A Prediction Method Based on Extreme Gradient Boosting Tree Model and its Application

Yongchang Lao¹, Fangzhong Qi²*, Jiakai Zhou² and Xiaobao Fang¹

¹Institute of Economics and Technology, State Grid Zhejiang Electric Power Co., Ltd, Hangzhou, China
²School of Management, Zhejiang University of Technology, Hangzhou, China

*Corresponding author email: fz_qi@zjut.edu.cn

Abstract. Improving the accuracy of financing risk prediction is of great significance to the healthy development of grid enterprises. Taking a provincial-level power grid company as the research object, the financing risk index system is constructed by considering multiple dimensions, and the monthly financing risk index RI of power grid enterprises from 2015-2018 is determined based on entropy weight and comprehensive index method, while the financing risk prediction model is constructed with the help of extreme gradient boosting tree model. The empirical results show that compared with support vector regression and BP neural network models, the financing risk prediction model constructed based on the extreme gradient boosting model has an excellent performance in terms of prediction accuracy and stability.

Keywords: Grid companies; Monthly financing risk; XGBoost model.

1. Introduction

The high investment and long return cycle will make financial institutions increase the interest rate of loans to grid companies, which will greatly increase the financing cost of grid companies. In addition, the high return cycle and the uncertainty of technology investment returns can also make the grid companies face the risk of overdue debt. Therefore, accurate and timely forecasting of grid company's financing risk is significant to enhance the capital risk prevention ability of grid company. Since financial data of grid companies seldom go public, less work has been done on financing risk prediction with grid enterprises. Therefore, the existing studies have mainly been conducted on small and medium enterprises (SMEs), Chinese listed companies and technology enterprises. Financing risk index system has a great influence on the accuracy of financing risk prediction. Most of the scholars integrate multiple aspects for the construction of the financing risk index system. Ebrahimnejad et al (2008) discovered effective standards for risk prediction and proposed a fuzzy MCDM model for project financing risk management in Iranian power plant industry. Some scholars think evolutionary, growth, cyclical, symbiotic and balance should be taken into consideration when constructing an index system of capital operation risk. Drew on the idea of ecology, some scholars combined internal and external financing ecology to construct an indicator system to analyze the financing constraints of Chinese public listed new energy enterprises. Others argued that when preventing financing risks technology-based enterprises should consider four aspects including market, environment, production operations, and finance. In terms of financing risk prediction techniques, traditional statistical models and machine learning models are two types of the most used techniques to conduct research by far. In the 1960's, Beaver (1966) proposed the use of a single financial ratio to predict a company's financial risk. Later, Altman (1968) used five financial indicators and
constructed multiple regression models to predict the business conditions of enterprises. With the development of technology, logistic regression was widely used for risk prediction (Zavgren, 1985; Tseng, 2005). Logistic regression has been used in understanding financing debt of high-quality manufacturing enterprises. Ohlson (1980) used a logistic regression model to construct an early warning model. Khemais et al (2016) adopted discriminant analysis and logistic regression to predict the default risk of Tunisian commercial banks. They found that the discriminant analysis method was slightly better than logistic regression. Hierarchical analyses have also been adopted in evaluating the current status of financing for science and technology-based SMEs in Dalian.

With the continuous development of machine learning technology, scholars have tried to apply some models such as Support Vector Machine (SVM) and BP neural network to solve problems of enterprise risk prediction and evaluation. Martin (1977) used a logistic regression model to predict the probability of bank insolvency. Logistic regression models assume that there is no non-linear functional relationship between the independent and dependent variables. And there is no multicollinearity between the independent variables. When the sample data does not meet the hypothesis, the prediction results will be affected. Hongtian et al (2019) verified that vector machines (SVMs) are suitable for working with small samples, non-linear and high-dimensional data. It is currently used for pattern recognition, classification and comparison, and regression estimation. Lin (2009) compared four models including neural network model, probit model, ogit model and discriminant analysis in the enterprise bankruptcy prediction task. It is found that the neural network model performed the best among the other models and presented high generalization ability. Rick and Sharda (1994) were the first to apply neural networks (ANN) to examine corporate financial distress. Using an equal number of bankrupt and financially healthy companies’ data as sample, they simulated the financial data one year before bankruptcy through the neural network prediction model. The prediction accuracy of the neural network was 79.5%. Some scholars optimized the BP neural network using a recursive genetic algorithm, which effectively improved the performance of the model in the financial warning task. However, the training process of neural network models require a large amount of data and there is no fixed design paradigm for the network structure. Therefore, SVM models are a better choice when predicting enterprise financing risk. Ding et al (2008) applied an SVM model to predict the financial status of Chinese manufacturing companies in high-tech field. The results showed the model was robust and have high prediction accuracy. To address the problem of differentiating sample data importance in support vector machines, Lin and Wang (2002) proposed the fuzzy support vector machine approach (FSVM). The main idea is to assign different weights to different positions of each sample point relative to the cluster centre, with higher weights given to points close to the cluster centre and lower weights given to the opposite. Huang et al. (2020) raised Kernel fuzzy twin support vector machine (KFT-SVM) for early warning systems (EWS), and combined with the resampling method of the synthetic minority over-sampling technique (SMOTE), studied the good ability of KFT-SVM to overcome the class imbalance problem. Some researchers constructed a sample-weighted SVM model to predict the financing risk of technology-based firms and found that given weight to the sample the model had better prediction accuracy for high-risk firms. Improved on the basis of the SVM model, the SVR model for regression tasks was also used in the prediction of risk. When SVR model is used to construct a financing risk prediction model, the accuracy of the SVR model is improved after the parameters are optimized by genetic algorithms. However, the prediction performance of SVR model is also related to the selection of parameters such as kernel function, which requires more subjective experience.

Although many scholars have conducted studies on financing risks, there is still some work left to be done. As far as the research subjects are concerned, the existing studies focus on SMEs, Chinese listed companies and technology companies. Grid enterprises were rarely used as the research object in financing risk problems. In addition, risk prediction is mostly on annual basis, with few scholars forecasting monthly risks for corporate finance. Some studies have found that the financing gap contains information that cannot be captured by financial indicators, which turns to be useful for the prediction of financing risk. The forecasting techniques adopted in the existing researches have certain limitations. For example, traditional models require high data quality. Besides, machine learning models such as SVR and BP neural networks models also have certain limitations. In order to solve
the problems, this article takes a power grid enterprise in the eastern coastal area of China as the research object and constructs a financing risk index system for power grid enterprises by considering seven aspects including debt servicing ability, cash flow, capital operation ability, profitability, development ability, financing scale and economic environment. The forecasting model of financing risk is constructed with the help of the extreme gradient boosting tree of the representative model in integrated learning and compared with the prediction accuracy of SVR and BP neural network models, which confirms the effectiveness of the prediction model constructed in this paper.

2. Construction of Monthly Financing Risk Model for Grid Companies

2.1. Construction of Financing Risk Indicator System

Systematicity and operability are concerned when constructing the financing risk index system. Seven aspects were examined including debt servicing capacity, cash flow, capital operation capacity, profitability, development capacity, financing scale, and economic environment, as shown in Table 1.

| Aspects                          | Indicator                                      | Nature of Indicator |
|---------------------------------|-----------------------------------------------|--------------------|
| Solvency                        | Current ratio/%                               | Positive           |
|                                 | Gearing ratio/%                               | Negative           |
| Cash Flows                      | Net cash flows from operations/¥              | Positive           |
|                                 | Total liabilities/¥                           | Negative           |
| Capital Operation Capability    | Current Asset Turnover Ratio / %              | Positive           |
|                                 | Total assets cash recovery rate / %           | Positive           |
|                                 | Total assets turnover/%                       | Positive           |
| Profitability                   | Operating Margin / %                          | Positive           |
|                                 | Return on Assets / %                          | Positive           |
| Development capacity            | Net profit growth rate/%                     | Positive           |
|                                 | Growth rate of operating profit /%            | Positive           |
|                                 | Growth rate of net asset /%                  | Positive           |
|                                 | Growth rate of total assets /%               | Positive           |
| Financing Scale                 | Financing gap/¥                               | Negative           |
| Economic Environment            | SHIBOR/%                                      | Negative           |

In this case, the financing gap is calculated according to the following formula, and the rest of the indicators are calculated according to the normal calculation of financial indicators. $C_0$, $C_i$ and $C_f$ refer to the net cash flows generated by operating activities, investing activities and financing activities. $W$ is the opening balance of cash and cash equivalents. $Z_0$ is the minimum cash holdings (safety provision).

1) if $(C_o + C_i + C_f + W) > Z_0$, Cash received on loan=Financing gap=0
2) else, Cash received on loan=Financing gap=$Z_0 - (C_o + C_i + C_f + W)$

2.2. Calculation of Financing Risk Index

There are many methods to determine the weight of financing risk indicators, including expert scoring, hierarchical analysis, entropy method and TOPSIS method. Expert scoring and hierarchical analysis are too subjective, as they require experts to be particularly familiar with the field in order to ensure the validity of the results. The entropy weighting method, on the other hand, can determine the weights according to the size of the information load of each indicator and quantify the relevant indicators. Thus, it has been widely used in many fields such as water resources evaluation and risk assessment. Therefore, this paper adopts the entropy weight method to determine the weights of each financing index. Meanwhile, it also combines the comprehensive index method to calculate the financing risk index (RI).
Assuming that there is a sample of \( m \) months of data to be evaluated, and the number of evaluation indicators is \( n \). Let \( a_{ij} \) represents indicator \( j \) of object \( i \), and the original evaluation indicator matrix is denoted as \( A = (a_{ij})_{m \times n} \). Then the specific financing risk index \( RI \) measurement process is described as below.

A. Standardizing indicators

Benefit-based indicators:

\[
b_j = \frac{a_{ij} - \min(a_{ij})}{\max(a_{ij}) - \min(a_{ij})}
\]  

\( j = 1, 2, \ldots, n \)  

Cost-based indicators:

\[
b_j = \frac{\max(a_{ij}) - a_{ij}}{\max(a_{ij}) - \min(a_{ij})}
\]  

\( j = 1, 2, \ldots, n \)  

B. Calculate the information entropy value \( E_j \):

\[
E_j = -\ln\left(\frac{1}{m}\sum_{i=1}^{m} p_{ij}\ln p_{ij}\right)
\]  

Then, \( p_{ij} = \frac{b_j}{\sum_{j=1}^{n} b_j} \).

When \( p_{ij} = 0 \), let \( \ln p_{ij} = 0 \), where \( E_j \) is the information entropy.

C. The weight \( w_j \) of indicator \( x_j \) is calculated as:

\[
w_j = \frac{1 - E_j}{n - \sum_{j=1}^{n} E_j}
\]  

D. Calculate the financing risk index \( RI \). Using the linear weighting method to calculate the financing risk index \( RI \) for grid enterprises, the formula is:

\[
RI = \sum_{j=1}^{n} w_j c_j
\]  

where \( RI \) denotes the grid enterprise financing risk index, \( w_j \) denotes the weight of indicator \( j \) and \( c_j \) denotes the normalized value.

2.3. Construction of a Financing Risk Prediction Model Based on XGBoostTrees

XGBoost is an improvement of the traditional GBDT (gradient boosting decision tree) algorithm, which is a typical representative of integration algorithms (Chen&Guestrin, 2016). The main idea is to continuously add new weak classifiers to fit the residuals of the previous weak classifier. Then obtain the scores of each sample after the training is completed. At last, sum up the scores of all weak classifiers as the final prediction result. The specific decision number learning process is shown below.

When there are \( t \) decision trees in the model:

\[
y_{i}^{(t)} = \sum_{t=1}^{t} f_{ik}(x_{i}) = y_{i}^{(t-1)} + f_{k}(x_{i}), f_{k} \in F, i \in n
\]  

where \( n \) represents the number of samples, \( f_{k} \) represents regression tree \( t \), and \( F \) represents the set of all regression trees, \( y_{i}^{(t)} \) represents the predicted value of sample \( i \) after completing \( t \) times of decision tree training.

The loss function of the model is:
\[ L^{(t)} = \sum_{i=1}^{n} l(y_i, y_i^{(t)}) + \sum_{k=1}^{T} \Omega(f_k) \]

\[ \Omega(f_k) = \gamma T + \frac{1}{2} \lambda \left\| w \right\|^2 \]

In equation (7), \( l \) represents the deviation between the true and predicted values, \( T \) represents the number of leaf nodes of the regression tree, \( w \) represents the weights of the leaf nodes, and \( \gamma \) and \( \lambda \) are the regularization factors.

Conducting a Taylor series expansion on the loss function yields the following equation:

\[ L^{(t)} = \sum_{i=1}^{n} [l(y_i, y_i^{(t-1)}) + f_i(x_i)] + \Omega(f_i) + \sum_{k=1}^{T} \Omega(f_k) \]

\[ = \sum_{i=1}^{n} [l(y_i, y_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_i) + \sum_{k=1}^{T} \Omega(f_k) \]

\[ = \sum_{i=1}^{n} [g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) + \gamma T + \frac{1}{2} \lambda w^2_i + C] \]

In Eq. (8), \( C \) is the constant term, \( g_i \) and \( h_i \) are the first and second derivatives respectively, which can be expressed as:

\[ g_i = \partial_{y_i^{(t-1)}} l(y_i, y_i^{(t-1)}) \]

\[ h_i = \partial^2_{y_i^{(t-1)}} l(y_i, y_i^{(t-1)}) \]

\[ g_i = \sum_{j=1}^{n} l(y_i, y_j^{(t-1)}) + \sum_{k=1}^{T} \Omega(f_k) \]

Let \( G_j = \sum_{i \in j} g_i, H_j = \sum_{i \in j} h_i \), derive the equation so that the optimal solution to equation (11) can be found as follows:

\[ w^*_j = \frac{G_j}{H_j + \lambda} \]

After a series of derivations of the formula, the optimal value of the objective function is:

\[ L^{(t)} = -\frac{1}{2} \sum_{j=1}^{n} \frac{G_j^2}{H_j + \lambda} + \gamma T + C \]

Based on the derivation of the above equation, we can understand the basic decision-making process of XGBoost. Firstly, all possible tree structures are enumerated and then solved to obtain the optimal value of the objective function. Based on equation (13), the score corresponding to each possible tree structure is calculated, and the tree structure with the lowest score is finally selected to calculate the predicted value for each leaf node. Moreover, when the enumeration of all possible tree structures is huge or even infinite, a greedy strategy is needed to generate the nodes of the decision tree in order to completing the complete training and learning process of the XGBoost decision tree.

3. Monthly Financing Risk Forecast for Grid Companies

3.1. Source of Sample

The financial data used in this paper was provided by a grid company in a provincial capital city on the eastern coast of China. The SHIBOR data is sourced from the official website of the People’s Bank of China. The data time range includes month-by-month data from January 2015 to December 2018, with a total of 48 data items. When the financing gap is zero or negative, an enterprise does not need financing and thus does not have financing risks. We therefore excluded the data with zero or negative
financing gap. Finally, there is 44 data items retained, among which 10% (5 sample points) were randomly selected as test set for the machine learning model, with the remaining 90% of the data used as training set.

3.2. XGBoost Model Parameter Settings
The model performance can be improved by adjusting the hyper-parameters of the model. The XGBoost model has a large number of parameters and can be tuned individually or by the GridSearch method or a combination of both. It takes a long time if all the parameters are searched by the GridSearch method. Therefore, individual parameters are searched individually in this paper. Combining the GridSearch method with a multi-parameter optimization strategy ensures the optimal combination of parameters and minimizes the training time. The number of base learners, n_estimators, is a relatively important parameter affecting the model. Combined with the K-fold cross-validation method (a 5-fold cross-validation method is applied here because of the small amount of data), the learning curve of n_estimators versus goodness-of-fit is shown in Figure 1. With an increased number of regression trees, the goodness of fit of the model no longer improves when n_estimators is greater than 35. Thus, 35 is chosen as the optimal value of n_estimators.

![Figure 1. Number of Estimators and Learning curve with goodness of fit R2.](image)

The remaining parameters were optimized using Grid Search and the optimal list of parameters is shown in Table 2.

| Parameter       | Value  |
|-----------------|--------|
| n_estimators    | 35     |
| min_child_weight| 4      |
| max_depth       | 6      |
| gamma           | 0.3    |
| booster         | gblinear |
| subsample       | 0.8    |
| colsample_bytree| 0.7    |
| learning_rate   | 0.38   |
| alpha           | 0      |

3.3. Model Accuracy Inspection
In this paper, a five-fold cross-validation method is used to train the model, find out the optimal parameters, and then test the accuracy of the model on the test set. The experimental test results for the three models, SVR, XGBoost and BP neural network, are given in Table 3. We optimize the parameters of SVR model and BP neural network model, so that the accuracy of the three models can be compared under the optimal performance. It can be seen from Table 3 that the Root Mean Square
Error (RMSE) and the Mean Absolute Error (MAE) of the three models are very small. Among them, the worst performing model is BP neural network, probably because the sample size is too small to exploit the advantages of BP neural network. In terms of specific values, compared with other data points, the absolute error between the predicted value and true value of data point 10 is the largest. The absolute error of XGBoost model is 0.0194, compared with 0.0585 for SVR and 0.1014 for BP neural network.

Table 3. Comparison of model prediction accuracy.

| Test Sets | SVR    | BP Neural Network | XGBoost |
|-----------|--------|-------------------|---------|
| 4         | 0.3851 | 0.3612            | 0.3875  |
| 10        | 0.3516 | 0.3945            | 0.3125  |
| 27        | 0.4201 | 0.4116            | 0.4322  |
| 30        | 0.3933 | 0.3772            | 0.3678  |
| 37        | 0.3948 | 0.3765            | 0.3839  |
| MAE       | 0.0228 | 0.0333            | 0.0068  |
| RMSE      | 0.0299 | 0.0482            | 0.0114  |

This paper ensures the stability of the XGBoost financing risk prediction model by gradually reducing the number of samples in the training set and increasing the number of samples in the test set. The experimental results are shown in Figure 2. It can be seen that the RMSE and MAE of the XGBoost model decreased slightly with the increase of the test set sample. From the error line, the decline is small and there is almost no difference. This indicates the high prediction accuracy and model stability of the grid enterprise financing risk prediction model built based on the XGBoost model.

4. Conclusion

With the deepening of the power system reform, enhancing the financing risk prediction capability of grid enterprises is a top priority to ensure the healthy and steady development of the enterprises. In this paper, we consider a variety of factors, combining entropy weighting and integrated index method to calculate the monthly risk index RI of grid enterprises and construct financing risk index system for a provincial grid enterprise. According to the characteristics of the data, the extreme gradient boosting tree (XGBoost) is used to predict the financing risk of grid enterprises. After comparing with the SVR and BP neural network models, the results show that the XGBoost-based model has excellent performance in terms of accuracy and stability. Therefore, the XGBoost-based financing risk prediction model can accurately predict the monthly financing risk of grid enterprises based on
relevant financial data. It can effectively assist enterprise managers to improve the enterprise's risk management ability.

Acknowledgments
This research is funded by the State Grid Corporation's Science and Technology Project (No. 5211JY18000X).

References
[1] Ebrahimnejad S Mousavi S M and Mojtahedi S M H 2008 A fuzzy BOT project risk evaluation model in Iranian power plant industry Proc, 5th IEEE Int, Conf on IndEngEng Management (Singapore) pp1038-1042.
[2] Beaver W H 1966 Financial Ratios as Predictors of Failure Journal of Accounting Research vol 4(1) pp 71-111.
[3] Altman E I 1968 Financial ratios discriminant analysis and the prediction of corporate bankruptcy Journal of Finance vol 13(4) pp 589-609.
[4] Zavgren C V 1985 Assessing the vulnerability to failure of American industrial firm: a logistic analysis Journal of Business Finance & Accounting vol 12(1) pp 19-45.
[5] Tseng F Lin L 2005 A quadratic interval logit model for forecasting bankruptcy Omega vol 33(1) pp 85-91.
[6] Ohlson J A 1980 Financial ratios and the probabilistic prediction of bankruptcy Journal of Accounting and Research vol 18 (1) pp109-131.
[7] Khemais Z Nesrine D and Mohamed M 2016 Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression International Journal of Economics and Finance vol 8(4) pp 39-53.
[8] Martin D 1977 Early Warning of Bank Failure: A Logit Regression Approach Journal of Banking & Finance vol 1(3) pp 249-276.
[9] Chen H T Jiang B Chen W and Yi H 2019 Data-driven Detection and Diagnosis of Incipient Faults in Electrical Drives of High-speed Trains IEEE Transactions on Industrial Electronics vol 66(6) pp 4716-4725.
[10] Lin T H 2009 Across model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models Neurocomputing vol 72(17) pp 3507-3516.
[11] Wilson R L and Sharda R 1994 Bankruptcy Prediction Using Neural Networks Decision Support Systems vol 11(5) pp 545-557.
[12] Ding Y Song X and Zen Y 2008 Forecasting financial condition of Chinese listed companies based on support vector machine Expert Systems with Applications vol 34(4) pp 3081-3089.
[13] Chun F L and Sheng D W 2002 Fuzzy Support Vector Machines IEEE Transactions on Neural Networks vol 13(2) pp 464-471.
[14] Huang X and Guo F Y 2020 A kernel fuzzy twin SVM model for early warning systems of extreme financial risks International Journal of Finance & Economics vol 26 pp 1459-1648.
[15] Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco) pp 785-794.