Hindawi
Complexity
Volume 2021, Article ID 9222617, 12 pages
https://doi.org/10.1155/2021/9222617

Research Article

Application Analysis of Credit Scoring of Financial Institutions Based on Machine Learning Model

Yi Wu and Yuwen Pan

Emlyon Business School, Ecully Cedex 69134, France

Correspondence should be addressed to Yi Wu; 15121403107@stumail.sdut.edu.cn

Received 13 August 2021; Accepted 22 September 2021; Published 22 October 2021

Academic Editor: Shaohui Wang

Credit score is the basis for financial institutions to make credit decisions. With the development of science and technology, big data technology has penetrated into the financial field, and personal credit investigation has entered a new era. Personal credit evaluation based on big data is one of the hot research topics. This paper mainly completes three works. Firstly, according to the application scenario of credit evaluation of personal credit data, the experimental dataset is cleaned, the discrete data is one-HOT coded, and the data are standardized. Due to the high dimension of personal credit data, the pdC-RF algorithm is adopted in this paper to optimize the correlation of data features and reduce the 145-dimensional data to 22-dimensional data. On this basis, WOE coding was carried out on the dataset, which was applied to random forest, support vector machine, and logistic regression models, and the performance was compared. It is found that logistic regression is more suitable for the personal credit evaluation model based on Lending Club dataset. Finally, based on the logistic regression model with the best parameters, the user samples are graded and the final score card is output.

1. Introduction

With the rapid development of big data technology, the credit information system has entered a new era. Credit reporting enterprises make profits from credit business, but they also bear certain credit risks. Constructing and perfecting credit information market is of great significance to economic development and market economy reform. On the one hand, as an economic means of national macrocontrol, the personal credit information system is related to the improvement of the future credit information system of the motherland and reduces the possibility of credit crimes. On the other hand, opening up a new personal credit system is conducive to financial institutions such as banks to improve their efficiency and handle credit business more accurately for different groups of people.

Due to the huge and changing volume of social media data (SMD) [1], the traditional data warehouse is no longer applicable. Starting from the new framework for analyzing twitter data, the researchers tested the scalability of the proposed framework by detecting data of different sizes and effectively processed a large amount of data. Query, obtain, and analyze unstructured data and data flow under certain memory resources. Big data play an important role in predicting and evaluating economic credit. Through the test of lending institutions, it is proved that [2] a special test within lending institutions is 18.4% lower than the big data credit score forecast for evaluating whether individuals will default on loans, and it has great advantages in predicting people who have never had a borrowing record and correcting financial reporting errors. The importance of credit evaluation score is becoming more and more prominent, and any institution needs to avoid some risks through a high-performance credit scoring model [3]. Put aside the technical research on credit scoring technology; adopt alternative data sources to improve the performance of statistical and economic models. In this way, the big data source calls the network, and then, the feature selection increases the profit value and uses the dataset to create the score of users applying for credit cards. Big data have been applied

and recognized in various fields. On the premise of protecting privacy, people hope that the big data-driven model can be applied more efficiently in financial credit scoring. A projected gradient-based scorecard learning method (FL-LRBC) can make the positive coefficients of the model form [4], and it has abundant data, especially excellent performance in statistics, and it has been applied in some financial and credit businesses. In the financial retail credit scoring, we need to prove the performance of machine learning (ML) model in the face of complex ML algorithm [5]. Besides reanalyzing the defects of traditional models, we must also put forward new viewpoints on structure selection and sensitivity, thus proposing a model criticism based on reverse generated counterfactual examples. Credit score directly affects the profitability of the industry, and it is also a key research topic. The credit score classification of Spiking Extreme Learning Machine (SELM) has been verified in many European countries [6]. It is a new spike energy function. Compared with some traditional spike neuron models, such as classification accuracy and area under curve (AUC), the accuracy and execution speed of SELM have been greatly improved, which has statistical significance for credit score datasets, indicating that the combination of biological spike function and ELM can better classify. The ability of lenders to repay is particularly important for banks, which reflects the value of credit scores, which can assess whether loans are risky or not. When banks approve loans, they will give priority to loans with high credit scores. Besides scoring, they also involve digital verification of customers' credentials, which is tedious. Researchers introduce a framework and create a digital identity and use block chain to calculate credit score [7]. In a distributed user network, details are transparent, which improves the trust level and plays a fair role in scoring. The banking industry pays special attention to customer recognition and provides personalized services for customers, and artificial intelligence runs through it from the front desk to the middle desk to the back office [8]. Peer-to-peer lending has gradually emerged in recent years, and researchers have discussed and analyzed this phenomenon. Research is carried out through open source data and the applicability and effectiveness of credit scorecards is checked. Scorecard is based on P2P loan open dataset, so it contains repayment records and loan functions. In order to improve credit management and the stability of credit score, a method is used. First, the user's data are obtained, and the data are input into the neural network configuration module for processing [9], so as to obtain a credit probability value and the score. In actual cases, this method is obviously stable, which greatly improves the user experience. It is very necessary for financial institutions to predict customers' default. At present, a conservative credit scoring model is established by using Kruskal–Wallis nonparametric statistics [10]. This model belongs to low cost and high accuracy. It is applied to the credit score of nontraditional data sources and changes according to the number of online applications of customers.

2. Brief Introduction and Preprocessing of Experimental Dataset

The personal credit evaluation model in this paper uses experimental data from Lending Club. Lending Club dataset is adopted in this paper because the dataset is public and can be evaluated. The dataset can reflect the evaluation standard of credit data. Through the Lending Club dataset algorithm, the design can show the advantages and better performance of the proposed algorithm.

The credit data provided by the Company in the first quarter of 2017 have a total of 96,781 samples and 145 characteristics, which mainly include the financial data of Lending Club Company, but do not include the credit rating. The dataset feature loan_status contains seven types of debit and credit states, as shown in Table 1.

For the convenience of follow-up research, we take loan_status feature as the target vector. Do not study the data of "repayment in progress," regard "full repayment" as users with good credit, and classify the data of other states as users with poor credit. Among them, there are 23,381 users with good personal credit and 9,088 users with poor personal credit, totaling 32,469.

2.1. Data Cleaning. Data cleaning is the first step in the process of data preprocessing; the main purpose of which is to fill the missing values and deal with outliers. For missing samples in datasets, the missing value processing method in this paper follows the following principles:

(1) If more than half of the data in a sample is null, the set of data is discarded. In the environment of big data, there are deficiencies in datasets due to differences in dataset collection. It is reasonable for a large-scale dataset to have NULL values. If there are multiple missing data, it can be deleted directly. If there is a large amount of data, deleting some will not affect the evaluation of the whole dataset.

(2) Usually, the missing value of the sample is filled by the mode of the feature. Mode filling principle is aimed at the value with the highest number of occurrences in the data, which is based on the maximum probability filling method and can improve the efficiency of dataset integration. Therefore, the preference of the whole dataset is better, which accords with the effect of data filling.

(3) If the missing feature of a sample exceeds 95%, the feature will be discarded. In order to ensure the integrity of data as much as possible, this paper converts the format of discrete data before filling in missing data, and the missing data are shown in Table 2.

Among them, the missing ratio is the ratio of the number of missing data of this feature to the number of rows of the dataset, which is 32469. Through data format conversion and mode filling, the dataset is complete. The processing of outliers is simpler. It only needs to judge whether the data

Table 1: Lending Club dataset lending status.

Loan_status	Description	Number
Current	Repayment in progress	64310
Fully paid	Full settlement	23381
Charged off	Bad debts	5740
Late (31–120 days)	Anticipation (31–120 Days)	1872
In grace period	Grace period	878
Late (16–30 days)	Anticipation (16–30 Days)	473
Default	Default	125

Table 2: Missing data and data conversion.

Characteristic	Deficiency ratio	Original data format	Current data format
dti	0.000277	float64	float64
revol_util	0.000493	Object	float64
mths_since_rcnt_il	0.023284	float64	float64
il_util	0.132680	float64	float64
bc_open_to_buy	0.010872	float64	float64
bc_util	0.011611	float64	float64
mo_sin_old_il_acct	0.023284	float64	float64
mths_since_recent_bc	0.010564	float64	float64
mths_since_recent_inq	0.089655	float64	float64
percent_bc_gt_75	0.010872	float64	float64

are within the reasonable data range, and then, delete the sample data with outliers. The Lending Club dataset used in this article found no outliers in the dataset. Finally, when the data cleaning is completed, there are 32469 samples and 74 features in the dataset used in this paper.

2.2. Standardization of Data. When executing machine learning algorithm, the magnitude of features is the main reason for model overfitting. When calculating features of different orders of magnitude, the distance between variables of large-scale features and small-scale features is too large, which will lead to deviation in training of the model, thus degrading the performance and failing to meet the expected standard. Therefore, before training the model, we must ensure that each variable has the same measurement standard for the sample, that is, standardized treatment. The methods of standardization are different. According to the types of variables, they can be divided into continuous variable standardization and discrete variable standardization. Min-max normalization and z-score normalization are two commonly used methods of continuous variable normalization.

2.2.1. Min-Max Standardization. The idea of min-max normalization is to specify a minimum and maximum interval, map features to a given interval, or convert the absolute value of each feature to a unit size [11]. Through the linear transformation of the original data, the data are normalized to the middle of [0, 1]. The conversion function is

$$x^* = \frac{x - \min}{\max - \min}. \tag{1}$$

One drawback of min-max standardization is that adding new data to the data source may cause max and min to change, and each time the data is added or subtracted, it needs to be redefined.

2.2.2. z-Score Standardization. The idea of z-score standardization is to scale the data based on the arithmetic mean and standard deviation of the original data [12]. After z-score standardization, the data show standard normal distribution, that is, the mean value is 0, the standard deviation is 1, and the conversion function is

$$x^* = \frac{x - \mu}{\sigma}. \tag{2}$$

In view of the continuous characteristics of the dataset, this paper uses the min-max standardization method to deal with it.

Compared with the standardization of continuous variables, the standardization of discrete variables is more complicated. In order to measure the distance between samples, it is necessary to introduce dummy variables by means of coding. Common encoding methods are one-hot encoding and dummy encoding.

2.2.3. One-Hot Code. The basic idea of one-hot is to abstract different values of discrete features into a class of states, and N different values correspond to N different states [13]. Assuming that a feature has five states, one-hot encoding is shown in Figure 1. The one-hot encoding ensures that only one bit of each state is activated as 1, and all other state bits are 0. The advantage of this is to ensure that the distance between data is $\sqrt{2}$.

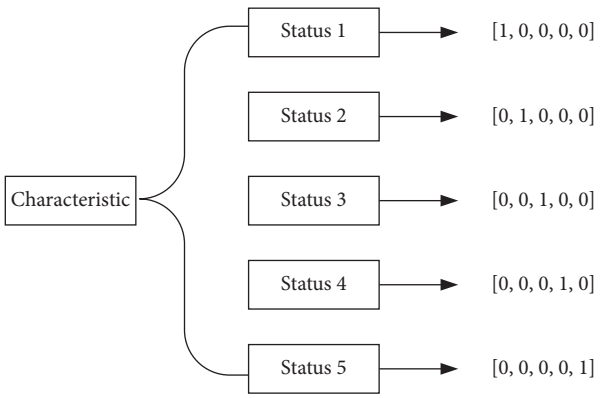

Figure 1: One-hot encoding.

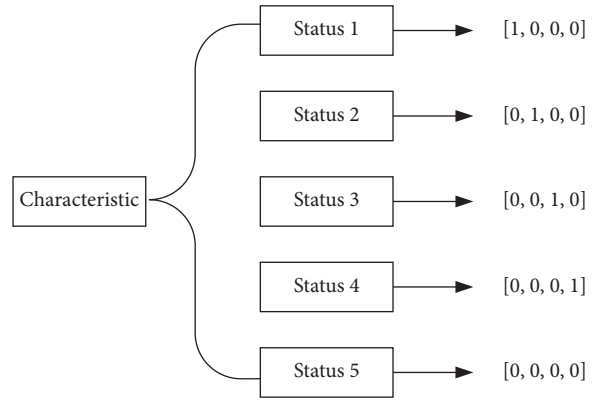

Figure 2: Dummy encoding mode.

2.2.4. Dummy Code. The basic idea of dummy variables is similar to one-hot encoding, except that dummy variables select a state in the feature and set all its state bits to inactive state [14]. Analogous to the encoding method in Figure 1, dummy variables can make the state vector one bit less, and its encoding method is shown in Figure 2.

As for the calculation of distance between samples, one-hot coding is more universal than dummy variables, and the distance between samples in different states is fixed, while dummy state samples in dummy variables will show differences in distance calculation. In this section, one-hot encoding is used to encode class variables.

2.3. Dimension Reduction of Experimental Dataset Based on Pdc-RF Algorithm. On the basis of data preprocessing in the previous section, this section is based on Python programming, and Jupyter Notebook data analysis platform has reduced the dimension of 32469 samples and 74 experimental datasets with characteristics. In Random Forest algorithm, test_size = 0.2 is chosen as the partition of the training set and the test set; the number of trees in the tree is set to n_estimators = 500, the depth of the tree is set to max_depth = 10, and other parameters are set as default. The features are sorted according to their importance, and then, the features whose importance is less than 0.01 are taken out as nonimportant features and eliminated. Finally, the first 41 features are selected as feature subsets. In this section, the feature selection and screening results of category features are counted, as shown in Table 3.

Then, based on Pearson correlation coefficient and distance correlation coefficient, the continuous features in the feature subset are further screened, and the correlation coefficients between some features are intercepted, as shown in Table 4.

It can be seen that the diagonal correlation coefficients of the matrix are all 1.0, and other coefficients correspond to Pearson correlation coefficients between the features of columns and columns, where the correlation coefficient between installation and loan_amnt is 0.95, far greater than the correlation coefficient between other characteristics. While installation represents the monthly amount of arrears, and loan_amnt represents the amount of loans applied by users. According to the formula, monthly arrears = (loan amount ÷ number of loan periods) + (loan amount-repaid amount) × interest rate; it is not difficult to conclude that the monthly amount of arrears is directly related to the loan amount, which is consistent with the correlation coefficient of 0.95. Based on the dimensionality reduction idea of preventing overfitting, loan_amnt can be eliminated. Finally, 21 features selected from Pdc-RF algorithm features are shown in Tables 5 and 6 .

In this section, min-max standardization and one-hot coding are used to standardize the continuous variables and discrete variables in the dataset, respectively, and the first step of the data preprocessing stage is completed at this time. Then, the improved pdc-RF algorithm is applied to standardized data, and the experimental data are extracted and reduced by features. Finally, the data are reduced from 74 dimensions to 21 dimensions, and the second step of data preprocessing is completed.

3. Construction of Personal Credit Rating Evaluation Model

In the previous section, based on the Pdc-RF feature selection algorithm, the dimensionality reduction of data is completed, and a dataset with 21 features and 32469 samples is obtained. On this basis, this section further builds a supervised personal credit scoring model. A complete personal credit evaluation model should include data processing, model building, model evaluation, and credit score card conversion. Based on the data preprocessing and dimensionality reduction in the previous section, the personal credit evaluation model of this paper starts with data partition, carries out box processing on the data, and completes the discretization of the data. Then, based on the related algorithms of machine learning, the samples are trained, and then, the performance of each machine learning algorithm is compared to obtain the best performance training model. Finally, the model is transformed into a scorecard to complete the construction of the whole credit evaluation model.

Table 3: Category features in screening based on importance.

Characteristic	Meaning	Importance	Yes or no
Term	The number of installments of payment for goods; the value is in monthly units, which can be 36 or 60	0.0124	√
Grade	Credit specifies loan grade	0.0611	√
home_ownership	The ownership status of the house provided by the borrower at the time of registration, including rent, self-ownership, mortgage, and others	0.0115	√
Verification_status	Indicate whether the common income or other income sources of co-borrowers have been verified	0.0119	√
Purpose	The purpose provided by the borrower for the loan application	0.0076	×
addr_state	Address (country) provided by the borrower in the loan application	0.0115	√
initial_list_status	Initial list status of loans, which may be W and F	0.0017	×
application_type	Indicate whether the lender applies individually or jointly with two co-borrowers	0.0008	×
disbursement_method	Payment method	0.0002	×

Table 4: Correlation coefficients between partial features.

	loan_amnt	int_rate	Installment	annual_inc
loan_amnt	1.000000	0.187009	0.955182	0.271570
int_rate	0.187009	1.000000	0.243790	0.061228
Installment	0.955182	0.243790	1.000000	0.251408
annual_inc	0.271570	0.061228	0.251408	1.000000
loan_status	0.093558	0.243346	0.102000	0.030061
dti	0.034353	0.112351	0.047161	0.099307
revol_bal	0.312685	0.011460	0.297680	0.222215
revol_util	0.127230	0.215067	0.145422	0.033434
total_acc	0.170489	0.036466	0.150041	0.143473

Table 5: Description of category variables.

Class variables	Meaning
Term	Number of installments of payment for goods
Grade	Credit specifies loan grade
home_ownership	Ownership of the borrower's house
verification_status	Revenue verification status

Table 6: Description of continuous variables.

Class variables	Meaning
int_rate	Interest rate
Installment	Monthly arrears
annual_inc	Annual income of borrower
dti	Ratio of monthly debt repayment to total debt
mths_since_rcnt_il	Number of months after the installment account was opened
il_util	Ratio of total user balance to high credit/credit limit
mo_sin_old_il_acct	Number of months since the earliest bank account was opened
mo_sin_old_rev_tl_op	Number of months since the earliest revolving account was opened
mths_since_recent_bc	Number of months since last online payment
mths_since_recent_inq	Number of months of last loan inquiry
num_rev_tl_bal_ge_0	Number of revolving accounts
pct_tl_nvr_dlq	Unknown
percent_bc_gt_75	Recent income-expenditure ratio
tot_hi_cred_lim	Credit card credit limit
total_bc_limit	Amount limit of online banking
total_il_high_credit_limit	Limit of overdue repayment amount
income_vs_loan	Payment of income and loan amount

3.1. Data Partition

3.1.1. Chi-Square Subbox. Before the establishment of the model, the continuous variables need to be discretized, which can avoid the influence of extreme values in features and reduce the risk caused by overfitting of the model. In this section, the chi-square box-dividing method is used to divide the continuous variables into boxes.

Chi-square box division depends on Chi-square test and adopts the bottom-up data discretization method. The idea of chi-square box dividing is to merge adjacent intervals with minimum chi-square value and cycle back and forth until the chi-square threshold condition is reached [15]. The steps of chi-square box dividing are as follows:

Step 1: preset a chi-square threshold and the maximum number of boxes

Step 2 (initialization): initialize and sort the values in the features, and each instance belongs to an initial interval

Step 3 (merging interval): calculate the chi-square value of adjacent intervals, and merge a pair of intervals with the smallest chi-square value according to

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}}, \tag{3}$$

where A_{ij} represents the number of instances of class j in the ith interval and E_{ij} represents the expected frequency of A_{ij}, which is calculated according to

$$E_{ij} = \frac{N_i \times C_j}{N}, \tag{4}$$

where N is the total number of samples, N_i is the number of samples in group i, and C_j is the proportion of samples in group J.

Step 4: repeat steps 2 and 3 until the chi-square value is greater than the chi-square threshold.

3.1.2. WOE and IV Values. Weight of Evidence is referred to as WOE, that is, evidence weight, which represents the possibility that a grouping in features is predicted as a bad sample. The calculation formula of WOE is shown in equation (5). The larger the WOE is, the greater the possibility that the grouping belongs to bad samples:

$$\mathrm{WOE}_i = \ln\left(\frac{py_i}{pn_i}\right) = \ln\left(\frac{\#y_i/\#y_T}{\#n_i/\#n_T}\right). \tag{5}$$

In personal credit evaluation, bad samples represent defaulting customers, while good samples represent non-defaulting customers. Then, in equation (5), py_i represents the proportion of bad samples in the grouping to bad samples of all data, pn_i represents the proportion of good customers in the grouping to good samples of all data, $\#n_i$ represents the number of good samples in the grouping, $\#y_i$ represents the number of bad samples in the grouping, $\#y_T$ represents the total number of bad samples of all data, and $\#n_T$ represents the total number of good samples of all data.

The full name of IV is information value, that is, information value (information amount). The IV value can measure the predictive ability of variables, that is, compare the difference between the distribution of good samples and bad samples in terms of the information value. For grouping i, the calculation formula of IV value is shown in

$$\begin{aligned}\mathrm{IV}_i &= \left(py_i - pn_i\right) * \mathrm{WOE}_i = \left(py_i - pn_i\right) * \ln\left(\frac{py_i}{pn_i}\right) \\ &= \left(\frac{\#y_i}{\#y_T} - \frac{\#n_i}{\#n_T}\right) * \ln\left(\frac{\#y_i/\#y_T}{\#n_i/\#n_T}\right).\end{aligned} \tag{6}$$

The IV value of the whole variable is the sum of the corresponding IV values of each group:

$$\mathrm{IV} = \sum_{i}^{n}\mathrm{IV}_i. \tag{7}$$

3.1.3. Exploration of Data Partition. Based on chi-square box and WOE coding, the method of calculating IV information degree is used to process the data. It is worth noting that, in the calculation process of WOE code and IV value, if the instance in a subbox is only a bad customer or a good customer, py_i or pn_i may be 0, and WOE and IV values may correspond to inf or-inf, which is not allowed. In the calculation process of the WOE value in this section, a weight of $1.0\ E-12$ is added to py_i and pn_i, which ensures that the data will not be overfitted. Through verification, the change of WOE and IV values before and after feature term does not exceed $1.0E-12$, which shows the robustness of adding weights. The "NONE" classification of home_owner ship was originally WOE inf, which is now 20.23. The IV value of home_owner, originally inf, is now 0.07, which confirms the effectiveness of adding weights. Generally speaking, IV value reflects the information value of features, and too small IV value is not conducive to the prediction of models. In this section, variables with IV values greater than 0.02 are selected, and then, WOE values are converted for each group. Some results are shown in Table 7.

3.2. Selection of Machine Learning Methods and Models.

Traditional credit evaluation models are often based on machine learning algorithms, such as risk control models of banks and securities companies. This paper is based on logistic regression, random forest, support vector machine, and three machine learning methods for feature training.

3.2.1. Logistic Regression Algorithm. Logistic regression is an efficient classification model, which is widely used to solve the problems of binary classification and multiclassification. Logistic regression is based on linear regression and usually calls the sigmoid function to fix the output to a specified interval. The expression for the sigmoid function is shown in

$$g(z) = \frac{1}{1 + e^{-z}}. \tag{8}$$

Table 7: WOE value conversion results.

	Term	Grade	home_owership	verification_status	int_rate
29767	0.140083	0.493626	0.257450	0.459424	0.611640
4420	−0.408107	−0.140569	0.257450	−0.002818	−0.188107
15942	0.140083	−0.545455	0.257450	−0.002818	−0.546382
945	0.140083	−0.545455	−0.302803	−0.002818	−0.546382
20793	0.140083	−0.140569	0.257450	−0.002818	−0.188107
16772	0.140083	−0.545455	−0.302803	−0.002818	−0.546382
24601	0.140083	0.493626	−0.302803	0.459424	0.232332
28313	0.140083	−0.649242	−0.302803	−0.354990	−0.718143
11367	0.140083	0.493626	0.257450	−0.002818	0.232332
13022	−0.408107	−0.140569	0.257450	−0.354990	−0.188107

The sigmoid function maps the inputs within the interval $(-\infty, +\infty)$ to (0, 1). Using the sigmoid function to apply linear regression equation $y = \theta^T x + b$, where $x = (x_1, x_2, \ldots, x_n)$, the coefficients $\theta = (b, \theta_1, \theta_2, \ldots, \theta_n)$ can obtain the core expression of logistic regression algorithm as

$$g(z) = f\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}}. \tag{9}$$

Thus, the output of $g(z)$ represents the probability that a sample is predicted to be positive, and $1 - g(z)$ represents the probability that a sample is predicted to be negative. How to find a W which makes the probability of all predictions of $G(Z)$ correct is maximized, and the cost function can be optimized by the gradient descent method or the quasi-Newton method, so as to estimate.

The cost function of the logistic regression model during training is shown in

$$J(\theta) = -\frac{1}{m}\sum_{1}^{m} Cost\left(f\left(\theta^T x\right), Y\right). \tag{10}$$

Among them,

$$Cost\left(f\left(\theta^T x\right), Y\right) = \begin{cases} -\log\left(f\left(\theta^T x\right)\right), & Y = 1, \\ -\log\left(1 - f\left(\theta^T x\right)\right), & Y = 0. \end{cases} \tag{11}$$

Logistic regression algorithm has obvious advantages in dealing with large-scale data, and the calculation time is effectively saved by the gradient descent method. After data training, the expression with real parameters can be obtained, and the model is intuitive and reliable.

The machine learning experiments in this paper are all based on Jupyter Notebook platform, using python programming language. Logistic regression algorithm calls Logistic Regression CV function in scikit-learn machine learning library to train the data in logistic regression. For the hyperparameters involved in the model, Randomized Search CV function is used for random grid search to optimize the parameters

3.2.2. Random Forest Algorithm. Random forest algorithm has some applications in previous feature selection. This paper introduces the random forest classifier in scikit-learn and its related parameters. Random forest generally uses CART decision tree as weak learner. The general decision tree step is to select the best feature among all n sample features on the node to divide the left and right subtrees of the decision tree. The algorithm idea of random forest is to randomly select some features on nodes by putting them back and select the best features according to the requirements of the learner. The best features are used as the watershed of the left and right subtrees and cycle back and forth until all the best features are selected. The idea of randomly selecting nodes in random forest improves the generalization ability of the model. The steps of binary classification in random forest algorithm are as follows:

Step 1: input: dataset X and sampling times k; output: the decision result of dataset X after K times sampling and training

Step 2: bootstrap is used to carry out k-round sampling, and m features are extracted from dataset X as training samples, and finally, a sample set S is formed

Step 3: classifying and training different samples in the sample set S

Step 4: output: for discrete variables, *Gini* coefficient is used as the evaluation standard for random forest, and the calculation is carried out according to

$$Gini = 1 - \sum(P(i) * P(i)), \tag{12}$$

where $P(i)$ represents the proportion of class I samples in the total samples of the current node of the decision tree, and the more average the class, the greater the Gini coefficient.

For large-scale credit data, random forest algorithm has good classification effect because of sufficient sampling. Through the division of decision tree, the anti-overfitting ability of the model is improved. The algorithm itself is very stable, applicable to continuous variables and discrete variables, not easily affected by outliers, and has high reliability.

3.2.3. Support Vector Machine Algorithm. It belongs to a classical binary classification model. The basic model of SVM is a linear classifier with the largest interval in feature space. The basic idea of SVM algorithm [16, 17] is to solve a

Table 8: Confusion matrix.

	Predicted value = 1	Predicted value = 0
True value = 1	TP	FP
True value = 0	FP	TN

Table 9: Common thresholds for AUC.

AUC value	Classification result
>0.7	There is a strong degree of discrimination
(0.6, 0.7]	Have a certain degree of discrimination
(0.5, 0.6]	Have a weak degree of discrimination
≤0.5	The discrimination is weaker than random guess

hyperplane which can divide the training dataset correctly and maximize the geometric interval between the data.

For the nonlinear classification problem, the input can be mapped into a linear classification problem in a certain feature space by nonlinear transformation, and the linearly separable support vector machine model can be solved in the high-dimensional linear space. Considering the huge scale of credit data and the nonlinear relationship of the whole data, this section uses nonlinear support vector machine algorithm. There are two very important concepts in the computation of nonlinear support vector machines: kernel function and penalty coefficient C. Kernel function is essentially a mapping relationship. Suppose X is a low-dimensional input space and H is a high-dimensional eigenspace. If there is a mapping $\varphi(x)$: $X \longrightarrow H$ such that, for all $x, z \in X$ functions $K(X, Z)$, equation (13) is satisfied:

$$K(x, z) = \varphi(x)\varphi(z). \tag{13}$$

Then, $K(x, z)$ is called kernel function. The advantage of the kernel function is that the classification in high-dimensional space can be mapped to low-dimensional space for calculation, which avoids the ultra-high computational complexity. Commonly used kernel functions include positive definite kernel function, linear kernel function, polynomial kernel function, Gaussian kernel function, and sigmoid kernel function. Penalty coefficient C is a superparameter produced to adjust the accuracy of classification, which is related to the generalization ability of the model. The higher the penalty coefficient C is, the lower the tolerance of the model to classification error is, and the overfitting phenomenon is easy to occur. If C is too low, it is easy to underfit. Nonlinear support vector machines (NSVMs) have a good effect on binary classification problem, but the disadvantage is that even if linear kernel function is used, the efficiency of processing large-scale data is still low.

3.3. Model Performance Evaluation. Because of the imbalance of positive and negative samples, the evaluation criteria of the scorecard model cannot be judged simply by the accuracy of the model. In this section, AUC and K-S values are used to evaluate, and ROC curves are made to distinguish and compare the three algorithms on the scorecard model.

Here, we first introduce a concept: confusion matrix. The confusion matrix refers to the results of the statistical classification model, and the observed values are represented as matrices. Confusion can be used to represent the classification results of the model. For the binary classification problem, its confusion matrix is shown in Table 8:

The true rate TPR = TP/(TP + FN) represents the proportion of real positive samples assigned to positive samples to all positive samples, and the false positive rate FPR = FP/(FP + TN) represents the proportion of negative samples wrongly assigned to positive samples to the total number of all negative samples; Precision = TP/(TP + FP), which represents the classification accuracy of positive samples, and Recall = TPR = TP/(TP + FN), which has the same meaning as the true rate.

The ROC curve is also known as the receiver operation characteristic curve. This curve was used in the field of radar signal detection in the early stage to distinguish signal from noise. At present, the ROC curve is often used to evaluate the predictive ability of models. The ROC curve is based on the confusion matrix. The abscissa of the curve is the false positive rate (FPR) and the ordinate is the true rate (TPR).

AUC is the area under the ROC curve. The larger the AUC, the stronger the identification ability of the model for positive and negative samples and the better the classification result. The calculation method is the calculus value of the ROC curve. Common thresholds for AUC are shown in Table 9.

The K-S value is calculated by the formula KS = max (TPR-FPR), that is, the difference between the recall rate and the false positive rate. The K-S value represents the difference between positive and negative samples of the model. The larger the K-S value, the better the prediction accuracy of the model. Generally speaking, KS > 0.2 means that the model has good prediction accuracy. The judgment threshold of K-S is shown in Table 10.

In this section, the dataset is divided into the training set and the test set according to the ratio of 7 : 3. Training is carried out on the training set, and then, the prediction results are given on the test set. Through grid search and

Table 10: K-S judgment threshold.

K-S value	Model accuracy
>0.3	Good
(0.2, 0.3]	Available
(0, 0.2]	Poor
≤0	Model error

Table 11: Evaluation results of three algorithms on personal credit data.

Algorithm	Confusion matrix	Precision rate	Recall rate	AUC value	K-S value
Logistic regression	[[6534 450] [2239 518]]	0.5351	0.1879	0.719	0.303
RandomForest	[[6181 803] [1985 772]]	0.4902	0.2800	0.719	0.304
SVM	[[6286 716] [2039 718]]	0.5007	0.2604	0.713	0.308

Table 12: Evaluation results of three algorithms on personal credit data after data reset.

Algorithm	Confusion matrix	Precision rate	Recall rate	AUC value	K-S value
Logistic regression	[[15452 945] [4959 1372]]	0.5921	0.2167	0.756	0.357
Random forest	[[14721 1676] [4382 1949]]	0.5377	0.3179	0.749	0.347
SVM	[[14939 1458] [4531 1800]]	0.5524	0.2843	0.748	0.354

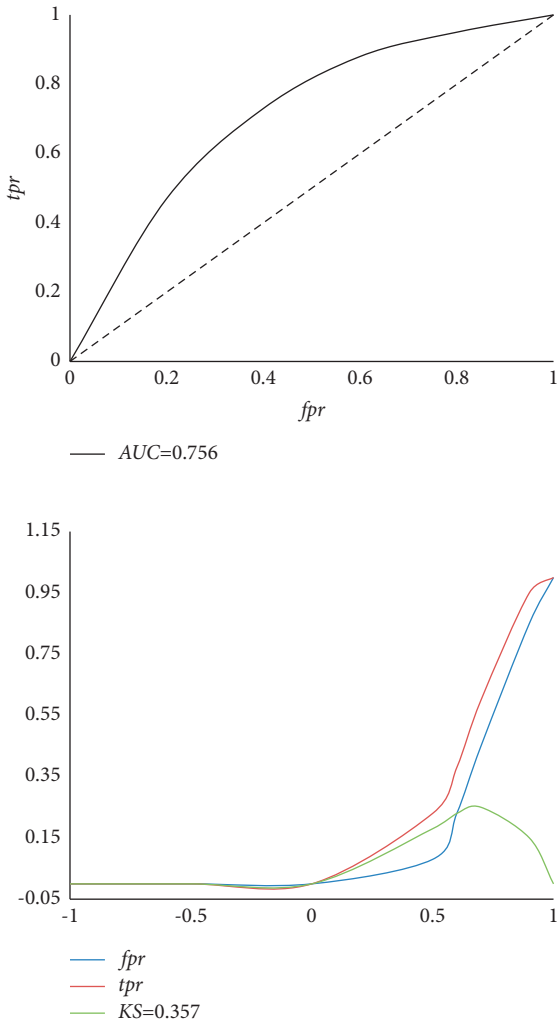

Figure 3: ROC curve and K-S curve of logistic regression.

parameter adjustment, this section obtains the evaluation results of logistic regression random forest and SVM algorithm, which are given in Table 11.

The data in Table 11 are the experimental data obtained after superparameter selection through random grid search. AUC and K-S indexes are within a reasonable range, but the discrimination of the model needs to be improved. Considering the imbalance of positive and negative samples in the confusion matrix, good customers are 7 times more than bad customers, which may cause prediction deviation. After resetting the experimental data, randomly select 20% samples from the whole sample to fill the data, and the experimental results are shown in Table 12 again.

It can be seen that the AUC and K-S values of the three models are improved at this time, and the logistic regression model performs best. For clearer comparison, ROC curves and K-S curves of logistic regression, random forest and SVM algorithms are shown in Figures 3–5 .

Finally, based on the above experiments, we consider the stability and generalization ability of the model and give the credit score card in the logistic regression model.

3.4. Credit Scorecard Based on Logistic Regression. Before the evaluation, we screened the features with IV value greater than 0.02, finally built a logistic regression model based on 22728 samples in the training set, and called statsmodels library in python library, when the significance level is 0.05; all the significance tests are passed, in which coef represents the coefficient of logistic regression, all of which are negative and significant.

Based on the above inspection, the scorecard can be set next. The score scale of the scorecard can be defined by the linear expression of the logarithm of default and nondefault probability ratio (log (odds)):

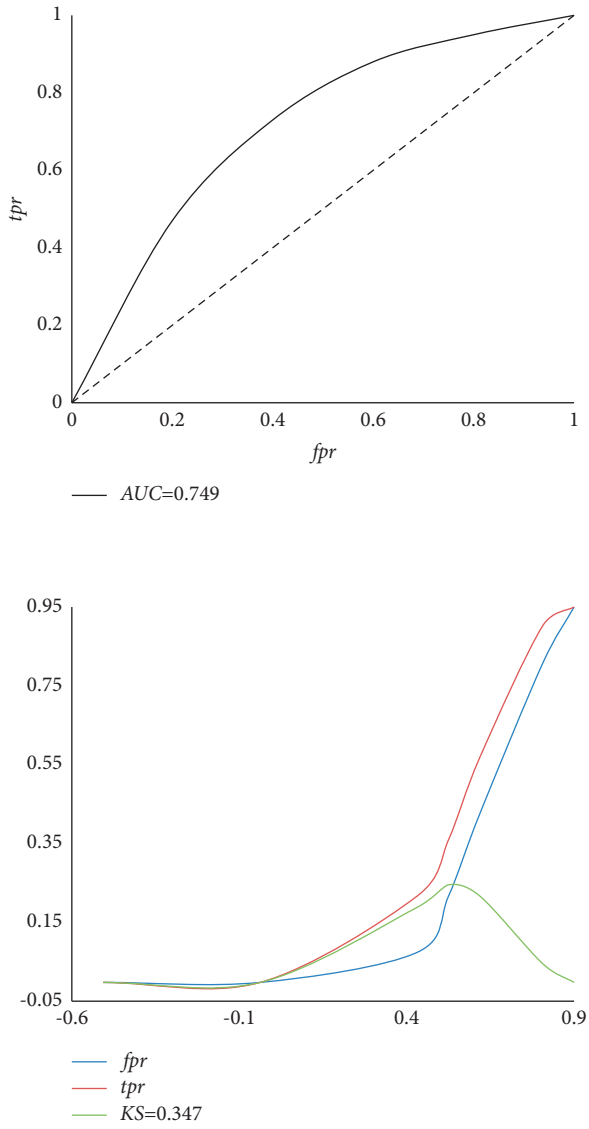

Figure 4: ROC curve and K-S curve of random forest.

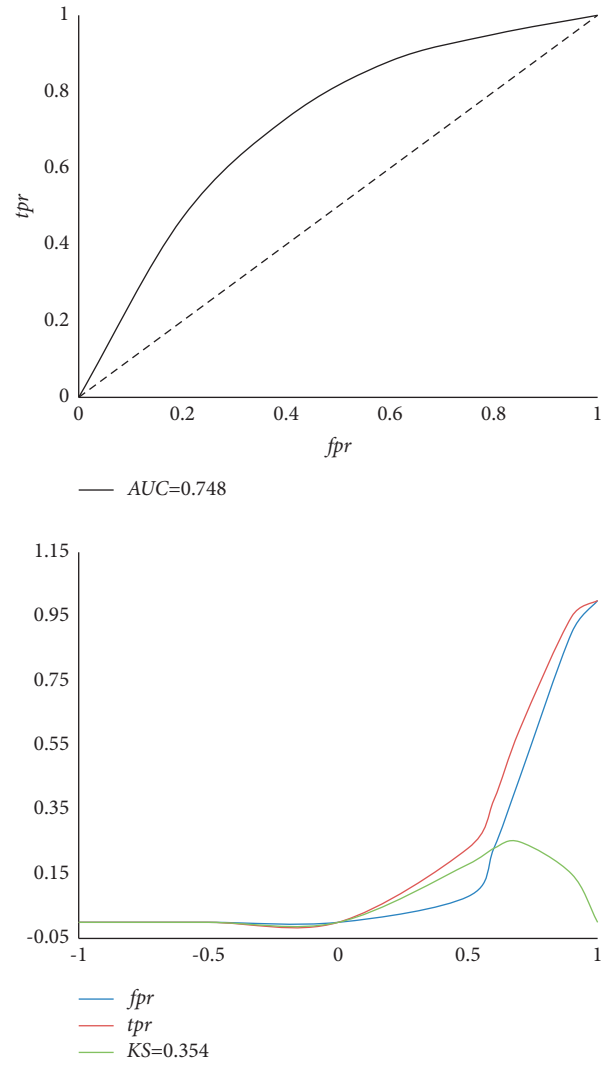

Figure 5: ROC curve and K-S curve of SVM.

$$\text{Score} = A - B\log(\text{odds}), \tag{14}$$

$$\text{odds} = \frac{\widehat{p}}{1-\widehat{p}}, \tag{15}$$

where the offset $A = \text{offset}$, the scale factor $B = \text{factor}$, and $\widehat{p}$ represents the default probability; the conversion process from log (odds) to score includes three commonly used parameters base: *bp*: (basepoint), *od*: odds corresponding to the reference score, *Po*: the score that the credit score increases when the odds doubles. The conversion formula between factor, offset, and *bp*, *od*, and *po* is

$$\text{factor} = \frac{po}{\log 2}, \tag{16}$$

$$\text{offset} = bp - \frac{po * \log od}{\log 2}. \tag{17}$$

Here, it is assumed that the score corresponding to odds 1/20 is 600 points, and the corresponding score is reduced by 60 points when odds is 2*od*, and $A = 481.86$, $B = 28.85$, and base score = 510. 0 are obtained by calculation. Based on the logistic regression model of the above optimal parameters and the setting of scorecard parameters, we get the basic scorecard of Lending Club and the scores corresponding to each sample, as shown in Tables 13 and 14 .

Based on the score of 510, samples below 510 are regarded as poor customers, and samples above 510 are regarded as good customers. According to the statistics based on data density, good customers are mainly concentrated in the interval [530, 570], while poor customers are concentrated in the interval [470–500], which accords with the distribution characteristics of the original data. Symmetrically, dividing the data according to the base score 510, we can obtain the following score grade (Table 15). Among the statistical users, the ratio of good customers to poor customers is close to 6 : 1, and only 8% of customers have scores lower than 450, which shows that there are very few customers with particularly poor credit, and most customers still have good credit performance.

Table 13: Basic scorecard part display.

No	Feature	Interval	Scores
0	Term	36 months	2.0
1	Term	36 months	−6.0
0	Grade	A	3.0
1	Grade	B	1.0
2	Grade	C	−0.0
3	Grade	D	−1.0
4	Grade	E	−2.0
5	Grade	F	−2.0
6	Grade	G	−2.0
0	Home_ownership	ANY	10
1	Home_ownership	MORTGAGE	5.0

Table 14: Lending Club final scorecard partial display.

num_rev_ti_bal_gt_0	Percebt_bc_gt75	tot_hi_cred_lim	income_vs_loan	Score
−3.0	1.0	1.0	−3.0	547.0
6.0	6.0	−2.0	−3.0	520.0
−7.0	6.0	2.0	−2.0	503.0
−3.0	−2.0	−3.0	−3.0	502.0
3.0	−2.0	3.0	3.0	548.0
−1.0	1.0	−3.0	−3.0	495.0
−3.0	−5.0	−3.0	1.0	516.0
3.0	6.0	−3.0	−2.0	524.0
−7.0	−5.0	3.0	1.0	526.0
6.0	6.0	3.0	−2.0	560.0
−3.0	6.0	3.0	−0.0	578.0
−7.0	−2.0	−2.0	3.0	533.0
−1.0	6.0	−2.0	−2.0	497.0

Table 15: Credit score rating scale.

Score	Credit rating	Possible credit risk
(570, 850)	AAA	Strong repayment ability and almost no risk
(530, 570)	AA	The repayment ability is satisfactory, and it may also be adversely affected by the external environment
(510, 530)	A	At present, there is enough ability to repay, but the deterioration of economic conditions may lead to the weakening of repayment ability
(490, 510)	B	There are risks, the current repayment ability is fragile, and it is easy to delay payment
(450, 490)	BB	The risk is high, and it is possible to default at present, so it is necessary to repay the debt with good external factors
(300, 450)	BBB	The risk is very high, there are cases of non-repayment on schedule or the current economic conditions do not support the loan

4. Conclusion

Focusing on the research goal of "providing scientific and reasonable credit evaluation to the wider people under the new situation," this paper establishes a series of complete credit evaluation systems including formula derivation, model building, and simulation analysis of the personal credit evaluation model according to advanced big data mining technology. In the research process, firstly, discrete variables are encoded to solve the problem that discrete variables and continuous variables cannot be considered uniformly. Then, aiming at the problem that high-dimensional data are easy to overfit, the data are reduced in dimension. Then, the data after dimensionality reduction is divided into data partitions, and the continuous variables and discrete variables are classified into partitions through variable discretization, and the WOE codes of each district are obtained. Then, the training of WOE datasets on the three models is carried out, respectively. After comparing the performance, it is found that logistic regression is more suitable for the personal credit evaluation model based on Lending Club datasets. Finally, based on the logistic regression model with the best parameters, the user samples are scored and the final scorecard is output.

Data Availability

The data 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.

References

[1] B. Tidke, R. G. Mehta, D. P. Rana, and H. Jangir, "Topic sensitive user clustering using sentiment score and similarity measures: big data and social network," *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), IGI Global*, vol. 15, no. 2, pp. 34–45, 2020.

[2] J. Jiang, L. Liao, X. Lu, Z. Wang, and H. Xiang, "Deciphering big data in consumer credit evaluation," *Journal of Empirical Finance*, vol. 62, no. 2, pp. 28–45, 2021.

[3] Óskarsdóttir, C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens, "The value of big data for credit scoring: enhancing financial inclusion using mobile phone data and social network analytics," *Applied Soft Computing*, vol. 74, pp. 26–39, 2019.

[4] S. S. Luvizan, P. T. Nascimento, and A. Yu, "Big data for innovation: the case of credit evaluation using mobile data analyzed by innovation ecosystem lens," in *Proceedings of the 2016 Portland International Conference on Management of Engineering and Technology (PICMET)*, pp. 925–936, Honolulu, HI, USA, September 2016.

[5] W. Wang, C. Lesner, A. Ran, M. Rukonic, J. Xue, and E. Shiu, "Using small business banking data for explainable credit risk scoring," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 8, pp. 13396–13401, 2020.

[6] V. Kuppili, D. Tripathi, and D. R. Edla, "Credit score classification using spiking extreme learning machine," *Computational Intelligence*, vol. 36, no. 2, pp. 402–426, 2020.

[7] N. Jain, T. Agrawal, P. Goyal, and V. Hassija, "A blockchain-based distributed network for secure credit scoring," in *Proceedings of the 2019 5th International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 306–312, IEEE, Solan, India, October 2020.

[8] C. Wang, D. Han, Q. Liu, and S. Luo, "A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM," *IEEE Access*, vol. 7, no. 99, p. 1, 2018.

[9] D. Chen, F. Lai, and Z. Lin, "A trust model for online peer-to-peer lending: a lender's perspective," *Information Technology and Management*, vol. 15, no. 4, pp. 239–254, 2014.

[10] A. Ashofteh and J. M. Bravo, "A conservative approach for online credit scoring," *Expert Systems with Applications*, vol. 176, p. 114835, 2021.

[11] J. Wu, Q. Zhang, Y. Zhao et al., "Radiomics analysis of iodine-based material decomposition images with dual-energy computed tomography imaging for preoperatively predicting microsatellite instability status in colorectal cancer," *Frontiers in oncology*, vol. 9, p. 1250, 2019.

[12] Z. Gao, L. Ding, Q. Xiong, Z. Gong, and C. Xiong, "Image compressive sensing reconstruction based on *z*-score Standardized group sparse representation," *IEEE Access*, vol. 7, no. 99, p. 1, 2019.

[13] L. Yuanqing, Y. Suying, X. Jiangtao, and G. Jing, "A self-checking approach for SEU/MBUs-hardened FSMs design based on the replication of one-hot code," *IEEE Transactions on Nuclear Science*, vol. 59, no. 5, pp. 2572–2579, 2012.

[14] J. Li, L. Ren, B. Wang, and G. Li, "Probabilistic load forecasting of adaptive multiple polynomial regression considering temperature scenario and dummy variables," *Journal of Physics: Conference Series*, vol. 1550, no. 3, pp. 1–16, 2020.

[15] L. Teng, H. Li, S. Yin, and Y. Sun, "A new chi-square distribution de-noising method for image encryption," *International Journal on Network Security*, vol. 21, no. 5, pp. 804–811, 2019.

[16] G. Chen, Q. Pei, and M. M. Kamruzzaman, "Remote sensing image quality evaluation based on deep support value learning networks," *Signal Processing: Image Communication*, vol. 83, p. 115783, 2020.

[17] G. Chen, X. Xie, and S. Li, "Research on complex classification algorithm of breast cancer chip based on SVM-RFE gene feature screening," *Complexity*, vol. 2020, Article ID 1342874, 12 pages, 2020.