Asian Business Research; Vol. 4, No. 3, 2019
ISSN 2424-8479 E-ISSN 2424-8983
Published by July Press

Fintech Application in Banking Operations - Application of Machine Learning in Mitigating Bank Derivatives Counterparty Risks

Tianshu Li[1]

[1] University of Melbourne, Sydney, Australia

Correspondence: Tianshu Li, U106 756 Sydney Rd Brunswick VIC Australia. Tel: 61-4-6906-5346.

Received: September 6, 2019 Accepted: September 20, 2019 Online Published: October 8, 2019

doi:10.20849/abr.v4i3.652 URL: https://doi.org/10.20849/abr.v4i3.652

Abstract

We all know that human has many psychological biases, including overconfidence, gender discrimination and so on. Although some genuine lenders may outperformance others, machine learnings have been utilized to solve this human psychological bias in many areas. By using machine learnings methods, people can make better financial decisions. This proposal tries to examine the effectiveness of several different machine learning models on predicting the ex-pose default risk, including BP neural network, decision tree, KNN, and random forest. I focus on loans on one electronic P2P lending platform, called "Paipaidai" in which lenders select and supply private loans to borrowers with different characteristics. I use machine learnings methods to predict the default risk and thus provides better ways for investors to select high-quality borrower. I will also further test how different machine learnings methods perform when there is soft information contained by using Prosper platform.

Keywords: machine learning, P2P, default risk, lending

1. Introduction

Information technology has changed our daily life in many aspects. The most important changes raise from the finance system. There is increasing research focusing on Fin Tech and machine learning (Bachmann et al., 2011; Chen, and Han, 2012; Lin, Prabhala, and Viswanathan, 2013; Emekter, Tu, Jirasakuldech, and Lu, 2015; Emekter, Tu, Jirasakuldech, and Lu, 2015). For example, Emekter, Tu, Jirasakuldech, and Lu (2015) evaluates Lending Club credit risk and how the P2P loan characteristics affect the loan performance. Jin, and Zhu (2015) explore the characteristics of loan and its applicant and use random forest to do the feature selection in the modeling phase by using data from the Lending Club. Lending market value is very massive, a worth of trillions of dollars in single country such as Americans. Many companies, including bank, P2P lending platform, small lending companies, are both looking for technology to improve companies' return on selecting better loans and improve their profits. Artificial intelligence and machine learnings will play a major role in this market due to its efficient, cost-saving and less human psychological bias. Although the desire to utilize machine learnings, the assessment of its effectiveness seems the priority thing to check. Mangasarian (1965) was the first to propose Linear Programing in pattern separation. Later, Wiginton (1980) utilize logistic regression (LR) to build credit scoring prediction models. Further, Then, Shi et al. (2002) extended it to multiple criteria linear programming and applied this approach in credit card portfolio problem. But at that time, this method has many disadvantages and the convergence rate is very slow. Odom and Sharda (1990) used neural network in bankruptcy prediction. Tam and Kiang (1992) compared linear classifier, logistic regression and neural networks models in predictive accuracy. Although there are many articles focusing on predicting credit risk for traditional financial and banking institutions, research that relates the credit risk of online P2P lending is very limited. Emekter et al.'s research (2014) investigates the determinants of loan default and uses LR to evaluate credit risk in online P2P lending. However, there is very limited research focusing on the effectiveness of different machine learnings model in default prediction. Given the practical value, I try to use one famous P2P lending in China to test the effectiveness of machine learning models. I also combine another US P2P lending platform, called Prosper to test how the effectiveness of machine learning models varies when they are soft information plays a role in investor decision.

The advantages of the lending market are due to its big data, which is the requirements for utilizing machine learning. Machine learnings can unitize all dimensions of borrowers' characteristics to assess the creditworthiness, assessing the history of individuals past debts history, estimating the value of the collateral, such as car, home, business, artwork, etc, and constructing a better model to selecting good quality borrowers. The beauty of machine learning is that machine learnings can incorporate macro/micro factors, such as future inflation, financial risk, and economic growth into the assessment model. More importantly, machine learnings can train the data and adapts to the changes gradually. Theoretically, it can analyze all of these data sources together to create a coherent decision.

Our research question is motivated by the existing literature that studies costs and benefits machine learning on financial market. A widely held view on the benefits of machine learnings is that machine learnings are much more time-saving and cost-saving, which enables firm cut down their human labor cost. However, utilizing machine learnings come at cost. Some research argues that human intelligence is more superior at gaining soft information compared to artificial intelligence. Inspired by the literature that documents the costs and benefits machine learning, we hypothesize that machine learnings can predict the default risk more accurately if there is more hard information available. To compare how different methods can predict the default risk, I will employ the leading provider of P2P lending platform data in China, called "Paipaidai".

P2P lending is an innovative form of financing, mainly providing money to small and medium-sized borrowers. The development of P2P lending largely alleviates the financial constraint for the market. However, the default risk of P2P lending is very high. In practice, P2P platform will analysis the borrowers' hard and soft information to predict their default risk. Previous literature mainly utilizes the logistic model or human intelligence to predict the default risk. In this proposal, I try to compare several mainstream machine learnings methods to predict the default risk, including BP neural network, decision tree, KNN (K-Nearest Neighbor) classification algorithm, random forest to predict borrower default risk and analyze each algorithms' accuracy rate.

This project will contribute to several strands of literature. First, the project will contribute to the utilization of machine learning on financial market literature by offering new evidence of the utilization on p2p lending platform. Despite more and more research has shown that how machine learnings can be utilized at bankruptcy prediction using models such as neural networks (Zhang et al., 1999), instance-based learners (Park and Han, 2002), Bayesian models (Sarkar and Sriram, 2001), rule learners (Thomaidis et al., 1999), decision trees algorithms (Mckee and Greenstein, 2000) and Support Vector Machines (Shin et al., 2005). Research focusing on p2p lending platform is still very limited. Besides, choosing a particular model may not be convincing given the strengths and weaknesses of each method. Thus, searching for best distress prediction models is still in progress. This study provides a critical analysis of most commonly used P2P default risk prediction models. I try to use a representative algorithm for each one of the most common machine learning techniques so as to investigate the efficiency of ML techniques in the setting of P2P lending. Second, I will also provide a more complete picture of the effect of machine learning methods on combing soft information and hard information. Testing the variation of each methods weakness and strength is also one of our innovation. Our study will focus on China where both the capital market and the P2P lending market sector are very large and developed. Nonetheless, the findings should also be of interest to regulators and policymakers in countries where P2P are the primary source for financing.

The following section describes the data set of our study and the research design process. Some elementary Machine Learning definitions and a more detailed description of the used techniques and algorithms are given in section 2. It is hoped that the brief introduction to methodological details of these models would be of great use to those with recent interest in this field. Section 3 will present the direction of cross-sectional variations of each method under different situation. Finally, section 4 discusses the conclusions and some future research directions.

2. Research Design

2.1 Data

The data comes from the auction of one of China's P2P lending platforms. I get loan-by-loan data from a private data vendor, called CNRDS. I get the information related borrower's certification status, occupation, gender, age, credit rating, registration time, historical repayments number, and historical default and pay off, the length of the loan, the amount, interest rate, the contract term and so on. The dependent variable is whether the borrower has default or not, a dummy variable, equal to 1 if default and 0 otherwise. I will divide the list of successful borrowings into 80% training sets and 20% of the test set, and use the training set samples to train each algorithm model. Besides, to clean the data, there are several traditional steps. First, due to the imbalance between the number of default and non-default sample size, I adopt a random sampling method to make sure that in both the test sample and training sample, the default rate is close or similar. Then, I normalized the continuous variables to ensure the accuracy of various machine learning algorithms methods. In the data collected by the online loan platform, the values of different indicators vary greatly, the magnitude of the variable is very sensitive to the result of the distance calculation. The method used in this paper is to x=(x- min)/ range, where range=max- min. So that each variable has a value between 0 and 1. Finally, I construct the default risk prediction model using different machine learning method.

2.2 Machine Learning Methods

2.2.1 BP Neural Network

The BP neural network is a multi-layer forward. It includes two parts in its learnings process: the forward propagation of the signal and the backpropagation of the error. First, I will put all the related borrowers' information into the first layer. The middle layer contains the hidden layer and is affected by the research design. The final layer is the outcome variable, default or not default. The output layer calculates the error between the actual output and the expected output, distributing the error to the neurons of each hidden layer, and calculates the adjusted weight by the formula until the error is less than a given threshold or the number of learning reaches the upper limit. In P2P lending, there are many variables affecting the default probability of the borrower. Therefore, the number of neurons in the input layer is larger, while the output layer has only one neuron, and 0 and 1 respectively indicating non-default and default. The number of neurons in the hidden layer is determined by an optimization algorithm such as an empirical formula or a genetic algorithm.

2.2.2 Decision Tree

In P2P network lending, lenders' investment decisions are multi-level. The lender will review the borrower's certification to make a reasonable investment only when the target borrowing amount is greater than 20,000. Second, the lender decides to continue his/her investment decisions only when the borrower's credit rating is AAA and above. This decision process is not reflected in the multiple logistical regression, but the decision tree method can be utilized to mimic this type of decision-making process. Decision Tree classifier is a supervised learning algorithm which builds a binary tree where each node level has an attribute value split. A Decision Tree is built top-down from a root node and involves partitioning the data into subsets that contain similar values. Entropy is a measure that is used to calculate the homogeneity. A completely homogeneous sample set will have entropy value of 0 and an evenly distributed sample will have the value 1 as its entropy. The decision tree is a top-down process. At the node, the test results are assessed using information gain, information gain rate, and Gini coefficient. Information gain at each level or attribute is calculated by using the entropy of the parent and the weighted sum of the entropy of its children. This information gain value acts as the split at each node. This tree represents the set of rules used for predicting. After this assessment, the best attributes are selected to divide the sample into several subsets. Finally, based on the entire sample, a complete decision tree is trained, and each node represents a combined rule.

2.2.3 KNN (K-Nearest Neighbor) Classification Algorithm

KNN is an instance-based classification algorithm. In the P2P lending setting, the relevant borrowers' information and output information are known in the training set. For the test sample, KNN methods will calculate the distance formula based on the input characteristics and find the nearest neighbors target sample. The distance between neighbors is a function of the similarity among different borrower's characteristics. In this process, I can set different weight based on the distance.

2.2.4 Random Forest

Random forest classifier is another tree-based ensemble supervised learning algorithm that generates multiple Decision Trees (forest) for random subsets of the data and predicts the class with highest frequency after running the sample on all the Decision Trees generated. Random forest is an expansion of the decision tree model. Random forests use multiple decision tree training samples, building by random selection of the data and features. Random forest helps in overcoming the over-fitting problem experienced in Decision Tree classification. It provides a significant increase in the accuracy of the model.

2.3 Model Comparison Methods

In this proposal, I select three measures to compare the effectiveness of each machine learnings method, including accuracy, Type-I error, and AUC. To assess the accuracy of each prediction model, there are four situations. Table 1 provides information for each combination.

Table 1. Prediction and actual combination matrix

	Prediction: default	Prediction: not default
Actual: default	TP	FN
Actual: not default	FP	TN

Accuracy is measured as (TP+TN)/ (TP+FN+FP+TN). The higher the value, the more accurate the model. Type-I error = FN/ (TP+FN). Type-I error refers to the probability that a defaulted borrow is predicted as not default. The lower the value, the better the model. AUC (Area Under roc Curve) is calculated as TP/ (TP+FN) and FP/ (FP+ TN). If the AUC is close to 1, the model is well-performed.

3. Cross-Sectional Design

In this section, I want to assess which model is better to predict the default risks of borrowers when there is more soft information available or when there is more hard information available. I choose another US platform, called Prosper, which offers a social network, in which borrowers and lenders can interact. Both, creditors and debtors, benefit from this network which helps to mitigate information asymmetry (see Freedman and Jin, 2008; Berger and Gleisner, 2009). The rationale behind this test is that human are recognized are more intelligent at handling soft information. I want to test the effectiveness of machine learnings methods when they are soft information available.

Prosper the interest rate of a loan used to be conducted by a Dutch auction process until December 19th 2010, when this procedure was replaced by a posted price mechanism. Lin, Prabhala and Viswanathan (2013) found that social network information help lenders lead to good judgments of borrowers from the largest online P2P lending platform, Prosper.com. Iyer et al. (2009) mainly studied how lenders in P2P lending markets judge the creditworthiness of borrowers. They found that though lenders consider more the standard banking “hard” information, the “soft” information like communication with borrowers online or have similar experience also play a role in the success of borrowing.I will utilize Prosper P2P lending platform and “Paipaidai” platform data. There is 27 hard information in “Paipaidai” related to borrowers’ characteristics including age, education, marriage, tenure, city, industry, firm size, income, credit level, credit amount, lending history, lending purpose, interest rate, and so on. Since the soft information is very hard to quantify. In Prosper, there is soft information involved. So I test how the model performs differently using the Prosper database.

4. Conclusion

In this proposal, I want to ask two questions, 1) how different machine learning model performs in terms of predicting the default risk of borrowers; 2) how the effectiveness of machine learnings model varies by hard information and soft information. By utilizing BP neural network, decision tree, KNN classification algorithm, and random forest, I propose to assess two P2P lending platform, “Paipaidai” and Prosper. By comparing the accuracy rate, type I error and AUC, I will get the conclusion which one is more suitable. This paper trys to combine the machine learning and the mind from genuine people to develop a better strategy to reduce the default risk.

References

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., ... Funk, B. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, *16*(2), 1.

Berger, S. C., & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *BuR Business Research Journal*, *2*(1). https://doi.org/10.1007/BF03343528

Chen, D., & Han, C. (2012). A Comparative Study of online P2P Lending in the USA and China. *Journal of Internet Banking and Commerce*, *17*(2), 1.

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics,* *47*(1), 54-70. https://doi.org/10.1080/00036846.2014.962222

Freedman, S., & Jin, G. Z. (2008). Do social networks solve information problems for peer-to-peer lending? Evidence from prosper.com. https://doi.org/10.2139/ssrn.1304138

Iyer, R., Khwaja, A. I., Luttmer, E. F., & Shue, K. (2009, August). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending?. In *AFA 2011 Denver meetings* *paper*. https://doi.org/10.2139/ssrn.1570115

Jin, Y., & Zhu, Y. (2015, April). A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending. In *2015 Fifth International Conference on Communication Systems and Network Technologies* (pp. 609-613). IEEE. https://doi.org/10.1109/CSNT.2015.25

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, *59*(1), 17-35. https://doi.org/10.1287/mnsc.1120.1560

Mangasarian, O. L. (1965). Linear and nonlinear separation of patterns by linear programming. *Operations research*, *13*(3), 444-452. https://doi.org/10.1287/opre.13.3.444

Odom, M. D., & Sharda, R. (1990, June). A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks* (pp. 163-168). IEEE. https://doi.org/10.1109/IJCNN.1990.137710

Shi, Y., Peng, Y., Xu, W., & Tang, X. (2002). Data mining via multiple criteria linear programming: applications in credit card portfolio management. *International Journal of Information Technology & Decision Making*, *1*(01), 131-151. https://doi.org/10.1142/S0219622002000038

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: the case of bank failure predictions. *Management science*, *38*(7), 926-947. https://doi.org/10.1287/mnsc.38.7.926

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, *15*(3), 757-770. https://doi.org/10.2307/2330408

Copyrights