Credit decision system based on combination weight and eXtreme Gradient Boosting algorithm

Chen Youlve¹, Bi Kaiyun² and Chen Jiangtian³
Tianjin University, Tianjin, China
*Corresponding author’s e-mail: chenyoulve@163.com

Abstract. According to the demand of bank credit business, a credit decision system is established. The system mainly uses principal component analysis and eXtreme Gradient Boosting algorithm to establish a credit risk assessment model. When calculating the index weight, it proposes a more accurate method: combination weight method, which combines the advantages of entropy weight method and expert scoring method to calculate the relatively accurate weight. Then, the credit risk of each enterprise is analyzed quantitatively. Through the results, we give the corresponding credit strategies of banks in different situations.

1. Introduction
In recent years, with the continuous development of the financial industry, bank credit business has developed rapidly[1]. In real life, for the sake of its own future development, banks usually provide loans to enterprises with strong strength, stable supply-demand relationship and good credit based on credit policy, transaction bill information of enterprises and influence of upstream and downstream enterprises, and can give partial interest rate preference to enterprises with high reputation and low credit risk. Firstly, banks evaluate the credit risk of small and medium-sized enterprises according to their strength and reputation, and then determine whether they can make loans and credit strategies such as loan line, interest rate and term according to many factors such as credit risk. Small and medium-sized enterprises lack mortgage assets. Therefore, it is of great significance to evaluate the credit risk of small and medium-sized enterprises according to their strength and reputation, so as to determine the credit strategy. From the perspective of banks, this paper uses Xgboost (eXtreme Gradient Boosting algorithm) and PCA (principal component analysis) to establish a credit risk assessment model, makes a quantitative analysis of the credit risk of various enterprises, and gives the corresponding credit strategies of banks in different situations through the results. Finally, compared with decision tree, SVM and other methods, our method is better than other methods. The establishment and use of the decision system are introduced according to the algorithm steps.

2. Algorithm

2.1. Data preprocessing
The data set used in this paper is from the bank credit loan situation released by the National Bureau of statistics in 2020. In order to facilitate the subsequent model establishment and data operation, the data should be preprocessed firstly.
- delete duplicate data:
  These errors may come from manual omissions, or failures in machine records.
delete outliers:
Remove outliers using the pauta criterion. For some unreasonable data, we assume that the background value range of a given index is limited to $X \pm 3\sigma$. So we think that in the interval $(x-3\sigma, x+3\sigma)$. At the same time, we eliminate the abnormal data. Then we use the boxplot to judge the value of extreme outliers accurately. For this part of data, we delete it and insert the corresponding value.

Quantification and normalization:
This is to eliminate the influence of measurement unit difference on measurement results. In order to eliminate the adverse effects of these factors and ensure the availability of statistical data, according to the law of numerical calculation, we use the Z-score normalization:

$$Z = \frac{x - \mu}{\sigma}$$  \hspace{1cm} (1)

So as to eliminate the influence of different amount of steel in the data itself, which is conducive to the establishment of statistical evaluation model under the same standard.

2.2. Principal component analysis
For the decision model, the first thing is to determine the model index. According to the data set and related information used in this paper, twelve indicators are selected to establish the credit risk assessment model, they are number of valid input invoices, number of invalid input invoices, total value and tax of valid input invoices, number of invalid input invoices, number of valid output invoices, number of invalid output invoices, total value and tax of valid input invoices, total value and tax of invalid output invoices, number of output enterprises, credit rating, default or not[2].

Because there are too many indexes, we use principal component analysis to reduce the dimension and extract the most relevant four indexes.

Principal component analysis (PCA) transforms a group of possibly correlated variables into a group of linearly uncorrelated variables through orthogonal transformation. The transformed variables are called principal components. PCA tries to make the best synthesis and simplification of multi variable interface data table under the principle of minimum loss of data information, that is to say, reduce the dimension of high-dimensional variable space[3].

correlation matrix $R = \left(R_{ij}\right)_{n \times m}$,

$$r_{ij} = \frac{\sum_{k=1}^{n} a_{ik} \cdot a_{kj}}{n-1}, i, j = 1, 2, \ldots, m$$  \hspace{1cm} (2)

Eigenvectors constitute $m$ new index variables:

$$y_1 = u_{11} \hat{x}_1 + u_{21} \hat{x}_2 + \ldots + u_{m1} \hat{x}_m$$
$$y_2 = u_{12} \hat{x}_1 + u_{22} \hat{x}_2 + \ldots + u_{m2} \hat{x}_m$$
$$y_m = u_{1m} \hat{x}_1 + u_{2m} \hat{x}_2 + \ldots + u_{mm} \hat{x}_m$$  \hspace{1cm} (3)

We extracted four principal components:

| Characteristic value | Contribution rate | Cumulative contribution rate |
|----------------------|-------------------|-----------------------------|
| 1                    | 5.094             | 42.454                      | 42.454                      |

Table 1. Results of principal component analysis.
The eigenvectors corresponding to the four eigenvalues are obtained, four principal components are obtained:

\[ y_1 = 0.3646\hat{x}_1 + 0.3572\hat{x}_2 + \ldots + 0.1627x_{12} \]
\[ y_2 = 0.2664\hat{x}_1 - 0.2082\hat{x}_2 + \ldots + 0.0543x_{12} \]
\[ y_3 = -0.1103\hat{x}_1 - 0.0698\hat{x}_2 + \ldots + 0.6975x_{12} \]
\[ y_4 = -0.2612\hat{x}_1 - 0.3458\hat{x}_2 + \ldots + 0.1304x_{12} \]

Taking the contribution rate of the four principal components as the weight, the principal component comprehensive evaluation model is constructed:

\[ Z = 0.4245y_1 + 0.2335y_2 + 0.1222y_3 + 0.0855y_4 \]  \( (4) \)

By substituting the four principal component values of each enterprise into the above formula, the comprehensive evaluation value and ranking of credit risk of each enterprise can be obtained. The higher the score of an enterprise is, the stronger the enterprise's ability to deal with risk is. The bank can lend more to the enterprise and reduce the interest rate.

2.3. eXtreme Gradient Boosting algorithm

Now we get the credit strategy. Then, we discuss the relationship between each indicator and credit rating.

We use xgboost algorithm for classification, the first step is to score the indicators. Because of the strong subjectivity of scoring, we use the combination weight method, which combines the advantages of entropy weight method and expert scoring method to calculate the relatively accurate weight[4].

The weight obtained by expert method is:

\[ W_1 = [\omega_{11}, \omega_{12}, \omega_{13}, \ldots, \omega_{1t}], \quad 0 \leq \omega_{1k} \leq 1, \sum_{k=1}^{t} \omega_{1k} = 1, k = 1, 2, \ldots, t \]  \( (5) \)

The weight obtained by using entropy weight method is:

\[ W_2 = [\omega_{21}, \omega_{22}, \omega_{23}, \ldots, \omega_{2t}], \quad 0 \leq \omega_{2k} \leq 1, \sum_{k=1}^{t} \omega_{2k} = 1, k = 1, 2, \ldots, t \]  \( (6) \)

Then, we can obtain the combined weight:

\[ W = aW_1 + bW_2 \quad 0 < W \leq 1, a + b = 1 \]

Xgboost is an integrated learning model, which integrates weak learners at the bottom[5-6]. Taking the weak learner of decision tree as an example, we use additive training to optimize the leaf node value of each tree for the objective function, and gradually optimize the function, that is, each time we add an optimal tree to the previous tree to minimize the objective function.

\[ \text{obj}^{(t)} = \sum_{i=1}^{n} \left[ 2(\hat{y}_i - y_i)f_i(x_i) + f_i^2(x_i) \right] + \Omega(f_i) + \text{constant} \]  \( (7) \)

Segment according to a certain attribute value, scan from left to right, use approximate algorithm to find the segmentation point, subtract \( \text{obj} \) from \( \text{obj} \) of two nodes after segmentation, and get the segmentation standard as follows,

\[ \text{Gain} = \frac{1}{2} \left[ \frac{G_j^2}{H_k + \lambda} + \frac{G_R^2}{H_R + \lambda} - \left( \frac{G_j + G_R}{H_k + H_R + \lambda} \right)^2 \right] \]  \( (8) \)
3. Empirical research
Using the above model, we test the data set, and the accuracy is 88%. The following is the characteristic
graph and confusion matrix.

![Feature importance](image1)
Figure 1. Figure with important features.

![Confusion Matrix](image2)
Figure 2. Figure with confusion matrix.

We also use other models to test and get their respective accuracy. The results show that the proposed
algorithm has the highest accuracy.

Table 2. Accuracy of different methods.

| Method                  | Accuracy |
|-------------------------|----------|
| Complex Tree            | 84.6%    |
| Medium Tree             | 84.6%    |
| Simple Tree             | 78.0%    |
| Linear SVM              | 78.9%    |
| Quadratic SVM           | 78.0%    |
| Cubic SVM               | 78.0%    |
| Fine Gaussian SVM       | 78.0%    |
| Medium Gaussian SVM     | 78.0%    |
| Coarse Gaussian SVM     | 78.0%    |
| Xgboost                 | 97.6%    |
4. Conclusion
The application of machine learning in big data analysis has broad prospects. Therefore, the model has great potential. In the future improvement, we can find a more suitable and more accurate machine learning algorithm, and combine the deep learning model with time series analysis, so that the production and operation status of enterprises can be tracked in real time and evaluated comprehensively, which is more conducive to the decision-making of bank credit policy.

References
[1] Simumba N, Okami S, Kohtake N. Credit decision tool using mobile application data for microfinance in agriculture[C]// 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.
[2] Sun Q, Guo W, X Luan. Integrated model and method of the credit sale decision based on choosing the optimal time[J]. Xitong Gongcheng Lilun yu Shijian/System Engineering Theory and Practice, 2017, 37(1):151-158.
[3] Moore, B. Principal component analysis in linear systems: Controllability, observability, and model reduction[J]. IEEE Transactions on Automatic Control, 1981, 26(1):17-32.
[4] Gallo M F, Lopez L M, Grimes D A, et al. Combination contraceptives: effects on weight[M]. John Wiley & Sons, Ltd, 2014.
[5] Chen T, Tong H, Benesty M. xgboost: Extreme Gradient Boosting[J]. 2016.
[6] Sheridan R P, Wang W M, Liaw A, et al. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships[J]. Journal of Chemical Information & Modeling, 2016:2353.