Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5608392, 10 pages
https://doi.org/10.1155/2020/5608392

Research Article

Personal Credit Default Prediction Model Based on Convolution Neural Network

Xiang Zhou, Wenyu Zhang, and Yefeng Jiang

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, China

Correspondence should be addressed to Wenyu Zhang; zhangwenyu8518@126.com

Received 9 May 2020; Revised 10 September 2020; Accepted 18 September 2020; Published 5 October 2020

Academic Editor: Carlo Renno

Copyright © 2020 Xiang Zhou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

It has great significance for the healthy development of credit industry to control the credit default risk by using the information technology. For some traditional research about the credit default prediction model, more attention is paid to the model accuracy, while the business characteristics of the credit risk prevention are easy to be ignored. Meanwhile, to reduce the complicity of the model, the data features need be extracted manually, which will decrease the high-dimensional correlation among the analyzing data and then result in the low prediction performance of the model. So, in the paper, the CNN (convolutional neural network) is used to establish a personal credit default prediction model, and both ACC (accuracy) and AUC (the area under the ROC curve) are taken as the performance evaluation index of the model. Experimental results show the model ACC (accuracy) is above 95% and AUC (the area under the ROC curve) is above 99%, and the model performance is much better than the classical algorithm including the SVM (support vector machine), Bayes, and RF (random forest).

1. Introduction

With the improvement in people living quality and the change in lifestyle, the loan consumption has been gradually accepted by the public [1]. Meanwhile, compared with the bank loan, the personal loan mode tends to be chosen by more and more borrowers because of the shorter approval procedure and time. In addition, to stimulate the economic growth, many developed countries such as Germany, Switzerland, Sweden, and Japan have entered the era of negative interest rate, and then, the frequent drops in the bank interest rate make many investors regard the personal loan as a managing finance tool. Especially since the first P2P (peer-to-peer) web lending platform, ZOPA, was established in 2005, the personal loan industry has grown rapidly throughout the world [2]. Lending Club founded in 2014, one of the first listed companies for online lending matching, declares in its website that the loan consumption market scale has exceeded 3 trillion dollars in the world, and the investors will have more and more broad space [3]. In China, the personal credit business has also entered a rapid development period since 2011. From 2011 to 2017, the number of Chinese online loan platforms increased from about 60 (among which there were only less than 20 active platforms) to 2325. However, the prosperity of the personal loan market brings a serious credit default problem, the main reason of which lies in lending platforms that relax the loan audit conditions and ignore the potential risks, in order to enhance their market competitiveness and get more benefits, which may cause great economic losses for the investors and the lending platform. Therefore, how to better control the economic risk caused by the customer loan default is the key to the stable development of the loan platform and the credit industry. So the experts and scholars of various countries tried to establish various credit default prediction models to control the risks. Nevertheless, these models had still not formed a uniform standard of the performance evaluation owing to different research concerns. In addition, because the loan data dimension is large, most scholars usually extracted the data features before modeling, to reduce the model complexity. However, at the same time, this feature extracting method led to human factors having a great influence on the model objectivity, and then, the ability of model was decreased to find the high-

dimensional correlation among the analyzing data, and at the last, the model prediction performance would be not unsatisfactory. In this paper, four models are built by the methods including SVM (support vector machine), Bayes, RF (random forest), and CNN (convolutional neural network); moreover, both ACC (accuracy) and AUC (the area under the ROC curve) are taken as the performance evaluation index of the model. Experiments show, when the proper network structure and parameters are determined by the contrast tests, that the CNN model ACC is above 95% and AUC is above 99%. And aiming to the problem of the personal credit default prediction, because of the strong ability of self-learning and data feature autoextracted, the prediction performance of the CNN model is much better than that of the other three traditional modals such as SVM, Bayes, and RF.

2. Related Work

In order to guarantee the sustainable, stable, and healthy development of the credit industry, the experts and scholars of various countries have studied the credit default prediction problem from different aspects. Based on Lending Club dataset, the paper [4] builds some models to distinguish which feature is important to predict the loan default and which kinds of borrowers may pay debts with interest on time. Moreover, when the ACC is only used as the model performance index, the paper finds that the RF model is the most suitable classifier to identify which borrowers may break a loan contract, and the DT (decision tree) model is the best choice to discriminate which customers may have the good credit. The paper [5] extracts firstly the features of the credit dataset using RBMs (restricted Boltzmann machines) and then establishes a LDA (linear discriminant analysis) prediction model. The contrast experiments show that ACC of the LDA model on the dataset of German credit is better than that of some models such as LR (logistic regression), ANN (artificial neural network), SVM, and RF, but it is only 76.5% which cannot meet the requirements of practical application. On the Lending Club loan dataset, the paper [6] builds a RF model for borrower status prediction, and the ACC of RF is 87%. The paper [7] proposes a credit scoring model using ANN, which classifies personal loan applications into default and nondefault groups. And the results show that the model can screen effectively those default applications. The above articles have done a lot of research in the field of the credit default prediction, but the performance of these prediction models is dissatisfactory. Moreover, two key problems are not considered in these research studies: one is that, besides ACC, generalization ability of the prediction model is also important, which can make the trained model keep stable on the unknown data; the other is that the manual features extraction for the dataset will reduce the high-dimensional correlation among the analyzing data, which may result in model performance degradation. This paper tries to establish a personal credit default prediction model using CNN, and in the model, not only both ACC and the stability can be improved but also artificial influence can be reduced.

3. Data Treating

In this study, the Lending Club loan dataset is used for modeling and testing, which includes 75 features such as current status, latest payment information, credit score, number of financial queries, and address. Some no-value data for modeling are deleted, which includes these records with incomplete current loan process and with seriously missing characteristic. The deleting ratio is 3.7%. Finally, 30000 relatively complete loan records are reserved as the initial dataset, which are divided into the default data group (25568 records) and nondefault data group (4432 records) according to the repayment status shown in Table 1. Although, these overdue (31–120 days) records should be regarded as the default data based on the Basel concordat, the paper focuses on prediction for the actual financial loss owing to the complete nonpayment of the borrowers, so these records are still in the nondefault data group.

Observing the records in the initial dataset, some problems are found including that some features of the record are missing and some features are not quantified, such as age and location, which will seriously affect the analysis process and prediction performance of the model, so in this study, these dirty data are firstly preprocessed by some suitable methods including that the missing feature value is filled according to the context and some features are quantified on direct coding. In theory, to denoise and balance, the data are helpful to improve the accuracy and the generalization ability of the model, while the study about the loan default prediction focuses more on the classification ability for small set samples.

In this paper, PCA (principal component analysis) method is used to extract the key information of the initial dataset and form a new extraction-feature dataset, and then, the influences on the modeling can be compared from the feature-extracted dataset and initial dataset. In the PCA method, the original data with some correlativity are recombined to form the small number of noncorrelation comprehensive data which not only contain most information of the original data but also better explain the economic implications because of the noncorrelation among the data. To determine whether the initial dataset is suitable for PCA, KMO (Kaiser–Meyer–Olkin) test and Bartlett's test are done. And the value of the KMO test is 0.778, which means the records of the initial dataset have strong correlation, and the value of Bartlett's test is 0.00, which shows the feature vectors of the initial dataset are not fully independent, so it is concluded that good results for descending dimension and feature extraction are got by using the PCA method on the initial dataset. In the paper, 26 principal components are got, which is described as follows:

$$\begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_m \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & \cdots & \cdots & a_{1n} \\ \vdots & \ddots & & & \vdots \\ \vdots & & a_{ij} & & \vdots \\ \vdots & & & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ xj \\ \vdots \\ x_n \end{bmatrix}, \quad (1)$$

Table 1: Classification table of default data and nondefault.

Loan status	Classification
Overdue (31–120 days)	Nondefault data
Overdue (16–30 days)	
In grace period	
Fully paid	
Charged off	Default data

where F_i is a principal component vector including 30000 records, $m = 26$ is the number of all principal components, x_j is the feature vector of the initial dataset, and $n = 75$ is the number of the features. $a_i = (a_{i1}, a_{i2}, \ldots, a_{in})^T$ is the ith eigenvector of the covariance matrix ($C = (1/k)\sum_{j=1}^{k} x^{(j)}x^{(j)^T}$, and $k = 30000$ is the number of the records) of the initial dataset, and the corresponding eigenvalue is written as λ_i. And H_i, the variance contribution rate of F_i, can be obtained as

$$H_i = \frac{\lambda_i}{\sum_{i=1}^{m}\lambda_i} \cdot 100\%. \tag{2}$$

The main result of principal component analysis is shown in Table 2.

In the Table 2, the first eight eigenvalues of the principal components are greater than 1, and these components contain 69% information of the initial dataset, so they are combined as a new dataset named as extraction-feature sample and the initial dataset named as full-feature sample. Both of these two samples are separately divided into the training set including 70% records and the test set including 30% records, and they are separately used to build models to compare the influences on the performance of the models.

4. Traditional Algorithm

In this subsection, we compare the predictive ability of three traditional models including SVM model, Bayes model, and RF model on loan data and carry out a research on the effect of extraction feature manually with the traditional model. Thus, we establish three computational models based on extraction-feature sample and full-feature sample. And the ACC and AUC are selected as evaluation indexes to judge the predictive performance of these models. The formula is as follows:

$$\mathrm{ACC} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN})}. \tag{3}$$

TP is true positive. FP shows false negative. TN indicates true negative. And FN reflects false negative. The system errors and statistical biases are marked as ACC. The AUC value generally represents the general predictive ability of the model. Usually, the AUC value is between 0.5 and 1.0, and the larger AUC represents a better performance [8]. The formula is defined as

$$\mathrm{AUC} = \frac{\sum_{i\in \mathrm{PositiveClass}} \mathrm{rank}_i - (M(1-M)/2)}{MN}, \tag{4}$$

Table 2: Table of total variance analysis.

Component	λ_i	H_i (%)	$\sum_i^m H_i$ (%)
F_1	7.527	28.952	28.952
F_2	2.373	9.128	38.080
F_3	2.008	7.725	45.804
F_4	1.452	5.584	51.389
F_5	1.241	4.774	56.162
F_6	1.186	4.562	60.725
F_7	1.148	4.417	65.141
F_8	1.005	3.864	69.005
F_9	0.972	3.740	72.745
...	...	...	...
F_{26}	0.002	0.008	100.000

where M is the number of positive samples and N is the number of negative samples. Rank represents the ordered set of probability values generated by the algorithm to classify a sample and $\mathrm{rank_i}$ represents the position of the ith sample in the rank. Besides, the ROC (receiver operating characteristic) curve combines sensitivity and specificity with the graphical method, which accurately reflects the relationship between specificity and sensitivity of the analysis method.

4.1. SVM Prediction Model. SVM model is a statistical learning method, which comes from the problem of optimal classification and separates the two classes of samples by using a hyperplane and achieves the goal of maximizing the classification gap [9]. In this work, we establish the SVM model based on extraction-feature sample and full-feature sample. For input dataset T, $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$, $x_i \in R^n$, $y_i \in \{0, 1\}$, where x_i is the loan data eigenvector of the ith sample and y_i indicates marking class of the ith sample. Due to the formula characteristics of the SVM model, the value range of y_i needs to be changed to $\{-1, 1\}$. When $y_i = +1$, the state of the sample is judged as default; when $y_i = -1$, the state of the sample is judged as nondefault. For any point (x_i, y_i) in the sample space and divisive hyperplane, $\omega^T \cdot x + b = 0$, where $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ is the normal vector including the weight of each eigenvalue, which decides the direction of a divisive hyperplane, and b indicates a constant presented displacement, which is the distance between the divisive hyperplane and origin. The distance between any point in the sample space and the divisive hyperplane can be expressed as

$$d_i = \left| \frac{\omega^T}{||\omega||} \cdot x_i + \frac{b}{||\omega||} \right|. \tag{5}$$

The model classifies the results of samples are denoted as

$$\begin{cases} \dfrac{\omega^T \cdot x_i + b}{||\omega||} \geq +1, & y_i = +1, \\ \dfrac{\omega^T \cdot x_i + b}{||\omega||} \leq -1, & y_i = -1. \end{cases} \tag{6}$$

The samples which make the equality established are called support vectors. Randomly taking two different types of support vectors, the sum of distance between them and hyperplane is r = (2/||ω||). The SVM model for solving the maximal segmentation hyperplane problem can be expressed as the optimal constraint:

$$\max_{\omega,b} \quad r$$
$$\text{s.t.} \quad y_i\left(\frac{\omega^T}{\|\omega\|}\cdot x_i + \frac{b}{\|\omega\|}\right) \geq r, \quad i = 1, 2, \ldots, N. \tag{7}$$

We make both sides of equation (7) divided by r. Because r and $||\omega||$ are scalar, we make $\omega = (\omega/||\omega||r)$ and $b = (b/||\omega||r)$ to simplify the formula. Equation (7) maximal is equivalent to make $(1/2)\|\omega\|^2$ minimal. So, equation (7) can be translated to

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2$$
$$\text{s.t.}\, y_i\left(\omega^T \cdot x_i + b\right) \geq 1, \quad i = 1, 2, \ldots, N. \tag{8}$$

Because equation (8) is the convex quadratic programming problem, it can be solved by adding a Lagrange multiplier $a_i \geq 0$. The Lagrange function is presented as follows:

$$L(\omega, b, a) = \frac{1}{2}\|\omega\|^2 + \sum_{i=1}^{N} a_i\left(1 - y_i\left(\omega^T \cdot x_i + b\right)\right). \tag{9}$$

It can get ω and b by solving equation (9). And we obtained the decision functions of the SVM model:

$$f(x) = \text{sign}\left(\omega^T \cdot x + b\right) = \text{sign}\left(\sum_{i=1}^{N} a_i y_i x_i^T x + b\right). \tag{10}$$

Moreover, the dimension of loan data is large, and the data space is more complex and not easy to split. In order to solve this problem, the polynomial kernel function and soft boundary are introduced. The formula of polynomial kernel function is as follows:

$$\varnothing(x) = \text{k}(x, x_i) = (x x_i + 1)^p. \tag{11}$$

So the decision functions of the SVM model can be translated to

$$\begin{aligned} f(x) &= \text{sign}\left(\omega^T \cdot \varnothing(x) + b\right) \\ &= \text{sign}\left(\sum_{i=1}^{N} a_i y_i \varnothing(x_i)^T \varnothing(x) + b\right) \\ &= \text{sign}\left(\sum_{i=1}^{N} a_i y_i k(x, x_i) + b\right). \end{aligned} \tag{12}$$

We add a penalty factor to the constraint and set its value to 10^2. The ROC curve contrast chart with two samples of the SVM model is obtained by the experiment shown in Figure 1.

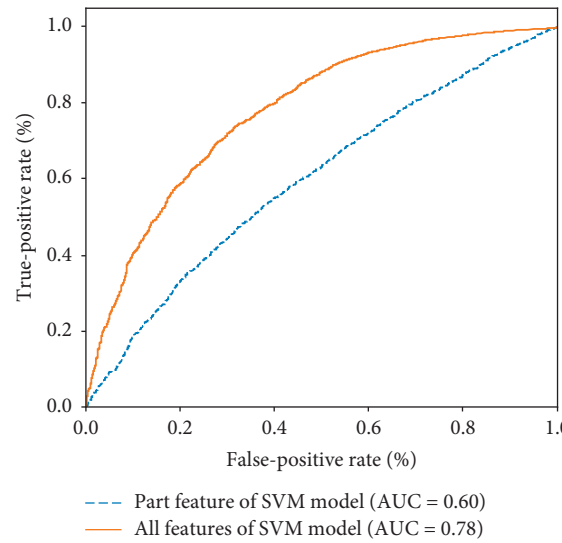

Figure 1: Comparison chart of ROC curve for different samples on SVM model.

It can be seen from the ROC curve comparison chart on SVM model that the AUC value of the SVM model with extraction-feature sample is 60%, which is 18% lower than that of 78% with full-feature sample.

In addition, it is found in Figure 2 that the ACC value of the SVM model with extraction-feature sample is 85%, in which the ACC value is lower than that of 88% with full-feature sample. The experimental results show that the ACC value and AUC value of the SVM model with extraction-feature sample is lower than that with full-feature sample. In addition, the SVM model based on extraction-feature sample is 296 fewer rightly judgments in default data than that based on the full-feature sample. In conclusion, the high-dimensional correlation is easy ignored, when the manual extraction-feature method is used in the SVM model, which affects the prediction performance of the model especially on default data.

4.2. Bayes Prediction Model. Bayes model is a classification model based on statistics which uses the prior probability to gain the posterior probability of one category and to judge what kind of data [10]. For input dataset T, $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}$, $x_i \in R^n$, $y_i \in \{0, 1\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$ is loan data eigenvector of the ith sample and y_i shows marking class of the ith sample; when $y_i = 0$, the state of the sample is judged as default; when $y_i = 1$, the state of the sample is judged as nondefault. The formula of Bayes model is defined as

$$\text{P}(y_i|x_i) = \frac{P(y_i)P(x_i|y_i)}{P(x_i)} = \frac{P(y_i)P(x_i|y_i)}{\sum_i^N P(x_i|y_i)P(y_i)}. \tag{13}$$

$\text{P}(y_i)$ is prior probability of y_i. $\text{P}(x_i|y_i)$ reflects the probability of x_i after y_i happened. N shows the number of data. Since the Bayes algorithm makes assumptions that the

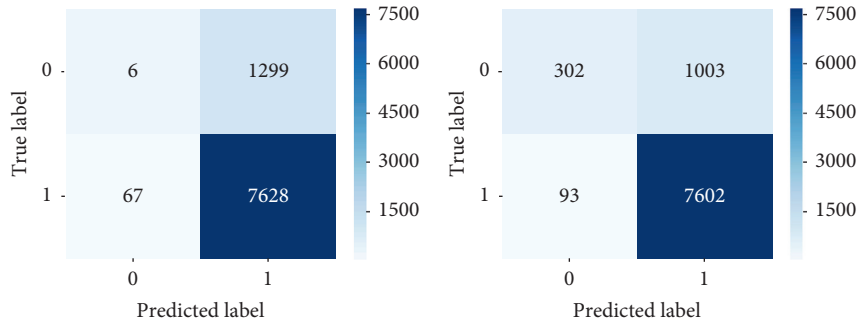

Figure 2: Comparison chart of confusion matrix on SVM model-based test set: (a) extraction-feature sample on SVM model; (b) full-feature sample on SVM model.

conditions are relatively independent, it can get the following:

$$P(x_i \mid y_i) = \prod_j^m P(x_{ij} \mid y_i), \quad j = 1, 2, \ldots, m, \tag{14}$$

where x_{ij} represents the jth feature in the ith sample. M is the number of features. So equation (13) can be translated to

$$P(y_i \mid x_i) = \frac{P(y_i)\prod_j^m P(x_{ij} \mid y_i)}{\sum_i^N P(y_i)\prod_j^m P(x_{ij} \mid y_i)} = P(x_i)\prod_j^m P(x_{ij} \mid y_i), \quad j = 1, 2, \ldots, m. \tag{15}$$

A category of data which has the highest probability is the ultimate classification of the data. So the Bayes classifier decision function is as follows:

$$f(x) = \arg\max_{y_i} p(x_i) \prod_j^N P(x_i^{(j)} \mid y_i), \quad j = 1, 2, \ldots, N. \tag{16}$$

Besides, there are 15% default data in the dataset. So the prior probability of default data is set as 0.15, and the prior probability of nondefault data is set as 0.85 in this work. The ROC curve comparison between two samples (extraction-feature sample and full-feature sample) of the Bayes model was obtained by experiment.

Figure 3 shows the AUC of the Bayes model based on extraction-feature sample is 60%. However, the AUC of the Bayes model with full-feature sample is 80%, which is 20% higher than that with extraction-feature sample.

Figure 4 shows ACC value of 88% based on extraction-feature sample, which is 6% higher than the ACC value of 82% with full-feature sample on Bayes model. The ACC value is calculated based on the better truncation value, which determines the cutoff value between positive and negative examples. Compared with the ACC value, the AUC value can synthesize the prediction performance of all truncation values [11]. Furthermore, the loan data are a kind

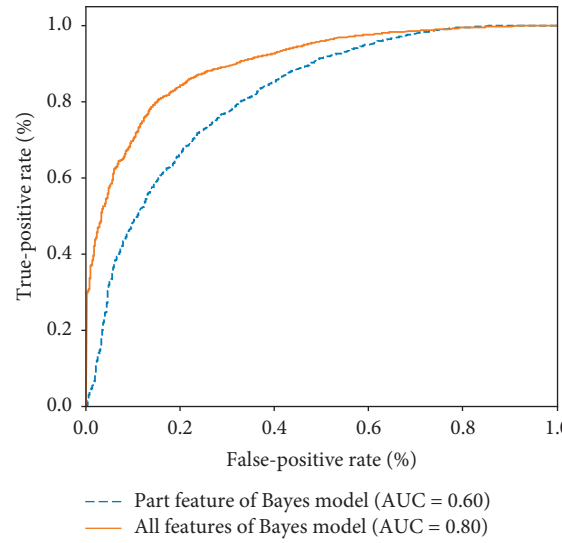

Figure 3: Comparison chart of ROC curve for different samples on Bayes model.

of skewed data. It influences not only the prior probability of nondefault data higher than default data but also the ACC value which is easily affected by large cardinality categories. Therefore, we priorly consider the AUC value, when the AUC value and the ACC value have different results in comparing the model performance. Moreover, the number of rightly judgments based on full-feature sample is more than that based on extraction-feature sample in default data. To sum up, estimated ability of the Bayes model with full-feature sample is better than with extraction-feature sample in the prediction of personal loan.

4.3. RF Prediction Model. RF model is an integrated machine learning model, which establishes multiple decision trees and synthesizes the result of each tree to obtain the final classification results [12, 13]. 100 subsets $(T_1, T_2, \ldots, T_{100})$ are randomly selected with replacement from input dataset T to establish 100 decision trees in the RF model, whose maximum sample is set 5000 and the maximum depth is set to 5. The Gini index was used in this paper to judge the

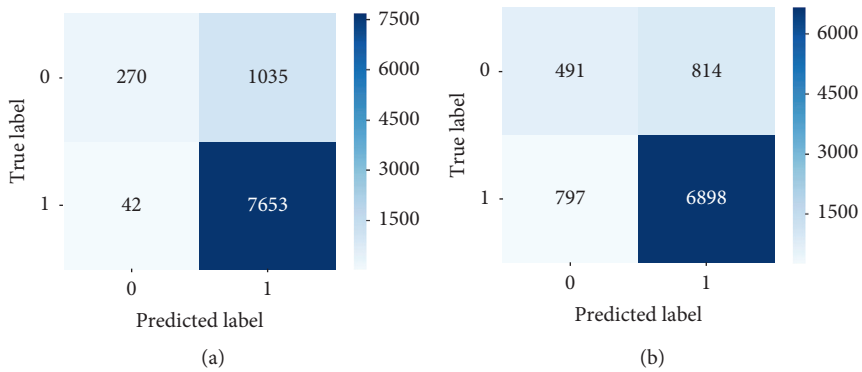

Figure 4: Comparison chart of confusion matrix on the Bayes model-based testing dataset: (a) extraction-feature sample on Bayes model; (b) full-feature sample on Bayes model.

optimal classification feature. The formula of Gini index can be expressed as

$$\text{Gini}(T) = \sum_{k=1}^{n} p^k\left(1 - p^k\right), \quad k = 1, 2, \ldots, N, \tag{17}$$

where p^k is the probability of the kth class and n is the number of categories in dataset T. So, after classification by a certain eigenvalue t, the Gini index is denoted as

$$\text{Gini}(T, t) = \sum_{k=1}^{n} \frac{\left|T^k\right|}{T} \text{Gini}\left(T^k\right). \tag{18}$$

The smallest $\text{Gini}(T, t)$ was selected as the segmentation point for classification.

As can be seen from Figure 5, the AUC value of 88% with extraction-feature sample is 6% lower than that of 94% with full-feature sample.

Besides, Figure 6 shows the ACC value of 88% based on extraction-feature sample is 3% lower than that of 91% with full-feature sample. And the RF model with extraction-feature sample has worse performance when predicting default data.

To sum up, the ACC value and AUC value of the RF model with extraction-feature sample are less than that with full-feature sample. It shows that manual extraction feature easily reduces the high-dimensional correlation in data, which affects the prediction performance of model. In addition, compared with three traditional models (SVM model, Bayes model, and RF model) in the two kinds of sample (extraction-feature sample and full-feature sample), the ACC value and AUC value of the RF model are higher than that of SVM model and Bayes model. Moreover, only when the RF model based on full-feature sample is used, both the ACC value and AUC value are above 90%. From the above studies, we find out the RF model has better prediction performance than SVM model and Bayes model in loan default. Unfortunately, the ACC value of three traditional models based on the two kinds of sample do not satisfy practical requirement of prediction effect in loan default.

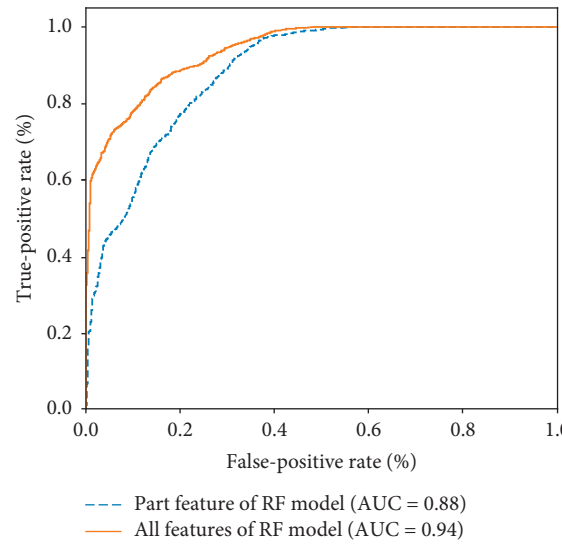

Figure 5: Comparison chart of ROC curve for different samples on RF model.

5. Personal Credit Prediction Model of CNN Algorithm

Because the lower ACC value and AUC value with three traditional models cannot satisfy the demand of practical application of loan risk prevention and control and the manual extracting feature influencing the performance of prediction model, we choose a CNN model to predict the loan default in order to obtain better the performance. CNN is a kind of feedforward neural network with convolutional computation and deep structure, which is one of the representative algorithms of deep learning. It is often composed of multilayer structure including convolution layer, pooling layer, activation layer, and full connection layer and adopts the gradient descent method which adjusts the parameters by a larger number of iterative training. The convolutional layer is the core part of the CNN model, which has the characteristics of local connection and weight-sharing [14]. In addition, CNN model can learn and map the relationship between input and output pairs autonomously from a large

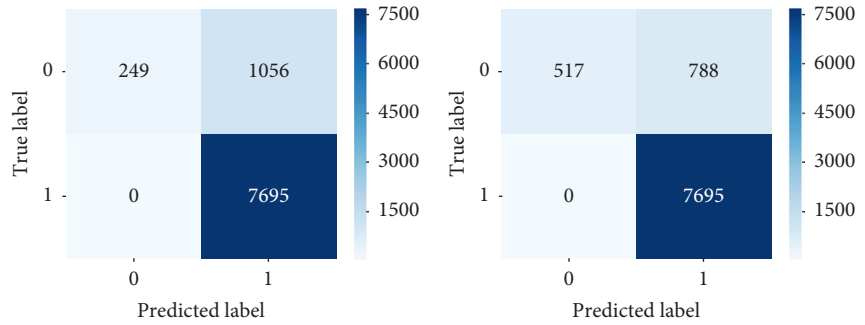

Figure 6: Comparison chart of confusion matrix on RF model-based testing dataset: (a) extraction-feature sample on RF model; (b) full-feature sample on RF model.

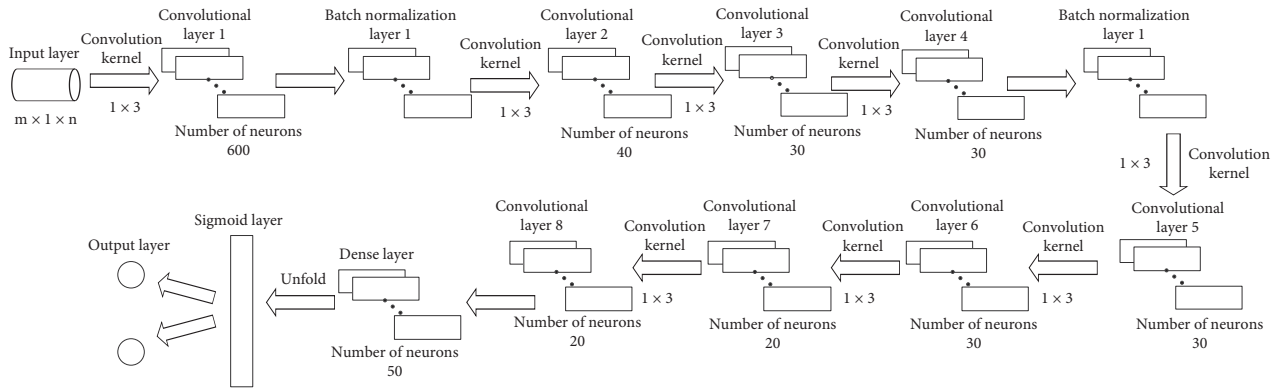

Figure 7: Structure chart of CNN model.

number of known data pairs without any precise mathematical expression between the input and the output.

Because the CNN model can extract the features autonomously, we only use full-feature sample to establish the 1D-CNN (1-dimensional convolutional neural network) model in this paper. Besides, due to not having clear standards at present, the model parameter and structure can only be determined through a series of repeated tests by the results of test dataset. Final, we establish a 1D-CNN model, which is composed of eight convolutional layers, through a lot of experiments. Figure 7 reflects the structure chart of CNN model.

We use $T = \{x_1, x_2, \ldots, x_n\}$ to denote the input data corresponding to feature vector of loan data. The form of T is $M \times 1 \times n$, where M represents the number of data and n represents the dimension of data. Every neuron in the convolutional layer has a small receiving domain for the input matrix, named as the convolution kernel, which obtains the output by convolution with a linear filter h_k. h_k from a window $x_{t:\, t+k-1}$ of the input vector is generated by a weight-sharing kernel tensor ω_k and a bias vector b_k:

$$h_k = \mathrm{conv}\left(u_{t:\, t+k-1}\right)\omega_k^T u_{t:\, t+k-1} + b_k, \tag{19}$$

where k is the kernel size. The calculation process of the CNN model is shown in Figure 8.

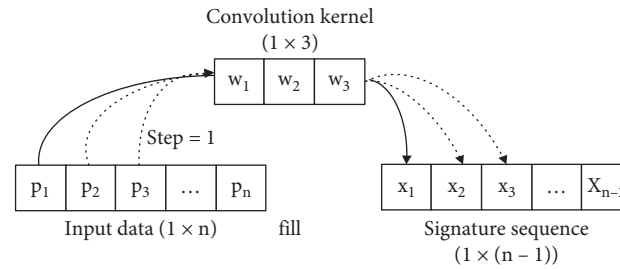

Figure 8: Kernel calculation process chart of CNN model.

The method of numerical calculation in feature sequence is denoted as

$$x_i = p_i \times w_1 + p_{i+1} \times w_2 + p_{i+2} \times w_3, \quad i = 1, 2, \ldots, n. \tag{20}$$

So the output of neuron is as follows:

$$f\left(h_k\right) = \sigma\left(h_k\right). \tag{21}$$

$\sigma$is the activation function of a neuron. If there is more than one channel in the input layer, the sum of the output after filtering is used as the output of the neuron. This paper uses the ReLU function as the activation function.

The formula is as follows:

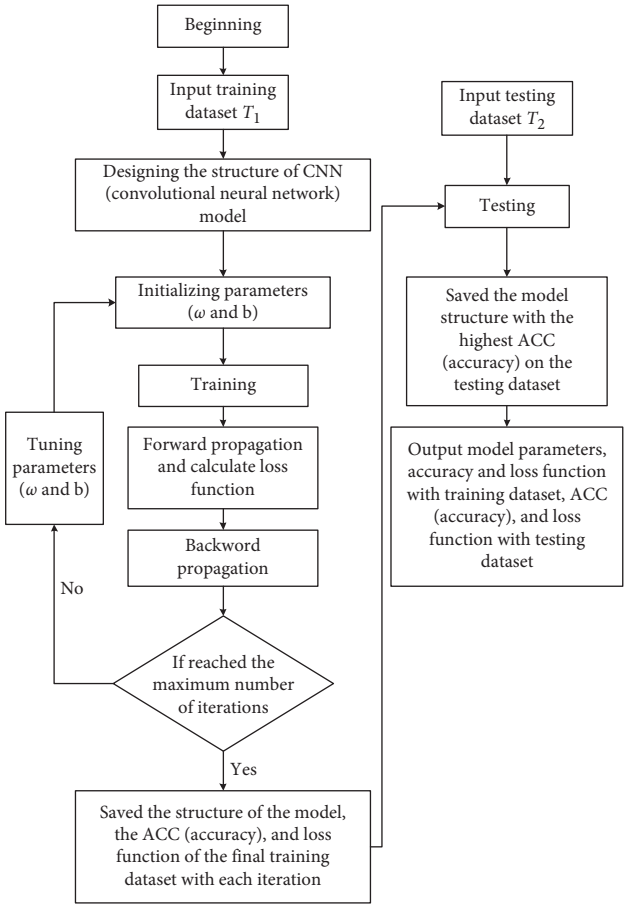

Figure 9: Workflow chart of CNN model.

$$\sigma(h_k) = \max\left(0, \omega_k^T x + b_k\right). \tag{22}$$

X is the input value. When the input x is less than 0, the output is 0. When x is greater than 0, the output is X. Because the ReLU function is a kind of unsaturated function, it effectively counters problems liking gradient disappearance and gradient explosion compared with the sigmoid function and makes the network converge more quickly. Before the output layer, multiple feature sequences shall be transposed and merged into the form of one-dimensional column vectors. Thus, the sigmoid layer users these vectors with the full connection layer. Finally, the output x_i of multiple neurons maps to between 0 and 1 by sigmoid. Thereby, we can get the final output a_i. The sigmoid function calculation formula is defined as

$$a_i = \frac{1}{1 - e^{x_i}}. \tag{23}$$

The mean square error is used as the loss function. We count the loss function as the following equation:

$$L = \frac{1}{n} \sum_{i=1}^{n} \left(O^{\text{real}} - O^{\text{outPut}}\right)^2. \tag{24}$$

N is the number of samples. O^{real} is the real vector, and O^{output} is the output vector. The learning rate of the convolutional neural network is set to 0.01, while the maximum number of iterations $E = 250$. Figure 9 reflects the flow chart of the CNN model.

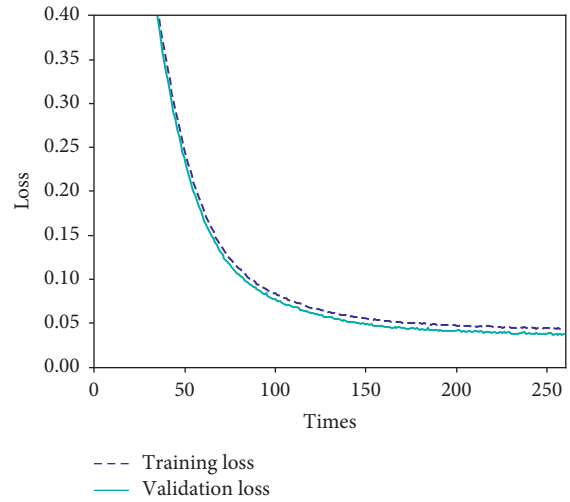

Figure 10: Loss function curve of CNN model.

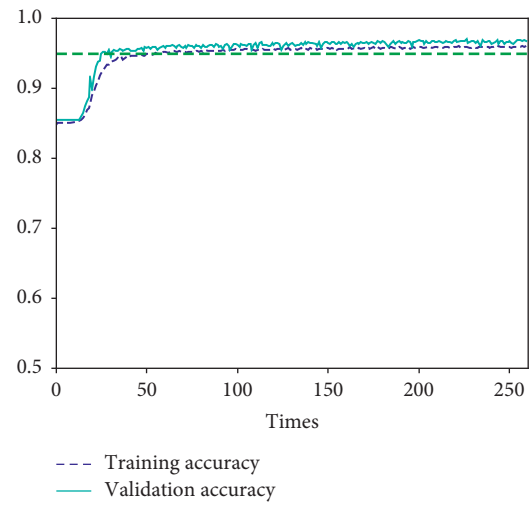

Figure 11: ACC chart of CNN model.

In addition, BN (batch normalization) layer is used to normalize arbitrary data in the construction of CNN model. Different from the traditional normalization methods only using input data, the BN layer realizes data normalization at any layer in the whole network. It not only accelerates the convergence speed of the CNN model but also relieves the gradient dispersion problem in the deep network. And it makes training of the CNN model easily and stably.

As shown in Figure 10, the loss function of the CNN model tends to be flat after 200 generations, which shows that the model can converge effectively.

As shown in Figure 11, after 100 generations, CNN model gets ACC value of 95% in training and testing which is higher than that of SVM model, Bayes model, and RF model. Moreover, according to Figure 12, the ACC value of the CNN model reaches up to 97% and only 289 wrong judgments in default data. In amount, the CNN model is good at predicting default data.

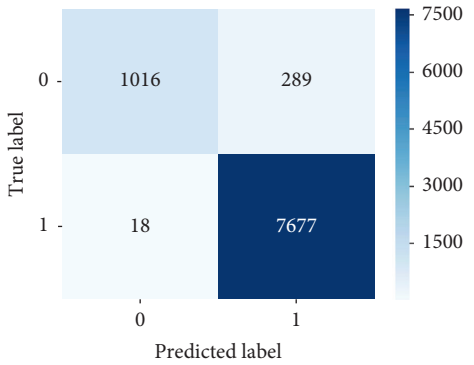

Figure 12: Chart of confusion matrix of CNN model-based testing dataset.

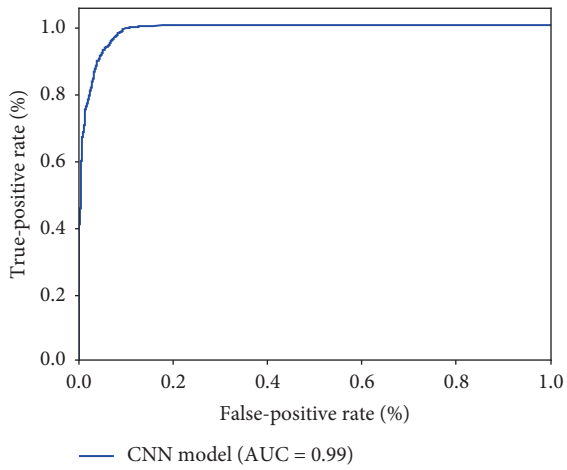

Figure 13: ROC curve chart of CNN model.

In addition, it can be seen from Figure 13 that the AUC value of the CNN model is up to 99%, whose AUC value is also higher than SVM model, Bayes model, and RF model. It shows that the generalization ability of the CNN model is more excellent, and its data prediction is more accurate.

6. Conclusions

It is important to the healthy development of credit industry by controlling the credit default risk. However, the ability of traditional models does not meet the practical application requirements. In this work, we establish a CNN model to predict default loan and compare with three traditional models including SVM model, Bayes model, and RF model based on extraction-feature sample and full-feature sample. The experimental result shows that the RF model has better performance than SVM model and Bayes model, whose ACC value and AUC value with extraction-feature sample and ACC value and AUC value with full-feature sample are as follows: 88%, 88%, 91%, and 94%. Besides, the CNN model has highest ACC value of 95% and AUC value of 99%. It indicates that the CNN model is superior to the three traditional models based on two samples. In addition, the extracting feature manually is often used in the traditional model. However, it decreases the objectivity and high-dimensional correlation, which affects the performance of prediction model, by comparing extraction-feature sample with full-feature sample on the same model. The CNN model effectively solves the problem cause by extracting feature manually through the autonomous extracting feature. To sum up, the CNN model has better performance and is more suitable for default prediction of personal loan data.

Abbreviations

P2P: Peer-to-peer
SVM: Support vector machine
RF: Random forest
CNN: Convolutional neural network
ACC: Accuracy
AUC: The area under the ROC curve
DT: Decision tree
RBMs: Restricted Boltzmann machines
LDA: Linear discriminant analysis
LR: Logistic regression
ANN: Artificial neural network
PCA: Principal component analysis
KMO: Kaiser–Meyer–Olkin
ROC: Receiver operating characteristic
1D-CNN: 1-dimensional convolutional neural network
BN: Batch normalization.

Data Availability

Some or all data, models, or code generated or used during the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of Liaoning Province of China (no. 20180551011).

References

[1] T. Wang and J. Li, "An improved support vector machine and its application in P2P lending personal credit scoring," *Institute of Physics Conference Series: Materials Science and Engineering*, vol. 490, no. 6, Article ID 062041, 2019.

[2] L. E. B. Ferreira, J. P. Barddal, H. M. Gomes et al., "Improving credit risk prediction in online peer-to-peer (P2P) lending using imbalanced learning techniques," in *Proceedings of the International Conference on Tools with Artificial Intelligence*, IEEE Computer Society, Boston, MA, USA, pp. 175–181, November 2017.

[3] H. Liu, H. Qiao, S. Wang, and Y. Li, "Platform competition in peer-to-peer lending considering risk control ability," *European Journal of Operational Research*, vol. 274, no. 1, pp. 280–290, 2019.

[4] L. Vinod Kumar, S. Natarajan, S. Keerthana et al., "Credit risk analysis in peer-to-peer lending system," in *Proceedings of the 2016 IEEE International Conference on Knowledge Engineering*

and Applications (ICKEA), pp. 193–196, Singapore, Singapore, September 2016.

[5] V. Ha, D. Lu, G. S. Choi et al., "Improving credit risk prediction in online peer-to-peer (p2p) lending using feature selection with deep learning," in *Proceedings of the 2019 21st International Conference on Advanced Communication Technology (ICACT)*, pp. 511–515, PyeongChang, Korea, December 2019.

[6] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Systems With Applications*, vol. 42, no. 10, pp. 4621–4631, 2015.

[7] A. Byanjankar, M. Heikkilä, and J. Mezei, "Predicting credit risk in peer-to-peer lending: a neural network approach," in *in Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence*, pp. 719–725, Cape Town, South Africa, December 2015.

[8] H. S. Liu, G. F. Ren, H. T. Chen et al., "Predicting lncRNA-miRNA interactions based on logistic matrix factorization with neighborhood regularized," *Knowledge-Based Systems*, vol. 191, Article ID 062041, 2020.

[9] X. Y. Zhang, "Click prediction for P2P loan ads based on support vector machine," *Journal of Physics*, vol. 1168, no. 3, Article ID 032042, 2019.

[10] M. Liu, M. Qu, and B. Zhao, "Research and citation analysis of data mining technology based on Bayes algorithm," *Mobile Networks and Applications*, vol. 22, no. 3, pp. 418–426, 2017.

[11] C. X. Ling, J. Huang, and H. Zhang, "AUC: a better measure than accuracy in comparing learning algorithms," *Advances in Artificial Intelligence*, vol. 2671, pp. 329–341, 2003.

[12] X. Ye, L.-A. Dong, and D. Ma, "Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score," *Electronic Commerce Research and Applications*, vol. 32, pp. 23–36, 2018.

[13] J. Luan, C. L. Zhanga, B. D. Xu et al., "The predictive performances of random forest models with limited sample size and different species traits," *Fisheries Research*, vol. 227, Article ID 105534, 2020.

[14] F. Lv, J. H. Huang, W. Wang et al., "A two-route CNN model for bank account classification with heterogeneous data," *PLoS One*, vol. 14, no. 8, Article ID 0220631, 2019.