borrowers to offer higher interest rates in order to attract funding.
8. The coefficient of Risk Preference is 0.5371. One possible reason for this is that individuals with a propensity to undertake greater investment risks may exhibit a willingness to embrace elevated interest rates in order to attain amplified returns.
9. The coefficients of the month variables indicate that compared to January, borrowers in February, March, April, May, June, July, August, October, November, and December may potentially observe fluctuations in their maximum loan interest rates. This may be due to changes in the economic environment and market demand in different months.

2.4 Weekday Test

In order to ascertain that the observed effect is not attributable to a particular day of the workweek but rather to the disparity between weekdays and weekends, the study conducted the following test. Firstly, we transformed the "weekday" variable into dummy variables, where the seven values of the "weekday" variable (1-7) were converted into seven dummy variables. "Weekday_6" and "weekday_7" represented Saturday and Sunday respectively, while the remaining variables represented Monday through Friday. We used the dummy variables for Saturday and Sunday as the baseline, and added these dummy variables to the original dataset.

Subsequently, a regression model was developed by solely incorporating the dummy variables for Monday through Friday, while keeping the remaining variables identical to the previous model. This methodology enabled us to evaluate the influence of weekdays and weekends on the outcome through the coefficients in the model. The findings from the regression analysis reveal that weekdays (Monday through Friday) have a significant impact on the Borrower Maximum Rate. When other variables are controlled, in comparison to Saturday and Sunday, the average Borrower Maximum Rate increases by 0.0153 to 0.0326 percentage points on weekdays ($p<0.001$). This indicates that there is a difference between weekdays and weekends, with weekdays exerting a stronger influence on the Borrower Maximum Rate.

3 Machine learning

This chapter applied machine learning methods[5] to predict the Borrower Maximum Rate. Predicting the Borrower Maximum Rate is of great practical significance, as it can reflect the credit risk of borrowers over a certain period of time. If the transaction is realized, the Borrower Maximum Rate can also to some extent reflect the risk preference of investors.

3.1 Single Regression Models

In this section, an empirical analysis utilizing three machine learning methods[10]: Random Forest Regression, Linear Regression, and XGBoost Regression[3][8] would be implemented. The dataset used in this analysis comprised 19 variables as features, such as daily_listings_num, Funded_per, daily_bid_number, is_homeowner, monthly_debt, income_log, is_weekend, RiskPreference, and month variables. The target variable was BorrowerMaximumRate. The dataset will be split into a training set and a test set using the train_test_split() function, with the test set comprising 30% of the total dataset.

Table 3. Single regression model prediction results

Regression Model	Mean Squared Error
Random Forest	0.4865
Linear Regression	0.6798
XGBoost Regression	0.6208

According to Table 3, by comparing the predictive performance of the three machine learning methods, it can be observed that both Random Forest Regression and XGBoost Regression outperform Linear Regression. Among them, the Random Forest Regression model has the best predictive performance[11].

3.2 Ensemble learning models

This chapter introduced two ensemble learning methods[2] in machine learning: Voting and Bagging[7], and compared their performance through empirical analysis. In ensemble learning, three base models: Random Forest Regression, Linear Regression, and XGBoost Regression were utilized.

Table 4. ensemble learning method prediction results

Ensemble Method	Mean Squared Error
Voting	0.5676
Bagging	0.5129

According to Table 4, by comparing the MSE of Voting and Bagging, it can be observed that their performances display significant parity. This suggests that Voting and Bagging are effective ensemble learning methods in machine learning. Moreover,

an additional benefit of Voting is pertinent to its expedited execution speed compared to Bagging, accentuating its practical utility in empirical applications.

4 Conclusion

This study aims to delineate conclusions and recommendations concerning risk management within P2P lending markets, through an in-depth analysis and modeling of loan data from the Prosper platform. The investigation underscores the significant impact of weekdays and weekends on risk preferences, thereby uncovering the crucial role of this temporal factor in risk assessment process. This revelation holds substantial pragmatic implications for P2P lending platforms and risk management professionals. These platforms are advised to calibrate their interest rate strategies and risk assessment measures according to the most significant interest rate fluctuations observed amongst borrowers during different time periods, so as to more accurately evaluate and control risk levels, thus bolstering the success rate of investments on P2P platforms. More specifically, it is suggested that P2P lending platforms adopt a strategy of endorsing borrowers who exhibit a propensity to accept lower interest rates during weekends to investors, thereby increasing the success rate of recommended orders on the platform.

Based on borrower data from the P2P platform, this study delves into the application of machine learning algorithms to assist in modeling. A continuous series of experimentation and parameter value adjustments were conducted across three different algorithms to procure the corresponding judgement outcomes. The predictive capabilities of different models in predicting default risks were mutually compared and analyzed using evaluation metrics. The resultant findings underscored the superior performance of the Random Forest model. Subsequently, ensemble learning models, using Voting and Bagging, were crafted integrating the Random Forest model, Linear Regression model, and Xgboost model. The application of machine learning algorithms to P2P lending prediction research, in conjunction with the development of fusion models, has the potential to provide auxiliary reinforcement for platforms seeking to recommend investable orders to investors.

References

1. Bavoso (Vincenzo). (2022). Financial Intermediation in the Age of FinTech: P2P Lending and the Reinvention of Banking. Oxford Journal of Legal Studies (1). doi:10.1093/OJLS/GQAB022.
2. Chen, S. , Wang, Q. , & Liu, S. . (2019). Credit Risk Prediction in Peer-to-Peer Lending with Ensemble Learning Framework. Chinese Control and Decision Conference.
3. A, X. M. , A, J. S. , B, D. W. , C, Y. Y. , A, Q. Y. , & A, X. N. . (2018). Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications, 31, 24-39.

4. Lin, X. , Li, X. , & Zheng, Z. . (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in china. Applied Economics, 1-8.
5. Setiawan, N. , Suharjito, & Diana. (2019). A comparison of prediction methods for credit default on peer to peer lending using machine learning. Procedia Computer Science, 157, 38-45.
6. Serrano-Cinca (C), Gutiérrez-Nieto (B), López-Palacios (L) (2015) Determinants of Default in P2P Lending. PLOS ONE 10(10): e0139427. https://doi.org/10.1371/journal.pone.0139427.
7. Kabari, L. G. , & Onwuka, U. C. . (2019). Comparison of bagging and voting ensemble machine learning algorithm as a classifier. International Journal of Computer Science and Software Engineering, 9(3), 19-23.
8. Jing Zhou, Wei Li, Jiaxin Wang... & Chengyi Xia. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A: Statistical Mechanics and its Applications(C). doi: 10.1016/j.physa.2019.122370.
9. Andre. (2010). Applied Panel Data Analysis for Economic and Social Surveys. Springer.
10. Yu, L. , Wang, S. , & Lai, K. K. . (2008). Credit risk assessment with a multistage neural network ensemble learning approach. Expert Systems with Applications, 34(2), 1434-1444.
11. Weicun Zhang. (2022). Compare Linear regression, Decision Tree Regressor, and Random Forest Regressor based on python, a restaurant company on Kaggle as a case.. (eds.) Proceedings of 2022 International Conference on Company Management, Accounting and Marketing (CMAM 2022) (pp.323-330).