P2P Investment Data Analytics: A Case Study of Lending Club

Bowen Yi[1, *]

[1]University of California San Diego, San Diego, the United States

*Corresponding author: boyi@ucsd.edu

Abstract. In this study, a forecasting model is constructed for default risk by applying several machine learning models, including logistic regression, random forest, gaussian naive bayes, artificial neural networks, support vector machine, and XGBoost. Cross validation and voting ensemble learning are also utilized for best suitable hyperparameters as well. Dataset comes from Lending Club and then cleaned, normalized and standardized it into an ideal input dataset which are splitting into training and testing ones later. Among the methods, XGBoost achieved the best performance.

Keywords: P2P lending; default risk; machine learning; supervised learning; Lending Club.

1. Introduction

1.1 Machine learning and finance

When referring to the early Artificial Intelligence (AI) systems, which Arthur Samuel introduced in 1959 while working for IBM, the phrase "machine learning" was primarily used to give examples of pattern recognition exercises that have a "learning" element. During that early period of time, wider AI system was the dominant field under which including the idea of Machine Learning systems. After that came the broaden of the range of practical applications of Machine, and jumped out of the limit which defined by former frame of AI.

The earliest instance of machine learning being used to address economic and financial issues dates back to 1974 [1]. While the first paper published not only applied Machine Learning models but also used them exclusively, of which the topic was about economics and finance came out in 1984 [2].

1.2 Lending club & related works

Lending Club, one of the top peer-to-peer (P2P) lending service providers, is headquartered in San Francisco, California, and is located in the United States. A P2P lending firm registered its loan offers as securities with the Securities and Exchange Commission for the first time at that time (SEC) that made it the pioneer of its peers. Lending Club is also a loan-trading service provider on the secondary market. Its main operation is allowing investors to buy notes secured by loan payments, enabling borrowers to access loans on the basis of an online platform. Given the convenience of searching and browsing each loan that listed on Lending Club website, the investors of loans can also select the loans they prefer to invest after their acknowledge of the information that contributed by the borrower, for example loan amount, the grade of their new loans, and why they apply for loans. The main income of the investors comes from the interest collected from each of these loans. And Lending Club charges both borrowers and investors fees which called origination fee and service fee separately.

P2P investment has its own advantages comparing with traditional banks. From borrower's perspective, the final production of the two organizations is the same which is loan. From lender's perspective, however, there exist difference between the business of the online platforms and the traditional banks. Thanks to explicit and implicit government guarantees which give benefit to banks which would result in rather lower funding-cost, while P2P platforms is a game-changer which in their operation they applied several new technologies that could lower costs of operation and risks in pricing and avoid legacy problems which can be aggravated by crisis. P2P lending platforms' business remove the intermediation of financial institutions [3] thus showing their innovations in both loan and investment field. For those who may not have availability to standard financial intermediaries,

P2P loans give the borrowers shortcuts to financing [4]. Why P2P lending platforms are able to compete with traditional lending markets as well as attracting so large number of investors, one of the main reasons could be its super ability that meet the demand of nowadays capitals [5].

Used historical data comes from the Lending Club website to analyze the operation of P2P company based on. The main risk of the project is default risk. A P2P company can make profit if it clearly knows the percentage of borrowers who will not pay the money back. When receiving a loan application, a decision has to be made by the company whether to approve the new loan or based on the information of loan appliers. Risks can be divided into two types in decision making. For the first one, if the loan is likely to be repaid, then a loss of business would come into exist to the company if the loan is not approved. One the other hand, if the loan is likely not to be repaid, then the borrower likely to default, then approving the loan can be a financial loss to the platform. Applications of ML principles are made in the implementation of a neural network with backpropagation to assess default risk, or the likelihood that a borrower will default [6]. The success rate of loans is inversely correlated with the borrowers' credit ratings [7]. Our classification models are intended to aid in the decision of whether or not to invest in P2P lending to a borrower.

Fig. 1. Lending Club's advertisement poster

1.3 Study overview

Our aim is to use models to forecast whether an applicant for loan is going to default or not. Dataset is obtained from Lending Club, and divided into training and testing datasets, then apply them to various machine learning models. Finally, performance of the models is evaluated.

2. Method and data processing

Figure 2 illustrates the structure of the methodology. The data comes from Lending Club, and dataset is preprocessed into a desirable one for machine learning models with few steps. Then having the processed data, which is divided into training and testing data of a 75/25 split. The training part is applied to various machine learning models. Performance of models is tested by applying testing data to them and then evaluate the models. Finally, summary of the result and output of data and models is presented.

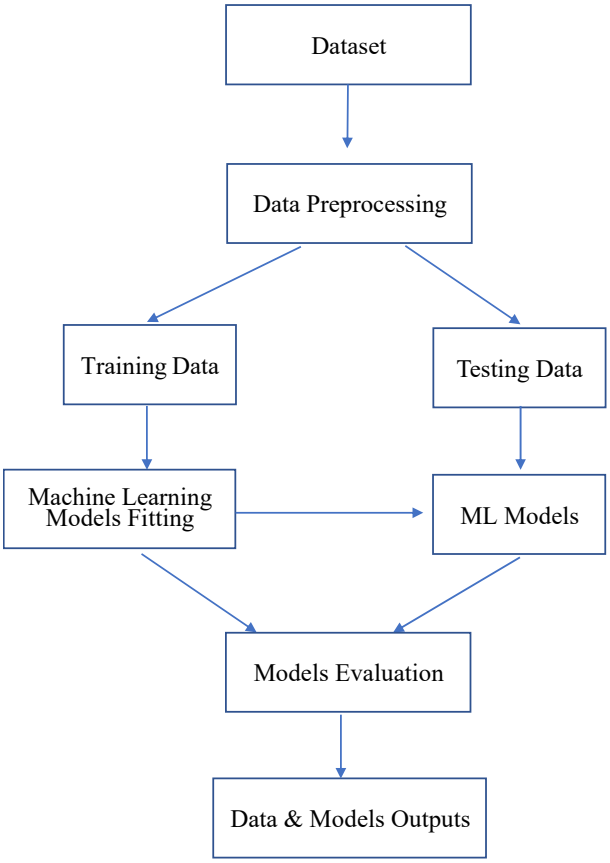

Fig. 2 Structure of the methodology

2.1 Dataset

The used dataset is collected from January 2007 to December 2018 which covers a very long time period and make sure our results are universally applicable. The original dataset contains 611780 observations and 27 features. Followed by data preprocessing procedure, the features are reduced to 23. Figure 3 illustrates the correlation heatmap of the features which are originally numeric and some features such as total payment and loan amount, total payment for investors and funded amount for investors, interest received to date and loan amount, principal received to date and funded amount. This could be because investors require higher payment and interest payment when they invest higher amount of money. Loan statues and issued numbers turn into loan volume of the dataset is in the Table 1.

As shown in Table 1, there are instances that borrowers fully paid their loans has significantly greater observation than instances that borrowers charged off. Our categorization models may have issues as a result of the two groups' imbalance because they have a tendency to favor the majority class which is Fully-Paid, and would result in models' overfitting. Table 2 shows the 19 features after reduction and their descriptions.

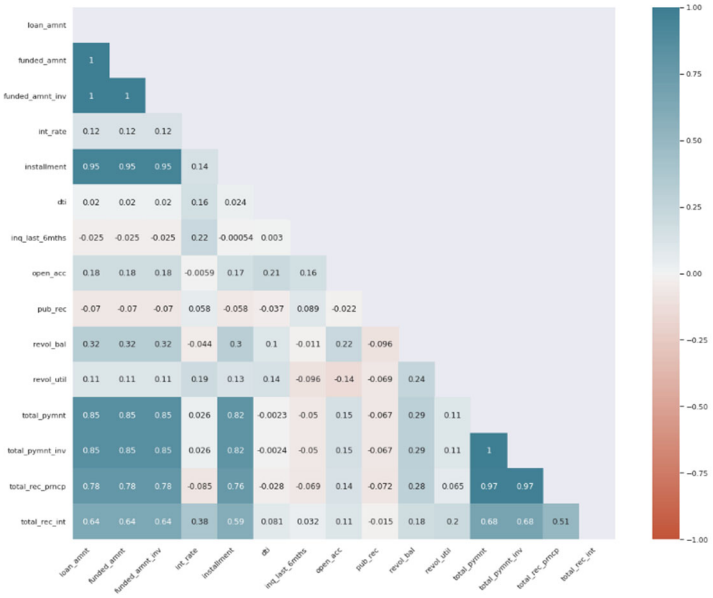

Fig. 3. The correlation matrix of features for the Lending Club dataset

Table 1. Loan volume broken down by loan status

Loan Status	Loan volume
Fully-Paid	482594
Charged-Off	129186
Total	611780

Table 2. Features and their descriptions.

Feature Variable	Description
Address state	State indicated in the loan application by the borrower
Dti	Total monthly payments made by borrower on outstanding debt obligations
Funded amount	The entire loan amount committed at the time
Funded amnt invt	The entire loan amount committed by investors at the time
Grade	Loan grade assigned by Lending Club
Home ownership	The borrower's declaration of home ownership during registration
Inq last 6mths	The quantity of creditors' queries over the last six months
Installment	The borrower's recurring monthly payment if the loan originates
Interest rate	Rate of interest for the loan
Loan amount	The borrower's indicated loan request's total amount
Open accounts	The quantity of credit-line that are open in the borrower's credit report
Out princpal	Principal that is still owing for the entire amount funded
Out princpal investor	Principal that is still owing for the entire amount funded funded by investors
Public record	Quantity of negative public records
Purpose	A classification of “why“ problem of the money borrower for the loans
Revol balance	Total outstanding revolving credit
Revolving utility	Revolving line use percentage
Term	A loan's total number of payments
Verification status	Indicates whether Lending Club has confirmed or not

2.2 Data preprocessing

For those features which are object, need be turned into numeric with mapping function, counting the unique values of each feature and then map them with integer starting from 0. Features including term, loan grade, home ownership, verification status, purpose, address state and loan status went

through this procedure. For the high correlation features, we need to drop them which includes total payment, total payment for investors, interest received to date and principal received to date. We also need to drop NA. Now, the mean and scaling to unit variance were removed from all of the numerical variables by using normalizing and standardizing. After that the dataset is divided into training and testing data which is 75/25. Our y is loan status which has binary values: '0' as Fully Paid and '1' as Charge Off, and X is the rest other features.

The loan status of the dataset is shown in Figure 4, and loan status distributions about other six features including term, credit grade, employment length, home ownership, verification status and loan purpose are shown respectively in Figure 5. For example, a greater proportion of 36-month loan in Fully Paid group is obtained, grade A, B, C are the majority in Fully Paid group, etc.

2.3 Machine learning models

The project uses a variety of machine learning models, such as logistic regression, random forest, gaussian naive bayes, and artificial neural networks, SVM, XGBoost. These are the typical supervised learning models. Cross validation and voting ensemble are also introduced.

In machine learning, a confusion matrix is among the best performance measures. The confusion matrix structure for binary classification is shown in Figure 4.

		Actual Values	
		Positive	Negative
Predicted Values	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**False Negative (FN)**

Fig. 3. Actual and predicted confusion-matrix

In terms of performance evaluation, accuracy, recall, precision, F1 score , and calculation formulas for the four parameters are listed as follow, from equation (1) to equation (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{4}$$

2.3.1 Random Forest

Random Forest is based on decision tree method, and a Random Forest has a number of decision trees that are organized using bagging, an ensemble learning technique. Finding the optimal split to subset the data is the goal of the decision trees in a Random Forest, and the data which would be trained through the Classification and Regression Tree (CART) algorithm. Metrics of Gini impurity are used to evaluate the quality of the split in this case.

Take a tree for instance, $W_j(x_i, x') = \frac{1}{k'}$, 0 unless if x_i is one of the k' points in the same leaf as x'. A Random Forest averages a set of m trees with unique weight functions' forecasts W_j, the predictions of which are as follow, equation (5).

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(x_i, x')y_i = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} W_j(x_i, x')\right)y_i \quad (5)$$

2.3.2 Artificial Neural Network

To explain the foundation of an Artificial Neural Network (ANN). It is a group of interconnected artificial neurons. A synthetic neuron receives impulses, analyses them, and then sends messages to neurons that are connected to it. The message delivered at such connection is always a real-number. The output of each neuron is determined by a nonlinear-function of the inputs' totals. Edges are the names for the connections between the layers. As learning occurs, the machine always changes the weight of neurons and edges each time. The purpose of wight is to modify whether to strengthen or weaken one of the many signals at a connection in the layers. The widely accepted industrial standard, logistic regression, is outperformed by artificial neural networks [8, 9].

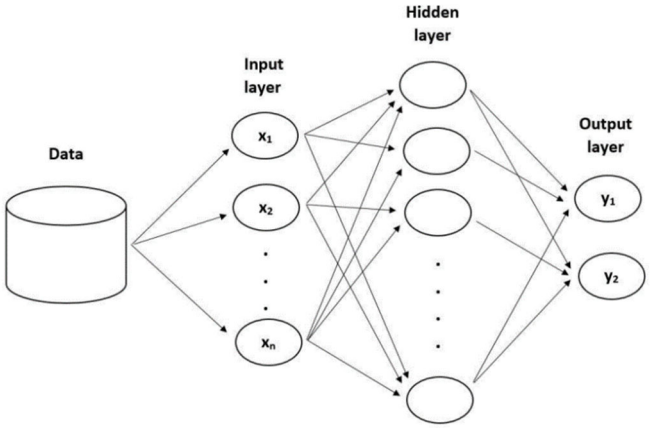

Fig. 4 ANN's multi-layer structure

The ANN algorithm was developed with inspiration from the human brain [10]. In an ANN, primarily, there are always three different kinds of layers. For the first layer which is so-called input-layer. Every input variable in this dataset will be represented by every neuron in the input-layer. Following that, the neurons in a new layer receive a modification that has an activation function from these changed neurons. There may be more than one Hidden layer, which is the name of the second-layer(s). The machine processed neurons from the hidden layer. After that, these final results are transferred to the top-layer. The output layer, the last layer, is made up of neurons that either belong to the final classes or carry prediction information, such probability predictions. Figure 5 presents a simple structure of ANN.

2.3.3 Support-Vector Machine

The Support Vector Machine (SVM) algorithm seeks to locate a hyperplane with n dimensions that clearly make data points classification. A binary linear classifier that assigns new instances to one category or the other is created by an SVM training method. In addition to linear classification, SVM can effectively do nonlinear classification by implicitly translating its input into a high-dimensional feature space.

2.3.4 XGBoost

A scalable and dispersed Gradient Boosting Decision Tree (GBDT) ML system is called Extreme Gradient Boosting (XGBoost). By examining a tree of "if-then-else" T/F feature questions, then calculating and finding out the bare minimum of number of these questions that needed to determine the chance of selecting the right answer. In such procedures, decision trees create a model which can predicts labels. Figure 8 shows the structure of a simple XGBoost model.

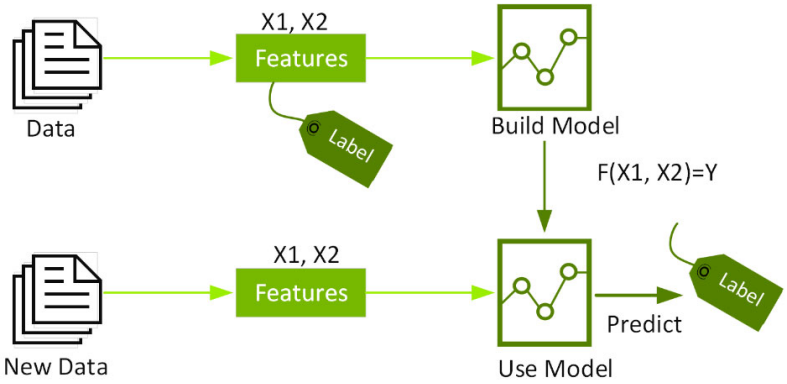

Fig. 5. XGBoost Structure [https://www.nvidia.com/en-us/glossary/data-science/xgboost/]

2.3.5 Voting Ensemble Learning

A technique for ensemble learning that integrates various algorithms is called ensemble models. It is not an algorithm itself, but rather a meta-algorithm in this sense. Methods of ensemble learning are useful because they can boost a predictive model's accuracy. The Voting Classifier uses a voting method and a list of various estimators as parameters. The majority rules mechanism and projected labels are used in the hard voting process. As an example, in Figure 6 the majority in voting section of the models for our classification is implemented.

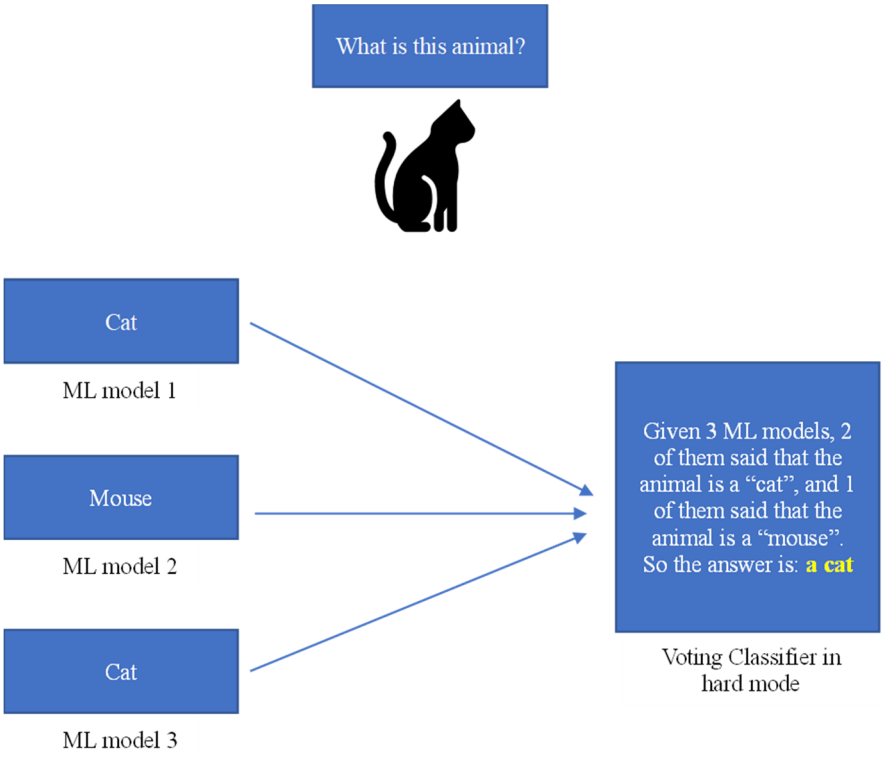

Fig. 6. Voting Classifier in "hard" mode

3. Model performance

Accuracy, recall, precision and F1 score are considered for each algorithm using the matching optimal hyperparameter set in order to evaluate the prediction models for default. Performance of the models are shown in Table 3 and Figure 8. For neural network, the hidden layer was set 30. The best performing model is XGBoost which gives the highest predicting accuracy. The six models were also subjected to a five-fold cross validation, the results of which are shown in Table 4. Table 5 gives the 10 trials of voting ensemble learning and the average accuracy of the prediction is 79.42% which is higher than 79.40% of XGBoost in cross validation, thus turns out that voting ensemble learning can help increase the accuracy of forecasting.

Table 3. Model performance

Algorithms	Accuracy	Recall	Precision	F1 score
Logistic_Regression	79.28%	79.28%	74.77%	73.67%
Random_Forest	79.10%	79.10%	74.36%	73.65%
Gaussian_Naive_Bayes	78.20%	78.20%	74.36%	73.65%
Neural_Network	79.38%	79.38%	75.11%	74.21%
SVM	78.99%	78.99%	62.39%	69.71%
XGBoost	79.38%	79.38%	75.10%	74.22%

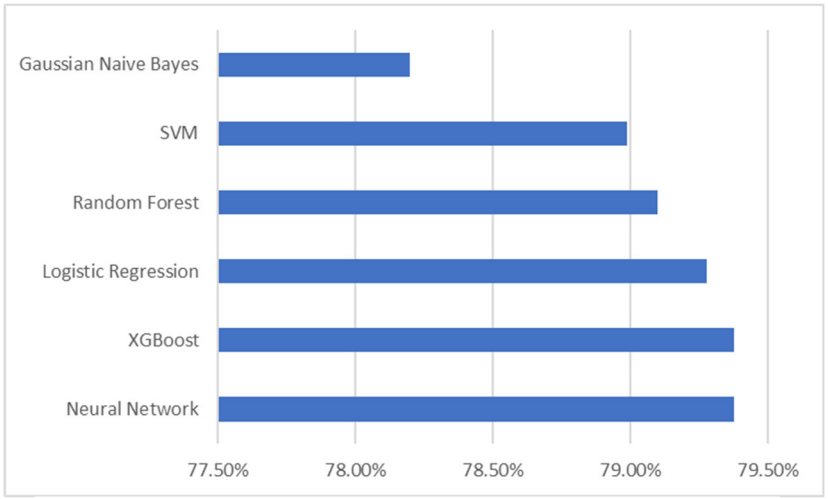

Fig. 7 Model accuracy

Table 4. Cross validation performance.

Algorithms	Accuracy
Logistic_Regression	79.16%
Random_Forest	79.04%
Gaussian_Naive_Bayes	78.27%
Neural_Network	79.29%
SVM	78.93%
XGBoost	79.40%

Table. 5 Voting ensemble

Trial	Accuracy
1	79.44%
2	79.47%
3	79.42%
4	79.34%
5	79.37%
6	79.36%
7	79.38%
8	79.38%
9	79.39%
10	79.52%
Average	79.42%

4. Conclusion

The study tried several machine learning classifiers as well as cross validation and voting ensemble in the classification case about credit risk within P2P market taking Lending Club as example. The main contribution of our study is finding the best model predicting whether a loan will default or not and ranking of different classification models and methods, which displayed in Tables 3, 4 and 5, based on our Lending Club dataset. We tried to find out the best performing model thus with given data of money borrower we can forecast whether he/she will default in the future or not which is very important in risk management of P2P loaning project.

Logistic regression, random forest, Gaussian Naive Bayes, Artificial Neural Networks, SVM and XGBoost models are used in our study, which also runs a five-fold cross validation method and a voting ensemble. The results show that among the models and methods, XGBoost in the best

performing model as it gives the highest accuracy in our evaluation testing. But it is also shown that the accuracy of the six models is very close to each other so that the advantage of XGBoost is not very obvious.

Currently, our own analysis has some limits. For example, as the dataset went through a rather wide range period of time including 2007-2008 global financial crisis during which the national and global economy was different from other years, the default risk forecasting models can be improved by incorporating financial performance of the macro economy for each year. There are huge differences of loan issuance between each year. The issuance of loans is related to global economy status which continue to climb all the way up since 2007 and topped in 2015, then dropped down in 2016 and 2017. Further research on how national and international economic data may aid in model optimization is warranted given that this market caters to borrowers in the lower income bracket. The exist of objective unscientificity in sample selection, and the scientificity and typicality of the sample selection are still inadequate, there is room for improvement in the sampling method, and there is still room for improvement in the sample size.

What's more, with the development of machine learning technology more and more new models come into exist these days. To study and put them into trial under the topic of economics and finance is necessary, and which indicates a future study direction. A higher accuracy is long-term goal, thus better machine learning models and development are always in great need in the field study as well as practice. And as mentioned above, an updated dataset with latest models comes a better forecast which conform the real situation well.

References

[1] Lee, Samuel C., and Edward T. Lee. "Fuzzy sets and neural networks." (1974): 83-103.

[2] Wang, H., Li, C., Gu, B., & Min, W. (1984). "Does AI-based credit scoring improve financial inclusion? Evidence from online payday lending". In 40th international conference on information systems. ICIS 2019.

[3] 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.

[4] Reddy, S., & Gopalaraman, K. (2016). Peer to peer lending, default prediction-evidence from lending club. The Journal of Internet Banking and Commerce, 21(3).

[5] Mills KG, McCarthy B (2016) The State of Small Business Lending: Innovation and Technology and the Implications for Regulation. HBS Working Paper No. 17-042.

[6] Zhang, Z., Gao, G., & Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. European Journal of Operational Research, 237(1), 335-348.

[7] Freedman, S., & Jin, G. Z. (2017). The information value of online social networks: Lessons from peer-to-peer lending. International Journal of Industrial Organization, 51, 185-222.

[8] Abdou, H., J. Pointon, & A. El-Masry (2008): "Neural nets versus conventional techniques in credit scoring in Egyptian banking." Expert Systems with Applications 35(3): pp. 1275–1292.

[9] Lessmann, S., B. Baesens, H. V. Seow, & L. C. Thomas (2015): "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." European Journal of Operational Research 247(1): pp.124–136.

[10] Wendler, T. & S. Grottrup (2016): Data Mining with SPSS Modeler. Springer International Publishing.Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.