Hindawi
Security and Communication Networks
Volume 2024, Article ID 9818032, 1 page
https://doi.org/10.1155/2024/9818032

Retraction

Retracted: Enterprise Financial Risk Analysis Based on Improved Model *C*-Means Clustering Algorithm

Security and Communication Networks

Received 8 January 2024; Accepted 8 January 2024; Published 9 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Sun and Y. Jiao, "Enterprise Financial Risk Analysis Based on Improved Model *C*-Means Clustering Algorithm," *Security and Communication Networks*, vol. 2022, Article ID 1109813, 12 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 1109813, 12 pages
https://doi.org/10.1155/2022/1109813

Research Article

Enterprise Financial Risk Analysis Based on Improved Model *C*-Means Clustering Algorithm

Jia Sun[1] **and Yanrong Jiao**[2]

[1]*Department School of Finance & Economics, Chongqing City Management College, Chongqing 401331, China*
[2]*School of Construction Management, Chongqing College of Architecture and Technology, Chongqing 401331, China*

Correspondence should be addressed to Jia Sun; cswusunj@163.com

Received 16 March 2022; Revised 27 April 2022; Accepted 3 May 2022; Published 12 July 2022

Academic Editor: Zhiping Cai

As a provider of loans to SMEs, banks should prudently examine loan risks while ensuring that they provide loans to SMEs from the perspective of cooperating with policy implementation and controlling their own risks. The existing loan risk measurement tools include multiple discriminant analysis models, multiple regression models, and machine learning methods. Most machine learning methods have higher prediction accuracy than traditional models when using historical data for calculation, but the existence of problems such as overfitting seriously affects the robustness of machine learning methods. A similar method is introduced into the loan default risk prediction of SMEs, and the mean clustering method is used to preset penalty items to reduce overfitting and high accuracy to help banks effectively identify the default probability of SMEs during the loan period. This study will use the mean clustering method to iteratively train 900,000 SME credit records published by the US Small and Medium Business Administration, with 27 dimensions of data provided by Small Business Administration (SBA) to provide partial guarantees. A regression tree evaluates the data, combining the scores of multiple regression trees to produce a final prediction of the probability of credit default on the input data. The research results show that the mean clustering method can effectively improve the prediction accuracy of traditional machine learning methods and multiple linear regression in the scenario of SME loan default prediction and reduce the overfitting and black-box properties. As a supplementary loan default risk measurement tool, it can strengthen the ability of commercial banks to control the risk of loan business and can also promote the development of small- and medium-sized enterprises and the market economy to a certain extent.

1. Introduction

The measurement and evaluation of corporate credit default probability are the key content of the Basel Accord internal rating method, and it is also one of the main independent variables of credit risk evaluation. The calculation of expected loss and value at risk according to the background and conditions of different credit applicants is the core work of commercial banks to determine capital and carry out loan business. Therefore, the accuracy of the credit default probability measurement has a significant impact on the bank's credit.

Banks can use some innovative risk prediction tools according to business characteristics, such as after recording the accuracy and error of reused historical data, or before credit business, using these tools to predict new loan needs, and then pilot departments to deploy them to other businesses after confirming the effectiveness of this method [1]. Testing and gray of these offline and online data may last for a long time. For small- and medium-sized enterprises, they should also actively cooperate with the data audit and disclosure work of banks and use the cloud platform to update and upgrade their own data statistics. In this way, with the synchronous efforts of loan demand and supply

side, we can expect the extensive and active application of innovative risk prediction method in small- and medium-sized enterprise credit [2].

At present, machine learning-based methods have a good performance in classification and prediction and have a relatively mature application framework in search, advertising, recommendation, risk control, and other scenarios [3]. Compared with traditional linear models, machine learning algorithms such as neural network and support vector machine generally have high prediction accuracy, but these traditional machine learning methods lack economic intuition and interpretability and have serious overfitting problems, which have higher risks than linear regression methods in the financial field.

New machine learning methods such as mean clustering and LightGBM released in 2016 have received extensive attention from computer academia. After improvement, these methods have become the best algorithms in current machine learning research under structured data. As for the mean clustering algorithm, the average clustering method in the loss function increases the number of tree structure complexity and L2 regular penalty term of leaf node weight, which makes the model training result more smooth. Compared with the ordinary machine learning method, the degree of overfitting is reduced, and the leaf node weight and sample division condition of each tree are highly readable, which is conducive to further test the rationality after model training. These features make the mean clustering method have a more stable performance and intuitive model results than the traditional methods, and it is more suitable for applications in such strict scenarios as credit risk control [4].

This study uses the mean clustering method to iteratively train the credit records of some guaranteed small- and medium-sized enterprises of the U.S. SME administration and the Small Business Administration (SBA) from 27 dimensions, and generate multiple regression trees that can be used to evaluate loan data, so as to generate the input data for the final prediction of credit default probability through the comprehensive multiple regression tree [5]. Using and improving the average clustering method, the performance of credit default prediction model is better than the traditional linear regression, and other nonlinear machine learning methods can provide new tools for bank risk control and improve the current work effect of credit risk assessment of commercial banks. In the rapid development of artificial intelligence and the active pursuit of the landing of the current time point, it also provides a new direction for the research and application of machine learning algorithms.

2. State of the Art

For the problem of classification prediction of credit default probability, many different methods have been developed in academia and industry, such as linear regression and generalized linear model and machine learning methods. There are also some special empirical models for the financial fields, such as Z-score methods [6]. The linear assumption of ordinary linear regression models lacks practicality in many practical scenarios, and the model accuracy is generally not high. Over the past 20 years, generalized linear models (GLMs) have been commonly used in the fields of actuarial and credit analysis. According to Nelder et al., in summary the generalized linear model is a traditional method to analyze the probability of credit default, but Chen (2020) believes that there are several limitations: 1. when the variables in the model are too large, there may be multiple collinearity resulting in model estimation distortion [7]; 2. They are incapable to use shrinkage estimation to effectively contract too many variables in the model; and 3. the assumptions and limitations of the model are still too strong compared with machine learning models, resulting in the results of generalized linear models being inferior to nonlinear models. In machine learning methods, when conducting supervised learning of historical data, random forest, support vector machine, neural network, and other algorithms have been relatively mature, and the prediction accuracy can reach 80% and 90% after jumping the whole parameters. However, machine learning methods are not completely better than GLM [8]. When the noise data are relatively high or the training data are not representative, there may be serious overfitting problems. The model performance difference between the training set and the test set is too large, which cannot be applied to the actual scene at all. For models of decision tree types, overcomplex tree structure often means overfitting, so the degree of overfitting can be suppressed to some extent, but unfortunately effective pruning strategies are also difficult to find. In addition, due to its black-box characteristics, the machine learning method is difficult to explain its parameter significance. Even if they have good accuracy in both the training set and the test set, there are certain risks when actually using it [9].

However, the significant advantages of machine learning methods in accuracy have still attracted the attention of many foreign scholars, and some scholars have already applied machine learning methods to the financial field. Xie et al. through comparison found that BP neural network (BPNN), support vector machine (SVM), and AdaBoost methods have higher prediction accuracy and relatively small variance compared with traditional GLM. According to Zhou et al., at this stage, more advanced machine learning methods such as delta clustering have also been applied in the financial field. According to Yang et al., the composite Poisson model is applied to the risk control system of insurance companies, and the cutting-edge machine learning method performs better compared with the existing methods [10]. Son reviewed the existing large number of machine learning algorithm in credit default prediction, summarized the prediction of enterprise credit default probability scenario that shows good results of machine learning algorithm, such as regression tree model, bagged algorithm, random forest, and neural network, and provided the unsupervised learning methodology applied in credit default prediction. Chen also compared the performance of the modified version of LightGBM based on the gradient lifting tree model with other traditional methods in the probability prediction of auto insurance claims and found that the accuracy and AUC indicators of LightGBM were

significantly higher than SVM, neural network, and GLM, reaching more than 90%, and also improved than the ordinary gradient lifting tree model. These literature studies show that the high accuracy of machine learning algorithms helps solve the reverse selection problem in the credit default prediction scenario [11].

On the other hand, algorithms for cluster types are developing rapidly, and Breiman proposed the concept of bagging. Bagging is a model training method, namely random sampling to train the classifier, and then, it combines the classifier into a higher accuracy classifier, such as random forest, which is a typical algorithm based on the bagging training mode. Freund et al. proposed a clustering algorithm relatively different from bagging [12]. The clustering algorithm will adjust the weight of each sample on the classifier according to the classification results of the last round. The greater the classification error rate, the greater the weight. The observation that this weight was modified was used to train the next classifier. Finally, the different classifiers were merged into the more precise classifiers [13]. Friedman also proposed the gradient rise (gradient boosting) method based on the clustering method, which focuses on reducing the loss function value of the objective function, and the reduction process is determined by the derivative of the objective function. Chen et al. (2016) further strengthened the training method of gradient rise algorithm and proposed the mean clustering method. The algorithm often achieves the classification and prediction of nearly 90% accuracy in high-quality datasets and shows higher performance than LightGBM in a large number of experiments. We can expect its performance in the risk assessment scenario [14].

Many researchers are still conservative about the application of machine learning algorithms to production in the financial sector. Comparing the application of generalized linear model, neural network, decision tree, and other algorithms in enterprise default prediction, Goyal et al. found that the neural network algorithm has the highest accuracy, but there is an obvious overfitting problem. When applying machine learning algorithms to economic and financial scenarios, Wuthrich believes that the main problem is that the extracted features lack economic intuition and low interpretability, even if the accuracy is higher than linear models, and it is still not convincing [9]. Meanwhile, according to Wuthrich et al. single-value decomposition and bottleneck network are used to extract the features and replace the features in the PCA model, which improves the interpretability of the model [15]. However, the critical value problem of how to effectively determine the prediction probability after this treatment is still pending, but in fact the mean clustering method has effectively improved the two main defects of machine learning algorithms in the financial field mentioned by these scholars: Chen et al. point out that the clustering algorithm is almost better than the ordinary gradient lifting algorithm, and the regular term in the loss function can effectively reduce the variance of the model; on the other hand, the regular term also punished the complexity of the decision tree, effectively completed the pruning work, and make the mean clustering method avoid the overfitting problem in many experiments [16]. It also retains the advantages of the gradient lifting tree algorithm: the rules of the structure of the decision tree are easy to extract, and it can also help people understand its training results through many visual analysis tools.

Based on the literature, that the cutting-edge machine learning methods have high availability of credit default risk data, and the overall effect is better than generalized linear models, SVMs, neural networks, and neural network models, which have low explanatory power [17]. Classification trees can be directly understood by us. The interpretability of the model is increased, the requirements for the dataset are lowered, and the degree of overfitting to the training data is reduced, and at the same time, the error rate is lower and the accuracy rate is higher, which is feasible in the scenario of credit default risk prediction [18].

3. Model and Data Processing

3.1. Data Description and Processing. The reasons for credit defaults of SMEs mainly include five categories, including business processes and basic conditions. The overall classification and detailed factors are shown in Figure 1. We need to evaluate the basic data of mid-performance companies in terms of these aspects.

After accessing the original data, it is necessary to clean the data according to the characteristics of the clustering algorithm. Because the basic processing of variables in the mean clustering method is discrete, and continuous variables are supported by default, a thermal code is used as a noncontinuous variable of computer recognizable variables [19]. Finally, the numerical value can be normalized again.

The specific steps are as follows: first invalid records (because the mean clustering method is sensitive to missing values, 96% of the data used in this study are complete, so they are selected to eliminate the records with missing values) are removed, all data variables are converted into standard time format, and then the variables of amount type into digital format are converted. Then, for discrete variables, one-hot coding is converted into a binary vector; i.e., for discrete variables with n different values, a binary number of length n represents each value, i-value is i-bit 1, and the remaining bit is 0. In this way, the feature number will be very sparse, to avoid being too difficult for the algorithm to identify the feature [20]. Finally, the value is normalized and gradient descent algorithm is avoided in the multidimensional plane because the variable numerical range difference is too large and difficult to find a steady drop path. At the same time, the distance between some algorithm samples is also set as the calculation reference classifier, which can speed up the calculation speed while avoiding being misled by the too large sample range. This study uses the most common normalization method, with zero mean centralization with standard deviation and mean, as shown in Figure 2.

In this study, the data quality and numerical distribution after data preprocessing were tested by variable correlation. From Figures 3 and 4, in addition to the label, there are many variables whose distribution is uneven and does not conform

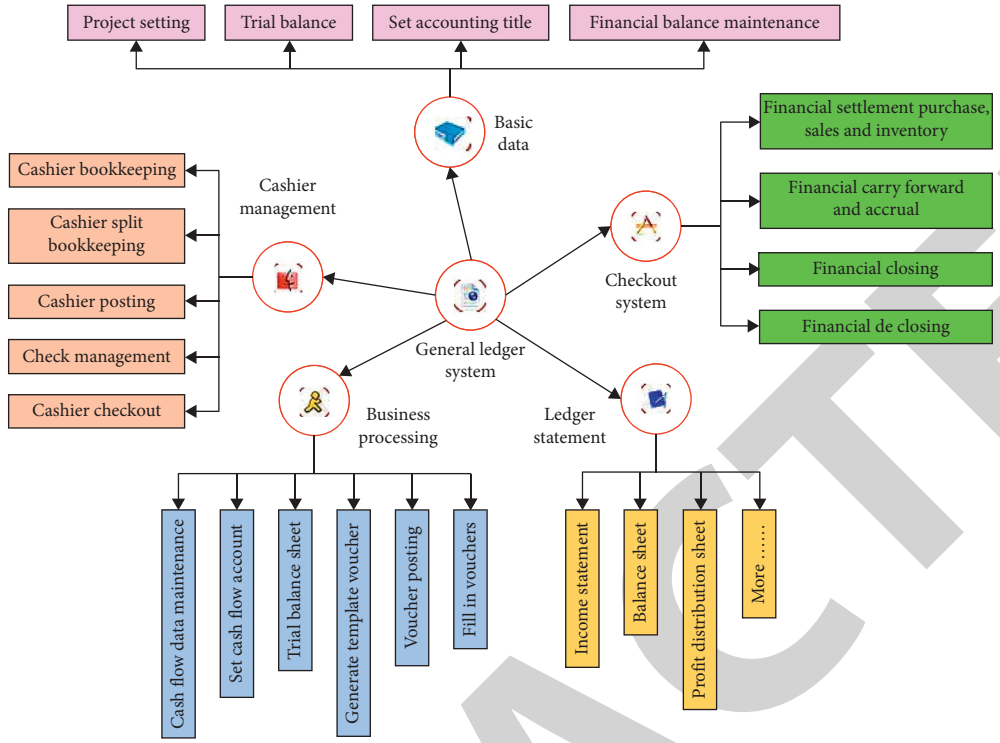

Figure 1: Overall classification and detailed factors for SMEs.

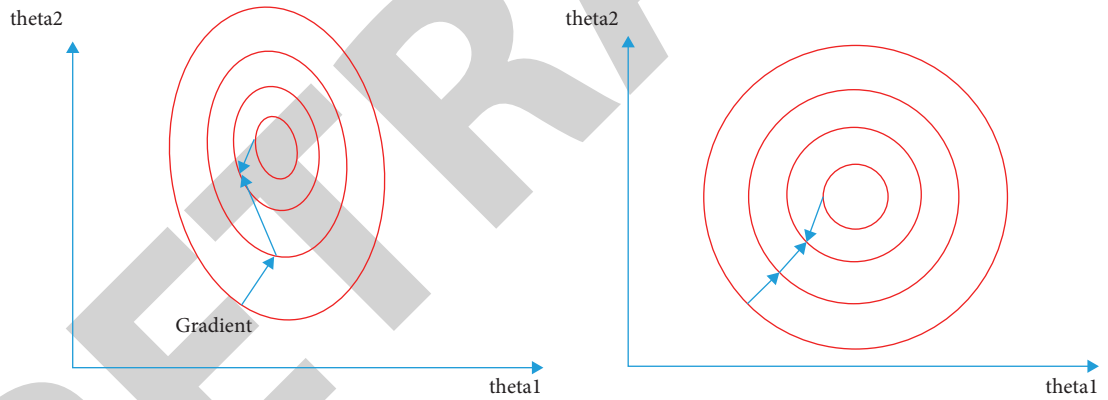

Figure 2: Effect of normalization on the model gradient descent process.

to the normal distribution, which is very unfriendly to many models, especially the linear regression model. For these biased distribution variables, this paper adopts the method of data processing to eliminate them. As can be seen from the correlation coefficient matrix (Figure 5), most variables are very weak (regional information is excluded from the variables, because there is too high correlation between this information, which may lead to some of the models used in this study). It can be judged that the now processed data quality can be applied to the empirical tests of multiple models.

3.2. Model Principle and Settings. The idea of clustering is to generate K regression trees through a certain training method, and then, the scores of these regression trees are added up to get an input final score. Under a certain weight value, the input samples are classified. If there are n samples, there are m features in total. q represents the structure of a regression tree, that is, the mapping of the sample to leaf nodes. T is the number of leaves of this tree.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} fk(x_i), f_k \in \Upsilon, \quad \text{where } T\Upsilon = \{f(x) = \omega_{q(x)}\}(q: R^m \longrightarrow T, \omega \in R^T). \tag{1}$$

We hope to find the optimal regression tree structure through a method, so we make the following constraints on the function in the above formula, so that it has the loss function L of regular penalty term. In addition, by minimizing the value of the loss function, the error between the predicted value and the actual value of the sample can be

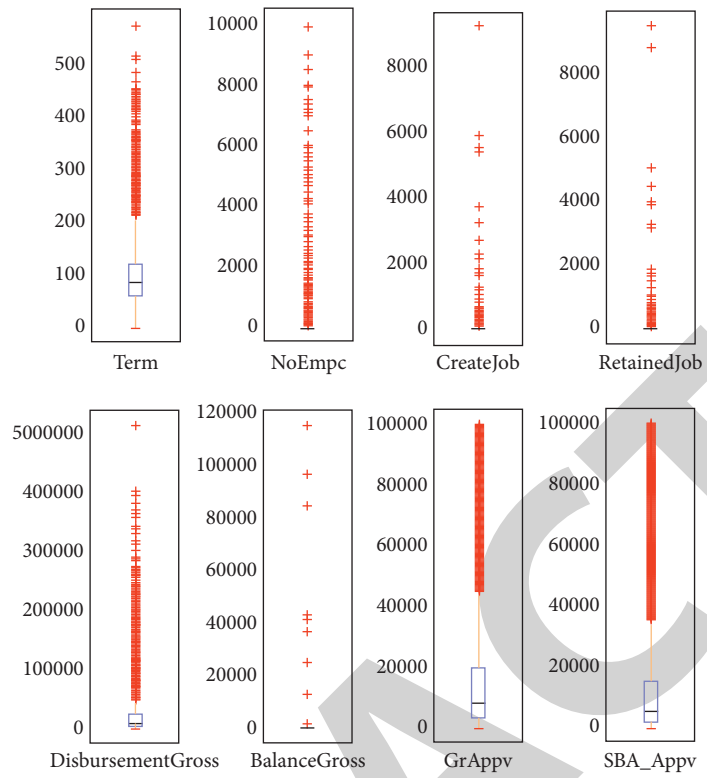

Figure 3: Box plots and violin plots of partial data distribution.

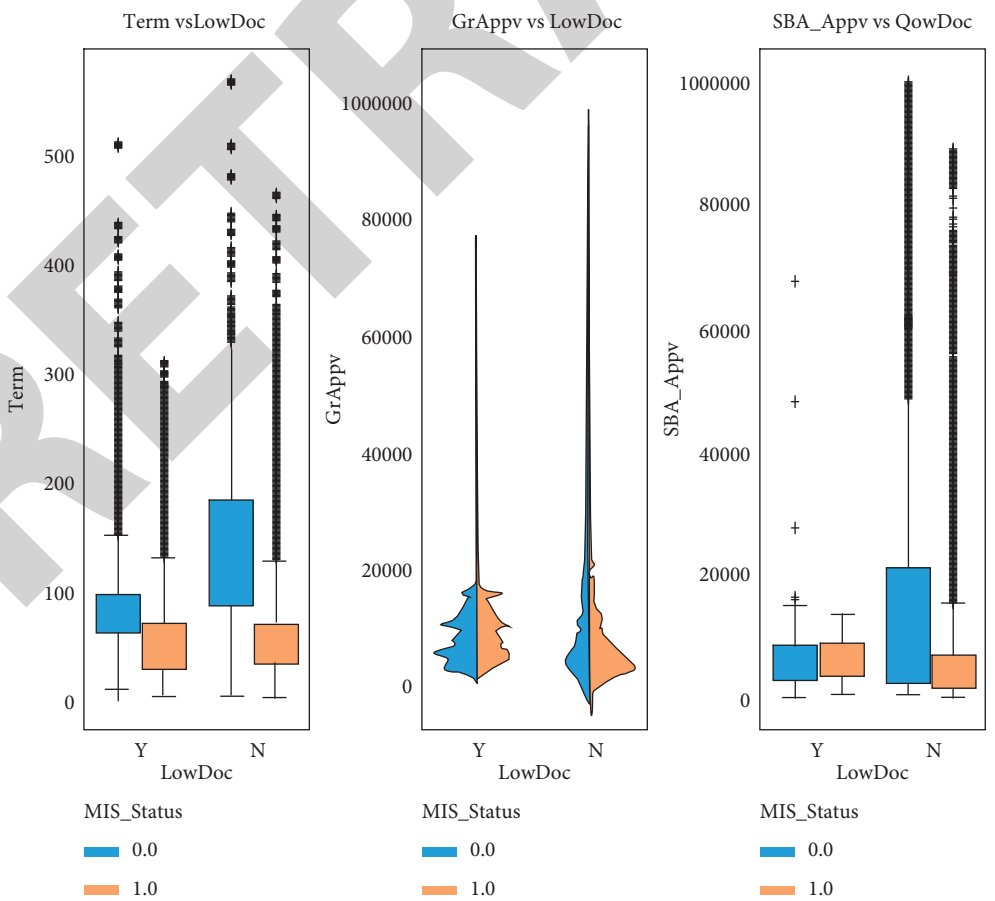

Figure 4: Box plots and violin plots of partial data distribution.

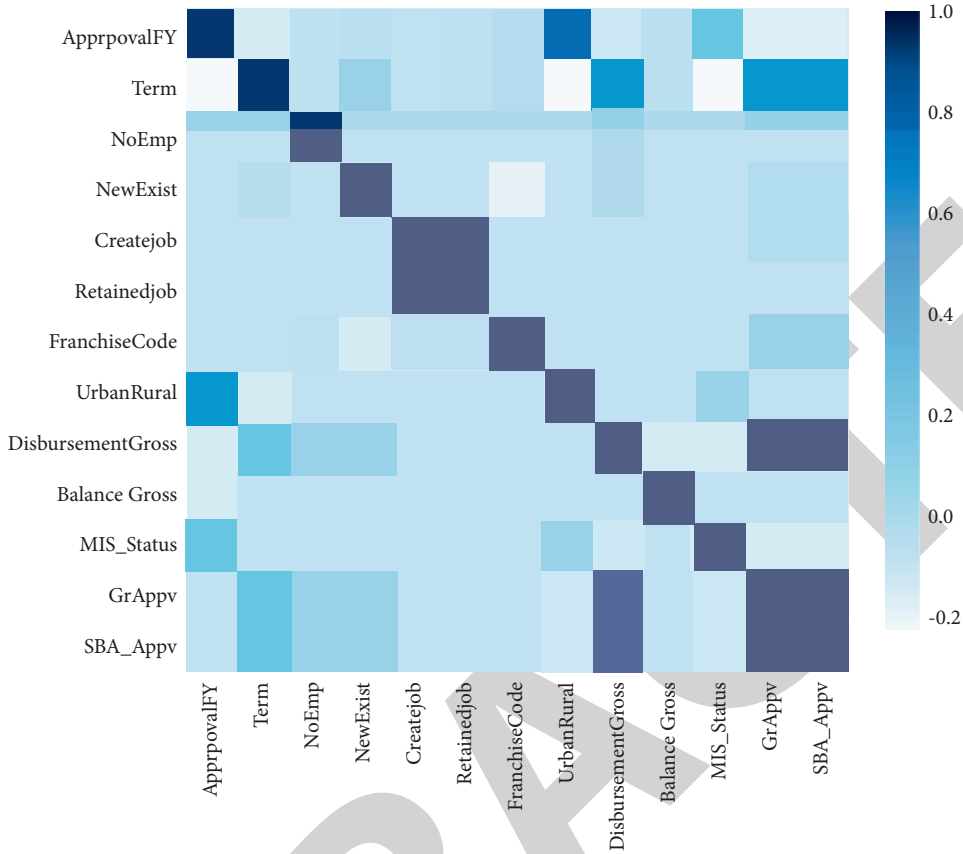

Figure 5: Correlation between the data variables.

continuously reduced, so as to find the optimal regression tree structure in theory.

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad \text{where } \Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2. \tag{2}$$

In 2.2, tree integration model cannot be optimized by the traditional method, so the mean clustering method tries to optimize from the process of iterative accumulation of tree set, from the t tree to the tree set; to minimize the loss function after joining the t tree, we expand the loss function on the first i 1 tree set and discard the remaining term. The constant term is removed to simplify the objective function.

$$\begin{aligned} \mathscr{L}(t) &= \sum_{i=1}^{n} l(y_i, \hat{y}_i(t-1) + f_i(x_i)) + \Omega(f_t), \\ \mathscr{L}(t) &= \sum_{i=1}^{n}\left[l\left(yi, \hat{y}^{(t-1)}\right) + g_i f_t(xi) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_i), \\ \text{where } g_i &= \partial_{\hat{y}^{(i-1)}} l\left(y_i, \hat{y}^{(t-1)}\right), \\ h_i &= \partial^2_{\hat{y}(t-1)} l\left(y_i, \hat{y}^{(t-1)}\right) \\ \tilde{\mathscr{L}}(t) &= \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t). \end{aligned} \tag{3}$$

Now, $I_j = \{i \mid q(x_i) = j\}$ is defined for all samples to fall on the set of the jth leaves so that we can rewrite the formula in 2.3,

$$\begin{aligned} \tilde{\mathscr{L}}(t) &= \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma T. \end{aligned} \tag{4}$$

For any given tree structure q (x), we can compute the optimal weight w of the jth leaf node by the following below$_j$ and calculate the optimal loss value at this time.

$$\begin{aligned} \omega_j^* &= -\frac{\sum i \in I_j g_i}{\sum_{i\in I_j} h_i + \lambda}, \\ \tilde{\mathscr{L}}(t)(q) &= -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T, \\ \mathscr{L}_{split} &= \frac{1}{2}\left[\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \sum_{j=1}^{T}\frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} + \sum_{j=1}^{T}\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \end{aligned} \tag{5}$$

2.6 can be used as an indicator to judge the current tree structure. As long as the above formula can be used in each

new tree, the global optimal tree structure set can be found, but because all the space of the possible tree structure space is infinite, so cannot really traverse to the optimal solution, so actually the mean clustering method uses a greedy strategy, let each tree split from a single node, each division as much as possible enum, compare when the tree structure split after 2.7 to judge whether the loss function value finally contraction, and choose the loss function value minimum, cannot find the tree structure when the iteration is completed. The empirical process of mean clustering method is the process of using sample data to continuously generate the tree structure in the above formula. After the training, the availability of the algorithm can be verified by detecting the accuracy and stability of the test set and checking the specific tree structure.

There are generally two approaches to find the approximate optimal splitting method from the mean clustering method: the feature number of splits based on all nodes and the feature-based total information gain.

The core of the method based on the number of feature divisions is used to calculate the total number of node divisions of each feature in the current node set of all the trees, and this number is used as the importance of this feature, defining the set of attributes in the data as $E = \{e_1, e_2, e_3, ...e_i\}^m$. The set of the regression trees is $F = \{f_1, f_2, f_3, ...f_i\}^k$, *Split* ($e_i$, f_j). For the number of times, the ith feature is divided on the jth tree. Then, the importance of this feature can be expressed as follows:

$$I_m(e_i) = \sum_{j=1}^{m} \text{Split}\left(e_i, f_j\right). \tag{6}$$

The key to the total information gain based on features is to calculate the sum of the information entropy of each feature in the set of nodes of all current trees and to judge the importance of the feature based on the sum of the information entropy. The greater the information entropy, the lower the information gain and importance of the feature. Set E = {e for all features1, e2, e3, ...ei}, for each feature, the value of ei. is contained in {ei 1, ei 2, ei 3, ...eij}. The probability of occurrence is the one of pij. According to the information entropy calculation formula, the information volume is as follows:

$$I\left(e_{ij}\right) = \log_2\left(\frac{1}{p_{ij}}\right) = -\log_2\left(p_{ij}\right). \tag{7}$$

$e_{i,}$ the information entropy, is as follows:

$$H\left(e_{ij}\right) = -\sum_{j=1}^{k} P_{ij}\log_2\left(P_{ij}\right). \tag{8}$$

X = (x) for datasets with a sample size of (ni, yi). Collection of values for different y is {y1, y2, y3, ...yi}. The proportion in the X is p_i. Then, the information entropy of dataset X is $I(X) = -)^{\sum k} p_i \log\ p_i$. If $k = 1$, the collection $X2$ will belong to only one category, and then, $I(X) = 0$. If the feature is e_iX was divided into k subsets $\{X_1, X_2, X_3, ...X_j\}^k$, among X_j, the number of samples included is n_j Then, the information entropy after the feature division is as follows:

$$E(e_i) = -\sum_{j=1}^{k} \frac{n_j}{n} I(X_i). \tag{9}$$

So, the feature ei on the regression tree fj is set. The information entropy on it can be expressed as follows:

$$\text{Gain}\left(e_i, f_j\right) = I(X) - E(e_i). \tag{10}$$

So, the importance of the feature ei can be expressed as follows:

$$\text{Gain}\left(e_i, f_j\right) = I(X) - E(e_i). \tag{11}$$

In this study, we will construct the mean clustering method model from the perspectives of information entropy and feature division number and analyze the economic meaning of variables according to the model's judgment of feature importance.

4. Model Demonstration and Analysis

4.1. Empirical of the Model

4.1.1. Comparison of the Results for the Multiple Models. After training these parameter set models on the dataset and repeating 5 cross-tests (the sample was randomly divided 5 times, of which 90% is the training set, and the rest was used to test the training model), the accuracy and AUC values of these models in the training set are shown in Figure 6.

The area under curve (AUC) value is to solve the problem that the accuracy of the model cannot truly reflect the model classification ability when the sample is extremely unbalanced. The two classification problem samples have positive and negative samples, and the judgment of the samples may be true or false, which constitutes the classification of true positive, true negative, false positive, and false negative. To truly reflect the accuracy of the model, two indicators of true-positive rate (TPRate) and false-positive rate (FPRate) are introduced.

$$\begin{aligned} \text{TPRate} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{FPRate} &= \frac{\text{FP}}{\text{TN} + \text{FP}}. \end{aligned} \tag{12}$$

By constantly taking the classification threshold of the model, the corresponding TPRate and FPRate are obtained, and then, the above points in the two-dimensional coordinate axis form the receiver operating characteristic (ROC) curve. The area of the ROC curve and the X-axis is the AUC value.

4.1.2. Model Stability Test. The stability test of machine learning models mainly focuses on whether the model performance varies greatly on different datasets, and the test method used in this study is fivefold cross-validation. Through five repeated experiments, 90% of the data of the dataset was used to train the model parameters, and 10% of the data were used to test whether indicators such as contrast

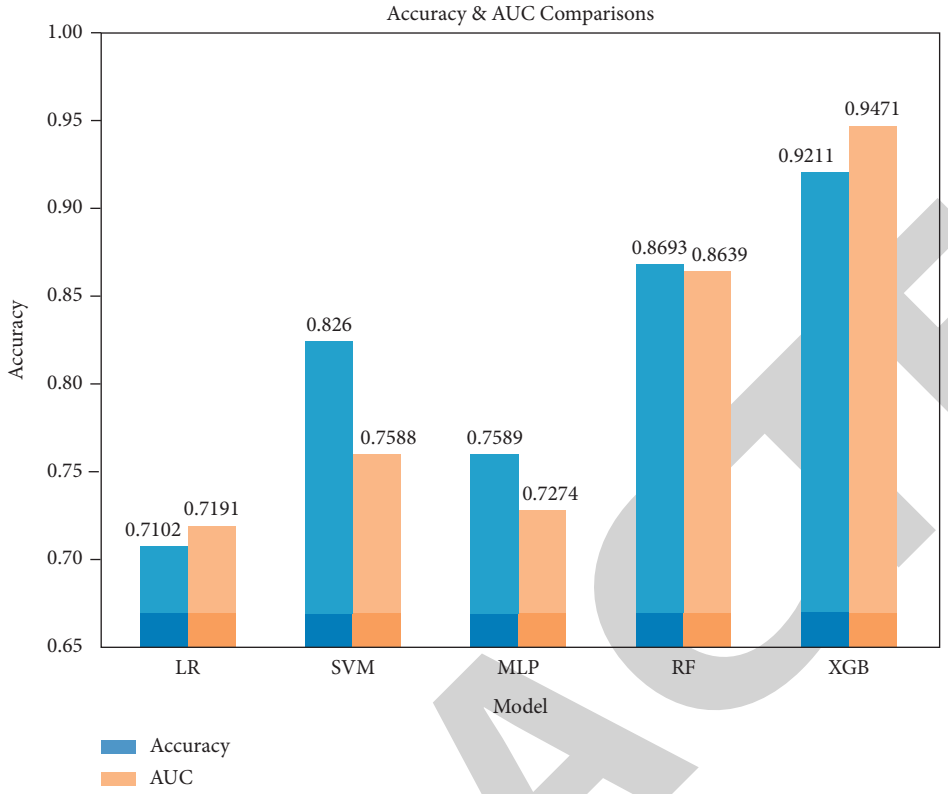

Figure 6: Model accuracy comparison.

Table 1: Cross-validation index of the mean clustering method.

Experimental number	Training set accuracy	Test set accuracy	Training set, AUC	Test set, AUC
1	0.929	0.9215	0.9689	0.9425
2	0.929	0.9093	0.9676	0.9507
3	0.9282	0.9169	0.969	0.9591
4	0.9258	0.925	0.966	0.9648
5	0.9323	0.9189	0.9708	0.9555

accuracy remained consistent. The results are shown in Table 1. Because the parameters of the model (the cluster-trained parameters are the base learner, or 100 regression trees) are completely determined by the training data, when the model is inconsistent in training data and test data, it is called overfitting. The indicator of the fivefold cross-validation conducted in this study can be considered stable and effective if the model does not show overfitting.

Five cross-validation ROC curve as shown below can be found that the ROC curve is very close to the line $y = 1$; although the test set prediction accuracy is about 2% lower than the training set, it still maintains above 90%, and the fluctuation range is small, the choice of data is not too big influence on the model, model stability is high, and the result is shown in Figure 7.

4.1.3. Optimization of the Classification Imbalance. As can be seen from the comparison of prediction accuracy and AUC values of different models, the mean clustering method is much higher than logistic regression and other methods of traditional machine learning methods. The average accuracy reached 94.71%, and the AUC value is close to the accuracy, and ROC curve is very close to $y = 1$, indicating that when processing these small- and medium-sized enterprise loan data samples, clustering method can have stable performance under different sample distribution, and from the perspective of prediction accuracy, average clustering method is the most suitable for predicting small- and medium-sized enterprise credit default risk studied in this study.

Lin et al. pointed out that in the model data sample category imbalance, even the ROC curve normal model has significant defects: the training process is not effective enough, most of the samples are unbalanced data samples, the proportion of the total loss function is too large, and such samples are generally easy to classify and did not provide enough useful signals for the model. Overall, simple positive samples can overwhelm the training, leading to model degradation. Only 20% of the sample in the loan data in this study is ultimately defaulted. The imbalance in the proportion of the loan sample and the default sample repaid on

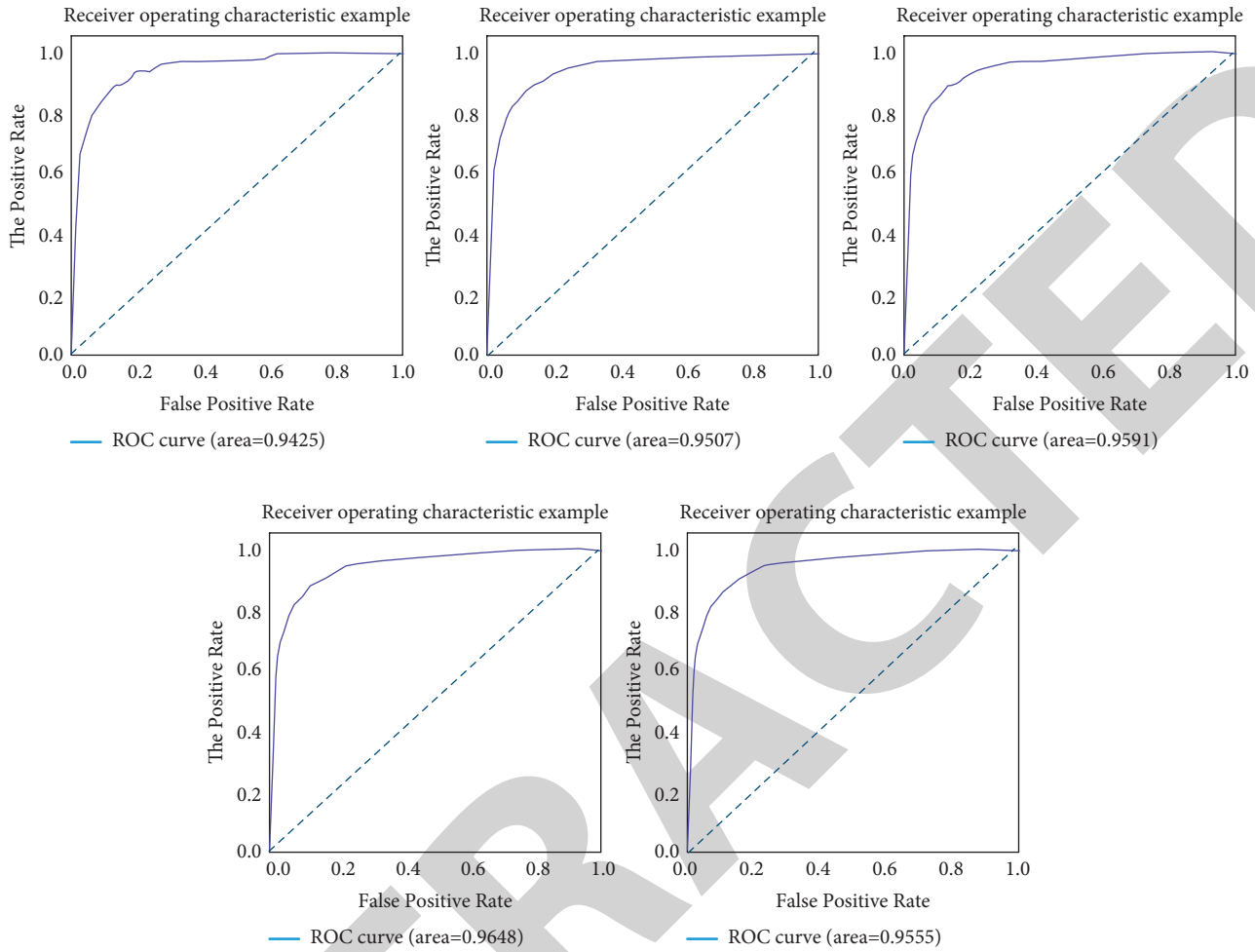

Figure 7: Fivefold cross-validation of the ROC curves.

time may limit the further improvement of the model performance, so this study will improve the ordinary clustering model from this aspect.

In past studies, the imbalance of data sample imbalance is relatively single. The mainstream approach is to increase a small number of samples in the dataset. The SMOTE method selects the nearest sample b for each small number of sample a. New samples were randomly generated between a and b. However, this approach is not applicable to the model shown in this study. Because, according to the empirical results of the model, certain important features such as term loan term volatility presents significantly different effects within different regions, then the samples generated by the SMOTE method may produce large amounts of noisy data. Another common treatment is online hard example mining (OHEM), which calculates the loss and then focuses on hard-to-classify data based on loss. Although the OHEM method increases the number of samples that are difficult to classify, on the other hand, it may cause the model to ignore the easily classified samples and may introduce new bias.

This article has decided to use Lin et al. after the comparison. The focal loss function is modified based on the ordinary cross-entropy loss function used on a single sample:

$$\begin{aligned}\text{Loss} &= -y_i \log f(x_i) - (1-y)\log 1 - f(x_i) \\ &= \begin{cases} -\log f(xi), & y = 1, \\ -\log(1 - f(xi)), & y = 0. \end{cases}\end{aligned} \tag{13}$$

In ordinary cross-entropy function, for each sample the weight is the same, in the classification imbalance data used are easy to ignore the influence of default samples, and focal loss function is to ordinary cross-entropy loss function increased a modulation coefficient, the size of the modulation coefficient depends on the parameter, when = 0. Focal loss function is ordinary cross-entropy function, but when more than 0, for easy to be classified samples, its f (xi). The value is large, which is controlled by the coefficient, the original loss is reduced, the impact on the overall loss is smaller, and the impact of samples that are not easily classified becomes larger. In this way, the processed loss function form can improve the identification ability of the average clustering method to identify a small amount of default data without changing the structure of the original data and introducing noisy data.

$$\text{FocalLoss} == \begin{cases} -(1 - f(xi))^{y} \log f(x_i), & y = 1, \\ -f(x_i)^{y} \log(1 - f(x_i)), & y = 0. \end{cases} \tag{14}$$

For the values of the parameters, Lin et al. found that with the increasing of the values, the importance of easily classified samples will gradually decrease, usually taking 2 to correct most data. After adding the mean clustering method of the focal loss function and comparing it with the previous model, it was found that the improved model had some improvement in the prediction accuracy and AUC value, indicating that the effect of the previous uneven sample distribution was indeed corrected by the modulation coefficient. Figure 8 shows the prediction results with different accuracies.

4.2. Analysis of Empirical Results. The data obtained here were fitted well by the mean clustering method. Although most machine learning algorithms have serious black-box properties, clustering can intuitively demonstrate its ability to analyze data features from two aspects. First, the model will calculate the ability to distinguish all features from samples in the process of generating the tree, and second, we can analyze the structural visualization results of the regression tree. The cluster method has two indicators to evaluate feature importance in dividing nodes: information gain and feature number. This study conducts experiments in these two ways and extracts the characteristic importance value accumulated during the calculation process. The results are compared as follows.

By observing the ranking results of two experiments, the absolute value of most variables is not high, and the ranking is relatively low. However, term, approval date, and disbursement date all showed a very high importance, especially because term was significantly higher than the other variables. This study will focus on the analysis of the term variables. In terms, they are called default and compliance variables.

Figure 9 shows the distribution in two cases. After observing the distribution in two cases (MIS_Status value is default), it is easy to find that the sample distribution of default is significantly concentrated in the short term, while the sample distribution of default is balanced in each time period. However, the term distribution under the two categories also coincides. If the classification is not considered, the term distribution of all samples is still more balanced.

This distribution suggests that the naive prior distribution is largely applicable to the analysis of term features and can only discuss its Bayesian probabilities after obtaining sample classification. The reason for the phenomenon of default samples concentrated in short-term loans may be related to the business model of small- and medium-sized enterprises. The capital demand of enterprises can be divided into long-term demand and short-term demand according to the length of time. These two demands are not only quite different in time but also have obviously different reasons for the demand.

For small- and medium-sized enterprises, long-term capital demand is more likely to come from the needs of fixed assets investment caused by the expansion of business development scale, while short-term capital demand is more likely to come from the shortage of funds needed for the operation of enterprises. In intuitive analysis of the enterprise operation and development information behind these two capital needs, the long-term capital demand is more related to the positive signal of the good development of the enterprise, while the short-term capital demand may be a sign that the enterprise fell into the predicament of poor management.

On the other hand, the interest rate of long-term loans will be higher than that of short-term loans under the same conditions. For enterprises with good operations that are less likely to default in the future, they will choose a reasonable loan cycle and amount according to their own needs. Enterprises with high uncertainty in the future are more likely to borrow from banks under the pressure of capital turnover and are more inclined to apply for short-term loans at high interest rates on long-term loans. Combined with long-term loans and short-term loans and the operating conditions of small- and medium-sized enterprises, the distribution of the data concentration loan cycle attribute is indeed in line with economic knowledge.

It can also be seen from the regression tree structure generated by the cluster training that the clustering gives a very high importance to the variable of term. In addition, the model is judged according to the different value range of term variables combined with other variables, rather than a simple linear simulation. Therefore, the scores of samples fluctuating within multiple intervals of term fluctuate, which also accords in line with the analysis of this study, so we cannot directly infer the default probability of loan application according to the size of term.

For the data variables used in this study, in addition to the attributes of the loan itself, it can basically be divided into macro-variables (time category) and enterprise characteristic variables. Judging from the model judgment and analysis of the importance of these variables, the importance of macro-variables is significantly higher than that of enterprise characteristic variables. The main reason for this distribution of importance is that the business scope of SMEs is relatively simple and the scale of capital is relatively small, resulting in low resistance to various risks. Not only large enterprises need to face this nonsystematic risk but also small- and medium-sized enterprises will encounter it under the influence of the economic cycle, because small- and medium-sized enterprises are vulnerable to the contraction of downstream demand and changes in upstream supply due to their small size. On the other hand, these loan applications have been reviewed by banks and SBA, and credit providers. The examiner reviewed the company's business and finances in advance and passed the sample. From a credit provider's perspective, detailed business information is more readily available when processing applications, while macroeconomic fluctuations are relatively unknown. These reasons make it easier to observe sample data.

Based on the empirical results of the above models, the clustering method has higher accuracy and consistent stability than other methods in predicting the credit default of small- and medium-sized enterprises, the error rate is lower, and the training method is simpler. For the data used in this study, the final cycle of loan default will have an impact, but the impact

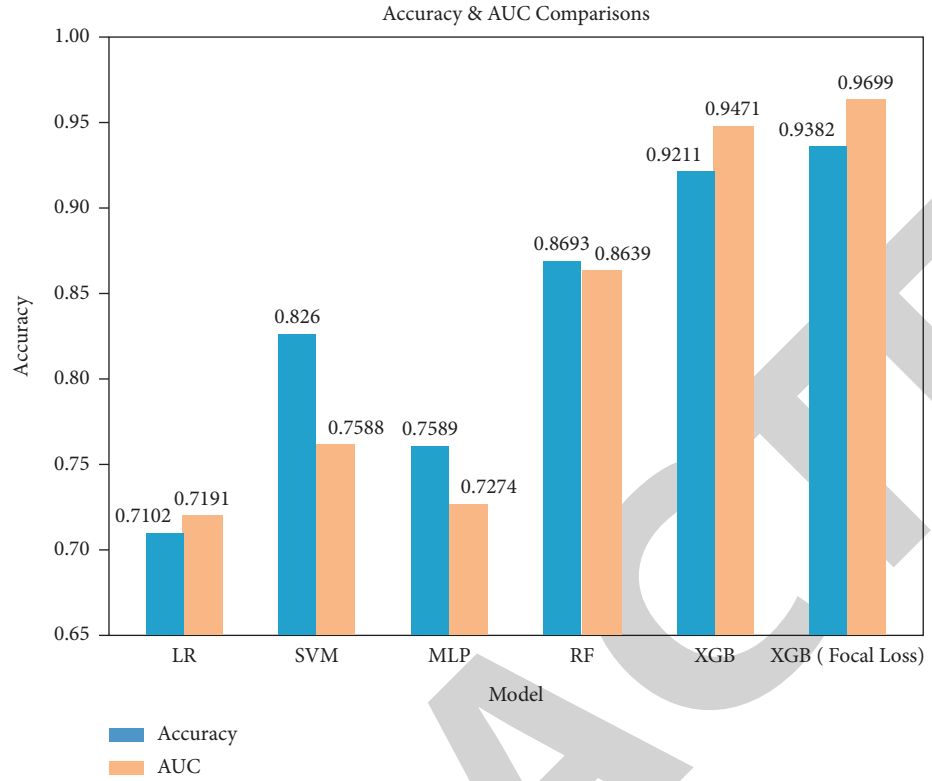

Figure 8: Comparison of prediction accuracy with the focal loss mean clustering method.

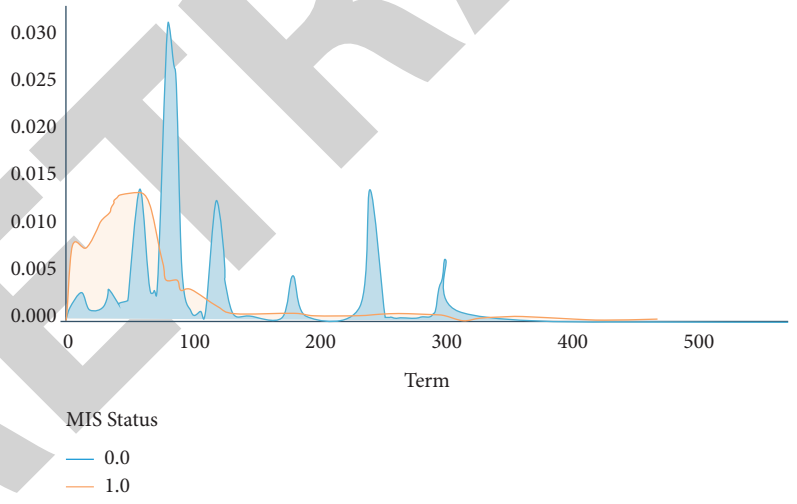

Figure 9: Scatterplot of the distribution of term variables in default and conservation cases.

is not linear, and it is still necessary to combine the characteristics of the macro- and company levels to effectively classify the sample. Combined with the historical background of the data, the model established in this study can further judge the loan default probability after analyzing and reviewing the loan, and the false reporting rate and false reporting rate are very low. As a supplementary means of bank credit risk control, the effect is very good. The downside is that while the models can more specifically analyze the impact of individual variables on defaults, they are not sufficiently powerful to explain how individual variables affect loan defaults.

5. Conclusion

This study analyzes and compares the credit records of SBA partially guaranteed by SVM, random forest, neural network, and logistic regression. The empirical results of the model show that the average clustering method has higher accuracy in predicting SME credit default scenarios than

other machine learning methods and ordinary multiple regression models. The mean clustering method showed higher prediction accuracy and AUC values than the other models. After many cross experiments, the accuracy of the model remains about 94%, which proves that the result of the model is relatively stable. By observing the ROC curve of the model, it can be predicted that when the positive rate is very low, the true positive rate can quickly converge to the maximum value, which proves that the mean clustering method can not only classifies the loans of small- and medium-sized enterprises but also stabilizes the prediction error of each category at a low level. The prediction accuracy of further optimizing the model according to the imbalance of loan classification reached 95%. By clustering model iteration in the process of the importance of data characteristic score derived, analysis model to the importance of variable judgment can be found that the loan cycle has the biggest influence on default probability, enterprise loan demand comes from long-term demand and short-term demand and presents significant differences in the term, and the root cause is the internal cause of loans is different. For small- and medium-sized enterprises, the long-term capital demand is strongly related to the fixed asset investment needs brought about by the expansion of business development scale, while the short-term capital demand is more related to the shortage of funds needed for enterprise operation. Enterprises with long-term capital needs tend to send positive signals of good development, and short-term capital needs are more likely to reveal that enterprises are in operating difficulties.

Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Chongqing City Management College.

References

[1] J. A. Nelder and Y. Lee, "Generalized linear model 2nd ed," *Applied Stochastic Models in Business and Industry*, vol. 7, no. 1, pp. 107–120, 1989.

[2] G. Sun, J. Li, and F. Lichen, "Positive study of — of Digital Credit based on three-stage Probit model," *Agricultural technology and economy*, no. 12, p. 18, 2021.

[3] W. Li, *Data-Driven Default Risk Forec Method of Consumer Finance*, Hefei University of Technology, Hefei, China, 2019.

[4] X. Li, "A personal credit default forecast study based on XGBoost," *Computer Knowledge and Technology: Academic edition*, vol. 15, no. 11X, p. 3, 2019.

[5] X. Qian, "Research on non-performing assets of Chinese commercial banks based on the view of credit trend [J]," *Western Finance*, no. 2, p. 5, 2019.

[6] International Finance, "Paint.enlightenment of archegos default event on counterparty credit risk management [J]," *International Finance*, vol. 6, no. 10, p. 6, 2021.

[7] X. Li, "Takes the perspective of the youth group," *Financial Observation*, vol. 1, 2018.

[8] H. Yue and Z. Jiao, "Countermeasures to reduce the default risk of real estate credit in shenyang," in *Proceedings of the Shenyang Science Annual Conference.Shenyang Association for Science and Technology*, Shenyang, China, April 2015.

[9] S. Dong, "From the credit practice of supply chain finance business of several insights," *Operation and Management of Commercial Banks*, no. 6, 2022.

[10] P. small group, "Research on credit risk evaluation of small and micro enterprises from the perspective of supply chain finance," *Business situation*, no. 2, p. 3, 2022.

[11] T. Wang, "On the P2P network credit under the background of Internet finance," *Journal of Financial Risk Management*, vol. 6, pp. 163–190, 2017.

[12] X. Yang and G. Duan, "Research on credit risk prevention and control of supply chain finance business of commercial banks from the perspective of financich," *Finance and Market*, vol. 6, no. 3, 2022.

[13] H. Ge and L. Zhang, "Study on risk pricing of financial credit ratings and credit default saps," *Credit investigation*, no. 11, p. 4, 2016.

[14] W. Jin, *Study on Loan ault Risk*, Central South University of Forestry and Technology, Changsha, China, 2016.

[15] Di. Liu, *Study on Risk Control of Supply Chain Finance Credit Business for Commercial Banks*, Yunnan University, Kunming, China, 2016.

[16] L. Wu, "Default risk of supply chain finance," *Inner Mongolia Coal Economy*, no. 23, p. 2, 2016.

[17] Le. Jiang, *Is Based on the Analysis of Small and Medium-Sized banks in S City*, Southwestern University of Finance and Economics, Chengdu, China, 2016.

[18] Ye. Qian, *Research on the Mechanism Influence and Counterbalance of Credit Rationing*, Huazhong University of Science and Technology HUST, Wuhan, China, 1981.

[19] X. Zhang, *A Credit Score Model Study Based on Machine Learning*, Southwest University, Juarez, Mexico, 2021.

[20] Y. Chong, W. Zhu, and T. Miao, "Bank effect maximization credit decision model based on *K*-means clustering," *Economics*, vol. 4, no. 3, pp. 30–32, 2021.