

Credit Risk Prediction of Listed Companies Based on Smote-RF-Catboost

Aohan Wang[#], Ziyi Li^{*, #}, Yan Zhuang[#]

Department of School of Economics and Management, Xidian University, Shaanxi, China

*Corresponding author: lzy769568190@163.com

[#]These authors contributed equally.

Abstract. The credit risk of listed companies is the focus of attention in the field of financial risk, and accurate identification and prediction of credit risk is of great significance to the healthy development of China's financial market and the smooth operation of the financial system. In this paper, a Smote-RF-Catboost method is proposed to study A-share listed companies in China, which selects 23 indicators that can comprehensively reflect the financial situation of enterprises to build an index system, combines Smote (Synthetic Minority Over-Sampling Technique) over-sampling to achieve sample equilibrium, and uses Random Forest feature importance to select features. Finally, the Catboost algorithm is used to predict the credit risk. The empirical results show that the prediction accuracy and AUC value of the test set are above 0.95, and the prediction performance is high, which provides a new way for financial institutions and related practitioners to analyze and predict credit risk.

Keywords: Listed company, Credit risks, Smote over-sampling, Random Forest, Catboost.

1. Introduction

Under the background of increasing downward pressure on the global economy, the monetary policies implemented by some countries in the world maintain a difficult balance between preventing financial risks and curbing inflation, and the risk of financial crisis is rising. With the continuous development of economic globalization and financial globalization, credit risk has become one of the main risks of economic development in various countries. With the continuous reform and development of China's socialist market economy, the default events of listed companies in China occur frequently. How to make a scientific and reasonable risk analysis of listed companies, help them identify, predict and prevent credit risk, and reduce the occurrence of default events, is a hot issue concerned and studied by scholars and financial institutions. According to the *Third Quarter Research Report on Credit Research of A-share Listed Companies in China in 2022*, the credit default risk of A-share listed companies in China is slightly higher than the average level in the past five years, and the overall default risk in the past year has remained basically stable, but it is still at a high level compared with history. Specifically, in the third quarter of 2022, 14 listed companies defaulted, and the overall default risk of A-share market increased slightly, still ranking second in the major stock markets of G20 economies (excluding the European Union). The default events of listed companies occur frequently, and the default risk is increasing, which seriously affects the healthy development of China's financial market and the smooth operation of the financial system. Therefore, how to accurately predict the credit risk of China's listed companies is particularly important.

Credit risk prediction can be divided into two main types of methods, statistical learning methods and machine learning methods. Statistical learning methods include Z-score model, KMV model and so on. On the basis of the traditional credit risk measurement model KMV, Wang Jia introduced the option pricing model with asset price jump. After determining the company's equity value and other indicators, he estimated the equity jump parameters through the maximum likelihood method, and then used the least squares method to estimate the asset value, jump parameters and other parameters, and then calculated the default point and default distance. Measuring Corporate Credit Risk via Distance to Default^[1]. With the development of artificial intelligence and computer technology, machine learning algorithms have more and more applications in the research of risk prediction and

early warning of listed companies. Machine learning methods include BP neural network, SVM, K-means clustering, boosting algorithm and so on. Zhou Ying and Su Xiaoting use the optimal index combination and the optimal training samples to establish a credit risk prediction model based on linear support vector machine, which is used to study the problem of enterprise credit risk prediction^[2]. Zhou Shugong and Li Juanyan proposed a credit risk early warning method for listed companies based on particle swarm optimization (PSO) and support vector machine (SVM) in order to obtain ideal credit risk early warning effect. In this paper, the credit risk early warning problem of listed companies is regarded as a multi-classification problem, and the credit risk early warning classifier of listed companies is established by using support vector machine to deeply analyze and mine the characteristics of credit risk changes of listed companies. Particle swarm optimization algorithm is introduced to optimize the parameters of credit risk early warning classifier of listed companies, and good prediction results are obtained^[3]. Yao Shuang and Gao Jiangbo used the FA-BA-BP neural network model to study the credit risk prediction of manufacturing listed companies. After selecting the index system, the factor analysis method (FA) was used to reduce the dimension. On the basis of BP neural network prediction, The bat algorithm is used for heuristic optimization, and a better prediction result is obtained^[4]. Lu Zhe and Zhang Jian proposed a combination algorithm SMOTETomek-RFE-MLP to predict the credit risk of listed companies, which uses SMOTETomek algorithm to solve the problem of unbalanced samples, and uses recursive feature elimination algorithm (RFE) to extract features. Then the multi-layer perceptron model (MLP) is used to predict the final risk, and compared with other models, the prediction effect is more ideal^[5]. Combination model can flexibly combine different analysis methods according to different data characteristics and different emphases of research problems to achieve the purpose of improving the accuracy of prediction, which is a better research idea of risk prediction of listed companies.

Drawing on the idea of this combination model, this paper proposes a Smote-RF-Catboost combination model method, which uses Smote over-sampling to achieve sample balance for class imbalanced data, and uses random forest feature importance to select features to reduce model complexity and over-fitting risk. Catboost algorithm is used to predict the credit risk. Through the selection of the index system and the application of Smote-RF-Catboost model, it can effectively predict the credit risk of listed companies, play an early warning role for companies, and provide new methods and ideas for relevant researchers, which is more conducive to the stable development of listed companies in China, and consolidate the stable operation of financial markets and financial systems.

2. Model construction

2.1. Index selection

In the process of constructing the credit risk prediction model of listed companies, the selection of indicators is very important for the rationality and scientificity of the final prediction results. A small number of indicators may ignore the impact of some key indicators on the prediction results, and a large number of indicators may weaken the impact of some key indicators on the prediction results.

This paper selects all listed companies in the A-share market as the research object. According to the listing rules of Shanghai Stock Exchange, Shenzhen Stock Exchange and GEM, ST refers to the stocks whose financial situation has deteriorated, which have suffered losses for more than two consecutive years and have been given special treatment. * ST refers to the stocks whose companies have suffered continuous losses for three consecutive years and have been warned of delisting risk. ST shares and * ST shares represent stocks with higher default risk and credit risk, and their companies also have higher credit risk.

Because the credit risk of listed companies is affected by many factors, in order to make the index system reflect the current credit situation of enterprises more reasonably and comprehensively, this paper selects 23 indexes which can reflect the financial situation of enterprises comprehensively on the basis of drawing lessons from the frame of credit evaluation index system of Enterprise Credit

Evaluation Index and relevant references^[6]. The index system covers many factors such as solvency, profitability, operational capacity, growth capacity, cash flow, enterprise scale and so on, focusing on the objectivity, effectiveness and availability of information.

This paper takes all A-share listed companies in China as the research object, selects the three-quarter report data of A-share listed companies in 2022 as the sample, and constructs the credit risk prediction model of listed companies. Among them, there are 5069 companies, including 135 companies with high credit risk and 4934 companies with low credit risk. The data used are from the iFinD database.

2.2. Model establishment

The establishment process of the model in this paper mainly includes data preprocessing, feature extraction, credit risk prediction and other steps. The data preprocessing includes Smote oversampling and other operations. The feature extraction uses random forest feature importance to extract. Catboost algorithm is used in the prediction. The model flow chart is shown in Figure 1.

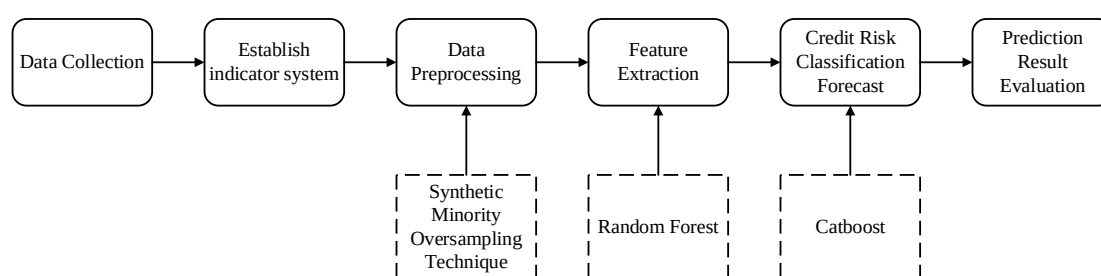


Figure 1. Flow chart of combined model

2.2.1 Smote oversampling

Imbalanced data refers to the situation that the number of data in each category is obviously unbalanced in the classification data. The machine learning classification algorithm is easily affected by the unbalanced data, which makes the accuracy of the classifier more inclined to the majority samples. While the overall accuracy is higher, it can not guarantee that the minority samples also have a more ideal classification accuracy^[7], in the data studied in this paper, there is a big gap between the number of companies with high credit risk and the number of companies with low credit risk, and there is a sample imbalance, so sample equilibrium is needed to optimize the prediction effect.

Oversampling is one of the ways to solve imbalanced data. By adding some minority samples, the number of samples in each category is close, and then classification learning is carried out. Smote (Synthetic Minority Oversampling Technique, composite minority oversampling technique) method is a widely used oversampling method. Its basic idea is to analyze minority samples and then synthesize new samples. The algorithm diagram is shown in Figure 2.

The algorithm flow is as follows:

(1) for each sample in the minority class x_i , respectively calculating the Euclidean distances between the sample and the samples of the same class, and selecting K nearest neighbors from the Euclidean distances;

(2) determining a sampling multiplying power n according to the unbalanced proportion of the samples; In the sample x_i YesRandomly selecting a sample y from neighbors, and repeating for n times;

(3) Sample x_i and y is shown as a formula (1), so that a new sample is construct, and finally n samples are synthesize.

$$x_{new} = x_i + (y - x_i) * rand(0,1) \quad (1)$$

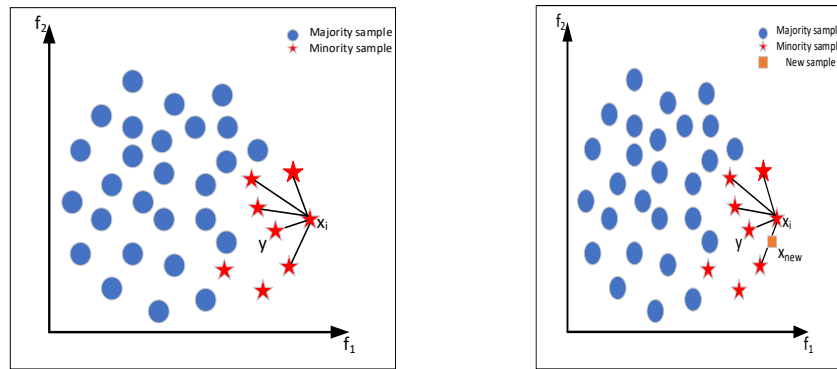


Figure 2. Schematic diagram of Smote oversampling algorithm

2.2.2 Random Forest Feature Importance Ranking

One of the characteristics of ensemble learning model is that it can output the importance of features, which can help us to filter the features to a certain extent, so that the model has better robustness. There are many indicators selected in this paper. In order to measure the contribution of 23 indicators to the change of credit risk and select the more important indicators to improve the prediction efficiency and reduce the complexity, the random forest classifier is considered to determine the importance of features.

The idea of feature importance evaluation in random forest is to calculate the importance of a feature X based on the out-of-bag data error rate, represented by FI , and sort the statistical feature importance according to the size of the FI value. The specific steps are as follows:

Step 1: Construct M decision trees

Step 2: When the decision tree $k_{tree}=1$, select the corresponding out-of-bag data OOB_k , and calculate the data error outside the bag, record it as E_k .

Step 3: The random perturbation of a statistical feature in OOB_k is written as OOB_k^i , calculate the error of out-of-bag data again, and record it as E_k^i .

Step 4: For each decision tree, $k_{tree} = 2, \dots, M$. Repeat steps 2 and 3.

Step 5: Calculate the importance of the feature X according to formula (2)^[8].

$$FI = \frac{1}{M} \sum_{k=1}^M (E_k - E_k^i) \quad (2)$$

2.2.3 Catboost prediction model

CatBoost is an improved GDBT algorithm based on symmetric decision tree, which is suitable for solving heterogeneous data, noise data and classification problems, and has been widely used in many disciplines.

The problem studied in this paper is a typical classification prediction problem. Compared with other classification algorithms, the main advantage of using CaBoost algorithm is that it can efficiently process the class features, solve the problems of prediction bias and gradient bias, thereby reducing overfitting, and improving the generalization ability and accuracy of the model^[9].

All classical lifting algorithms suffer from the overfitting problem caused by biased pointwise gradient estimation. In the traditional GBDT framework, building the next tree is divided into two phases: selecting the tree structure and calculating the values of the leaf nodes after the tree structure is fixed. Prediction shift is caused by the gradient bias in the first stage, so CatBoost is mainly optimized in the first stage to reduce the bias of gradient estimation and improve the generalization ability of the model.

In order to solve this problem, the CatBoost algorithm is improved. In the first stage, uses the unbiased estimation of the gradient step, the prior term and weight coefficient are introduced into the model by the Ordered Boosting method^[10]. The second phase is performed using the conventional GBDT scheme.

The Ordered Boosting algorithm that changes the traditional biased gradient estimation into unbiased gradient estimation: for each sample X_k , we train a separate model M_k , and the model is never updated using the gradient estimate based on the sample. We use M_k to estimate X_k and use this estimate to score the resulting tree. Conditional bias and gradient bias can be effectively avoided, so the algorithm can achieve high-precision prediction under small-scale training, and has strong generalization ability.

3. Empirical analysis

3.1. Data preprocessing

This paper selects the three-quarter report data of A-share listed companies in 2022 for empirical analysis, in which the total number of companies is 5069, including 135 companies with high credit risk, the category number is set to 1, and the category number of other companies is set to 0.

(1) Missing value processing: There are 426 companies with missing values in the data, which is smaller than the overall number and has little impact on the classification results, so they are directly deleted. After missing value processing, there are 4643 companies in total, including 108 companies with high credit risk. The ratio of positive and negative samples is 108:4535, and there is still a problem of sample imbalance.

(2) Sample equalization: The Smote oversampling method is used to process the unbalanced samples, and the number of samples with class number 1 is increased to generate 4427 data, and the samples reach the equilibrium state.

(3) Data standardization: The data dimensions of various indicators are different, and the dimensional influence is eliminated through standardization processing. In this paper, min-max standardization method is used for processing, as shown in formula (3).

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

(4) Sample division. The samples are divided into a training set and a test set according to the ratio of 7:3.

3.2. Feature selection

There are 23 indexes in the index system in this paper, and there are high-dimensional coefficient features. In order to reduce the complexity of the model and improve the efficiency, the random forest feature importance ranking algorithm is used to evaluate the feature importance of the index system. The ranking results are shown in Figure 3. The threshold of feature importance is set to 0.04, and 9 feature attributes are selected.

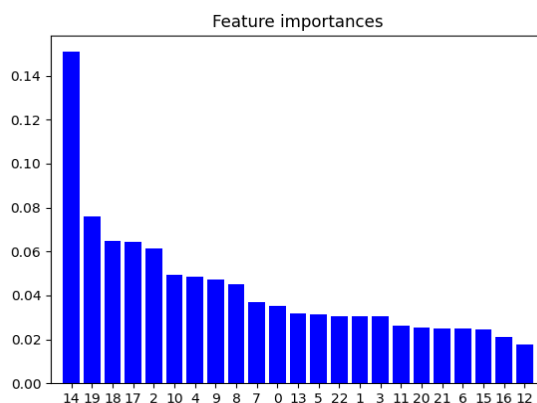


Figure 3. Random Forest Feature Importance Ranking

3.3. Model prediction results and analysis

(1) Default parameter classification prediction.

The default values of the parameters of the CatBoost Classifier are used for classification prediction. The evaluation index table of the classification results is shown in Table 1, and the ROC curve of the test set is shown in Figure 4.

Table 1. Evaluation Index of Classification Results

	Accuracy	F1-score	AUC
Training set	0.9986	0.9986	0.9986
Test set	0.9574	0.9573	0.9584

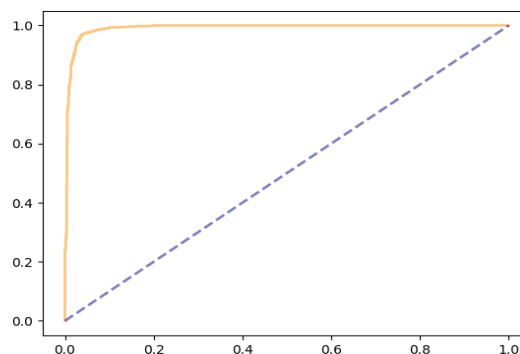


Figure 4. ROC curve of test set

The accuracy rate measures the proportion of the samples with correct prediction in the total samples. In the case of balanced samples, it has a certain reference value. The accuracy rate of the test set is more than 95%, and the prediction effect is good. F1-score is the harmonic mean of the recall rate and the precision rate. The closer to 1, the better the prediction effect is. The ROC curve can intuitively observe the accuracy of the method. The closer the curve is to the upper left, the more accurate the classification results of the model are. The AUC value is the size of the area under the ROC curve. The closer it is to 1, the better the performance of the model is. On the whole, the classification effect of the current model is great.

(2) Parameter tuning

Bayesian optimization algorithm in optuna, which is an automatic hyperparameter optimization software framework, is used to optimize the hyperparameters. Bayesian optimization is an effective global optimization algorithm, which helps to find the global optimal solution. The Catboost algorithm has many parameters, and the use of many parameters will increase the complexity and consume a lot of time. Therefore, this paper selects four important hyperparameters, namely iterations, learning rate, L2 regularization coefficient (L2 _ leaf _ reg) and depth of the tree, to optimize^[11], and the constraint ranges are shown in Table 2.

Table 2. Hyperparameter Constraint Range

Hyperparameters	Constraint range
Number of iterations	1000-1200, and is an integer
Learning rate	0.01,0.03,0.05,0.1,0.2,0.5
The L2 regularization coefficient	1-5, and is an integer
The depth of the tree	3-10, and is an integer

Taking the accuracy of the model as the objective function, the number of model training is 100, and the optimal accuracy of the test set reaches 96.77%, which is 1.03% higher than the prediction effect of the default parameter value. The training history results are shown in Figure 5.

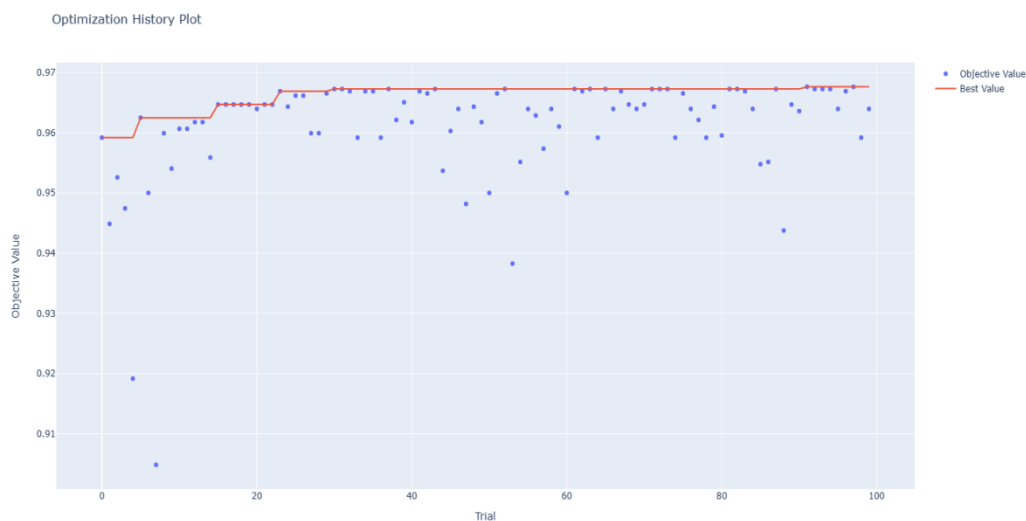


Figure 5. Parameter tuning training results

3.4. Research Summaries

Through the practical application of the model, this paper carries out Smote oversampling, data standardization and other operations for unbalanced experimental data, and carries out random forest feature importance ranking and feature extraction for 23 feature attributes, then uses Catboost model to classify and predict, and uses Bayesian optimization algorithm to optimize parameters. The following conclusions are drawn:

(1)When the default parameters of the Catboost classifier are used to predict the credit risk classification, the prediction results of the training set and the test set are ideal, and the prediction accuracy, F1-score and AUC values of the test set are all above 0.95, which indicates high prediction performance of the model.

(2)By adjusting the learning rate, regularization coefficient and other important parameters, the optimal prediction accuracy is 0.9677, which is higher than that of the classifier with default parameters.

4. Conclusions

This paper takes all listed companies in A-share market as the research object, and reasonably selects the measurement index to construct the index system. During data preprocessing, missing value processing, Smote oversampling and data standardization are carried out to optimize the prediction effect; The selected 23 feature attributes are sorted by random forest feature importance, and several more important indicators are selected to achieve the purpose of improving prediction efficiency and reducing complexity. Catboost algorithm is used for classification prediction and Bayesian algorithm is used for parameter optimization. Finally, the optimal prediction accuracy of the model reaches 96.77%, which provides a referential prediction model for the credit risk prediction of listed companies, and also provides new methods and ideas for relevant researchers, which is more conducive to the stable development of listed companies in China and the consolidation of the stable operation of financial markets and financial systems.

References

- [1] Wang Jia, Cao Qiongyu. Credit risk assessment of listed companies based on jump-diffusion KMV model [J]. Technical Economy, 2022, 41 (01): 160-168.
- [2] Zhou Ying, Su Xiaoting. Enterprise credit risk prediction based on optimal index combination [J]. Journal of Systems Management, 2021, 30 (05): 817-838.

- [3] Zhou Shugong, Li Juan. Credit risk early warning of listed companies based on particle swarm optimization and support vector machine [J]. Modern Electronic Technology, 2020, 43 (11): 72-75. DOI: 10.16652/J. ISSN. 1004-373x.2020.11.017.
- [4] Yao Shuang, Gao Jiangbo, Huang Weiqiang. Research on credit risk early warning of manufacturing listed companies based on FA-BA-BP neural network model [J]. Accounting Newsletter, 2022, No.896 (12): 136-140. DOI: 10.16144/J. CNKI. Issn1002-8072.2022.12.029.
- [5] Lu Zhe, Zhang Jian. Credit risk prediction of listed companies based on SMOTETomek-RFE-MLP algorithm [J]. Systems Science and Mathematics, 2022, 42 (10): 2712-2726.
- [6] GBT 23794-2015, Enterprise credit evaluation index, China, 2015.
- [7] Shi Hongbo, Chen Yuwen, Chen Xin. Survey of SMOTE oversampling and its improved algorithm [J]. Journal of Intelligent Systems, 2019, 14 (06): 1073-1083.
- [8] Qiu Yang, Li Sheng, Jin Liang, Zhang Mimi, Wang Jie. Identification method of abnormal monitoring data of bridge based on statistical feature mixing and random forest importance ranking [J]. Journal of Sensing Technology, 2022, 35 (06): 756-762.
- [9] Ma Shuo, Li Zhao, Zhao Jun. Explanatory study of TreeSHAP based on CatBoost credit prediction model [J]. Computer system application: 1-7.
- [10] Li Huayang, Tan Qiang, Zhu Shijie, Deng Jingen, Yan Ke, Zhang Junyue, Yu Haibing. Pore pressure prediction method based on CatBoost algorithm and its application in wellbore stability analysis [J]. Science and Technology of Work Safety in China: 1-7.
- [11] Zhu Weijun, Li Lin. Network security situation assessment method based on Bayesian optimization CatBoost [J]. Modern Computer, 2022, 28 (06): 11-17.