

A Comprehensive Investigation: Machine Learning for P2P Lending Prediction

Qichang Ma^a

School of Management, Shandong University, Jinan, China

Keywords: P2P Lending, Machine Learning, Default Prediction.


Abstract: Peer-to-Peer (P2P) lending, a new business format, has developed rapidly. However, there have also been many frauds, which have brought great challenges to investors and P2P lending platforms. This paper reviews the methods of P2P default prediction based on machine learning. Four common machine learning models are introduced: decision tree, support vector machine, deep neural networks and ensemble learning. For each model, the entire process of constructing the model in the literature is illustrated to describe how these models are applied in P2P default prediction. Three main challenges in using machine learning for P2P default prediction are proposed: because of the "black box" property in machine learning methods, P2P default prediction faces difficulties in interpretability; different models use data of different national sources, structures, features, etc., existing models cannot be directly applied to a specific P2P lending platform, so P2P default prediction faces difficulties in applicability; P2P default prediction involves a large amount of important user privacy and security data, so P2P default prediction also faces difficulties in privacy. Solutions to these difficulties are also proposed. This article provides a good review of using machine models for P2P default prediction, which can provide inspiration for managers of P2P platforms.

1 INTRODUCTION

The fourth industrial revolution has had a significant influence on the way that the financial system operates: The swift advancement of technology for the internet has promoted the further combination of internet technology and traditional finance. Various new financial development models have emerged, and one of a booming innovative financial model is called Peer-to-Peer (P2P) lending. P2P lending is referred to individuals or enterprises borrowing from other individuals or institutional lenders through technological platforms, and it does not involve traditional financial intermediaries (Lin et al., 2023). Due to its advantages such as low transaction costs and low financing thresholds, it has developed rapidly. However, P2P lending also comes with many drawbacks, such as imperfect credit reporting systems, incomplete credit assessment methods, and information asymmetry between borrowers and lenders (Chen, 2022). These drawbacks may cause some borrowers to fail to repay loans and interest, thereby restricting the development of P2P platforms

and bringing risks to investors. Therefore, it is necessary to accurately predict whether borrowers will default before lending occurs.

In recent years, technologies such as artificial intelligence have developed rapidly. Various algorithms have emerged, which have been applied in many fields. Similarly, artificial intelligence can also be applied to P2P lending default prediction. P2P lending has accumulated a large amount of customer data over the years of development, and many studies have conducted research on default risk prediction. One of the challenges faced by researchers in default risk prediction is the imbalanced data, that is, the number of fully paid loans is much larger than default loans. However, in default risk prediction, the prediction of default loans received more attention, but the classifier tends to predict the majority class. The ratio of fully paid loans and default loans in the data used by Chen et al. was 3.5:1, and they used under-sampling and cost-sensitive learning to solve

^a <https://orcid.org/0009-0009-2342-8716>

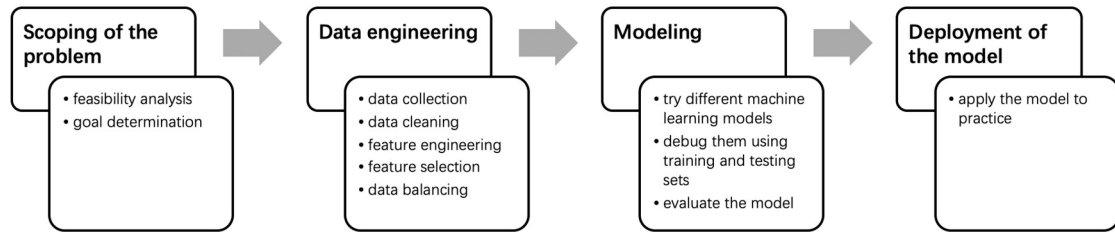


Figure 1: Workflow of building P2P default prediction machine learning model (Kampezidou et al., 2024).

the problem (Chen, et al., 2021). There are many researchers utilized several algorithms for analyzing the possibility of defaults. Yang et al. showed comprehensive data processing for default prediction, including data cleaning, feature engineering and feature selection, modelling and Stacking. The study tried many machine learning models and also ensemble algorithms, among which Decision Tree (bagging), Random Forest (bagging), Light GBM (bagging), XGBoost (bagging) had better fitting performance, with an amazing recall percentage of more than 87% (Yang, 2024). Turiel et al. used Linear Regression (LR), Support Vector Machine (SVM), and Deep Neural Networks (DNNs) when predicting the chance of defaults, using data from lending club. Among them, DNNs performs the best with recall score of 72% (Turiel et al., 2020). In addition to structured data, there is also many unstructured data about the borrower which contains a lot of effective information about credit. However, unstructured data has been rarely used in default risk prediction in the past. Kriebel et al. showed that text generated by eusers can significantly enhance credit default prediction. Deep learning and machine learning algorithms centered on word frequency ranges etc. were used on user-generated text from data of lending club, and it showed excellent performance (Kriebel et al., 2022). P2P lending brings vitality into economic development. Because of the development of this field, many new methods combining artificial intelligence algorithms for default risk prediction have emerged. Therefore, it is necessary to make a comprehensive review of this field.

The rest of the paperwork is structured as follows: The next part will briefly introduce the general process of machine learning first, and then summarize some methods for default prediction in previous studies. The third part will propose the challenges and future prospects of the P2P lending default prediction field. Finally, the fourth part will summarize the whole paper and put forward conclusions.

2 METHOD

2.1 Introduction of Machine Learning Workflow

In P2P lending default prediction, it is required to develop a supervised model which includes the loan status as the dependent variable. Generally speaking, a complete supervised learning process is shown below (Kampezidou et al., 2024), and also in Fig. 1.

- Scoping of the problem: this includes feasibility analysis, goal determination, etc. This is the beginning of a project.
- Data engineering: In this scenario, it generally includes: (1) data collection: download public data from P2P platforms such as Lending Club or other channels as needed. (2) data cleaning: fill in missing fields, treat outliers and remove exclusions or other required, to make the data cleaner and more complete. (3) feature engineering: the most critical step in the process of machine learning data preparation, including feature transformation, feature generation, etc. (4) feature selection: filter features to decrease the number of features, to circumvent the dimensionality curse. (5) data balancing: in P2P user data, compared to bad loans, the quantity of good loans is significantly higher., so the data needs to be balanced. Through these steps, the structured data is generated.
- Modeling: In this process, try different machine learning models, debug them using training and testing sets, and finally evaluate the model.
- Deployment of the model: the process of applying the model to practice, which is generally the task of software developers.

2.2 Decision Tree

Decision tree divides the whole into branch-like parts, which with root, internal, and leaf nodes, make up a

reversed tree. It can be used to establish a classification system or a dependent variable prediction algorithm. Decision tree can be used in many fields of data mining, among which prediction is one of its most important uses (Song et al., 2015).

Kumar et al. using data from Lending Club, ignored loans with "loan status" as "Issued", and removed features that contained similar information or had too many missing values. Then they used the median to impute fields with 10-15% missing value and deleted some excessive and least important fields. After that, they used three models including decision tree. When using decision tree, "Interest rate" was selected as the primary split because it was the most information-gaining, succeeded by "Debt to income ratio", "Installment", "Annual income" and "Loan amount". Finally, the results are compared by comparing the precision and accuracy of the three models and calculating the Area Under Curve (AUC) (Kumar et al., 2016).

2.3 Support Vector Machine

A classification technique called support vector machine (SVM) divides d-dimensional data to two categories by locating a hyperplane. Sometimes, the data is linearly inseparable, so SVM puts the data into a space with multiple dimensions where it can be separated using a technique called "kernel induced feature space" (Boswell, 2022).

The model built by Shan et al. used data from Lending Club. Shan et al. first used K-Nearest Neighbors (KNN) to impute missing fields, that is, to interpolate using the average of the k nearest data points, and the paper used the default value of 5. Then they used the first Factor Analysis of Mixed Data (FAMD) to select features. This method is a combination of Principal Component Analysis (PCA) for qualitative features and Multiple Correspondence Analysis (MCA) for quantitative features. Features were selected according to the percentage of contribution to the data variance exceeding a certain threshold. In this study, loan status is regarded as the independent variable Y. The original data had 7 categories. Shan et al. balanced the data by reclassifying the loan status. Then he used a variety of models including Support Vector Machine, and compared the model results mainly based on accuracy rate and ROC, as well as Confusion Matrix (Shan et al., 2018).

2.4 Deep Neural Networks (DNN)

The DNN is made up of numerous parallel-operating,

basic processing units (that resemble neurons) layered into one another. Two layers make up the most basic neural network: a layer of input and a layer of output. The network is referred to as a deep network as the number of layers rises. DNN is trained by learning to perform specific tasks to enhance the connections between units, so that the model can perform the same task on new inputs, which is similar to the human brain (Cichy et al., 2019).

Turiel et al. collected the data of Lending Club, and then deleted the features with more than 70% missing values, imputed the missing values with the mean, balanced the data by down sampling the training set. In the process of using DNN to build the model, categorical features were excluded, because hot encoding could generate too many columns which consumed more training time. A tanh activation function was chosen, and the adaptive moment estimation (Adam) optimization method, which is designed for neural networks, was used for optimization. "Softmax cross entropy" served as the loss function, and Dropout was chosen for regularization. Then the network structure was adjusted by performing empirical grid search on multiple network configurations and evaluated by stratified fivefold cross-validation to prevent the training or testing set from being shrunk. Finally, the AUC-ROC scores as well as recall ratings were used to test the performance (Turiel et al., 2020).

2.5 Ensemble Learning

The purpose of ensemble learning is to integrate multiple machine learning algorithms into one model, extract features by performing multiple projections on the data, produce weak prediction results, and combine the outcomes with different vote methods to outperform each member model individually (Dong et al., 2020).

Li et al. developed a default model for prediction founded on heterogeneous ensemble learning using data from a P2P network in China. The model was trained using the number of missing data as a new feature, and the numerical features were dispersed using the equal size division approach. To select features, the model-based feature ranking method was used. When pre-training with the Extreme Gradient Boosting model, it was tested 30 times in a loop to obtain the feature importance score and order. Three member algorithms: XGBoost, Deep Neural Networks as well as Logistic Regression were trained for 36 iterative loops, and ten-fold cross validation was adopted during the algorithms' training process. The article used three methods for hyperparameter

optimization and compared them to obtain the optimal hyperparameter value range. Finally, the three member algorithms were combined through the liner weight ensemble strategy and AUC was used to evaluate the model (Li et al., 2020).

3 DISCUSSIONS

3.1 Limitations and Challenges in This Field

3.1.1 Interpretability

Interpretability and transparency are very important for credit management process models, including default prediction. Although many machine learning methods mentioned above show high accuracy in P2P lending default prediction, most algorithms lack interpretability due to the “black box” property in machine learning methods. Creating interpretable algorithms is an essential subject of study for predicting defaults, particularly for more open domains like peer-to-peer lending: in the context of P2P lending, due to the lack of participation of traditional intermediaries, information asymmetry is more serious, and higher requirements are placed on the interpretability of default prediction (Ariza-Garzón et al., 2020). P2P lending default prediction should be helping in the creation of credit regulations and offer details to investors, administrators, and borrowers, otherwise it is difficult to be trusted.

3.1.2 Applicability

It can be found that each research in this field usually uses different data sources: different policies in the countries where the data is located, different sizes and structures of data, different features contained in the data, etc., which makes the above research methods only applicable to the specific data and situations they are based on. When the data and situations are changed, these methods may not be applicable and cannot produce satisfactory results. Therefore, when an entity P2P lending platform wants to predict default behavior, it cannot copy the above methods directly. If the platform wants to train a model by itself, it will be difficult because of high costs and poor accuracy.

3.1.3 Privacy

With the development of AI technology, people are increasingly concerned about data privacy and

security issues. P2P lending default prediction involves a large amount of very important customer privacy data, which may include asset status, work status, family income, etc. If this private information is leaked, it will bring significant threats to customers and bring moral and legal risks to the P2P lending platforms.

3.2 Future Prospects

3.2.1 Solutions to Interpretability

Many researchers have concentrated on the topic of Explainable Artificial Intelligence (XAI) and created some techniques for explaining machine learning models in response to the increasing need for more explainable machine learning models. These methods can be basically divided into four categories (Linardatos et al., 2020). When building a P2P lending default prediction model, researchers can apply these methods. Ariza-Garzón et al., for instance, employed SHapley Additive exPlanations (SHAP) values as a tool for interpretability (Ariza-Garzón et al. 2020).

3.2.2 Solutions to Applicability

When solving the problem of applicability, the application of domain adaptation can be considered. Domain learning is a subfield of machine learning a unique example of transfer learning. It seeks to address the issue of disparate data distributions by modifying the distinctions among domains in order to increase the trained algorithm's applicability. In domain adaptation, only the domain is different, while the task remains the same (Farahani et al., 2021). Therefore, using domain adaptation can allow P2P lending platforms to use models that have been trained in the past.

3.2.3 Solutions to Privacy

In order to solve the privacy leakage problem of machine learning, federated learning can be used. Federated learning is a technology and also a business model. It allows different companies to train data uniformly without sharing data, and establish a unified model, which greatly reduces the risk of data leakage, improves data privacy security, and also provides personalized and targeted services (Yang et al., 2019).

4 CONCLUSIONS

In this paper, several machine learning methods for P2P default prediction are discussed, and challenges in this field as well as solutions are proposed.

In method section, four machine learning methods are introduced, namely decision tree, support vector machine, deep neural networks and ensemble learning, and it also includes how the researchers start from data cleaning, go through data balancing, feature selection and other steps, and finally build a model.

Then, the three challenges faced in the field of P2P lending default prediction are discussed, namely interpretability, that is, most machine learning algorithms lack interpretability and make the model untrustworthy; applicability, that is, a model cannot be directly applied to a specific P2P lending platform; and privacy, that is, user data privacy issues. And then corresponding solutions are proposed. When facing the interpretability problem, four approaches for explaining machine learning algorithms can be used; when facing the applicability problem, domain adaptation (a machine learning method that only the domain is different, but the task remains unchanged) can be applied; when facing the privacy problem, federated learning can be used to improve data privacy security.

REFERENCES

- Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. J. 2020. Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access*, 8, 64873-64890.
- Boswell, D. 2002. Introduction to support vector machines. Department of Computer Science and Engineering University of California San Diego, 11, 16-17.
- Chen, X. R. 2022. Master of Research on Default Prediction of P2P Online Loan Users Based on Integrated Learning (Dissertation, Chongqing University). Master
- Chen, Y. R., Leu, J. S., Huang, S. A., Wang, J. T., & Takada, J. I. 2021. Predicting default risk on peer-to-peer lending imbalanced datasets. *IEEE Access*, 9, 73103-73109.
- Cichy, R. M., & Kaiser, D. 2019. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4), 305-317.
- Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14, 241-258.
- Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. 2021. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877-894.
- Kampezidou, S. I., Tikayat Ray, A., Bhat, A. P., Pinon Fischer, O. J., & Mavris, D. N. 2024. Fundamental Components and Principles of Supervised Machine Learning Workflows with Numerical and Categorical Data. *Eng*, 5(1), 384-416.
- Kriebel, J., & Stitz, L. 2022. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, 302(1), 309-323.
- Kumar, V., Natarajan, S., Keerthana, S., Chinmayi, K. M., & Lakshmi, N. 2016. Credit risk analysis in peer-to-peer lending system. In 2016 IEEE international conference on knowledge engineering and applications (ICKEA) (pp. 193-196). IEEE.
- Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. 2020. Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. *World Wide Web*, 23(1), 23-45.
- Lin, Y. Y., Zhang, P. P., & Hou, X. P. 2023. P2P online lending platform under the background of financial technology: progress and prospects. *Shanghai Management Science* (06), 56-61.
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 18.
- Shan, Q., & Nilsson, M. 2018. Credit risk analysis with machine learning techniques in peer-to-peer lending market. *International Journal of Risk Management*, 4(2), 1-43.
- Song, Y. Y., & Ying, L. U. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130.
- Turiel, J. D., & Aste, T. 2020. Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society open science*, 7(6), 191649.
- Yang, Q., Liu, Y., Chen, T., & Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 1-19.
- Yang, R. 2024. Machine Learning-Based Loan Default Prediction in Peer-to-Peer Lending. *Highlights in Science, Engineering and Technology*, 94, 310-318.