Financial Engineering and Risk Management (2021) 4: 1-6
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2021.040301
ISSN 2523-2576

Personal Credit Assessment Based on Machine Learning Methods

Yihan Peng

School of management, HeFei University of Technology, Hefei, Anhui 230000

Keywords: credit evaluation, machine learning, logistic regression, empirical analysis

Abstract: Machine learning plays an increasingly important role in credit evaluation. Compared with other methods, it can deal with more complex credit evaluation problems and improve the accuracy of prediction results. There are many kinds of research in the field of credit evaluation using machine learning methods. However, most of them combine independent machine learning classifiers, and few studies compare the impact of independent classifiers on the prediction results. In this paper, six different machine learning classifiers are used to do empirical research on credit data, and the binary prediction results of each classifier are analyzed and compared. Through experiments, it is found that the gradient boosting decision tree (GBDT) classifier performs best, with an accuracy of 94%. This study finds out the best method model and gives the binary prediction results to provide the bank decision-makers with a powerful basis for decision-making, thus promoting the construction of personal credit.

1. Introduction

Lending money is a traditional process. [1] The critical factor in this process is whether the borrower can repay on time, that is, the credit evaluation of the borrower. Since 1960, the first mock exam has been designed to obtain information about personal repayment behavior through the design of the credit scoring process. This pattern has developed over time. Among them, the process calculates the percentage of loan risk: the possibility of the customer repaying the loan to the lender at a specific time. Moreover, it is known that giving a personal credit score before the loan to observe the risks involved.

There are many algorithms for credit scoring modeling, and machine learning algorithm is an important trend. The current study mainly includes artificial intelligence algorithms. However, most of these studies put forward a hybrid model, which combines the advantages of independent models and improves the accuracy of prediction, without comparing the influence of different independent machine learning methods on the prediction results. Although the existing studies involve comparing different machine learning algorithms in the field of credit evaluation, these studies are few and fail to cover more machine learning algorithms.

In this paper, six different machine learning models, GBDT, random forest, support vector machine (SVM), logistic regression, K-Nearest Neighbor (KNN), and naïve Bayes, are used to train data sets in the same environment and compare the prediction results of the six methods in order to explore which independent model performs best. In addition, the paper also uses the mean square error (MSE) as the evaluation index to make the results more accurate and have a specific statistical significance.

The rest of the study is organized as follows: Section 2 reviews the work in credit evaluation. Section 3 details the modeling process of the six methods used for comparison. Section 4 describes the experimental design, including data sets, evaluation indicators, and comparison, and experimental results. Finally, section 5 summarizes the prospect of this study and future work.

2. Literature review

In recent years, machine learning methods have been widely used in the field of credit evaluation. Although the logistic regression model has some limitations in model prediction, it has a decisive superiority in variable interpretability and stability. [2] Therefore, the logistic regression method plays an essential role in the application of credit evaluation. For example, Gang Dong et al. (2020) proposed logistic regression with random coefficients method to build credit scorecards, and the experiments show that the proposed method can improve the accuracy of prediction [2].

In 2006, Li Xusheng et al. proposed a personal credit evaluation model based on a naive Bayesian classifier for the first time and tested it on the German-Australian credit data set. The comparison showed that the naive Bayesian classifier has a low classification error.[5] Among the machine learning methods, the SVM method is widely used in the field of credit evaluation. For example, N. Malini et al. (2017) combined the KNN algorithm and outlier detection and analyzed credit card fraud identification technology. [7] PawełPławiak et al. (2019) proposed combining with support vector machine to predict the Australian credit scoring. [4] Also, GBDT, KNN, and random forest have been applied in this field. For instance, Cai Wenxue et al. (2019) applied the combination of GBDT and logistic regression model to personal credit risk assessment. Extracting useful combination features from original data by the GBDT model significantly improves the accuracy of prediction. [8] In 2020, Cong Junrao et al. established a 2-stage Syncretic Cost-sensitive Random Forest model to evaluate the credit risk of the borrowers and improve the classification or prediction accuracy. [6]

3. Methodology

3.1 Modeling

3.1.1 GBDT classifier

GDBT is an iterative decision tree algorithm composed of multiple decision trees, and the conclusions of all trees are added up to make the final answer. In recent years, the machine learning model used for search sorting has attracted much attention. In the experiment, the parameters were set, and the learning rate was 0.1, and the maximum depth was 1.

3.1.2 Random Forest classifier

Due to a large amount of data, the data contains more dimensions, more similar samples, and features that do not need to reduce dimensions. The default value is well, so the use of random forest can be better handled. In building the model, we make a dichotomous prediction on the appropriate number of random numbers in the decision tree.

3.1.3 SVM classifier

The purpose of the SVM is to draw a line that "best" distinguishes the two categories of points so that if new points are created in the future, the line will also make a reasonable classification. In this paper, the problem of the credit dichotomy is that bank borrowing, or not borrowing, is predicted by using the SVM model and calculating the predicted value.

3.1.4 Logistic Regression classifier

In a common dichotomy problem, the classification variables for this data set are two classes of variables: numerical variables, that is, labeled 1 or 0. [3] Each observed object is independent of the other and does not interfere with each other. In the process of building the model, we regularize the data in order to solve the overfitting problem. Finally, a logistic regression model is established.

3.1.5 KNN classifier

Due to the sufficient sample data and high dimension, it is convenient to use the Euclidean distance calculation formula to calculate the distance between test data and training data. The Euclidean distance calculation formula is as follows:

$$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Therefore, the KNN algorithm is adopted. This set of data features satisfies the intersection of class domains and overlaps even more. We set only 1 and 0 labels, whether the bank loans or not, which makes the prediction accuracy higher through binary classification.

3.1.6 Naive Bayes classifier

In Bayesian statistical reasoning, we can assume that the different measures of this set of data are independent of each other and calculate their conditional probabilities. Then, it also considers some evidence or background related to the event, considers the conditional probability, calculates the posterior probability, and finally calculates the probability of the random event, that is, whether the bank lends money or not. Therefore, we pretreated the samples and used feature screening.

4. Experimental analysis

4.1 Experimental Dataset

The data comes from the Give Me Some Credit project of Kaggle (https://www.kaggle.com/). In the experiment, twelve characteristic variables of borrowers are regarded as independent variables .In addition, the result of dichotomy, that is, 0 or 1, is taken as the dependent variable. When the dependent variable is 0, it means that the borrower cannot borrow, and when it is 1, it means that the borrower can borrow.

4.2 Evaluation Metrics

Credit-lending forecasting, the mean square error (MSE) is the most common indicator. This indicator can be used to evaluate experimental results. MSE can be defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

4.3 Experimental Procedure

4.3.1 Data preprocessing

(1) import data

Import the sample data, knowing that the sample number is around 250,000, and there are 12 characteristic variables.

(2) Processing defect data

It is used to deal with missing values with the number of filling methods.

(3) Divide training data and test data

The experiment divides the project data into the training set and test set. The training set (cs-training.csv) contains 150,000 sample data, including 12 variables. The test set (cs-test.csv) is a sample of 101503 pieces of data containing 12 variables.

(4) Normalized data

Standardize data to ensure that each dimension's variance of characteristic data is 1 and the mean value is 0 so that some eigenvalues with too large dimensions will not dominate the prediction results.

4.3.2 Model test

Six different machine learning methods (GBDT, Random Forest, SVM, Logistic Regression, KNN, and Naive Bayes) were used to predict and compare the results.

4.4 Empirical results

4.4.1 Experiment results

Using six different methods to train the training set and adjust the parameters, we get the score of the sample data in the test set and the binary prediction result, that is, to judge whether a borrower can borrow according to his credit status. In addition, we also used MSE evaluation indicators to evaluate the results, as detailed in Table 1,2,3.

Table 1: Score situation of method comparisons

Method	Score
GBDT	**0.9364**
Random Forest	0.9348
SVM	0.9340
Logistic Regression	0.9325
KNN	0.9323
Naïve Bayes	0.9305

According to different machine learning methods, the corresponding credit score is calculated. The specific data is shown in Table 1.

Table 2: Score situation of method comparisons

Method	Binary classification prediction results
GBDT	array ([0, 0, 0, ..., 0, 0, 0], d type=int64)
Random Forest	array ([0, 0, 0, ..., 0, 0, 0], d type=int64)
SVM	array ([0, 0, 0, ..., 0, 0, 0], d type=int64)
Logistic Regression	array ([0, 0, 0, ..., 0, 0, 0], d type=int64)
KNN	array ([0, 0, 0, ..., 0, 0, 0], d type=int64)
Naïve Bayes	array ([0, 0, 0, ..., 0, 0, 0], d type=int64)

The dichotomy prediction results obtained by different machine learning methods are shown in Table 2.

Table 3: Test results of method comparisons

Method	MSE
GBDT	**0.0636**
Random Forest	0.0652
SVM	0.0660
Logistic Regression	0.0675
KNN	0.0677
Naïve Bayes	0.0695

According to the mean square error formula, the MSE values of the six machine learning methods are respectively calculated, and the obtained results are shown in Table 3.

4.4.2 Results discussion

Compare the scores with the experimental results of mean square error, and draw a bar chart, as detailed in Fig 1, 2.

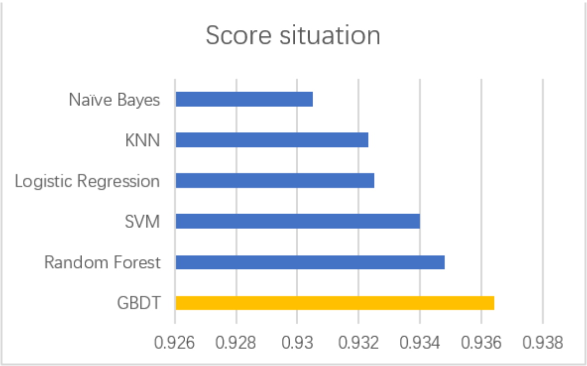

Figure 1

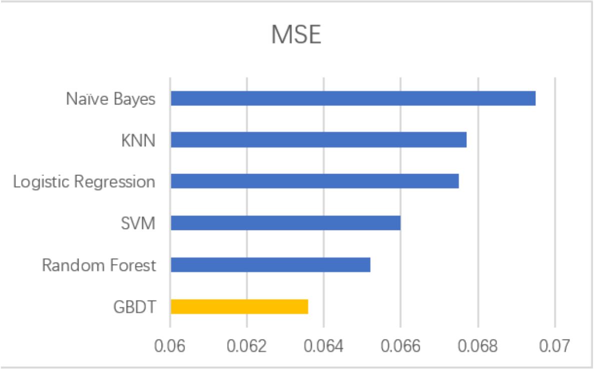

Figure 2

Drawing the results of different credit scores into a statistical bar chart is conducive to observing the specific differences between different methods and drawing relevant conclusions. The error results are also plotted as a statistical bar graph, as shown in Fig 2.It can be seen from the experimental results that the GBDT method has the best performance, with the highest scoring value and the lowest mean error. The results of the six methods are ranked as follows (from the best to the worst):$GBDT > Random\ Forest > SVM > Logistic\ Regression > KNN > Naive\ Bayes$.

5. Conclusion and Prospect

Credit evaluation plays an essential role in accurately identifying credit defaulters and predicting whether borrowers can borrow or not. This study successfully implemented the goal by comparing six different machine learning classifiers (GBDT, naive Bayesian, SVM, KNN, and random forest). On the whole, through the comparison of the results, it can be found that the GBDT of the machine learning method performs best, reaching 94% accuracy. Additionally, this paper demonstrates the two classification prediction results and credit score results of different classifiers and finds that the results of each model have a small gap, getting more precise classification prediction results within the allowable error range.

Equally, as in other studies, this study also has some limitations. Only the credit data of the Kaggle website is collected in the study. However, the scope of testing other credit data on established forecasting models is broad. In future research, larger data sets and more complex machine learning classifiers can also be used to identify credit score predictors. Moreover, other data analysis methods can be combined to predict the results.

References

[1] Shrawan Kumar Trivedi, A study on credit scoring modeling with different feature selection and machine learning approaches, Technology in Society 63 (2020) 101413.

[2] Yuelin Wanga, Yihan Zhanga, Yan Lua, Xinran Yua, A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data, Procedia Computer Science 174 (2020) 141–149.

[3] Gang Dong, Kin Keung, Lai Jerome Yen, Credit scorecard based on logistic regression with random coefficients, International Conference on Computational Science, ICCS 2010.

[4] PawełPławiak, MoloudAbdar, U.Rajendra Acharya, Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring.

[5] Li Xusheng, Guo Yaohuang, Personal credit evaluation model based on Naive Bayes classifier [J]. Computer Engineering and Applications, 2006, 42(30): 197-201.

[6] Congjun Rao, Ming Liu, Mark Goh, Jianghui Wen, 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers, Volume 95, October 2020, 106570.

[7] N. Malini, M. Pushpa, Analysis on credit card fraud identification techniques based on KNN and outlier detection.

[8] Cai Wenxue, Luo Yonghao, Zhang Guanxiang, Zhong Huiling, Risk assessment model of individual credit based on GBDT and Logistic regression fusion and empirical analysis [J]. Management Modernization, 2017, 37(02): 1-4.