Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 162 (2019) 503–513

Procedia Computer Science

www.elsevier.com/locate/procedia

A study on predicting loan default based on the random forest algorithm

Lin Zhu[a], Dafeng Qiu[a] , Daji Ergu[a,*] , Cai Ying[a], Kuiyi Liu[b]

[a]*College of Electrical & Information Engineering, Southwest Minzu university, Chengdu 610041,China*
[b]*key Laboratory of Electronic and Information Engineering, State Ethnic Affairs Commission, Southwest Minzu University,Chengdu 610041,China*

Abstract

Recently, with the advance of electronic commerce and big data technology, P2P online lending platforms have brought opportunities to businessmen, but at the same time, they are also faced with the risk of user loan default, which is related to the sustainable and healthy development of platforms. Therefore, based on the Random Forest algorithm, this paper builds a loan default prediction model in view of the real-world user loan data on Lending Club. The SMOTE method is adopted to cope with the problem of imbalance class in the dataset, and then a series of operations such as data cleaning and dimensionality reduction are carried out. The experimental results show that: Random Forest algorithm outperforms than logistic regression, decision tree and other machine learning algorithms in predicting default samples.

Peer-review under responsibility of the scientific committee of the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

Keywords: Random Forest; loan default; P2P lending

1. Introduction

In recent years, with the tremendous development of big data and the Internet financial, the P2P network lending industry continues to mature, amounts of P2P online lending platforms constantly springing up in the internet. Peer-to-peer (P2P) lending as a typical form of the Internet financial application, is a service which directly connects to the individual investors and loan borrowers to establish credit relationships and complete the transaction procedures through online platform, without the intermediaries of commercial banks. P2P lending has gradually become an important channel for small and medium enterprises and individuals to loan. Lending Club, the world's largest online financial platform for borrowers and investors, the amount of borrowers has

* Corresponding author.
E-mail address: ergudaji@163.com.

Peer-review under responsibility of the scientific committee of the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019)
10.1016/j.procs.2019.12.017

contributed to 3 million and invested over 50 billion USD. It based the Internet to build a marketplace which keeps costs lower and investment return greater than traditional commercial banks. It has created a novel mechanism that make borrowing and investing simple and easy for everyone.

With opportunities have challenges, P2P lending meets the needs of China's current economic development to a large extent, but the risks are following. Financial risks are mainly reflected in the liquidity risk caused by insufficient liquidity of funds, the credit risk brought by information asymmetry, the operational risks, and the legal risks brought about by imperfect laws and regulations related to Internet finance. In brief, compared with traditional finance, the characteristics of Internet financial risks are more complex. Due to the technical and virtual characteristics brought by Internet technology, it also has some special risks based on traditional financial conventional risks. Such as the internet financial risk is more sudden and contagious, besides the risk-destructive enhancement is more serious and less controllable. So risk aversion is one of the interesting and important discussions among cult of investors, policy makers, researchers, financial practitioners [1].

These risks will significantly increase the possibility of default of the borrowers and then P2P is confronted with the credit risks due to the default. Loan defaults are unfavourable to the profits of investors and the development of P2P lending platforms. Therefore, domestic and foreign scholars have conducted numerous researches on loan evaluation, which is an effective tool for the P2P platform to carry out credit risk assessment and control.

P2P lending platforms generally use a credit scorecard which is based on their business needs to construct the loan evaluation model, e.g., the LC score and FICO [2]. A credit scorecard can easily and quickly give each loan a credit score, but it can't effectively distinguish non-defaulters from defaulters [3].

Nowadays, with the maturity of big data technology, machine learning and artificial neural network are widely used in risk management controls of the financial field. Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Deep Neural Networks (DNNs) are widely used for prediction of stock prices and its movements [4, 5]. These models have been proved to be effective in the forecasting of financial time series. Scholars have found that the performance of Random Forest used in loan evaluation is better than other methods in P2P lending [3]. Based on their research, some scholars proposed to optimize the combination of decision trees in a parameter-optimized Random Forest by using genetic algorithm[6].In view of the current research, the Random Forest algorithm is adopted to construct a loan default prediction model based on Lending Club's loans of the first quarter of 2019,and four different approaches are conducted and compared with Random Forest in further testing. Our research is of great significance to improve the performance of loan evaluation and further facilitate the promotion and healthy development of P2P lending.

The rest of this paper is divided into following five sections. Section 2 presents a brief literature review of the work that has been conducted on loan evaluation and credit risk assessment. Section 3 describes the details of the Lending Club and the Random Forest. Then, experiments and the results are discussed in Section 4. Finally, Section 5 concludes.

2. Literature Review

At present, the research on P2P platform at home and abroad mainly place emphasis on the load defaults and credit risk assessment by using machine learning methods.

Research on the prediction of load default: Serrano-Cinca et al. used loan sample data from the Lending Club to account for default factors by adopting single factor mean test and survival analysis [7].Advanced-support vector regression (SVR) techniques are applied to predict loss given default of corporate bonds by Yao et al., the results show that versions of SVR techniques perform better than other methods [8]. Malekipirbazari and Aksakalli proposed a Random Forest (RF) based classification method to identify high-quality P2P borrowing customers, by comparing with different machine learning methods, the results indicate the RF-based method is significantly preferred than the FICO credit scores as well as LC grades in identifying the good borrowers. Emekter et al.

constructed a logistic regression (LR) model to predict the default probability of the borrower in the Lending Club, and the empirical evidence suggests that credit grade, debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults [9] .Bagherpour used KNN, SVM, Random Forest and Sand Factorization Machines (FM) algorithms for predicting loan default on a large set of data [10]. Kvamme et al. based on Convolutional Neural Networks (CNN) to predict loan default by taking into account the time series data related to customer transactions in current accounts, savings accounts and credit cards. The research showed that CNN model outperforms Random Forest classifier [11]. Kim et al. proposed a method combining label propagation and transductive support vector machine (TSVM) with Dempster-Shafer theory to accurately predict the default of social lending with unlabelled data [12].

Research on the credit risk assessment: Tang et.al proposed a model of trust spiral and applied it to the study of credit risk issues within the lending relationship between banks and small businesses [13]. Moradi and Mokhatab Rafiei train an adaptive network-based fuzzy inference system (ANFIS) using monthly data from a customer profile dataset and then using the newly defined factors and their underlying rules, a second round of assessment begins in a fuzzy inference system. Thus, they produce a table of bad customers on a monthly basis and creating a dynamic model based on the table for assessing the credit risk of the customers [14]. Brown, I. et al. obtained that Random Forest and Gradient Boosting classifiers perform outstanding in a credit scoring context and are able to cope preferably with pronounced class imbalances in these data sets by empirical study [15]. Li qualitatively analysed the possibility of loan defaults of borrowers who loan in the lending club, based on the borrowers' loan purpose, income level, residential address and work seniority, then via logistic regression model for predicting the default probability of borrowers so as to calculate the credit score of borrowers [16]. Zhang et al. adopted Multiple Instance Learning (MIL) to build a novel credit scoring model by using the socio-demographic and loan application data as well as the transaction history data of the applicant [17]. Djeundje B. V. and Crook used GAMs with cubic B-splines to estimate credit card survival models , the results show that GAMs outperform in improving the accuracy of predictions and so on[18]. Masmoudi K et al. adopted a discrete Bayesian network with a latent variable to model the loans subscribers who have default payment behaviour. The model is constructed to evaluate credit risk and cluster loans subscribers [19]. Papouskova and Hajek used heterogeneous ensemble learning to build a two-stage consumer credit risk model which adopts class-imbalanced ensemble learning and regression ensemble for predicting credit scoring and exposure at default respectively [20]. LightGBM, XGBoost, Logistic Regression and Random Forest are used by Ma et al. [21] and Coser et al. [22] to establish a series of prediction models for evaluating the probability of a customer's loan default. Cho et al. proposed an investment decision model in P2P lending market based on the instance-based entropy fuzzy support vector machine (IEFSVM) classification [23].

Through there are a large body of work on predicting the loan defaults in the Lending Club, the existing works which used Random Forest method for research can be further improved by our research.

3. Related knowledge and theory

3.1 Lending Club

Since the establishment of Lending Club, it has laid special emphasis on risk management. Only 10% of all loan applications can be approved. Lending Club classifies loans into seven grades, A-G, based on risk. Different levels indicate different degrees of risk and corresponding returns.

The overall operation mode of Lending Club is that it acts as the intermediary between the borrower and the investor, determines the loan interest rate according to the borrower's credit rating and the term of the loan amount, and then provides the loan demand to investors. The business pattern of Lending Club is shown as Fig.1. However, in fact, investors' money is not directly paid to the borrower but used to buy Lending Club bonds, so investors need to consider that once the Lending Club platform has other liabilities or collapse, investors may face the

credit or platform risk of Lending Club.

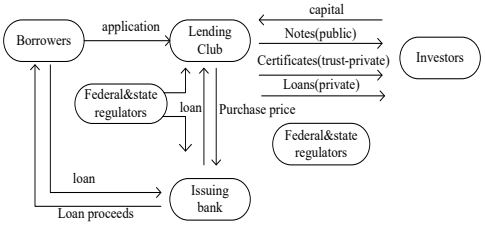

Fig. 1. The business pattern of Lending Club

3.2 Random forests (RF)

Random forest belongs to supervised learning algorithm, is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or means prediction of the individual trees.

The decision tree is a tree structure (which can be a binary tree or a non-binary tree). Each of its non-leaf nodes corresponds to a test of a feature, each branch representing the output of the feature attribute over a range of values, and each leaf node storing a category. The decision tree starts with a root node, tests the corresponding feature attributes in the category to be classified, and output branches are selected according to their values until the leaf node is reached, finally the category stored by the leaf node is regards as the decision result.

A random forest is a collection of decision trees in which each decision tree is unrelated. Selection metrics we used for splitting attributes in the decision tree is Gini index, and the number of levels in each tree branch depends on the algorithm parameter d [24].

The Gini Index at an internal tree node is calculated as follows: For a candidate (nominal) split attribute X_i, denote possible levels as $L_1;\ldots;L_j$. Gini Index for this attribute is calculated as:

$$G(X_i) := \sum_{j=1}^{J} \Pr(X_i = L_j)(1 - \Pr(X_i = L_j)) \qquad (1)$$

$$= 1 - \sum_{j=1}^{J} \Pr(X_i = L_j)^2$$

The reason why we choose Random Forest algorithm over other machine learning algorithms is that Random Forest has the following advantages:

- It runs efficiently on large data bases and is unexcelled in accuracy among current algorithms;
- It can fix well with errors in class population unbalanced data sets;
- It supports an effective method for estimating missing data, which can maintain accuracy even in the case that a large scale of the data is missing.

4. Experiment

4.1 Dataset characteristics

In this paper, the dataset we used is derived from the Lending Club for the first quarter of 2019. It contains more than 115,000 original loan data of users with 102 attributes. After that, the missing values are filled in by means of mode interpolation, and the duplicate or meaningless attributes are deleted, finally we have retained 15 attributes and table 1 shows the attributes for the experiment.

4.2 Feature engineering

At first, in our work, we define a new feature named ‘installment feat’, which represents the monthly repayment expense of a user as a percentage of monthly income. The greater the value of ‘installment feat’, the lender's debt will more stressful and more likely to default.

Secondly, feature abstraction. We encode loan statuses ‘Current’, ’Fully paid’, ‘Issued’ as normal = 0,encode ‘Default’ , ’Charged off’, ‘In Grace Period’, ‘Late(16-30days)’ and ‘Late (31-120days)’ as default = 1. Then we can visualize the loan status as shown in Fig.2 below. Samples with a “normal” loan status accounted for 98.47%, but the “default” samples were only 1.53%, indicating that a serious category imbalance in the dataset. Meanwhile, we abstract the feature ‘emp_length’ and ’grade’ , and the remaining features are one-hot encoded.

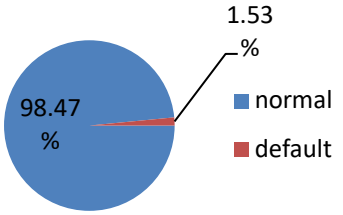

Fig. 2. Percentage of each loan status

Thirdly, after feature scaling, we need to carry out feature selection, giving priority to the features with high relevance to the target and removing irrelevant features can reduce the difficulty of learning. We used the Recursive Feature Elimination method to select 30 features with the strongest correlation with the target variable, and eliminated the features step by step to achieve the first dimensionality reduction, with the independent variable reduced from 102 to 30. We plotted the Pearson correlation graph of the 30 features, as shown in Fig. 3. On the basis of the first dimension reduction, redundant features are selected and eliminated by Pearson correlation graph, the dimension of features is reduced from 30 to 15 according to Fig. 4 .

Population correlation coefficient is defined as the covariance and standard deviation between two variables:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{2}$$

Estimate the covariance and standard deviation of the sample to obtain the Pearson correlation coefficient of the sample:

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{3}$$

Finally, we adopt the Random Forest algorithm to rank the importance of features and reduce the learning difficulty to achieve the purpose of optimizing the model calculation.

Table 1. The selected attributes and pre-treatments.

Feature name	Description and Pre-treatment	Type
loan_amnt	The listed amount of the loan applied for by the borrower.	Numeric
installment	The monthly payment owed by the borrower if the loan originates.	Numeric
grade	LC assigned loan grade.	Numeric
open_acc	The number of open credit lines in the borrower's credit file.	Numeric
total_pymnt	The total payment of the borrower.	Numeric
total_rec_int	The interest on loans for the borrower.	Numeric
home_ownership_ANYor_MORTGAGE	The home ownership status provided by the borrower during registration.	Nominal
		Nominal
verification_status_Not Verified	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified.	Nominal
application_type_Individual	Indicates whether the loan is an individual application or a joint application with two co-borrowers.	Nominal
purpose_major_purchase	A category provided by the borrower for the loan request.	Nominal
purpose_renewable_energy		Nominal
purpose_small_business		Nominal
purpose_vacation		Nominal
term_ 36 months	The number of payments on the loan. Values are in months and can be either 36 or 60.Pretreatment: binary discretization.	Nominal

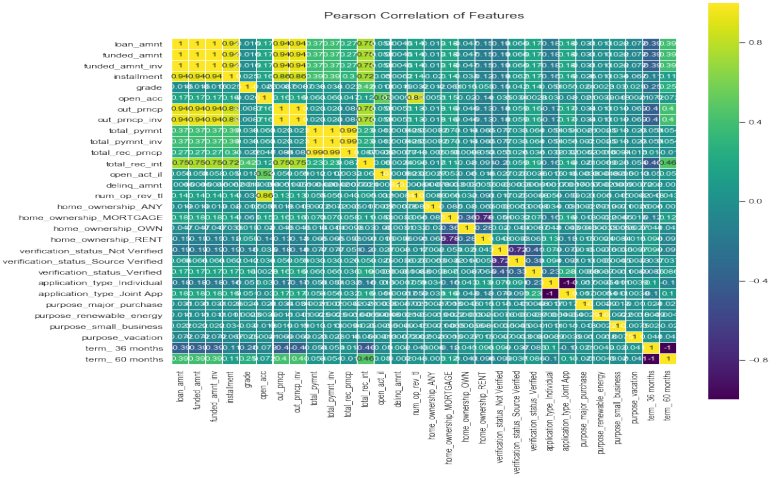

Fig. 3. Person correlation of 30 features

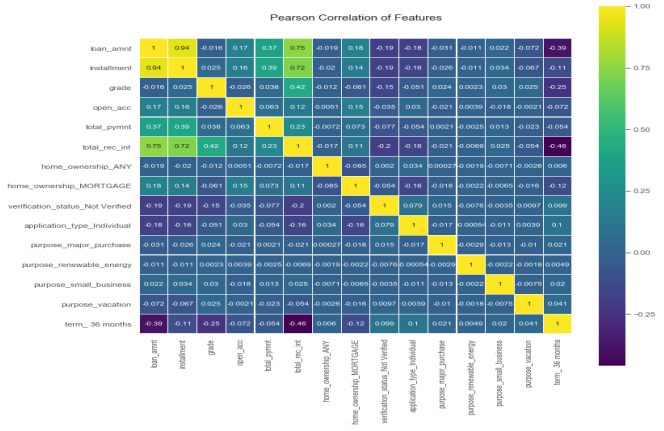

Fig. 4. Person correlation of the selected 15 features

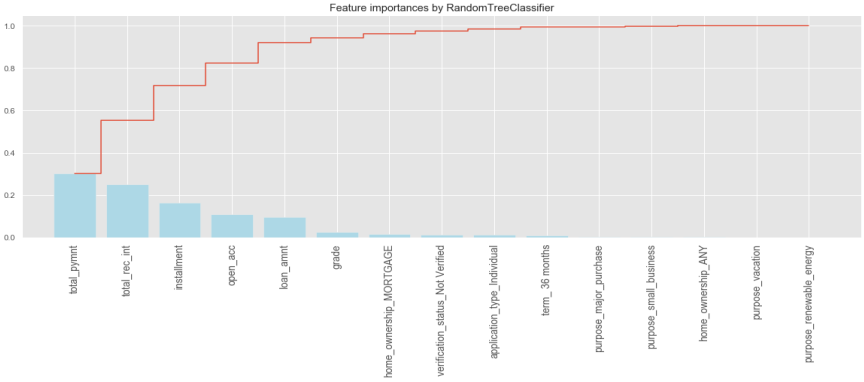

Fig. 5. The ranking of the importance of features

4.3 SMOTE method

As mentioned above, the target variable 'loans status' has a large difference in the number of normal and default categories, which will cause trouble to model learning. The method of oversampling is used to handle sample imbalance problem, we adopt SMOTE (Synthetic Minority Oversampling Technique) method in this paper. The rationale of SMOTE is:

- The nearest neighbor algorithm is adopted to calculate the K nearest neighbours of each minority sample with Euclidean distance as the standard
- Setting a sampling proportion according to the unbalanced proportion of samples, and for each sample x_i of the minority class, randomly select several samples from its k-nearest neighbours.
- Assume that the selected neighbour is x_n . For each randomly selected neighbour x_n , new samples are constructed according to the following formula with the original samples respectively.

$$x_{new} = x_i + rand(0,1) * | x - x_n | \qquad (4)$$

By iterating each samplex_i , the original sample size of the minority class is eventually expanded to an ideal ratio.

4.4 Model evaluation and assessment

In this paper, we mainly focus on the three performance evaluation criteria for classifier comparison of accuracy, AUC and ROC.

- Accuracy

Accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set. The formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

- F1-score

F1-score, also called a balanced F Score, is defined as the balanced average of Precision and recall.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{6}$$

- Recall

Recall is the fraction of all positive (default) instances the classifier correctly identifies as positive. It is also known as the True positive rate [14].

- ROC（receiver operating characteristic curve）

In statistics, a receiver operating characteristic (ROC), is a two dimensional graphical plot that illustrates the performance of a binary classifier system. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC curve can intuitively represent the performance of classifier.

$$FPR = \frac{FP}{FP + \mathrm{TN}} \tag{7}$$

$$TPR = \frac{TP}{TP + \mathrm{FN}} \tag{8}$$

- AUC value

AUC represents the area under the Receiving Operating Characteristic (ROC) curve in the testing dataset.

Assume that ROC curve is formed by sequential connection of points with coordinates of $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_m, y_m)\}$,AUC can be estimated as,

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \bullet (y_i + y_{i+1}) \tag{9}$$

where AUC values range from 0.5 to 1.0, and a classifier with larger AUC value has better performance.

4.5 Results and discussion

In this paper, the Random Forest algorithm is used to predict its performance (accuracy, AUC, F1-Score and recall), and compared with another three machine learning methods namely the decision tree, the logistic regression and the SVM. The obtained results are displayed in Table 2. The results show that, the performance of random forest and decision tree have comparable performance than that of support vector machine and logistic regression, but the random forest still performs the best, with an accuracy of 98%, higher than the decision tree with an accuracy of 95%. The precision and recall of the prediction model based on random forest are all above 0.95, indicating that the model has strong ability of generalization.

Figure 8 shows the ROC curves of these four methods. The closer the ROC curve is to the upper left corner, the higher the recall rate of the model will be. The point on the ROC curve closest to the upper left corner is the best threshold with the least classification errors, and the total number of false positive examples and false negative examples is the lowest. We can conclude obviously that the random forest algorithm outperforms the other three approaches.

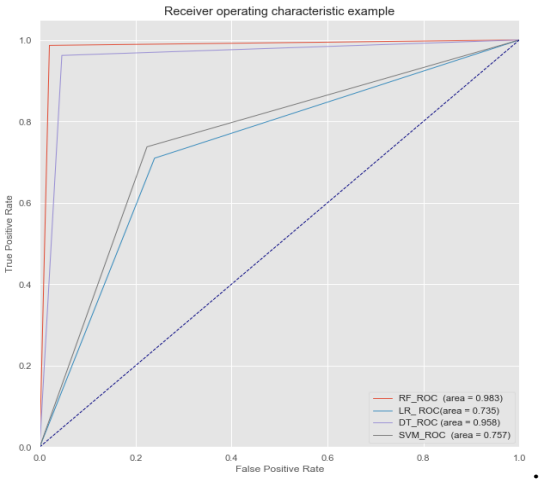

Fig. 6. ROC performance comparison of the four classifiers

Table 2. Evaluation metrics comparison of the four techniques

Rank	Classifier	Accuracy (%)	AUC	F1-score		Recall	
				0	1	0	1
1	Random Forest	98%	0.983	0.98	0.98	0.98	0.99
2	Decision Tree	95%	0.958	0.96	0.96	0.95	0.96
3	SVM	75%	0.757	0.76	0.75	0.78	0.74
4	Logistic Regression	73%	0.735	0.74	0.73	0.76	0.71

5. Conclusion

In this paper, the random forest algorithm is adopted to build a model for predicting loan default in the lending club and the results are compared with other three algorithms of logistic regression, decision tree and support vector machine. The experiment shows that the random forest algorithm performs outstanding than the other three algorithms in the prediction of loan default and has strong ability of generalization. In further study, we will try to conduct experiments on larger data sets or try to tune the model so as to achieve the state-of-art performance of the model.

Acknowledge

This work was supported by grants from the National Natural Science Foundation of China #U1811462, #71774134 and #71373216, in part by the Innovation Scientific Research Program for Graduates in Southwest Minzu University (No. CX2018SZ158).

References

[1] Challa, M.L., Malepati, V. & Kolusu, S.N.R. Financ Innov (2018) 4: 24. https://doi.org/10.1186/s40854-018-0107-z
[2] Briceno Ortega,Ana Cecilia and Frances Bell. “Online social lending: borrower generated content [C].”AMCIS 2008 Proceedings, 2008. 380
[3] Malekipirbazari M , Aksakalli V . Risk assessment in social lending via random forests[J]. Expert Systems with Applications, 2015, 42(10):4621-4631.
[4] Selvamuthu, D.， Kumar, V.&Mishra, A. Financ Innov(2019)5:16 https://doi.org/10.1186/s40854-019-0131-7.
[5] Zhong, X. and Enke, D. Financ Innov（2019）5：24 https://doi.org/10.1186/s40854-019-0138-0
[6] Ye, Xin, Dong, Lu-an, Ma, Da. Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score[J]. Electronic Commerce Research and Applications, 32:23-36.
[7] Serrano-Cinca, C., Gutiérrez-Nieto, B., López-Palacios, L., 2015. Determinants of default in P2P lending. PLoS One 10, 1–22.
[8] Yao X , Crook J , Andreeva G . Support vector regression for loss given default modelling[J]. European Journal of Operational Research, 2015, 240(2):528-538.
[9] Emekter, R., Tu, Y., Jirasakuldech, B., Lu, M., 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl. Econ. 47, 54–70.
[10] Bagherpour, A. (2017),Predicting Mortgage Loan Default with Machine Learning Methods; University of California, Riverside;
[11] Kvamme,H. et al. (2018), Predicting Mortgage Default Using Convolutional Neural Networks; Expert Systems With Applications, 102, pp.207-217;
[12] K. Aleum and S.B. Cho, “An ensemble semi-supervised learning method for predicting defaults in social lending,” Eng. Appl. Artif. Intell., vol. 81, pp. 193–199, May 2019
[13] Tang, Y., Moro, A., Sozzo, S. et al. Financ Innov (2018) 4: 19. https://doi.org/10.1186/s40854-018-0105-1
[14] Moradi, S. & Mokhatab Rafiei, F. Financ Innov (2019) 5: 15. https://doi.org/10.1186/s40854-019-0121-9
[15] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39, 3446-3453.
[16] Y.W. Li, Research on credit score of P2P online lending [D]. Harbin Institute of Technology, 2018.
[17] Zhang, T. et al. (2018), Multiple Instance Learning for Credit Risk Assessment with Transaction Data;Knowledge-Based Systems, 161, pp.65-77.
[18] Viani Biatat Djeundje, Jonathan Crook, Identifying hidden patterns in credit risk survival data using Generalised Additive Models., European Journal of Operational Research (2019), doi: https://doi.org/10.1016/j.ejor.2019.02.006
[19] Masmoudi K , Abid L , Masmoudi A . Credit risk modeling using Bayesian network with a latent variable[J]. Expert Systems with Applications, 2019, 127:157-166.
[20] Monika Papouskova, Petr Hajek , Two-stage consumer credit risk modelling using heterogeneous ensemble learning. Decsup (2019), https://doi.org/ 10.1016/j.dss.2019.01.002
[21] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, X. Niu, 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 (2018), doi:https://doi.org/10.1016/j.elerap.2018.08.002

[22] Coser, A , Maer-matei, MM, Albu, C .Economic Computation & Economic Cybernetics Studies & Research. 2019, Vol. 53 Issue 2, p149-165. 17p.
[23] Cho, P., Chang, W., Song J.-W. Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision, IEEE Access, 7, (2019),doi:https://doi.org/10.1109/ACCESS.2019.2896474
[24] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.