Towards Personal Credit Default Prediction Method Based on Data Mining

Yiran Xue

xueyiran1109@163.com

Faculty of Engineering, The University of New South Wales, Sydney, Australia

Abstract—This project focuses on assessing the user's personal credit risk based on data mining techniques. This research is designed to help financial institutions predict whether borrowers will be able to repay all their loans within a given period, thereby reducing the financial losses caused by deviations between the risk assessment and the actual situation during the lending process. This study has collected a number of user data, and used five types of algorithms, such as DT (decision tree), NB (Naive Bayes) and LR (logistic regression), to build personal credit default prediction model respectively. Meanwhile, ACC (accuracy), AUC (area under the ROC curve) and KS values were selected as the model evaluation metrics. The experimental result shows that the DT model is the most suitable for personal credit default prediction.

Keywords—credit default, financial loss, prediction model, evaluation metrics, data mining

1 INTRODUCTION

With the rapid expansion of the financial market, financial services have become more prosperous, and the concept of inclusive finance has become deeply rooted among the people, but it has also revealed many problems, among which the credit default problem is particularly prominent [1]. Firstly, due information lag, financial institutions are not able to obtain comprehensive and timely information about the borrower's financial situation, repayment ability and repayment willingness. This lack of information may cause financial institutions to lose money during the process of lending. Secondly, to improve market competitiveness and earn a higher profit, many lending institutions have relaxed their credit assessment criteria and ignored potential risks, which may result in significant financial losses to the lending institutions. Therefore, it is an effective way for financial institutions to reduce their own losses by assessing the risk levels of user and making credit decisions through multi-faceted data [2].

Data mining techniques are the optimal approach to assessing a user's credit risk [3]. The main feature of data mining technology is the parsing, extraction, transformation, analysis, and modelling of large amounts of data, thereby extracting potentially valuable and useful information from heterogeneous multi-source data. The application of data mining technology enables credit institutions to know the assets and liabilities of the borrower effectively and promptly and to evaluate the income level and repayment ability of the customer, so that the credit institution can better determine its lending amount and minimize its own risk.

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334599

At present, credit risk has become an unavoidable risk for credit institutions in their operations, and various risk prediction models for credit risk under different lending situations have been developed by scholars. However, due to the complexity of financial business, the diversity of financial rules and the significant differences in data selection, there is still no unified performance evaluation standard for personal credit default prediction models. In this research, we established three models using DT (decision tree) algorithm, NB (Naive Bayes) algorithm and LR (logistic regression) algorithm respectively, then compared and analyzed the ACC (accuracy), AUC (area under the ROC curve) and KS values of these models. According to the experimental results, the DT model has an ACC of 0.87, an AUC of 0.83 and a KS value of 0.51, which are higher than the corresponding values of the NB model and the LR model. Therefore, in terms of predicting credit default problems, the DT model is significantly better than the NB and LR models.

2 RELATED WORK

To ensure the continued and stable development of the credit industry in the financial markets, a large amount of work has been carried out in the field of credit default prediction by domestic and international researchers who want to find more efficient, more effective and less costly methods in terms of credit risk assessment. Based on the Loan Club's public loan application data, research uses data pre-processing techniques and machine learning, involving DT (decision tree), RF (random forest) and Bagging (Bootstrap Aggregating), to finally determine which features are important in predicting loan defaults and which borrowers can repay their loans on time with have good credit [4]. The result shows that the RF model is better at identifying defaults, and the DT model is more powerful at looking for the good credits [4]. In recent years, there have also been some researchers who used CNN (Convolution Neural Network) algorithm for credit default prediction. The paper by Zhou, X., Zhang, W. and Jiang, Y. developed a CNN model for credit risk assessment and also compared it with three traditional models, involving SVM (support vector machine), Bayes and RF (random forest), based on extracted feature samples and full feature samples. Consequently, the ACC and AUC of the three traditional models with full feature samples are higher than those after extracting feature samples. The CNN model has the highest ACC and AUC of 95% and 99% respectively, which is better than the three traditional models. Therefore, they concluded that the CNN model is more suitable for predicting personal loan defaults [5]. The research by Jing Gao, Wenjun Sun and Xin Sui. collected user information from a small commercial bank and compared the performance of the XGBoost model and the XGBoost-LSTM model in terms of predicting personal loan risk. The outcome of this experiment indicates that the XGBoost-LSTM model reduces the number of sample misclassifications and achieves better test accuracy compared to the XGBoost model [6].

3 DESIGN & IMPLEMENTATION

3.1 Data preparation

Data pre-processing. Firstly, read the original data set from the csv-file. Secondly, to minimize the negative impact of dirty data on the model, we have removed the samples that were severely missing from the original dataset. Thirdly, the problem of sample imbalance may cause the

prediction results of the model to be biased towards the category with a high number of samples, resulting in a high ACC, but the model itself is meaningless [7]. Therefore, it is necessary to determine whether there is a sample imbalance problem in the data set. We conducted data statistics for the two labels 0 and 1, representing no late payments and late payments respectively. The statistical result indicates that there is a large difference between the two datasets, with a ratio of positive samples to negative samples of almost 13:1. Regarding the study of sample imbalance, it is generally considered that 'imbalance' means that the number of minority categories is less than 10% of the total sample. To resolve the sample imbalance problem, we randomly sampled 47003 items from the positive sample, and merged the positive sample set after random sampling with the negative sample set, so that the percentage of negative samples reached approximately 15%. Finally, the 55360 items were retained as the initial data set.

Dataset splitting. We split the initial dataset in a 7:3 ratio, with one part (X_train and Y_train) being used to learn and train the model, and the other part (X_test and Y_test) being used to examine the accuracy of the model [8]. It is important that the splitting is done in such a way that the proportion of 0 and the proportion of 1 in the training set and testing set is the same as the proportion in the initial data set.

Model selection. Three models have been built in this experiment, which are DT model, NB model and LR model. We trained three models with different parameters and selected the best parameters to get the optimal state for each model. Finally, by comparing the overall rating of the three models, we chose the model that was best suited to assess the credit risk of the users.

Data standardization. Since different features have different properties and different orders of magnitude, training the model without data standardization will weaken the impact of lower order of magnitude features on the model. Therefore, we need to standardize the training datasets and testing datasets to a common scale prior to training the model [9]. In this experiment, we chose z-score standardization. Z-score standardization can scale data of different orders of magnitude down to the same interval, thus reducing the impact of differences in terms of size, characteristics, and distribution on the model. After z-score standardization, both the training datasets and the testing datasets satisfy a standard normal distribution. The transformation function is:

$$x^*=(x-\mu)/\sigma. \tag{1}$$

Evaluation metrics selection. In the classification training process, evaluation metrics play a crucial role, and appropriate evaluation metrics can help us to discriminate and obtain the optimal classifier [10]. As this experiment involves a binary classification problem, we selected ACC, AUC, and KS values as the evaluation metrics.

Accuracy is defined as the number of samples that are correctly predicted by the model as a percentage of the total number of samples. The calculation formula is：

$$ACC=\frac{TP+TN}{TP+TN+FP+FN}, \tag{2}$$

where TP represent true positive cases, FP represent false positive cases, TN is true negative cases as well as FN reflects false negative cases.

The ROC curve is considered to be an important metric for the evaluation of the merits of a binary classifier. By examining the ROC curve, we can compare the difference in terms of precision of

the classification between two or more classifiers [11]. The horizontal axis of the ROC curve represents the proportion of misclassified positive samples out of all negative samples, and the vertical axis is the proportion of correctly categorized positive samples out of total positive samples. The AUC value is higher meaning that the model performs better. Assuming that there are M positive samples and N negative samples, the AUC is calculated according to the following formula:

$$\mathrm{AUC}=\frac{\sum_{i\in \mathrm{PositiveClass}} \mathrm{rank}_i-(M*(1-M)/2)}{M*N}. \tag{3}$$

The KS value is an efficiency evaluation metric used to check the degree of separation between positive and negative samples predicted by the model and is commonly used in financial markets [12]. The range of the KS value is between 0 and 1, with the higher values indicating better discrimination between positive and negative samples. The formula is define as:

$$\mathrm{KS}=\max(\mathrm{TPR}-\mathrm{FPR}), \tag{4}$$

where TPR is the true positive rate and FPR is the false positive rate.

3.2 Decision Tree

DT is one of the machine learning algorithms based on tree models, including ID3, C4.5 and CART. Since this experiment involves binary classification algorithm, we adopted the CART model. CART are suitable for both classification and regression, in which the classification tree uses the Gini index minimization criterion for performing feature selection, and then recursively generates a binary tree.

In a classification problem, suppose that there is a data set $D= \{(x_1, y_1), (x_2, y_2), ... (x_m, y_m)\}$, and its output class is $C_1, C_2, ..., C_k$. We also assume that the probability that a sample point is from the k^{th} category is p, then the Gini index of the probability distribution is given by:

$$\mathrm{Gini}(D)=\sum_{k=1}^{K} p_k(1-p_k) =1-\sum_{k=1}^{K} p_k^2. \tag{5}$$

In fact, the Gini(p) represents the probability that two randomly sampled sample points belong to two distinct categories. According to the mathematical meaning of the Gini value, we know that a smaller Gini value represents a higher purity of the data set D. For the dichotomous classification problem, the above formula can be simplified as:

$$\mathrm{Gini}(D)=2p(1-p). \tag{6}$$

For a data set D with the number of samples |D|, we assume that the sample size for the k^{th} category is $|C_k|$, then the Gini index of D is:

$$\mathrm{Gini}(D)=1-\sum_{k=1}^{K} \left(\left|\frac{C_k}{D}\right|\right)^2. \tag{7}$$

During the search for the optimal partitioning attribute, D is partitioned into $|D_1|$ and $|D_2|$ based on an attribute value a of feature A. The calculation of the Gini index for the data set D and the attribute a is as follows:

$$\mathrm{Gini\_index}(D, a)=\sum_{v=1}^{V} \frac{|D_v|}{D} \mathrm{Gini}(D). \tag{8}$$

In data set A, different attributes may have different or the same Gini values, and the attribute with the smallest Gini value is selected as the best divided attribute:

$$a_* = \arg\min \text{Gini\_index}(D, a). \quad (9)$$

The CART algorithm takes the root node as the starting point and then processes each node recursively. Firstly, the Gini coefficient Gini(D, a) is calculated for each attribute under each feature of the present node. Secondly, the Gini(D, a) values for different attributes under the same feature are compared. The attribute with the lowest Gini(D, a) and the corresponding feature are chose as the optimum feature and the optimum attribute. Thirdly, two child nodes are created from the present node. The dataset D is divided into |D1| and |D2| based on the optimum features and the optimum attributes, as well as |D1| and |D2| are assigned to the two new child nodes. Next, the above 3 steps are repeated continuously until the stopping condition is satisfied. Finally, a decision tree is generated.

The CART model is generated based on the 'DecisionTreeClassifier' function in the 'sklearn' library. We obtained the best CART model by continuously adjusting the parameters of the function. Fig. 1 shows the ROC curve of the CART model, with AUC of 83%.

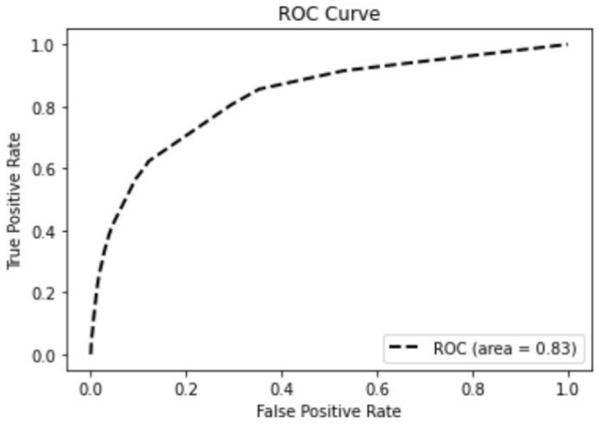

Fig. 1. ROC curve of CART

The evaluation results of the individual credit default model using the CART algorithm are displayed in Table I.:

TABLE I. EVALUATION RESULTS OF CART ALGORITHM

Algorithm	**Confusion matrix**	**ACC**	**AUC**	**KS value**
CART	[[13545, 556], [1527, 980]]	0.87	0.83	0.51

3.3 Naive Bayes

The Naive Bayes classification is an essential and widely used classification method in machine learning, and its classification principle is based on Bayes' theorem [11]. We assume that there are m data samples, and each sample has n features, which can be expressed as $\{(x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)}, y_1), (x_1^{(2)}, x_2^{(2)}, ... , x_n^{(2)}, y_2), ..., (x_1^{(m)}, x_2^{(m)}, ... , x_n^{(m)}, y_m)\}$. The feature category is $C_1, C_2, ..., C_k$.

First of all, the Naive Bayes prior probability distribution is:

$$P(Y=C_k)\ (k=1,2,3,..., K). \tag{10}$$

Secondly, we need to calculate the conditional probability distribution:

$$P(X=x|Y=C_k)=P(X_1=x_1,X_2=x_2,...,X_n=x_n|Y=C_k). \tag{11}$$

The Naive Bayes model assumes that each attribute is independent of each other：

$$P(X_1=x_1,X_2=x_2,...,X_n=x_n|Y=C_k)=\prod_{j=1}^{n} P(X_j=x_j|Y=C_k). \tag{12}$$

Then, the joint distribution of X and Y can be obtained according to the Bayesian formula:

$$P(X,Y=C_k)=P(Y=C_k)P(X_1=x_1,X_2=x_2,...,X_n=x_n|Y=C_k). \tag{13}$$

Finally, the function of the Naive Bayes classifier is:

$$C_{result}=\arg\max P(Y=C_k)\prod_{j=1}^{n} P(X_j=X_j^{(test)}|Y=C_K). \tag{14}$$

The output result C_{result} is the type of instance X(test).

The Naive Bayes model is built according to the 'GaussianNB' function in the 'sklearn' library, and the best Bayes model is obtained by continuously adjusting the parameters in the function. Fig. 2 is the plot of the ROC curve for the Naive Bayes, and the AUC is 0.71.

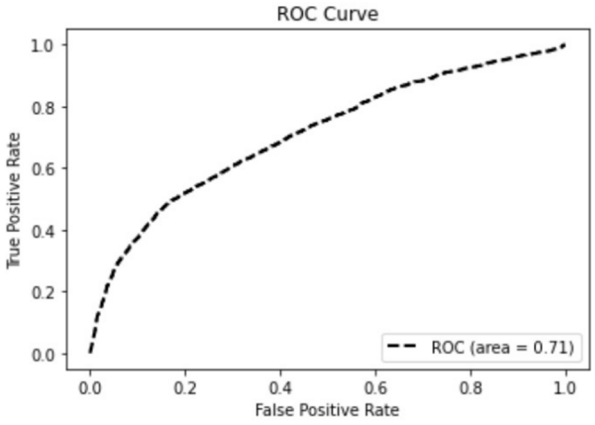

Fig. 2. ROC curve of Bayes

The results of the evaluating the Naive Bayes model are presented in Table II.:

TABLE II. EVALUATION RESULTS OF NAIVE BAYES

Algorithm	Confusion matrix	ACC	AUC	KS value
Naive Bayes	[[14041,60], [2427,80]]	0.85	0.71	0.32

3.4 Logistic Regression

Logistic regression is a type of classification algorithm in machine learning that has the advantage of simplicity and efficiency. Logistic regression algorithms are often used to solve binary classification problems, such as identifying spam, sick or not, financial fraud, etc. Logistic regression algorithms generally recall a sigmoid function that maps the output of linear regression to the interval [0, 1]. The sigmoid function is:

$$g(z)=1/(1+e^{-z}). \tag{15}$$

The basic logistic regression model can be derived from the sigmoid function:

$$g(z)=f(\theta^{T}x)=1/(1+e^{-\theta^{T}x}). \tag{16}$$

It is assumed that there are m samples, and each sample has n features. The loss function is derived by maximum likelihood estimation:

$$\text{cost}(f(\theta^{T}x),y)=\begin{cases} -\log\left(f(\theta^{T}x)\right) & \text{if } y=1 \\ -\log\left(1-f(\theta^{T}x)\right) & \text{if } y=0 \end{cases}, \tag{17}$$

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}\text{cost}(f(\theta^{T}x),y). \tag{18}$$

We solve the loss function J(θ) using the method of gradient descent, that is, finding the most optimal value of the parameter θ and minimizing the loss function. As the function is the convex function, it has a unique answer.

The logistic regression algorithm invokes the 'LogisticRegression' function in the 'sklearn' library. It is necessary to try various parameters in the 'LogisticRegression' function. Fig. 3 shows a graph of the logistic regression model's ROC curve, with AUC of 70%.

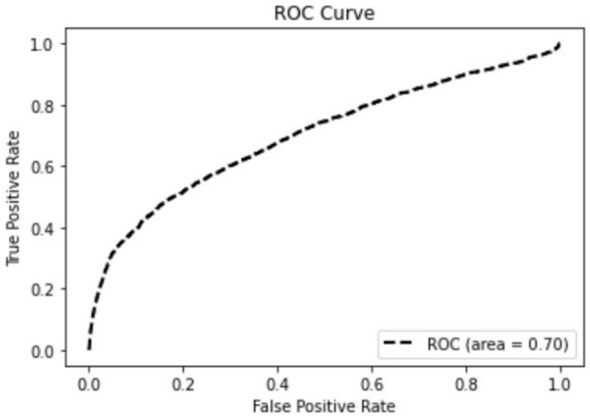

Fig. 3. ROC curve of Logistic Regression

The results of the individual credit default model scoring under the logistic regression algorithm are presented in Table III.:

TABLE III. EVALUATION RESULTS OF LOGISTIC REGRESSION

Algorithm	Confusion matrix	ACC	AUC	KS value
Logistic Regression	[[14025,76], [2304,203]]	0.86	0.70	0.32

4 EXPERIENTIAL RESULTS & DECISION

Credit loan risk assessment is an important means to effectively reduce financial losses for financial institutions. This report examines three risk assessment models in terms of formula derivation, model construction and result analysis. The KS values of all three models are greater than 0.3, with the CART model having the highest KS value of 0.51. In fact, KS values above 0.3 are considered to have a good ability to separate positive and negative samples. In addition, the CART model had an ACC of 87% and an AUC of 83%, which are both higher than the Naive Bayes and Logistic Regression models. Therefore, we believe that the CART model has a better performance in assessing the credit risk of users.

5 CONCLUSION

This study concentrates on the problem of user credit default in the financial market, proposes the design idea of building a personal credit default prediction model to classify the user's risk level, and uses data mining technology to develop and implement the model. The best results

were achieved with the decision tree model. This experiment provides a reference for banks to prevent default risk.

REFERENCE

[1] Wu, Y. and Pan, Y. (2021) "Application analysis of credit scoring of financial institutions based on machine learning model," Complexity, 2021, pp. 1–12.

[2] Madaan, M., Kumar, A., Kesri, C., Jain, R. and Nagrath, P. (2021) "Loan default prediction using decision trees and Random Forest: A comparative study," IOP Conference Series: Materials Science and Engineering, 1022(1), p. 012042.

[3] Jayasree, V. and Balan, R.V.S. (2013) A review on data mining in banking sector, American Journal of Applied Sciences. Science Publications.

[4] Vinod Kumar L et al. (2016) "Credit risk analysis in peer-to-peer lending system," 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA).

[5] Zhou, X., Zhang, W. and Jiang, Y. (2020) "Personal credit default prediction model based on Convolution Neural Network," Mathematical Problems in Engineering, 2020, pp. 1–10.

[6] Gao, J., Sun, W. and Sui, X. (2021) "Research on default prediction for credit card users based on XGBoost-LSTM model," Discrete Dynamics in Nature and Society, 2021, pp. 1–13.

[7] Hoens, T.R. and Chawla, N.V. (2013) "Imbalanced datasets: From sampling to classifiers," Imbalanced Learning, pp. 43–59.

[8] Wang, C. and Xiao, Z. (2022) "A deep learning approach for credit scoring using feature embedded Transformer," Applied Sciences, 12(21), p. 10995.

[9] Chou, T. and Lo, M. (2018) "Predicting credit card defaults with deep learning and other machine learning models," International Journal of Computer Theory and Engineering, 10(4), pp. 105–110.

[10] M, H. and M.N, S. (2015) "A review on evaluation metrics for Data Classification Evaluations," International Journal of Data Mining & Knowledge Management Process, 5(2), pp. 01–11.

[11] Krichene, A. (2017) "Using a naive bayesian classifier methodology for loan risk assessment," Journal of Economics, Finance and Administrative Science, 22(42), pp. 3–24.

[12] Gonçalves, E.B. and Gouvêa, M.A. (2021) "Credit risk analysis applying logistic regression, neural networks and genetic algorithms models," International Journal of Advanced Engineering Research and Science, 8(9), pp. 198–209.