A Literature Review on the P2P Lending Loan Default Prediction Using Machine Learning Methods

Lew Xin Yi

Razak Faculty of Technology and Informatics
lewyi@graduate.utm.my

Article history

Received:
xx xxx 2019

Received in revised form:
xx xxx 2019

Accepted:
xx xxx 2019

Published online:
xx xxx 2019

*Corresponding author
lewyi
@graduate.utm.my

Abstract

The P2P lending market saw a rapid growth around the world in the past few years due to its attractive interest rate, return, and less regulations. In respect to the expansion of the P2P lending platform, the problem of high default risk remained unsolved. Machine learning algorithms which are especially useful in predicting the borrowers with high loan default risk can help to effectively save costs and minimize negative impact brought by the default risk. This study aims to review previous literature related to the P2P lending loan default prediction model using machine learning algorithm to understand the fundamental components in building a comprehensive prediction model. There are four section of main findings which include data preprocessing techniques, machine learning scheme, evaluation metrics, and feature importance. This study also proposed a complete data mining process that suit the loan data obtained from the P2P lending platform by using the CRISP-DM methodology as the reference.

Keywords: *data mining, machine learning, P2P lending, Peer-to-Peer, data preprocessing, literature review, CRISP-DM, loan default, default risk*

1. Introduction

The term Peer-to-peer (P2P) lending has become more popular and familiar in the financial market since it was established in 2005. In the past ten years, the P2P lending industry has seen rapid growth globally, especially in developed countries like the United States (US) and China [1]. P2P lending refers to an online platform that allows individual borrowers and lenders to make deals without the intervention of financial intermediaries. The largest benefits of the P2P lending industry are the attractive interest rate for lenders and lesser loan approval boundaries for borrowers. Although the decentralized system in the P2P lending industry allows the borrower to borrow money easily, it has caused the platform to associate closely with high default risk and low-quality borrowers [2]. Loan default is an issue that arises when borrowers do not pay off their debt payment or interest. Therefore, there is a need for effective risk management to minimize the loan default risk that is increasing in line with the expansion of the P2P lending market [3]. The platform needs to improve on the identification of borrowers' creditworthiness to ensure customers

* *Corresponding author. masaini@abc.xxx.xxx*

who are more likely to default will not be granted a loan. One way to effectively predict borrowers with high default risk is through a machine learning algorithm. The machine learning algorithm will train the loan default prediction model by using previous loan data from the P2P lending platform and learn to predict the loan defaulter. This enables the platform to improve the loan approval process and reduce costs effectively. This study aims to review the previous literature on P2P lending default risk prediction to develop an understanding on the important features that caused loan default, data preparation techniques, and machine learning models used.

2. Overview

a. **Loan default prediction of Chinese P2P market: a machine learning methodology** [1]**:** Previous studies found on the P2P lending market mainly focus on the application of statistical models like probit and logit models on predicting loan default and fewer studies evaluate the feature's importance. By using the loan data from the largest P2P lending platform in China, borrower identity and asset certification are the most important positive features that reduce loan default. Meanwhile, borrowers that are more likely to default are those that pass the verification of mobile phone, job, residence, or education level. The result shows that random forest outperforms the extreme gradient boosting tree (XGBT), neural network (NN), and gradient boosting model (GBM). All the machine learning algorithms used have high accuracy of more than 90% in predicting loan default.

b. **Default prediction in P2P lending from high-dimensional data based on machine learning** [2]**:** One major concern about using P2P loan data is that it is highly imbalanced and distributed sparsely. Besides, the statistical models and machine learning methods do not predict the probability of loan default accurately and precisely. Therefore, the authors proposed a decision tree model based on a heterogeneous ensemble that involved XGBT, gradient boosting decision tree (GBDT), and light gradient boosting machine (LightGBM) to predict the loan default probability and solve the mentioned problems. The model is then compared with other benchmark machine learning models based on evaluation metrics like accuracy and precision. The result shows that the proposed model can solve the imbalance of data issues and effectively predict the results.

c. **Predicting default risk on Peer-to-Peer lending imbalanced datasets** [4]**:** The P2P lending data is highly imbalanced which means the number of fully paid loans and default loans is not equal. The issue of imbalanced data is caused by information asymmetry where lenders evaluate the loan approval only based on the information provided by the borrowers. In effect, normal machine learning models are not familiar with the highly imbalanced dataset and the result might be less accurate and biased toward the majority class. Therefore, the authors proposed to implement re-sampling and cost-sensitive learning on the loan data to solve the issue prior to the training of machine learning models. After balancing the dataset using different methods, the random under-sampling method provides the highest performance than the others.

d. **Loan default prediction with machine learning techniques** [5]**:** Feature extraction is one of the methods to improve model accuracy with include only the important features. This study applies machine learning algorithms which include AdaBoost, XGBT, random forest, K Nearest Neighbors (KNN), and Multilayer Perceptrons (MLP). The authors select only the most relevant feature by removing columns with many missing values and duplicating columns with the same information. For example, the validity period of bank cards is extracted as one feature. The result shows that the AdaBoost model has a 100% complete accuracy as compared to the others in predicting loan default.

e. **Credit risk assessment of P2P lending platform towards Big Data based on BP Neural Network** [6]**:** The loan data used are obtained through web crawling. The authors focus on using BP neural network-based algorithm to predict lending risk in the P2P lending platform. The proposed BP neural network prediction model is then compared with the logistic regression model to outcast the benefit of the neural network algorithm. Result shows that BP neural network model achieved higher performance than the conventional logistic regression model.

3. Synthesis of Findings

The application of machine learning algorithms to train a predictive model can effectively improve the model's performance as compared to traditional statistical methods. Besides the normal machine learning algorithms, various types of proposed ensemble models from previous studies can enhance the prediction of loan default. The data preprocessing phase is also one of the most important steps to better train the model and increase the model's accuracy. The features that are important to the default risk changes in the P2P lending platform can be obtained through the data mining process. Based on the findings from previous literature reviewed, this section can be separated into four subsections that aim to build a model that best suits the P2P lending loan default prediction.

1. Data Preprocessing Techniques
   Most studies conduct feature selection, data encoding, and resampling to clean the P2P loan data before using it to train the model.
   a. Feature selection or feature engineering is to choose the relevant and important features that affect the default risk most. Previous studies found that loan maturity and loan purpose is among the most significant factors that affect the probability of loan default [7] [8] [9]. For example, borrowers with longer loan maturity periods have a higher chance to default on the loan. Feature selection helps to filter out irrelevant features like those with missing values or duplicated columns thus increasing the data reliability. In effect, machine learning algorithms can

learn from more accurate data to provide a high-quality prediction model.

b. Data encoding refers to the process of converting categorical data into numerical data. Each of the categories involved in the column will be transformed into a number representation. One of the most popular techniques used in encoding is the one-hot encoder. Data encoding is needed due to most of the machine learning algorithms cannot work with categorical data (label) directly where it read only numeric data.

c. The P2P lending loan data is usually dimensional and highly imbalanced where the number of default loans and fully paid loans are unequal. Some of the resampling methods are under-sampling and over-sampling where under-sampling will select a random subset sample from the majority class (fully paid loan) to balance the data while over-sampling will increase the sample size of the minority class by duplicating random the minority class's data. Resampling is important to avoid the result of models trained biased toward the majority class.

2. Machine Learning Scheme

The machine learning algorithms that are widely used in developing a loan default prediction model for the P2P lending platform can be discussed in two subcategories which are normal machine learning algorithms and proposed ensemble models.

a. The machine learning algorithm that is most frequently used in predicting P2P lending loan default is the neural network, random forest, and boosting based algorithm [1] [2] [4] [10]. The machine learning algorithms are then used to train the prediction model to classify whether the data belongs to a fully paid or default loan.

b. Another type of machine learning applied in constructing the prediction model is the proposed ensemble model by some of the authors that aims to improve the model accuracy and solve problems like overfitting and multi-observational data. For example, a study done by [2] proposed a heterogeneous ensemble learning-based loan default prediction model which involved GBDT, XGDT, and LightGBM to further improve the prediction accuracy. The main purpose of the proposed model is to solve the issues of high dimension and imbalanced data in the P2P lending market. The result shows that the proposed model achieved high accuracy and enhance operational efficiency.

3. Model Evaluation

After training and testing the machine learning-based default prediction models, evaluation needs to be conducted to find the model with the best performance in predicting loan default probability. Some of the most popular metrics used to access the model's performance are confusion matrix, AUC values, accuracy, precision, recall, and F1-score.

a. The confusion matrix enables the users to check the True Positive, False Positive, True Negative, and False Negative data point predicted by the model.

b. Models with high accuracy, precision, and recall indicate better performance.
c. AUC value which is the area under the ROC curve is a calculation that is used to make a comparison between two prediction classes.
d. The F1-score is the harmonic mean between two common metrics: precision and recall.

4. Important Features that Affect the Default Risk in P2P Lending

   By training the machine learning-based default prediction model, important features that affect the default risk can be identified easily. The identification of feature importance enables the platform to understand the root factor of high default risk and thus take appropriate action to fix the specific aspects. The features can be divided into two categories which are micro level factor (individual) and macro level factor (platform and macroeconomic).

   a. In terms of micro-level factors, the characteristic of individual borrowers like yearly income, credit history, current housing situation, and indebtedness are variables that affect the default risk in the P2P lending platform [11]. Borrowers with higher annual incomes are less likely to default on the loan. Personal loan purpose is another important feature that affects the default risk where an individual that borrows for a wedding tend to have a lower default risk while those that borrow for business purpose has the largest default risk. By understanding the micro-level factor, the platform's officer can focus more on the screening and verification of business borrowers.
   b. Macro-level factors that affect the loan default risk in the P2P lending platform consists of the loan characteristic provided by the platform and the economic changes. The interest rate given by the platform is correlated positively with the default risk where a higher interest rate charged on the loan will cause the borrower to have a higher chance to default. From the view of the economy, the age of the company and sector type have a high impact on the default risk changes in the P2P lending platform. Companies that are younger have a higher chance to default on the loan. Meanwhile, customer services, automotive, transport and logistics, and retail are sectors that are more likely to default [9]. In the event of inflation where the price of goods continues to increase over time, the number of borrowers that default on the loan is higher. The unemployment rate also affects the default risk in the P2P lending market positively. In contrast, the income per capita and home price index is negatively correlated with the P2P lending default risk [12].

4. Proposed Work

After reviewing the previous literature on the P2P lending default risk prediction model based on machine learning, this study proposed a best practice to fill the literature gap by constructing a machine learning-based loan default prediction model using the Cross Industry Standard Process for Data Mining (CRISP-DM) framework.

a. Business Understanding

The main objective is to construct a loan default prediction model that can effectively detect potential borrowers with a high probability of defaulting on a loan to minimize the platform's default risk. The development of a loan default prediction model by using machine learning algorithms with high accuracy can help to improve the platform's business performance and attract more lenders and borrowers to make the transaction with it.

b. Data Understanding

The loan data that consists of personal loan information of the borrowers are publicly available on the P2P lending platform website. For instance, large P2P lending platform like Lending Club in the United States provided their loan data for public use in their database without any restriction. Exploratory data analysis prior to modeling can be conducted to ensure data reliability.

c. Data Preparation

i. Data preprocessing is one of the most important tasks in the data mining process to ensure the data is highly reliable and clean for training machine learning models.

ii. Since the raw loan dataset usually involved more than 100 variables, feature selection is the activity that must be taken to remove duplicated or less relevant variables. Feature selection through ranking help to identify the most important features that affect the default risk and discussion will be given to each top N feature.

iii. Data resampling is another important task that must be conducted to prepare the data for building a more accurate model. Under-sampling which randomly selects a subset sample from the fully paid loan to ensure the data between fully paid loan and default loan are equal will be conducted. The resampling method prevents the result obtained from the prediction model to biased toward the majority class.

d. Modeling

Since the ensemble learning algorithm works well with the high dimensional and imbalanced loan data from the P2P lending platform yet still provides a high accuracy performance, this study proposed to build the prediction model using an ensemble learning algorithm. An ensemble learning algorithm refers to the combination of several different machine learning algorithms to build an optimal loan default prediction model. For instance, CatBoost which is a decision tree-based heterogeneous ensemble learning default prediction model introduced by [2] can enrich the feature dimension through the combination of category features as well as avoid overfitting to happen. Some benchmark classifiers like logistic regression and neural networks will be constructed to make a comparison with the proposed model. A comparison of the model's performance is needed to evaluate the effectiveness of the proposed ensemble loan default prediction model.

e. Evaluation

Model evaluation is an essential phase to access the model’s performance of each machine learning algorithm used in predicting the loan default. Evaluation metrics like confusion matrix, accuracy, precision, recall, and AUC value will be applied to measure the model’s performance. The machine learning-based model with the highest accuracy will be chosen as the most suitable model to predict the P2P lending loan default. Findings from all the previous steps will be discussed to ensure the data mining process is conducted smoothly. The platform can utilize the prediction model to reject loan applications by low-quality borrowers or borrowers that are most likely to default on the loan.

f. Deployment

Once the loan default prediction model with the highest performance is chosen, the complete code from the data preparation stage to the model building stage needs to be deployed into the platform system. This helps the P2P lending platform to automatically digest new data coming in and make prediction immediately.

5. Conclusion

Data mining techniques are useful and have been widely implemented by large companies around the world due to their benefits. There is various type of machine learning algorithms that serve different function and deals with different types of data input. The loan default risk is the largest problem that arises in the P2P lending market and destabilizes the platform's performance. Default risk can be minimized through the utilization of machine learning algorithms in training the loan default prediction model which aims to predict the borrowers that are most likely to default. The loan default prediction model enables the P2P lending platform to reject loan applications of low-quality borrowers and thus effectively reduce the cost incurred on default loans. In order to produce an effective and comprehensive loan default prediction model, a well-planned methodology based on commonly used framework like CRISP-DM need to be prepared. The in-depth literature review on machine learning-based loan default prediction in the P2P lending platform provides valuable information that helps the users to understand well the fundamentals related to the topic. This study also proposed a CRISP-DM methodology that is suitable for the P2P lending platform loan data and aims to improve the model performance.

References

[1] J. Xu, Z. Lu and Y. Xie, "Loan default prediction of Chinese P2P market: a machine learning methodology," *Scientific Reports,* no. 18759, pp. 1-19, 2021.

[2] J. Zhou, W. Li, J. Wang, S. Ding and C. Xia, "Default prediction in P2P lending from high-dimensional data based on machine learning," *Physica A: Statistical Mechanics and its Applications,* vol. 534, no. 122370, pp. 1-11, 2019.

[3] J. Y. Lee, "Prediction of Default Risk in Peer-to-Peer Lending Using Structured and Unstructured Data," *Journal of Financial Counseling and Planning,* vol. 31, no. 1, pp. 115-129, 2020.

[4] Y. R. Chen, J. S. Leu, S. A. Huang, J. T. Wang and J. I. Takada, "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets," *IEEE Access,* vol. 9, pp. 73103-73109, 2021.

[5] L. Lai, "Loan default prediction with machine learning techniques," *2020 International Conference on Computer Communication and Network Security,* pp. 5-9, 2020.

[6] Y. Guo, "Credit Risk Assessment of P2P Lending Platform towards Big Data based on BP Neural Network," *Journal of Visual Communication and Image Representation,* vol. 71, 2020.

[7] C. Croux, J. Jagtiani, T. Korivi and M. Vulanovic, "Important factors determining Fintech loan default: Evidence from a lendingclub consumer platform," *Journal of Economic Behavior & Organization,* vol. 173, pp. 270-296, 2020.

[8] M. Gao, J. Yen and M. T. Liu, "Determinants of defaults on P2P lending platforms in China," *International Review of Economics & Finance,* vol. 72, pp. 334-348, 2021.

[9] B. Xu, Z. Su and J. Celler, "Evaluating Default Risk and Loan Performance in UK Peer-to-Peer Lending: Evidence from Funding Circle," *Journal of Advanced Computational Intelligence and Intelligent Informatics,* vol. 25, no. 5, pp. 530-538, 2021.

[10] P. C. Ko, P. C. Lin, H. T. Do and Y. F. Huang, "P2P Lending Default Prediction Based on AI and Statistical Models," *Entropy,* vol. 24, no. 801, pp. 1-23, 2022.

[11] C. Serrano-Cinca, B. Gutierrez-Nieto and L. Lopez-Palacios, "Determinants of Default in P2P Lending," *PLoS ONE,* vol. 10, pp. 1-22, 2015.

[12] Y. Yoon and Y. Feng, "Factors affecting platform default risk in online peer-to-peer (P2P) lending business: an empirical study using Chinese online P2P platform data," *Electronic Commerce Research,* vol. 19, pp. 131-158, 2019.