Loan Default Prediction Based on Machine Learning Methods

Yuran Zhou
201911011130@mail.bnu.edu.cn

School of Statistics, Beijing Normal University, Beijing, China

Abstract—Loan default prediction helps institutions predict whether a borrower will default on a loan and decide whether to lend, thereby reducing losses. We investigate the performance of different machine learning models in predicting customers' loan defaults. Four machine learning models: Logistic Regression, Decision Tree, Random Forest and XGBoost, are used to predict the loan default, considering dependent variables such as the value of all the assets, living status and yearly income. Our results show that XGBoost is the best model with the highest *Recall* of 0.35 and AUC of 0.832. This study is expected to help lending institutions identify potential default users, and then decide on who to accept or refuse for a loan.

Keywords-Loan Default Prediction; Machine Learning; Extreme Gradient Boosting Tree

1. Introduction

Customer's failure to pay the loans on time may cause losses to lending institutions, and meanwhile, the inability of individuals to pay their debts is one potential trigger for a financial crisis. Therefore, from the perspective of lending institutions and the whole country, loan default prediction is of great importance in practice. Loan default prediction is to forecast the probability of default based on the information already available about the loan applicant, and to determine whether to release the loan. Default prediction is critical for financial institutions and investors. For example, banks use default probabilities to select target borrowers, measure the terms of new loans, and control the risks arising from lending activity [1]. To conduct the prediction, information about the borrower is needed before the loan, including the borrower's moral quality, financial strength, repayment ability, guarantee conditions, etc. This paper focuses on data about such information of the customers of *The Grant Group of Companies*. This information can help banks and institutions construct credit management mechanisms to predict the default of loans.

However, the credit management mechanism of banks and institutions is not perfect. Loan management not only is too subjective in the process of loan approval and disbursement but also lacks scientific evaluation in both the pre-loan investigation and the loan qualification review process. Many qualitative methods have been used for credit analysis, but there was a lack of systematic statistical analysis to evaluate credit risk.

The techniques of machine learning (ML) have seen tremendous growth over the past two decades and they have been applied to improve the accuracy of credit risk assessment [2]. ML is a technique that computationally learns from experience to improve system performance and it has been widely used in the problem of prediction. For example, in marketing applications, ML

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China

DOI 10.4108/eai.2-12-2022.2328740

can be applied to predict a customer's propensity to buy a product or discontinue an order; in economics, it can be used to predict the probability that a person will choose to enter the labour market.

This study focuses on the use of Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Extreme Gradient Boosting Tree (XGBoost) to predict the likelihood that a borrower will default on a loan. Results show that XGBoost has the highest *Recall* and AUC, so we consider that it is the best method. Previous studies have focused on applying these methods to classical datasets with the goal of achieving the highest accuracy. Our contributions are that firstly, we use these methods on data that had barely been studied. Our conclusions demonstrate that these methods are useful in solving the problem of loan default prediction even using such data, and further, ML methods have great potential in the field of prediction and can be extended to the study of problems in other similar fields. Secondly, unlike most studies, we measure the performance of the methods mainly by *Recall* instead of *Accuracy*, because the inability to detect the potential default is more costly than mistaken the identified default. Thirdly, we analyze the importance of factors influencing the probability of loan default, which is ignored by many studies.

The rest of this paper is organized as follows. Section 2 gives a brief literature review of the work that has been conducted on loan default and the ML models. Section 3 describes the loan default prediction models and the measurement for model evaluation. Section 4 presents the dataset and discusses the results. Finally, Section 5 gives the conclusion.

2. Related Work

2.1 Default prediction

Various methods have been used to predict loan default, such as Linear Probability model and Probit model, and one of the most commonly used models is Linear Discriminant Analysis (LDA) [3]. Lugovskaya used LDA to predict defaults of Russian small and medium enterprises (SMEs) and found that this method is very effective in this case [4]. Mileris used LDA together with a Simple Bayesian Classifier to measure the default probability of clients, which improved the correct classification rate and rating scale [5]. Krichene used the Naive Bayesian Classifier algorithm to evaluate the loan risk of a commercial bank, and results showed that the classification accuracy on the validation set was 58.66%, but the type I and type II errors were relatively high [6].

2.2 ML methods

Wildly used techniques are Artificial Neural Networks (ANN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting Tree (XGBoost).

Kumar & Bhattacharya compared the prediction performance of ANN and LDA in the study of credit rating prediction, and they presented that ANN method outperformed LDA when dealing with missing data [7].

Ciampi used LR to build social enterprise default prediction models, and he showed that the model considering financial ratios and corporate governance characteristics performed better than the model with only financial ratios [8]. Agbemava et al. mentioned that default risk models

using LR are relatively efficient and cost-effective, so they suggested microfinance institutions use this model to determine loan default levels [9]. Khemais et al. compared the effectiveness of LDA and LR for default risk prediction and they found that both classifiers worked well, but LR was slightly better than LDA [10]. Further, Sheikh et al. pointed out that the customer default prediction model using LR has the advantage of identifying the right target customers since LR can easily evaluate the probability of loan defaults [11].

RF and DT are another two methods widely used in loan default prediction. For example, Zhu et al. constructed a loan default prediction model using user loan data from Lending Club and concluded that RF outperformed LR and DT [12]. Madaan et al. conducted a comprehensive comparative analysis of both RF and DT algorithms in predicting loan defaults in the banking sector and found that RF has a higher accuracy [13] . Xu et al. applied RF, XGBoost, Gradient Boosting Model and Neural Network to predict loan defaults by using data from Renrendai.com in China, and they concluded that RF has higher accuracy and kappa value than other models [14].

Chang et al. used a Support Vector Machine to predict the probability of default rate of Lending Club borrowers to maximize returns [15]. Wang et al. used K-Nearest Neighbor Classifier to evaluate personal credit information [16]. Coşer et al. used classifiers such as LightGMB and XGBoost to build predictive models to predict the probability of customer loan default [17]. Duan proposed a Deep Neural Network-based decision-making method to assess P2P lending risk and found that the method achieved 93% accuracy in classifying the test data [18].

3. Modeling

This section gives a detailed description of model and the measurements used to evaluate the models.

3.1 Models

3.1.1 Logistic Regression

Logistic Regression is a powerful multi-category classification method. In addition to category label information, it provides explicit classification probabilities. And it also allows analysis of prediction values for all types of data, such as continuous variables, discrete variables, and dummy variables. It has the advantage that we do not need to make restrictive assumptions about the normal distribution or equal dispersion matrix of the predictor variables or consider the failure of prior probabilities.

The data is divided into training and test sets. For a given training set denoted by $D = \{x_i, y_i\}_{i=1}^{n}$, $x_i \in R^n$ denotes predictor measurements, and $y_i \in \{0, 1\}$ denotes the discrete outcomes. We maximize the log-likelihood function

$$L(\beta_0, \beta) = \sum_{i=1}^{n} y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0 + \beta^T x_i}), \quad (1)$$

where β_0 and $\beta = (\beta_1, \ldots, \beta_n)^T$ are the unknown parameters.

3.1.2 Decision Tree

DT is a predictive model that represents a mapping relationship between object attributes and object values. It is a very common classification method where each sample has a set of attributes and a predefined category. A classifier is obtained by learning which is able to give the correct classification for the emerging objects.

Here, we denote a sample of training examples by S, and the proportion of positive examples in S by p. Then the entropy of S, denoted by $H(S)$, can be expressed as

$$H(S) = -p \times log_2 p - (1-p) \times log_2 (1-p). \quad (2)$$

Entropy measures purity at each node, information gain looks at all nodes together and the expected drop in entropy after split. We first use the entropy of the remaining tree as our measure to prefer one attribute over another, and then choose the attribute that gives us the biggest information gain. The expected reduction in entropy due to sorting on A, denoted by $Gain(S, A)$, is given by

$$Gain(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v). \quad (3)$$

3.1.3 Random Forest

RF is a modified form of bagging that creates ensembles of independent decision trees. It randomly selects a subset of variables and trains hundreds of trees to determine the classification result. RF is a collection of decision trees, where each decision tree is unrelated, so one of the main benefits of RF over DT is it can prevent over-fitting.

When splitting attributes, we choose splits to maximize node purity, that is, to minimize Gini index. For a candidate split attribute L, denote possible levels as $l_1, \ldots, l_J$. The Gini Index for L, denoted by $G(L)$, is calculated as

$$G(L) = \sum_{j=1}^{J} Pr(L = l_j)(1 - Pr(L = l_j)), \quad (4)$$

where $Pr(L = l_j)$ denotes the probability that the candidate split attribute is at the jth level.

RF can also be applied to rank the importance of variables. The Gini index on the out-of-bag samples (observations left out of the bootstrap sample) for each tree and the Gini index after randomly permuting the data for column $m$in the out-of-bag samples are recorded. The decrease in Gini as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable m. The indicator is called Mean Decrease in Gini (MDG). The higher the MDG, the more important the variable in the model.

3.1.4 XGBoost

XGBoost is an improvement of the Gradient Boosting Algorithm. Based on Gradient Boosting Algorithm, it uses the Newton's method to expand the loss function, and also adds a regularization term to it. The objective function consists of two parts: the first part is the Gradient Boosting Algorithm Loss and the second part is the regularization term. Here we use function $l(y_i, \hat{y}_i)$ to

denote the difference between the observed truth value,y_i, and the predicted value of the ith observation, $\hat{y}_i$. f_t is the weak classifier formed by the t iteration and $\Omega(f_t)$ is a regular function that reduces over-fitting of the model. The objective function is given by

$$min\{\sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t) + C\}, \tag{5}$$

where C is a constant.

3.2 Model Measurements

The Receiver Operating Characteristics (ROC) curve is one of the most commonly used metrics to measure performance of models. ROC curve allows us to find the classification threshold that gives the best trade-off between False Positive Rate (FPR) and True Positive Rate (TPR). AUC (short for the area under the ROC curve) is the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one. AUC ranges from 0 to 1, since FPR and TPR ranges from 0 to 1. The random classifier has an AUC of 0.5. The closer the AUC to 1, the better the model. The perfect classifier has an AUC of 1.

We use *Recall*, *Precision*, *Accuracy*, and AUC to evaluate the trained models. We first give some notations and simple explanations. As shown in Table 1, TP denotes the number of true positives, and FN denotes the number of false negatives. FP is the number of false positives, and TN is the number of true negatives.

Table 1 Confusion Matrix

	Positive	**Negative**
True	TP	TN
False	FP	FN

Recall is the fraction of loans that are actually positive (non-default) that were predicted as positive by the model. The higher the *Recall*, the more positive loans are correctly predicted by the model. It is defined as

$$Recall = TP/(TP + FN). \tag{6}$$

Precision is the fraction of loans that were predicted as positive by the model that are actually positive. The higher the *Precision*, the more negative loans are correctly predicted by the model. It is defined as

$$Precision = TP/(TP + FP). \tag{7}$$

Accuracy is the ratio of loans that were correctly predicted. The higher the *Accuracy* of the model, the more loans are correctly predicted by the model. It is defined as

$$Accuracy = (TP + TN)/(TP + TN + FP + FN). \tag{8}$$

4. Experiment

This section is the empirical part including data description, data processing, and data analysis.

4.1 Data Description and Processing

The data for this study is about customers information of *The Grant Group of Companies*, and the data is obtained from Kaggle. This dataset has 87,501 observations and 30 variables.

Firstly, we remove missing data and obtain 68,505 observations. Then, we normalize all numerical variables. We use 28 of the 30 variables in our study and classified these variables into 3 categories: categorical, continuous, and discrete. The two variables are not selected because they are too diverse and cannot be generalized. Table 2 lists the selected variables and explanations.

Table 2 Variables and Explanations

Variable Name	Type	Meaning
Asst_Reg	Continuous	value of all the assets registered under the borrower's name
GGGrade	Categorical	Grant Group Grade
Validation	Categorical	validation status of the borrower
Yearly_Income	Continuous	total yearly income of the borrower
Home_Status	Categorical	living status of the borrower
Unpaid_2_years	Discrete	number of times the borrower has defaulted in last two years
Debt_to_Income	Continuous	debt to income ratio
Lend_Amount	Continuous	total funded amount to borrower
Usage_Rate	Continuous	ratio of processing charges on the loan amount
Inquiries	Discrete	inquiries in last 6 months
Present_Balance	Continuous	current balance in the borrower's account
File_Status	Categorical	status of the loan file
State	Categorical	state which the borrower belongs to
Duration	Categorical	duration for the amount funded to the borrower
Unpaid_Amount	Continuous	unpaid balance on the credit card
Reason	Categorical	reason for loan application
Due_Fee	Continuous	charges incurred if the payment on loan amount is delayed
Default	Categorical	Target Variable
Experience	Categorical	total year of work experience of the borrower
Already_Defaulted	Discrete	number of other loans the borrower was default
Designation	Categorical	designation of the borrower
Deprecatory_Records	Discrete	an entry that may be considered negative by lenders because it indicates risk and hurts the borrower's ability to qualify for credit or other services
Interest_Charged	Continuous	interest charged on the total amount
Gross_Collection	Continuous	the gross amount payable by way of settlement or judgment in respect of the claims, excluding any costs
Sub_GGGrade	Discrete	sub Grant Group Grade
Account_Open	Discrete	total number of open accounts in the name of the borrower
Total_Unpaid_CL	Continuous	unpaid dues on all the other loans
Claim_Type	Categorical	among all application type what is the borrower's claim type I - Individual Account J - Joint Account

We divide the data into training set and test set. Here, we use 60% of the data (e.g., 41,103 samples) as the training set and 40% (27,402 samples) is used as the test set. The test set is used to verify the accuracy of the models.

4.2 Data Analysis

4.2.1 Logistic Regression

Table 3 shows results obtained using LR. We consider variables are significant if corresponding *p*-value is less than 0.05. Then we have 14 variables left which means that they have significant effects on the loan default.

In particular, *Inquiries*, *Reason_renewable_energy*, *Reason_house*, *Yearly_Income* and *Present_Balance* show positive impacts. This is because home or car purchases tend to cause borrowers to fail to make scheduled payments, and credit inquiries imply that borrowers borrow frequently. Meanwhile, Grant Group may have lenient requirements for high-income and high-deposit borrowers.

Asst_Reg, *Due_Fee*, *Home_Status_Rent*, *Duration_5years*, *Unpaid_2_years* and *Debt_to_Income* show negative impacts. For *Asst_Reg* and *Due_Fee*, it can be explained that more property implies a lower probability of default and a higher default penalty prevents people from defaulting by making them more cautious. *Home_Status_Rent* is a dummy variable and the results show that borrowers who rent are less likely to default compared to those who mortgage their houses. Additionally, it is possible that *Duration_5years* gives borrowers enough time to pay for the loan. And for *Unpaid_2_years* and *Debt_to_Income*, it might because some borrowers perform well in other aspects, even they have some bad loan behaviors.

Table 3 Maximum Likelihood Estimates of Logistic Model for Loan Default Prediction

	Dependent variable:
	Default
Asst_Reg	-1.362***
	(0.020)
Yearly_Income	0.055***
	(0.019)
Unpaid_2_years	-0.059***
	(0.017)
Debt_to_Income	-0.111***
	(0.018)
Lend_Amount	-0.055***
	(0.020)
Usage_Rate	-0.216***
	(0.021)
Inquiries	0.049***
	(0.015)
Present_Balance	0.092***
	(0.021)
Due_Fee	-0.149***
	(0.022)
GGGrade_II	0.152**

	(0.066)
Home_Status_Rent	-0.106***
	(0.040)
File_Status_whole	-0.063**
	(0.031)
Duration_5years	-0.617***
	(0.042)
Reason_house	0.533**
	(0.270)
Reason_renewable_energy	1.194**
	(0.532)
Observations	41,103
Log Likelihood	-15,026.580
Akaike Inf. Crit.	30,299.170
Note:	*p<0.1; **p<0.05; ***p<0.01

4.2.2 Decision Tree

We build a DT and set the complexity $cp = 0.002$. Results show that the division is based on 6 variables, namely *Asst_Reg*, *Gross_Collection*, *Duration*, *Yearly_Income*, *State* and *Inquiries*. In the first and second step, the samples are divided into 2 groups based on *Asst_Reg*, in specific, whether their *Asst_Reg* ≥ -0.89 and whether their *Asst_Reg* ≥ -1.5, respectively. In the third step, the samples are divided based on whether their *Gross_Collection* ≥ -0.085 (the red box) and whether their *Gross_Collection* ≥ -0.084 (the blue box). And in the following steps, samples are divided based on the values of indicators *Duration*, *Yearly_Income*, *State* and *Inquiries*. Results are shown in Fig. 1.

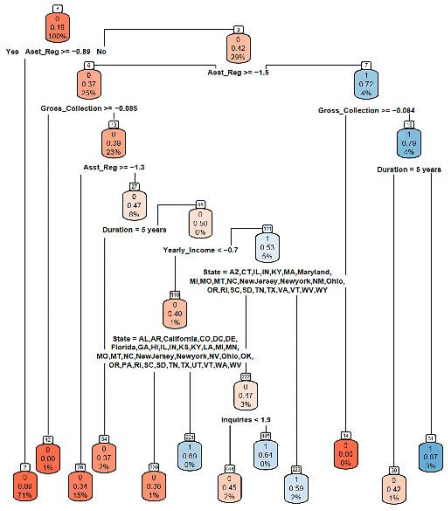

Figure 1. Decision Tree for Loan Default Prediction

4.2.3 Random Forest

We set the number of trees grown for prediction $ntree = 400$ and the number of variables randomly sampled as candidates at each split $mtry = 5$.

Table 4 shows MDG in RF. We can see that *Asst_Reg* is the most important variable with MDG of 197.531, followed by *Gross_Collection*, with MDG of 87.53. Besides, *Unpaid_Amount*, *Total_Unpaid_CL*, *Yearly_Income* and *Present_Balance* are also relatively important. The MDG of Claim_Type is 0, because more than 99.95% of the samples have Claim_Type of I.

Table 4 MDGs of Variables in RF

	Mean Decrease Gini
Asst_Reg	197.531
Yearly_Income	42.143
Unpaid_2_years	5.770
Already_Defaulted	-1.731
Debt_to_Income	21.411
Lend_Amount	22.485
Deprecatory_Records	0.804
Interest_Charged	-1.676
Usage_Rate	32.104
Inquiries	9.886
Present_Balance	39.019
Gross_Collection	87.530
Sub_GGGrade	29.319
Account_Open	21.235
Total_Unpaid_CL	43.876
Unpaid_Amount	44.106
Due_Fee	14.601
GGGrade	22.130
Experience	-0.093
Validation	7.375
Home_Status	13.633
File_Status	20.009
State	1.601
Duration	27.355
Reason	7.872
Claim_Type	0
Designation	6.374

4.2.4 XGBoost

We set the maximum depth of a tree $max\_depth = 7$, learning rate $eta = 0.6$ and the maximum number of iterations $nrounds = 106$. We obtain the logarithmic loss of 0.141955. We can see from Fig. 2 that logarithmic loss decreases with increasing iteration.

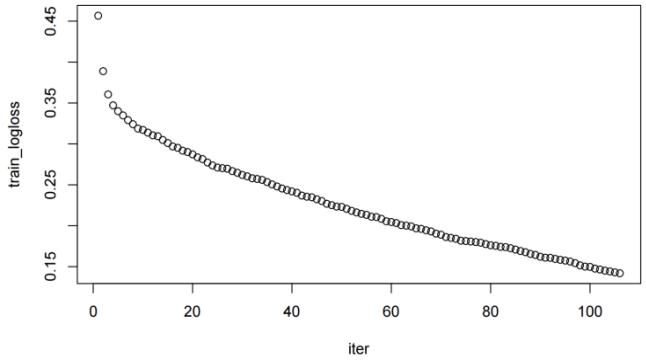

Figure 2. Scatter Plot of Logarithmic Loss and Iteration in Xgboost

4.3 Comparison

Table 5 compares the *Accuracy*, *Recall*, *Precision* and AUC of the four methods. It shows that DT has the highest *Precision* but the lowest AUC, and RF has the highest *Accuracy* but the lowest *Recall*. We can also see that XGBoost obtains the highest *Recall* and AUC but the lowest *Accuracy* and *Precision*.

We expect to catch a relatively large number of potential defaulters as we focus on the problem of loan default prediction. XGBoost achieves the highest *Recall* and AUC as well as a not much lower *Accuracy* than other methods. Taking all these factors into consideration, we consider XGBoost is the best method.

Table 5 Evaluation Metrics Comparison of The Four Methods

Method	***Accuracy***	***Recall***	***Precision***	**AUC**
Logistic Regression	0.8335158	0.3208669	0.6210097	0.828
Decision Tree	0.8351945	0.2171078	0.7228608	0.597
Random Forest	0.8371287	0.2520138	0.7000533	0.613
XGBoost	0.8217283	0.3546222	0.5488275	0.832

5. Conclusions

The objective of this paper is to use ML methods to predict whether a person will default on a loan based on his or her behavior and compare the performance of these models. We investigate LR, DT, RF, XGBoost models on the customer information from *The Grant Group of Companies* and use *Recall*, *Precision*, *Accuracy* and AUC to measure the performance of the four ML models. Results show that DT has the highest *Precision* but the lowest AUC, and RF has the

highest *Accuracy* but the lowest *Recall*. Additionally, XGBoost obtains the highest *Recall* and AUC but the lowest *Accuracy* and *Precision*.

In further study, we will focus on solving the problem of the uneven dataset. First, we will try to conduct experiment on larger datasets and tune the models to achieve the better performance of the model. Second, we will consider using balancing techniques, such as ROSE (Random Over Sampling Examples) and SMOTE (Synthetic Minority Oversampling Technique) methods, to generate synthetic data. These methods have achieved good results on unbalanced datasets.

References

[1] M. Moscatelli, F. Parlapiano, S. Narizzano, and G. Viggiano, "Corporate default forecasting with machine learning," Expert Systems with Applications, vol. 161, p. 113567, 2020.

[2] S. Bhatore, L. Mohan, and Y. R. Reddy, "Machine learning techniques for credit risk evaluation: A systematic literature review," Journal of Banking and Financial Technology, vol. 4, pp. 111–138, 2020.

[3] E. I. Altman and A. Saunders, "Credit risk measurement: Developments over the last 20 years," Journal of Banking and Finance, vol. 21, pp. 1721–1742, 1997.

[4] L. Lugovskaya, "Predicting default of Russian SMEs on the basis of financial and non-financial variables," Journal of Financial Services Marketing, vol. 14, pp. 301–313, 2010.

[5] R. Mileris, "Estimation of loan applicants default probability applying discriminant analysis and simple Bayesian classifier," Ekonomika ir vadyba, vol.15, pp. 1078-1084, 2010.

[6] S. Bhatore, L. Mohan, and Y. R. Reddy, "Machine learning techniques for credit risk evaluation: A systematic literature review," Journal of Banking and Financial Technology, vol. 4, pp. 111–138, 2020.

[7] K. Kumar and S. Bhattacharya, "Artificial neural network vs linear discriminant analysis in credit ratings forecast," Review of Accounting and Finance, vol. 5, pp. 216–227, 2006.

[8] F. Ciampi, "Corporate governance characteristics and default prediction modeling for small enterprises. an empirical analysis of Italian firms," Journal of Business Research, vol. 68, pp. 1012–1025, 2015.

[9] E. Agbemava, I. K. Nyarko, T. C. Adade, and A. K. Bediako, "Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana," European Scientific Journal, vol. 12, p. 175, 2016.

[10] Z. Khemais, D. Nesrine, and M. Mohamed, "Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression," International Journal of Economics and Finance, vol. 8, p. 39, 2016.

[11] M. A. Sheikh, A. K. Goel, and T. Kumar, "An approach for prediction of loan approval using machine learning algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490–494.

[12] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," Procedia Computer Science, vol. 162, pp. 503–513, 2019.

[13] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and Random Forest: A comparative study," IOP Conference Series: Materials Science and Engineering, vol. 1022, p. 012042, 2021.

[14] J. Xu, Z. Lu, and Y. Xie, "Loan default prediction of Chinese P2P market: A machine learning methodology," Scientific Reports, vol. 11, no. 1, 2021.

[15] S. Chang, S. D. Kim, G. Kondo, “Predicting default risk of lending club loans,” Machine Learning, pp. 1–5, 2015.
[16] Y. Wang, Y. Zhang, Y. Lu, and X. Yu, “A comparative assessment of credit risk model based on machine learning —— a case study of bank loan data,” Procedia Computer Science, vol. 174, pp. 141–149, 2020.
[17] A. Coşer, M. M. Maer-matei, C. Albu, “Predictive models for loan default risk assessment,” Economic Computation and Economic Cybernetics Studies and Research, vol. 53, pp. 149–165, 2019.
[18] J. Duan, “Financial system modeling using Deep Neural Networks (DNNs) for effective risk assessment and prediction,” Journal of the Franklin Institute, vol. 356, pp. 4716–4731, 2019.