Early Warning of Companies' Credit Risk Based on Machine Learning

Benyan Tan, College of Economics and Management, China Three Gorges University, China
Yujie Lin, College of Economics and Management, China Three Gorges University, China*

ABSTRACT

With the advent of the big data era, information barriers are gradually being broken down and credit has become a key factor of company operations. The lack of company credit has greatly and negatively impacted the social economy, which has triggered considerable research on company credit. In this article, a credit risk warning model based on the XGBoost-SHAP algorithm is proposed that can accurately assess the credit risk of a company. The degree of influence of the characteristics of a company's credit risk and the warning threshold of important characteristics are obtained based on the model output. Finally, a comparison with several other machine learning algorithms showed that the XGBoost-SHAP model achieved the highest early warning accuracy and the most comprehensive explanatory output results. The experimental results show that the method can effectively provide a warning of the credit risk of a company based on the historical performance of the company's historical characteristics data. This method provides positive guidance for companies and financial institutions.

KEYWORDS

dishonest civil debtor, machine learning, SHAP, web crawler, XGBoost

INTRODUCTION

Credit risk warning is an integral aspect of financial activity that helps financial institutions to earn profits, predict financial risks, and decrease the probability of default (Tang et al., 2021). In an effort to curb the severe impact of default problems on the economic market, the Chinese government has developed several systems to restrain defaults. For example, to stimulate the debtor to consciously fulfil the obligations determined by legal documents in force and to promote the construction of the social credit system, the Supreme People's Court of China has formulated relevant regulations for announcing the dishonest civil debtor to society, the social activities of a dishonest civil debtor are restricted, and the purpose of credit discipline is achieved.

According to statistics, by the end of March 2022, there were nearly 7.27 million dishonest civil debtors in China, including defaulted individuals and companies. It is worth noting that the development of listed companies is related to the healthy operation of the capital market and the

DOI: 10.4018/IJITSA.324067

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

quality of economic transformation. By the end of March 2022, 4,782 companies were listed in China, with a total market capitalisation of 80.7 trillion yuan, which ranks second in the world in terms of market capitalisation, and the amount of taxes paid by these companies accounted for one-quarter of the national tax revenue. Listed companies are a key factor in developing China's national economy, and any default of listed companies impacts the capital market. Against this background, historical data of listed companies were matched with government credit data to establish an effective early warning model of listed company credit risk. This model not only provides a basis for commercial banks' lending decisions but also has practical significance for the development of listed companies and the regulation of the financial industry.

The above discussion clearly shows that corporate credit is an important influencing factor in financial activities, which has led to the emergence of a considerable body of research on corporate credit risk. Early researchers used statistical methods to evaluate credit risk, and the most representative research was presented by Altman (1968), who constructed a linear discriminant model based on financial indicators and proposed an effective corporate bankruptcy discriminant tool, the Z-SCORE model. In addition, statistical methods, such as logistic regression (Costa e Silva et al., 2020) and probit regression (Chi et al., 2016) are often used to evaluate credit risk, but statistically based credit evaluation methods have the problem of low-discriminatory accuracy. With the application of computer technology in multiple fields and disciplines, scholars have widely used credit risk evaluation methods based on machine learning to construct more accurate credit risk early warning models. Prominent examples are machine learning algorithms, such as K-nearest neighbour (Ata & Hazim, 2020), decision trees (Teles et al., 2020), support vector machines (Shi et al., 2013), neural networks (Zhao et al., 2015; Bekhet & Eletter, 2014) and integrated learning algorithms. Of these, integrated learning algorithms have become a popular avenue of credit risk research in recent years. For example, Wang & Ma (2011) proposed an integrated RS-boosting algorithm combining boosting and random subspaces to predict corporate credit risk. The empirical results showed that RS-boosting achieved the highest prediction accuracy compared to seven kinds of logistic regression analyses, decision trees, ANN, bagging, boosting and random subspaces. Zhu et al. (2017) used data from Chinese-listed companies and found that integrated machine learning offers significant advantages in predicting credit risk for small- and medium-sized enterprises (SMEs). Machine learning methods have greatly improved the accuracy of credit risk assessment; however, the method is prone to problems of insufficient interpretation and weak causality (Lei et al., 2022).

Because machine learning algorithms can only output the validity of a certain indicator system but cannot provide explanatory analysis of individual indicators, researchers have turned to the interpretability of indicators for machine learning, and related explanatory algorithms have been developed rapidly. For example, Lundberg & Lee (2017) proposed the SHapley Additive exPlanations (SHAP) algorithm based on the cooperative game theory concept of Shapley values, which can calculate the contribution of model output features. Since the SHAP algorithm was proposed, it has been widely used in various fields. In the field of credit risk, Bussmann et al. (2021) applied correlation networks to Shapley values and proposed an eXplainable Artificial Intelligence model to measure the credit risk of SMEs, explain their credit scores and predict the future direction of the enterprise. Gramegna & Paolo (2021) used both the SHAP algorithm and the LIME algorithm to calculate the credit risk of Italian SMEs, and their comparison showed that the SHAP algorithm performed better.

The above discussion of credit risk shows that machine learning methods are widely used in the field of credit risk, but the black-box nature of these machine learning methods has resulted in a lack of explanatory models. To overcome this problem, the present paper proposes an early warning model of company credit risk that combines the XGBoost algorithm with the SHAP algorithm. The results show that the XGBoost algorithm outputs the model with the highest accuracy, and the SHAP algorithm compensates for the shortcoming of the XGBoost algorithm (i.e., insufficient interpretation). The SHAP algorithm analyses the main factors that affect the credit risk of Chinese-listed companies based on the perspective of nonlinear effects and is able to define the credit risk early warning intervals for important characteristics. Therefore, this paper proposes that the XGBoost-SHAP model can be widely applied for credit risk early warning of companies. Credit risk research judgments are made based on relevant index data of companies.

RESEARCH METHOD

Introduction of the XGBoost Algorithm

The learning algorithm XGBoost combines both a linear scale solver and a CAtegorical Regression Tree (CART) proposed by Chen & Guestrin (2016). It is a boosting-based model (as shown in Figure 1) that performs well in a variety of practical application scenarios. When the classification performance of a single weak learner is poor, the underlying concept of the XGBoost algorithm is that the mistakes of previous learners are corrected by continuously training new weak learners. That is, the later learners fit the residuals of previous learners to obtain the prediction objective function, and the performance of the model is improved by iteration.

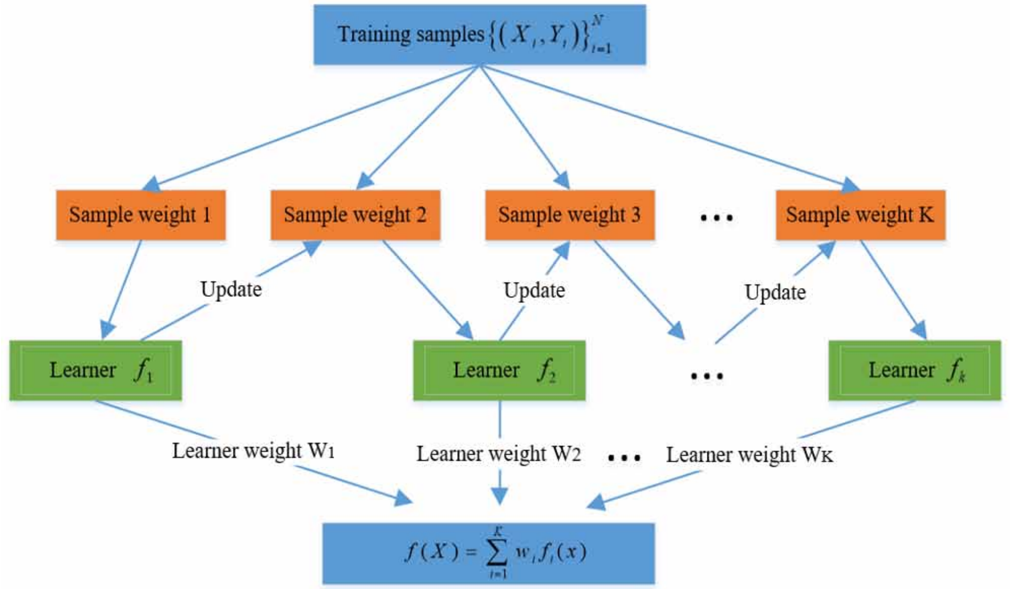
In this paper, the XGBoost model is constructed by Python 3.9.7. First, the sample data are divided into a training set and a test set at a ratio of 7:3, and 9,665 sample data were assigned to the training set, and 4,143 sample data were assigned to the test set. Second, the XGBoost model is constructed using the data of the training set. Finally, the test set data are substituted into the trained XGBoost model to obtain results.

The specific principle of the XGBoost algorithm is described as follows:

Given a sample set with n samples and m features, the sample set is represented as follows: $D = \left\{\left(x_i, y_i\right)\right\}\left(\left|D\right| = n, x_i \in R^m, y_i \in R\right)$. The boosted tree model uses the results of K iterations as the output to obtain the predicted value of the $i-$th sample x_i, where y_i is denoted by Equation (1).

$$\hat{y}_i = \varphi\left(x_i\right) = \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in F \, . \tag{1}$$

Figure 1. Flowchart of the boosting algorithm



K is the number of CARTs, and the set of all CARTs is represented as follows: $F = \left\{ f\left(x\right) = w_{q\left(x\right)} \right\} \left(q : R^m \rightarrow T, w \in R^T\right)$. q denotes the structure of each tree, which maps the samples to corresponding leaf nodes; T corresponds to the number of leaf nodes of the tree; $f\left(x\right)$ corresponds to the structure of the tree q and the leaf node weights w . The XGBoost algorithm keeps the predictions from the previous $t-1$ rounds unchanged at each round of model training by adding the new function f_t to the model, expressed by Equation (2).

$$\hat{y}_i^t = \hat{y}_i^{\left(t-1\right)} + f_t\left(x_i\right). \tag{2}$$

Equation (2) represents the prediction result of the $i-$th sample in the $t-$th training. Assuming that the errors of base learners are independent of each other, find f_t, the objective function is minimised with regular terms, and the $t-$th round objective function is presented in Equation (3).

$$L^{\left(t\right)} = \sum\nolimits_{i=1}^{n} l\left(y_i, \hat{y}_i^{\left(t-1\right)} + f_t\left(x_i\right)\right) + \Omega\left(f_t\right) + constant \tag{3}$$

$$\Omega\left(f_t\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 . \tag{4}$$

The first term in Equation (3) is a loss function, the second term is a regular term (the purpose of which is to control the complexity of the tree and prevent overfitting) and the third term is a constant. Equation (4) is an expansion of the second term in Equation (3), which is the regular term's penalty function to control the model's complexity, γ is the complexity parameter and λ is a fixed coefficient. The expansion of Equation (3) is immediately followed with the second-order Taylor formula. The expanded equation is shown as Equation (5).

$$L^{\left(t\right)} \cong \sum\nolimits_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{\left(t-1\right)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) \right] + \Omega\left(f_t\right) + constant . \tag{5}$$

In Equation (5), $g_i = \partial_{\hat{y}^{\left(t-1\right)}} l\left(y_i, \hat{y}_i^{\left(t-1\right)}\right)$ and $h_i = \partial^2_{\hat{y}^{\left(t-1\right)}} l\left(y_i, \hat{y}_i^{\left(t-1\right)}\right)$. To solve Equation (5), iterations on the tree model are transformed into iterations on the leaf nodes of the tree, the optimal leaf node fraction is identified, and the optimal values of leaf nodes are brought into the objective function. The final objective function takes the form of Equation (6).

$$L^{\left(t\right)}\left(q\right) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum i \in I_j g_i\right)^2}{\sum i \in I_j h_i + \lambda} + \gamma T . \tag{6}$$

Equation (6) can be used as a score function to measure the quality of the tree structure q , where the better the tree structure q , the smaller the function and the smaller the score. Then, the XGBoost algorithm uses Friedman's greedy algorithm (Friedman, 2001) to obtain the optimal tree structure by continuously enumerating different tree structures using the score function.

Introduction of the SHAP Algorithm

SHAP was proposed by Lundberg & Lee (2017) and originated from the cooperative game theory concept of Shapley values. The core concept is to calculate the marginal contribution of features to the model output and then interpret the black-box model both globally and locally. The SHAP algorithm is a typical post hoc interpretation method. The SHAP value is obtained by calculating the mean of the absolute values of the degree of influence of the features on the target variable. Its output can reflect the positive or negative relationship of the influence of certain features on the target variable while maintaining consistency. The main principles are explained as follows:

Assuming that the $i-$th sample is x_i, the $j-$th feature of the $i-$th sample is x_{ij}, the predicted value of the model for the $i-$th sample is y_i and the mean value of the target variable for all samples is y_{base}; then, the SHAP value of x_{ij} obeys Equation (7).

$$y_i = y_{base} + f\left(x_{i1}\right) + f\left(x_{i2}\right) + \cdots + f\left(x_{ik}\right). \tag{7}$$

$f\left(x_{ij}\right)$ is the SHAP value of x_{ij}. $f\left(x_{i1}\right)$ represents the contribution value of the first feature in the $i-$th sample to the final prediction value y_i. $f\left(x_{i1}\right) > 0$ indicates that the feature has a positive impact on the predicted value, and $f\left(x_{i1}\right) < 0$ indicates that the feature has a negative effect on the predicted value. The main advantage of SHAP values beyond the traditional feature importance generated by the integrated algorithm itself is that these values can reflect the influence of features in each sample, displaying both their positives and negatives. SHAP values contribute each feature sum to the final result, demonstrating that the SHAP interpretability algorithm eliminates the interpretative discrepancies resulting from structural differences between different models.

RESEARCH DESIGN

Designing a Credit Risk Early Warning Indicator System

In this paper, a credit risk indicator system for listed companies is constructed, applying the index selection principles of relevant scholars (Chi et al., 2021). If a company has credit risk, its financial statements will show abnormal fluctuations, which is why financial indicators are the main indicators to disclose whether a company has credit risk. In addition to financial indicators that directly reflect the company's abnormal fluctuations, nonfinancial data are also important indicators that affect the company's credit risk. External macroeconomic changes in China's economic market will also impact the company's credit risk. Still, undeniably, the company cannot solely rely on individuals to change the general economic environment. Consequently, from the perspective of listed companies, in this paper, financial and nonfinancial indicators are selected that companies can control and a reasonable system of company credit risk indicators is constructed. This system has practical significance for company decision-makers to control company credit risk and for market investors to identify such company credit risk.

The original combination of features selected in this paper consists of 41 indicators and one default label. The indicators are divided into two dimensions, including financial and nonfinancial indicators of listed companies. The four major financial analysis capabilities of solvency, profitability, operating capacity and growth capacity measure financial indicators. Nonfinancial indicators are measured by the two major characteristics of chairman characteristics and company characteristics. The company's credit risk early warning indicator system and its description are shown in Table 1.

Table 1. Company credit risk early warning indicator system

First-level index	Second-level index	Third-level index (Codes for the index)	Description of the index
Financial indicators	Solvency	Current ratio (X1)	Current assets/current liabilities
		Quick ratio (X2)	(Current assets - inventory)/current liabilities
		Cash ratio (X3)	Cash ending balance/current liabilities
		Working capital to borrowing ratio (X4)	(Current assets - current liabilities)/borrowings
		Interest cover multiplier (X5)	EBIT/finance costs
		Cash flow interest cover multiple (X6)	Net cash flow from operating activities/finance costs
		Net cash flow / current liabilities (X7)	Net cash flow from operating activities/current liabilities
		Gearing ratio (X8)	Total liabilities/total assets
	Profitability	Return on assets (X9)	(Total profit + finance costs)/total assets
		Return on net assets (X10)	Net profit/shareholders' equity balance
		Net operating margin (X11)	Net profit/operating income
		Operating profit margin (X12)	Operating profit/operating income
		Expense ratio during sales (X13)	(Selling expenses + administrative expenses + financial expenses)/operating income
	Operating capacity	Accounts receivable turnover ratio (X14)	Operating income/accounts receivable balance
		Inventory turnover ratio (X15)	Operating costs/inventory ending balance
		Current assets turnover ratio (X16)	Operating income/current assets balance
		Fixed assets turnover (X17)	Operating income/net fixed assets
		Capital intensity (X18)	Total assets/operating income
		Shareholder equity turnover (X19)	Operating income/end balance of shareholders' equity
		Total assets turnover (X20)	Operating income/total assets
	Growth capacity	Operating profit growth rate (X21)	(Operating profit for the current period - operating profit for the previous period)/operating profit for the previous period
		Total profit growth rate (X22)	(Total profit for the current period - total profit for the previous period)/total profit for the previous period
		Net profit growth rate (X23)	(Net profit for the current period - net profit for the previous period)/net profit for the last period
		Growth rate of total assets (X24)	(Total assets for the current period - total assets for the previous period)/total assets for the previous period
		Growth rate of total operating income (X25)	Growth rate of total operating income for the current period over the previous period
		Growth rate of owner's equity (X26)	Growth rate of owner's equity for the current period over the same period of the previous year
		Growth rate of basic earnings per share (X27)	(EPSt-EPSt-1)/EPSt-1

continued on the following page

Table 1. Continued

First-level index	Second-level index	Third-level index (Codes for the index)	Description of the index
Nonfinancial indicators	Chairman characteristics	Gender (X28)	0 = Female; 1 = Male
		Age (X29)	Split the box into four parts. 0 = 23.0–37.5 years; 1 = 37.5–52.0 years; 2 = 52.0–66.5 years; 3 = 66.5–81.0 years
		Education (X30)	0 = College and below; 1 = Bachelor and above; 2 = Other
		Whether both chairman and general manager (X31)	0 = No; 1 = Yes
		Whether both chairman and CEO (X32)	0 = No; 1 = Yes
		Whether the workplace of independent directors and listed companies are the same (X33)	1 = Same, 2 = Different, 3 = Uncertain
	Company characteristics	Concentration of equity of the top 10 outstanding shareholders (X34)	The sum of the shareholdings of the top 10 outstanding shareholders of the company
		Concentration of equity interests of the top 10 shareholders (X35)	The sum of the shareholdings of the top 10 major shareholders of the company
		Whether the top 10 shareholders are related (X36)	1 = No association; 2 = Associated; 3 = Uncertain
		Whether the company is special treatment) (X37)	0 = No; 1 = Yes
		Whether the company discloses the internal control audit report (X38)	0 = No; 1 = Yes
		Audit opinion (X39)	1 = Standardly unqualified opinion; 2= Reserved opinion; 3 = Unable to express an opinion; 4 = Unqualified opinion plus matter paragraph; 5 = Reserved opinion plus matter paragraph
		Number of board meetings (X40)	
		Number of shareholder meetings (X41)	
Default indicator		Whether the company is listed as a defaulted company	0 = No; 1 = Yes

Design of Credit Risk Early Warning Model

To explore the important factors and characteristic performance affecting the credit risk of a company, and based on the credit risk index system presented in the previous section, this paper uses machine learning methods to construct a credit risk early warning index system. This system can provide companies, investment institutions and regulators with accurate credit risk early warning information, which is conducive for maximising their benefits and minimising their losses. For example, for the company itself, the generation of credit risk can cause a decline in reputation and lead to a decrease in the value of company investments. However, using the machine learning model constructed in this paper can issue a warning regarding the credit risk of the company in advance and thus reduce the risk of loss of benefits for the company. The framework of the credit risk early warning model in this paper is described below.

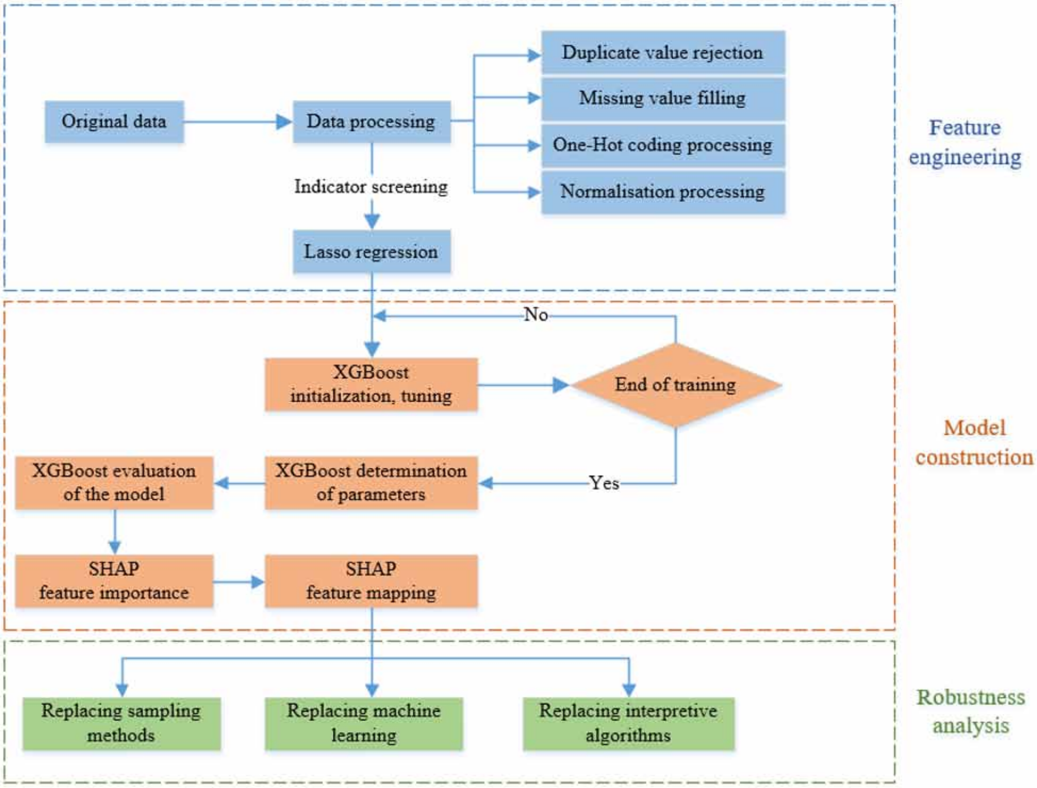
First, the original data are processed (including duplicate value rejection, missing value filling, one-hot coding and normalisation processing) to obtain a complete credit risk early warning indicator system. To eliminate data redundancy and reduce data dimensionality, lasso regression is used to identify the indicator system with strong discriminatory ability. Second, this paper integrates XGBoost and SHAP algorithms to construct an early warning model for the credit risk of listed companies. Although the XGBoost algorithm can train an early warning model to have high prediction accuracy of credit default, the XGBoost model is a black-box model with insufficient explanation. Therefore, the contribution of the output features of the SHAP algorithm are used to increase the interpretability of the model. Third, three methods (i.e., replacing machine learning, replacing sampling methods and replacing interpretive algorithms) are used to verify the robustness of the early warning model, combining XGBoost and SHAP algorithms. The results show that the credit risk early warning model constructed by this combination of XGBoost and SHAP algorithms achieves the best credit risk evaluation effect. The specific construction process of the company credit risk early warning model is shown in Figure 2.

RESULTS AND ANALYSIS

Data Source

The indicator data in this paper were obtained from the database of Guotai ’an, and 41 financial and nonfinancial indicators disclosed in the annual reports of 2,874 listed companies from 2012–2019

Figure 2. Flowchart of the constructed company credit risk early warning model


Original data
Data processing
Duplicate value rejection
Missing value filling
One-Hot coding processing
Normalisation processing
Indicator screening
Lasso regression
Feature engineering
No
XGBoost initialization, tuning
End of training
Yes
XGBoost evaluation of the model
XGBoost determination of parameters
SHAP feature importance
SHAP feature mapping
Model construction
Replacing sampling methods
Replacing machine learning
Replacing interpretive algorithms
Robustness analysis

were selected. Because of the number of indicators, X1, X2, X3,… , and X41 are used as substitutions for indicator names. The label data were obtained from the company default data from the China Executive Information Disclosure Network. This paper takes the existence of default records of listed companies as a measure of company default. Credit data of listed companies from 2014–2021 were obtained, involving a total of 950 default records of 116 listed companies. The specific company default records are shown in Table 2, indicating the following:

(1) Defaulted listed companies are mainly distributed in the manufacturing industry. Their default frequency accounts for 54% of the total default frequency, indicating that the default probability of manufacturing companies is higher compared to other industries.

(2) The default frequency of listed companies shows an annually increasing trend, especially after 2018 when the default frequency of listed companies increased significantly.

Construction of Credit Risk Early Warning Index System

Data Processing

In this paper, data processing mainly included duplicate value rejection, missing value filling, one-hot coding processing and normalisation processing.

(1) Duplicate value rejection. For companies with multiple default records, this paper takes only one default record of a company in a specific year. Finally, a total sample of 13,808 companies was obtained, including 201 default samples and 13,607 nondefault samples.
(2) Missing value filling. In referring to existing literature on the treatment of missing values, mean padding was used for continuous-type features, and plural padding was used for category-type features.
(3) One-hot coding processing. In this paper, 10 category-based features were selected, five of which were dichotomous variables, and the other five were multicategorical variables. Dichotomous variables were represented by 0–1 variables, and the multicategorical variables were treated by one-hot coding. Because there was no quantitative computational relationship between the selected multicategorical variables, representation by N 0–1 variables can more effectively portray the relationship between multicategorical variables with N states.

Table 2. Defaults of listed companies in china according to industry classification

Year	Agriculture, forestry, animal husbandry and fishery	Manufacturing	Electricity, heat, gas and water production and supply industry	Wholesale and retail trade	Information transmission, software and information technology services	Real estate industry	Other industries	Total
2014	0	0	0	7	0	0	0	7
2015	0	1	0	7	0	0	0	8
2016	0	5	0	4	0	2	0	11
2017	0	1	0	2	2	1	1	7
2018	0	24	7	4	7	11	2	55
2019	5	99	14	20	27	8	11	184
2020	19	189	15	14	64	4	29	334
2021	10	194	8	4	52	17	59	344
Total	34	513	44	62	152	43	102	950

(4) Normalisation processing. To standardise the magnitude, continuous-type features were processed by Z-score normalisation. Z-score normalisation standardises data based on the mean ($\overline{x}$) and standard deviation (σ) of the original data, which are processed to have a mean of 0 and a standard deviation of 1. The transformation equation is shown as Equation (8).

$$x^{*} = \frac{x - \overline{x}}{\sigma} . \tag{8}$$

Finally, 54 features and one label were obtained after data processing.

Indicator Screening

The developed credit risk early warning index system is screened based on lasso regression (Tibshirani, 1996). This cluster indicator screening approach can identify the indicator system with the best discriminatory ability and is more applicable to high-dimensional indicator downscaling (Meinshausen & Bühlmann, 2006). Specifically, lasso regression was implemented with Python and five-fold cross-validation was conducted to select the best regularisation coefficient of » =0.00017. The results of lasso regression model coefficients obtained with the best regularisation coefficients were screened, and the 22 indicators with coefficients of 0 were deleted. Eventually, the best early warning indicator system for credit risk was obtained, and the screened indicators are shown in Figure 3. The vertical coordinates in Figure 3 represent the indicator codes (see Table 1 for indicators), and the horizontal coordinates are the lasso regression indicator coefficient weights.

The original combination of indicators selected in this paper has 41 indicators with one default label, which is expanded to 54 indicators with one default label after the one-hot coding process. In this paper, to reduce and remove information redundancy, the best combination of indicators is obtained based on lasso regression screening. Indicator screening retained 32 indicators and one default label, where the indicators after the one-hot coding process behave as indicator codes with suffixes.

Empirical Tests and Analysis of Results

XGBoost Model-Based Parameter Tuning

The screened indicator data were divided into a training set and a test set according to a ratio of 7:3, with 9,665 sample data in the training set and 4,143 sample data in the test set. The XGBoost model was trained using data of the training set. When the parameters were not adjusted, the model had a recall of 0.3594 on the test set, indicating that the model was not sensitive to capture discredited samples. The model parameters are adjusted to improve the model's capture rate of the loss-of-trust samples. Grid search is one of the most common ways to tune parameters, and this paper uses grid search and a five-fold cross-validation method to adjust parameters. The specific grid search list and the optimal parameter results are shown in Table 3. After parameter adjustment, the XGBoost model performs well on the test set as recall increases from 0.3594 to 0.9063, and the area under the curve (AUC) decreases from 0.9891 to 0.9269. These results indicate that the model has a lower classification effect but a substantially better capture rate of discredited samples; therefore, the parameter adjustment is ideal.

XGBoost Model-Based Tuning Evaluation Results

Based on the characteristics of imbalanced samples of listed companies, this paper uses the five evaluation indexes of accuracy, recall, specificity, AUC and G-mean to evaluate the XGBoost model. A confusion matrix is used to visually represent the evaluation results of the XGBoost model. After tuning the parameters, the confusion matrix is shown in Figure 4 according to the model's performance

Figure 3. Lasso Model Feature Importance

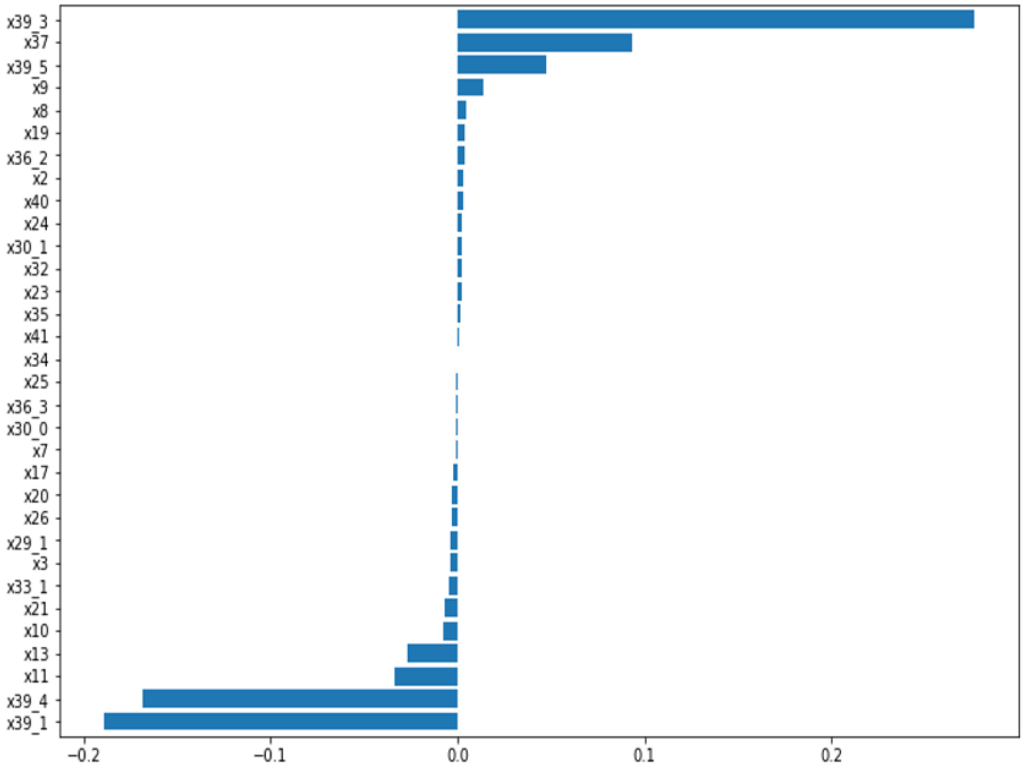
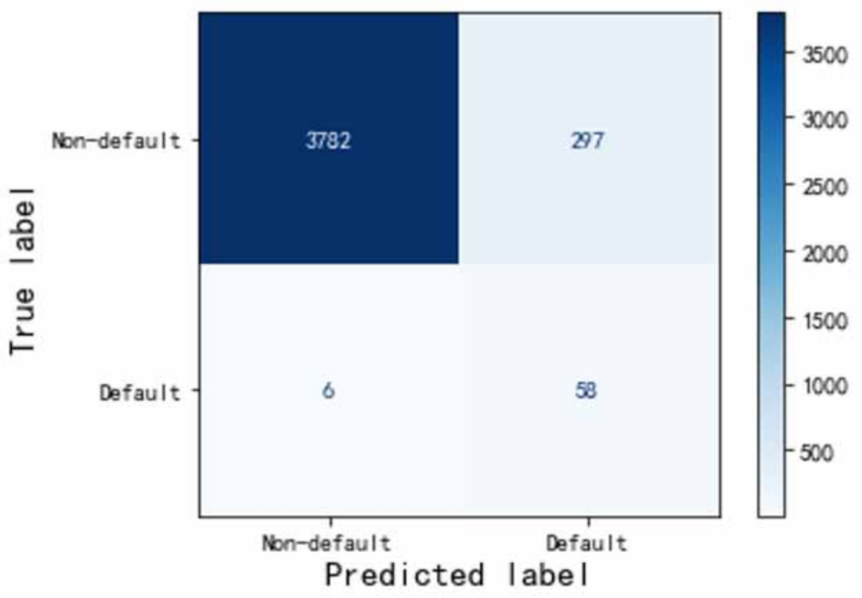


Table 3. List of XGBoost model parameters and optimal parameter values

Parameter name	Parameter meaning	Optimal parameter values
learning_rate	Learning rate in integration	0.1
n_estimators	Number of weak classifiers	175
max_depth	Maximum tree depth of weak classifier	2
min_child_weight	Minimum sample weights required at a leaf node	3
subsample	Proportion of sampling from the sample	1
colsample_bytree	Proportion of features randomly sampled out of all features when constructing each tree	1
scale_pos_weight	Handling sample imbalance in labels	900

for the test set. According to Figure 4, there are 4,143 samples on the test set, and the XGBoost model predicts 355 default samples and 3,788 nondefault samples. From the confusion matrix, it can be calculated that accuracy, recall, specificity, AUC and G-mean are 0.9269, 0.9063, 0.9272, 0.9584 and 0.9167, respectively. The overall prediction accuracy exceeds 90%, indicating that the XGBoost model has a good prediction effect after adjusting parameters.

Figure 4. Confusion matrix of XGBoost model after parametric adjustment



SHAP Algorithm-Based Feature Importance Analysis

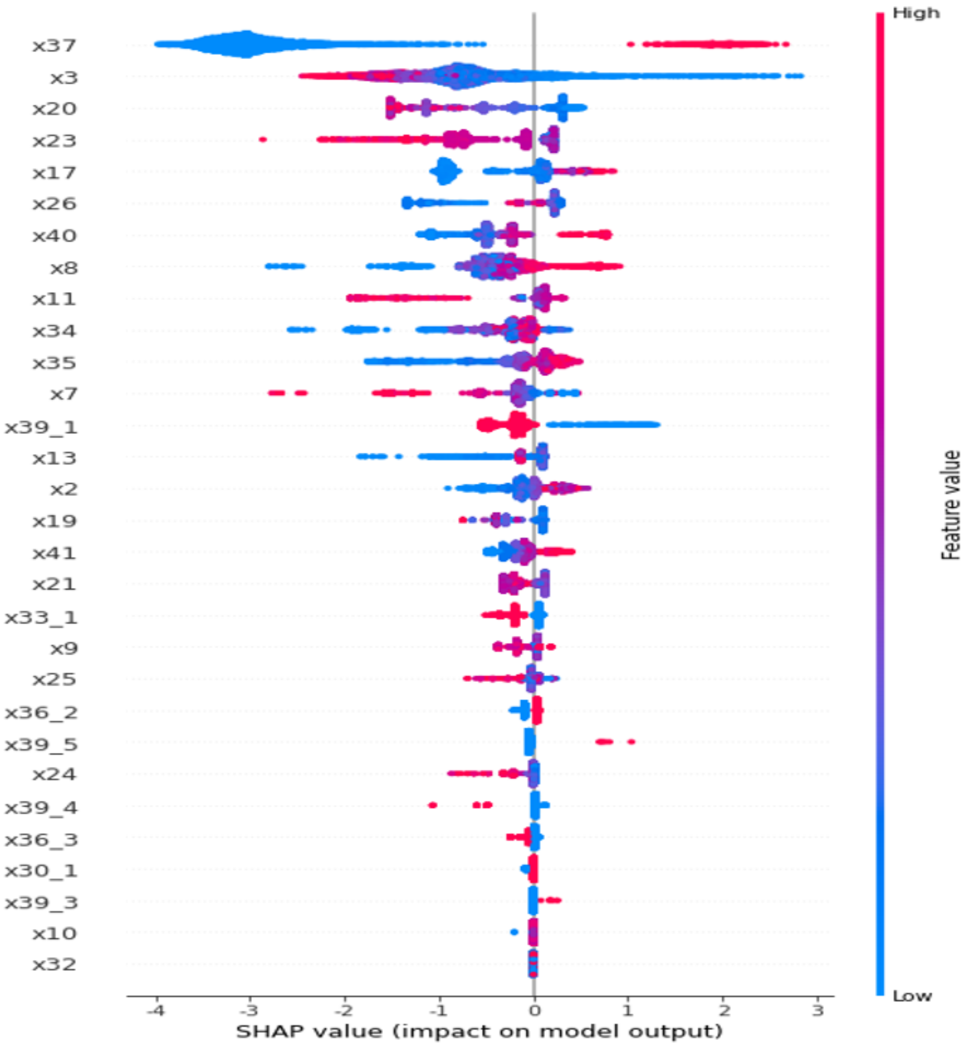
Based on the XGBoost model outputting high-precision prediction results, to arrive at a more intuitive and comprehensive understanding of the relationship between features and SHAP values, this paper uses the SHAP algorithm to output the top 30 features in terms of feature importance. Then, a SHAP summary graph of these 30 features is drawn as shown in Figure 5. In this SHAP summary chart, the horizontal coordinate is the SHAP value. The left vertical coordinate reflects the importance ranking of features, where features are listed in decreasing order of importance from top to bottom. The colour change of the right vertical coordinate represents the value of the feature itself, where the redder the colour, the higher the value of the feature, and the bluer the colour, the lower the value of the feature.

According to the 30 features shown in Figure 5, the following results can be summarised:

(1) There are 16 financial indicators. As an example, among the four solvency indicators screened, cash ratio (X3) and net cash flow/current liabilities (X7) show a left red and right blue state on both sides of the dividing line of a SHAP value of 0, indicating that the smaller the value of these two indicators, the higher the risk of company default. In contrast, quick ratio (X2) and gearing ratio (X8) show a left blue and right red state, where the larger the value of these two indicators, the higher the risk of company default.
(2) There are 14 nonfinancial indicators. For example, among company opinion indicators, whether the company is ST (X37) presents a left blue and right red status. This status indicates that the company is judged as ST by the China Securities Regulatory Commission, which will increase the company's default risk. In contrast, whether the company's audit opinion is a standardly unqualified opinion (X39_1) presents a left red and right blue status, indicating that the company's audit opinion is not a standardly unqualified opinion, which increases the company's default risk.

From a global perspective, the SHAP summary chart presents a comprehensive interpretation of the relationship between indicators and default status, where the importance of indicators and each indicator's positive/negative relationship with the company's credit risk are output. In addition, the

Figure 5. SHAP summary chart (total sample)



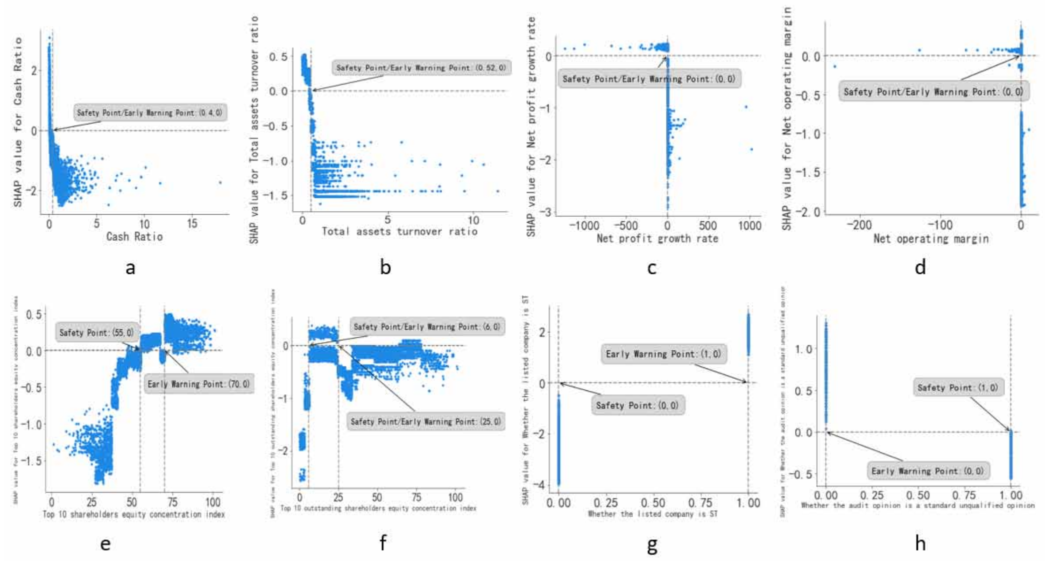
SHAP algorithm is not required to evaluate the linear relationship between indicators. Nonetheless, its output SHAP values of indicators can establish their default properties, which have relevance for financial activities.

SHAP Algorithm-Based Analysis of the Main Features of Early Warning

To further understand how the features affect company default, eight features are selected as the primary research objects. These are cash ratio, total asset turnover ratio, net profit growth rate, net operating margin, top 10 shareholders' equity concentration index, top 10 outstanding shareholders' equity concentration index, whether the listed company is ST, and whether the listed company's audit opinion is a standardly unqualified opinion. The SHAP value mapping relationship is analysed for these eight indicators. The corresponding mapping relationships are shown in Figure 6.

(1) Impact of financial indicators on the company's credit risk

Figure 6. Mapping relationship of eight important features of SHAP values



Based on the previously identified hierarchy of financial indicators, one indicator is selected from the solvency, operating, growth and profitability indicators used in this paper, namely, cash ratio, total asset turnover, net interest rate growth and net operating margin, respectively (Figures 6a–d).

1) In Figure 6a, the cash ratio grows from 0. When the cash ratio exceeds 0.4, the mapped SHAP values are negative, indicating a significant decrease in the company's credit risk. The cash ratio is one of the solvency indicators and represents the company's ability to liquidate.
2) In Figure 6b, the total asset turnover ratio grows from 0. When the total asset turnover ratio exceeds 0.52, the mapped SHAP values are negative, representing a decrease in the company's credit risk. The total asset turnover ratio is one of the indicators of operating capacity, which is a measure of the efficiency of the company's asset operations. Combined with Figure 5, the cash ratio and total asset turnover ratio rank second and third in importance in the SHAP summary graph, respectively, indicating that these two indicators are important and affect the company's credit risk. Therefore, when a company's cash ratio is below 0.4, or its asset turnover ratio is below 0.52, the company should focus on reducing and controlling its credit risk.
3) In Figures 6c and d, when the net profit growth rate and net operating margin exceed 0, the mapped SHAP value is negative, and the company's credit risk is reduced. The net profit growth rate is one of the growth capability indicators, and the net operating margin is one of the profitability indicators, representing the company's profitability. Therefore, when the profitability level of a company is lower than its profitability level of the previous year, the company should pay attention to credit risk reduction and control.
(2) Impact of nonfinancial indicators on the company's credit risk

Based on the previous description of the features of nonfinancial indicators, four company characteristics indicators are analysed in depth in this paper. These are two continuous-type indicators for the top 10 shareholders' equity concentration index and the top 10 outstanding shareholders' equity concentration index, and two category-type indicators for whether the listed company is ST and whether the company's audit opinion is standardly unqualified opinion (Figures 6e–h).

1) In Figure 6e, the mapped SHAP value is negative when the top 10 shareholders' equity concentration index is below 55%, indicating a lower credit risk for the company; the mapped SHAP value is positive when the top 10 shareholders' equity concentration index exceeds 70%, indicating a higher credit risk for the company.
2) In Figure 6f, the SHAP value is negative when the top 10 outstanding shareholders' equity concentration index is either lower than 6% or higher than 25%. When the top 10 outstanding shareholders' equity concentration index is below 6%, the listed company faces less pressure to restructure and reduce the equity of outstanding shares, which is conducive to reducing the company's credit risk. The top 10 outstanding shareholders' equity concentration index is higher than 25%, indicating that the secondary market has a positive attitude towards the listed company's market value and the company's credit risk is low. Therefore, when adjusting the equity structure, the company should reasonably control the shareholding ratio of both shareholders and outstanding shareholders to reduce the company's credit risk.
3) In Figure 6g and f, whether the listed company is ST and the company's audit opinion is standardly unqualified are dichotomous variables. If the China Securities Regulatory Commission has determined that the listed company is ST, the mapped SHAP value is positive, and the company has a high credit risk. When the company's audit opinion is standardly unqualified, the mapped SHAP value is negative, and the company's credit risk is low. Therefore, the company should pay attention to the evaluation criteria of third-party institutions to reduce its credit risk.

The SHAP algorithm can output a SHAP value mapping relationship graph of individual indicators, which visually represents the distribution of SHAP values of indicators. Based on the analysis of the above representative indicators, the change in the SHAP value mapping relationship of indicators summarises the law. It innovatively delineates the default risk warning threshold of company indicators, which can provide an effective warning of the emergence of a company's credit risk. Therefore, the reference values of credit risk warnings for the above eight important features are shown in Table 4.

Robustness Analysis

Replacing Machine Learning Algorithms

Following the sample data filtered by lasso regression, the training and test sets were divided at a ratio of 7:3. Five different algorithms were used to construct the models, namely, XGBoost, Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and LightGBM. The evaluation results of each model (Table 5) were obtained after parameter adjustment. The results show that the XGBoost algorithm has the highest recall, AUC and G-mean values, which proves that

Table 4. Reference values of credit risk warning for eight important features

Feature name	Range of values	Early warning reference value
Cash ratio	[0,17.92]	0.40
Total asset turnover ratio	[0,11.42]	0.52
Net profit growth rate	[–1242.56,1018.99]	0.00
Net operating margin	[–231.26,9.61]	0.00
Concentration of equity of the top 10 shareholders	[1.31,100]	70.00
Concentration of equity of the top 10 outstanding shareholders	[0.04,100]	[6,25]
Whether the listed company is ST	0/1	1
Whether the audit opinion is a standardly unqualified opinion	0/1	0

the XGBoost model is optimal for the fitted samples and illustrates the robustness of the XGBoost algorithm used in this paper.

Replacing Sampling Methods

The Synthetic Minority Oversampling Technique (SMOTE) is an improved algorithm that is based on the random oversampling algorithm. It uses the method of simply copying samples to increase minority class samples, which is prone to the problem of model overfitting. The basic idea of the SMOTE algorithm is to analyse minority samples and add new samples to the data set by artificially synthesising them (based on the minority samples) so that the samples can be balanced by adding further minority samples.

In this paper, the sample imbalance ratio is about 1:68, which is an extremely imbalanced sample. To test whether the sample imbalance will impact the model, the SMOTE oversampling method is used to balance the sample in advance. In contrast, the five algorithms presented above are still used to construct the model. The evaluation results of each model (as shown in Table 6) are obtained after parameter adjustment. The results show that in the model training comparison after SMOTE, the recall value of the SMOTE-XGBoost model is 0.906, which far exceeds the recall values of the other four algorithms, proving that the SMOTE-XGBoost model has the best sensitivity for capturing default samples. A comparison of the contents presented in Table 5 shows that the recall of the XGBoost model is also 0.906 and that its accuracy, specificity, AUC and G-mean values exceed those of the SMOTE-XGBoost model. This result shows that the model trained by using the XGBoost algorithm is more effective.

Replacing the Explanation Method

(1) Feature importance

The XGBoost model features a plot_important that can output the feature importance ranking, and Python can calculate that the XGBoost model outputs the top 20 features with 94.8% importance. In

Table 5. Comparison of evaluation results between XGBoost and other machine learning algorithms

Name of algorithm	Accuracy	Recall	Specificity	AUC	G-mean
XGBoost	0.927	**0.906**	0.927	**0.958**	**0.917**
LR	0.988	0.375	0.997	0.949	0.612
SVM	0.921	0.828	0.923	0.914	0.874
RF	0.937	0.734	0.940	0.905	0.831
LightGBM	0.990	0.422	0.999	0.943	0.649

Table 6. Comparison of the evaluation results of xgboost and other machine learning algorithms after SMOTE

Name of algorithm	Accuracy	Recall	Specificity	AUC	G-mean
SMOTE-XGBoost	0.810	**0.906**	0.809	0.948	0.856
SMOTE-LR	0.931	0.859	0.932	0.951	0.895
SMOTE-SVM	0.946	0.703	0.949	0.938	0.817
SMOTE-RF	0.928	0.813	0.930	0.947	0.870
SMOTE-LightGBM	0.981	0.719	0.986	0.969	0.842

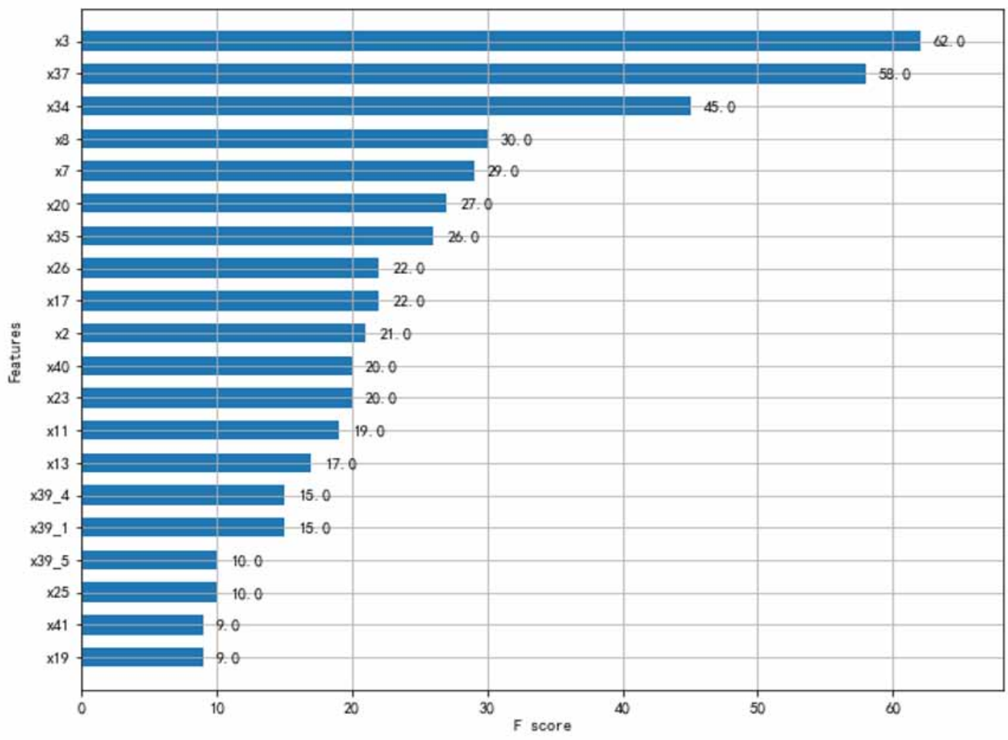
this paper, the top 20 features based on the XGBoost algorithm feature importance are output using Python, and the results are shown in Figure 7. A comparison of the top 20 feature importance rankings presented in Figures 7 and 5 shows that although there are differences in the feature importance rankings, 17 feature names overlap, indicating that the two algorithms output approximately the same indicator importance. Because the SHAP algorithm can output specific SHAP values, its interpretation is much higher than that of plot_important, making it more robust when using the SHAP algorithm.

(2) Partial dependence plot

A partial dependence plot (PDP) is introduced based on the XGBoost model to analyse the influence patterns of individual features on the prediction results. Following the eight features selected from the SHAP mapping map, as shown in Figure 6, the PDP is drawn as shown in Figure 8. In Figure 8, the *x*-axis represents the values of the features, the *y*-axis represents the change in the model prediction compared to the baseline value or the rightmost value and the blue-shaded part represents the confidence interval. According to Figure 8, feature values above the baseline have a positive effect on the model prediction results (default), and feature values below the baseline have a negative effect on model prediction results (no default).

1) In terms of financial indicators, according to Figure 8a–d, when the cash ratio, total asset turnover ratio, net profit growth rate and operating profit are all greater than 0, the impact on the credit default of listed companies is negative. At this time, the credit risk of listed companies is low.
2) In terms of nonfinancial indicators, Figure 8e indicates that when the top 10 shareholders' equity concentration index reaches 40%, it begins to affect the prediction results; when it reaches 50%,

Figure 7. Feature Importance ranking based on the XGBoost algorithm



it has a very significant positive impact on the prediction results, in other words, it significantly increases the credit risk of the listed company. As shown in Figure 8f, when the top 10 outstanding shareholders' equity concentration index exceeds 0, the credit risk of the listed company increases; when it reaches 10%, the credit risk of listed companies increases significantly. According to Figure 8g and h, the credit risk of companies increases dramatically when they are classified as ST or when the audit opinion is not standardly unqualified.

A comparison between Figures 8 and 6 shows that the trend of each feature in the two plots is almost the same when the feature value increases, but there is a difference in their early warning threshold values. The credit risk early warning reference values output by SHAP are more accurate, indicating that the SHAP algorithm used in this paper is more robust. The risk warning reference values based on the eight features of the SHAP mapping and PDP plots are presented in Table 7.

Figure 8. PDP Chart of eight important features

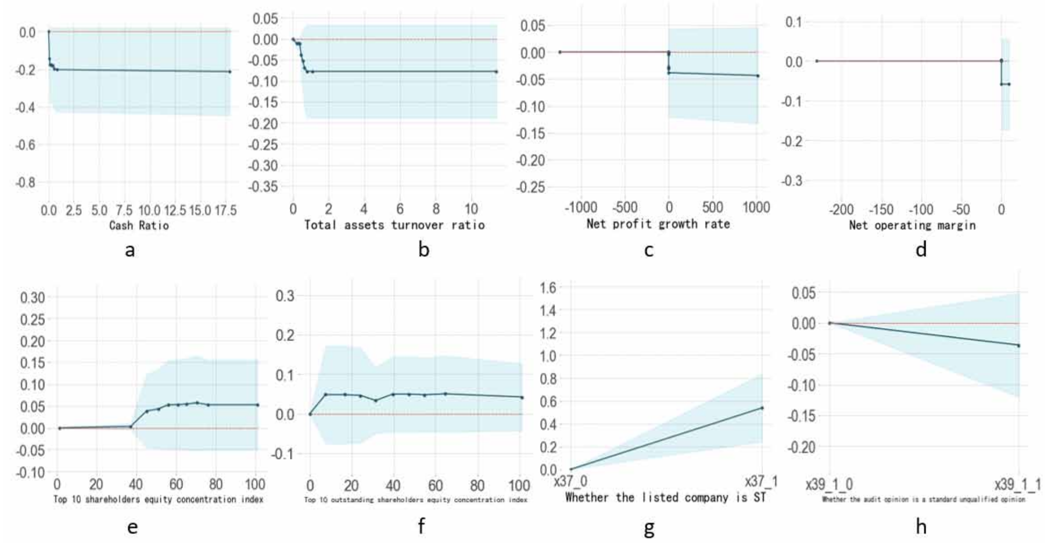


Table 7. Comparison of SHAP mapping and PDP mapping risk warning reference values

Feature name	SHAP mapping chart warning reference values	PDP chart warning reference values
Cash ratio	0.40	0
Total asset turnover ratio	0.52	0
Net profit growth rate	0.00	0
Net operating margin	0.00	0
Concentration of equity of the top 10 shareholders	70.00	50
Concentration of equity of the top 10 outstanding shareholders	[6,25]	10
Whether the listed company is ST	1	1
Whether the audit opinion is a standardly unqualified opinion	0	0

CONCLUSION

This paper presents a credit risk early warning method for companies based on XGBoost and the SHAP algorithm, which provides a strong credit risk prevention guarantee for companies and financial institutions. The method is based on the XGBoost algorithm for evaluating the credit risk of listed companies, and its comprehensive evaluation indexes AUC and G-mean are 95.8% and 91.7%, respectively. These values are higher than the evaluation results of LR, SVM, RF and LightGBM algorithms, achieving high-precision data fitting. However, the black-box nature of the XGBoost model means that the decision process is not transparent, and the output results are not interpretable. To overcome these limitations, this paper innovatively introduces the SHAP algorithm to identify the important features that affect the credit risk of a company and delineates the credit risk warning thresholds of these important features. The above research has substantial reference value for companies to reduce their own credit risk and for banks and other related financial institutions when making major lending decisions.

Although the empirical presented in this paper has obtained meaningful results, there are certain limitations which provide avenues for future research. On the one hand, the construction of the indicator system is not sufficiently comprehensive, and on the other hand, the model algorithm does not have sufficient depth. Therefore, in future work, the team will strive to expand the index system and improve the model algorithm to further develop the research breadth and depth of credit risk warning.

AUTHOR NOTE

The data used to support the findings of this study are included within the article.

The authors declare that there is no conflict of interest regarding the publication of this paper.

This research received no external funding.

REFERENCES

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. doi:10.1111/j.1540-6261.1968.tb00843.x

Ata, O., & Hazim, L. (2020). Comparative analysis of different distributions dataset by using data mining techniques on credit card fraud detection. *Tehnicki Vjesnik (Strojarski Fakultet)*, *27*(2), 618–626. doi:10.17559/TV-20180427091048

Bekhet, H. A., & Eletter, S. F. K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, *4*(1), 20–28. doi:10.1016/j.rdf.2014.03.002

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*(1), 203–216. doi:10.1007/s10614-020-10042-0

Chen, T. Q., & Guestrin, C. (2016, August). Xgboost: *A scalable tree boosting system* [Conference session]. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery. https://doi.org/ doi:10.1145/2939672.2939785

Chi, G. T. (2021). Credit index construction and forecasting of Chinese listed companies, *Newspaper of Financial Times in China*, 11. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=JRSB202107190110&DbName=CCND2021

Chi, G. T., Zhang, Y. J., & Shi, B. F. (2016). The debt rating for small enterprises based on Probit regression. *Journal of Management Sciences in China*, *19*(6), 136–156. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=JCYJ201606010&DbName=CJFQ2016

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics*, *47*(13–15), 2879–2894. doi:10.1080/02664763.2020.1759030 PMID:35707418

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. doi:10.1214/aos/1013203451

Gramegna, A., & Paolo, G. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558. doi:10.3389/frai.2021.752558 PMID:34604738

Lei, X. N., Lin, L. F., Xiao, B. Q., & Yu, H. H. (2022). Re-exploration of default characteristics of micro and small enterprises: A machine learning model based on SHAP interpretation method. *Journal of China Management Science*, *2022*, 1–13. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=ZGGK2022031600I&DbName=DKFX2022

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Vol. 30, pp. 4766–4775). Advances in neural information processing systems. Curran Associates., https://doi.org/arXiv:1705.07874v2

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, *34*(3), 1436–1462. doi:10.1214/009053606000000281

Shi, J., Zhang, S. Y., & Qiu, L. M. (2013). Credit scoring by feature-weighted support vector machines. *Journal of Zhejiang University SCIENCE C*, *14*(3), 197–204. doi:10.1631/jzus.C1200205

Tang, J. J., Li, J., Xu, W. Q., Tian, Y., Ju, X., & Zhang, J. (2021). Robust cost-sensitive kernel method with Blinex loss and its applications in credit risk evaluation. *Neural Networks*, *143*, 327–344. doi:10.1016/j.neunet.2021.06.016 PMID:34182234

Teles, G., Rodrigues, J. J. P. C., Saleem, K., Kozlov, S., & Rabelo, R. A. L. (2020). Machine learning and decision support system on credit scoring. *Neural Computing & Applications*, *32*(14), 9809–9826. doi:10.1007/s00521-019-04537-7

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, *58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Wang, G., & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. *Expert Systems with Applications*, *38*(11), 13871–13878. doi:10.1016/j.eswa.2011.04.191

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Lui, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, *42*(7), 3508–3516. doi:10.1016/j.eswa.2014.12.006

Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing & Applications*, *28*(1), 41–50. doi:10.1007/s00521-016-2304-x