SHS Web of Conferences **218**, 02016 (2025) https://doi.org/10.1051/shsconf/202521802016
ICDDE 2025

Research on Credit Risk Evaluation System for Small and Medium-Sized Enterprises Based on Machine Learning Model

Liwei Ou*

School of Public Finance and Taxation, Southwestern University of Finance and Economics,
Chengdu, Sichuan, 610074, China

Abstract. Corporate credit ratings are crucial for enterprise development, reflecting financing capabilities, business expansion potential, tax payment capacity, and corporate reputation. Credit is a fundamental prerequisite for enterprises to engage in market activities. A favorable credit standing can facilitate the establishment of stable cooperative relationships and lay a foundation for business operations and decision-making. This study focuses on credit risk evaluation for small and medium-sized enterprises (SMEs) and employs machine learning models to predict credit ratings using public datasets. The research looks into applying multiple machine learning techniques, such as Logistic Regression, Decision Tree, Random Forest, LightGBM, CatBoost, and Multilayer Perceptron (MLP), to determine the crucial factors that impact credit risk. The results show that the LightGBM-based SME credit rating model is more adaptable than other models. Analysis of the contribution degrees of evaluation indicators reveals that firm size, performance, and employee welfare stability are crucial in credit risk assessment. This helps financial institutions and enterprises make more precise decisions by accurately evaluating SME credit and identifying risks.

1 Introduction

Corporate credit rating is an important basis for measuring corporate financing capabilities, business expansion, tax payment capabilities and reputation. Good credit status helps companies establish stable cooperative relationships with customers and provides support for business decisions. Accurate prediction of credit risk can reduce bank loan risks, reduce bad debts and improve financial security.

In the past, the analysis and evaluation of corporate credit risk have been dominated by the three major American rating agencies: Standard & Poor's, Moody's, and Dun & Bradstreet. The construction of corporate credit evaluation systems initially focused solely on financial indicators such as debt repayment capacity, profitability, operational efficiency, and growth potential. Altman E.J. developed an indicator system based on key accounting information, such as the interest coverage ratio and total asset turnover rate. Since then,

---

* Corresponding author: 42203050@swufe.edu.cn

credit evaluation models based on financial statement data have continued to evolve [1]. Lin S. M. categorized numerical financial ratios to achieve more effective credit prediction results [2]. Gupta et al. found that operating cash flow was effective in credit risk modeling for British firms [3]. Cultrera L. et al. identified that certain financial ratios, such as operating profit margin, current ratio, and total asset turnover rate, were excellent predictors for Belgian companies [4]. Consequently, the financial analysis indicators widely recognized by the academic community primarily encompass four key dimensions: debt repayment capacity, profitability, operational efficiency, and growth potential.

Because small and medium-sized enterprises are not publicly listed, it is challenging to obtain authentic and reliable financial data. Arslan and Karan (2009) utilized a Logistic model to investigate the common determinants of credit risk for both domestic and international small and medium-sized enterprises (SMEs) in Turkey [5]. Locke et al. argued that firm size, age, and business type are associated with credit risk [6]. Zhou et al. performed a comprehensive examination of 80 credit risk metrics across 500,000 firms in Henan Province. They created a system that combines financial and non-financial information. The research showed that the random forest model had the best F1 score performance [7]. Zhong et al. compared the effectiveness of four learning algorithms, including Back propagation Neural Network and Support Vector Machine, in corporate credit rating and found that neural network algorithms performed well in credit evaluation [8]. Qiu and He utilized China UnionPay credit data, selected common credit models such as Support Vector Machine (SVM), random forest, and decision tree to test the effectiveness of credit risk indicator selection [9]. Their approach, based on a two-stage feature selection method of Lasso-RF, optimized credit indicator selection and improved the accuracy of traditional classifier models [9]. As a vital part of China's market economy, SMEs contribute significantly to economic growth, improving people's livelihood and expanding employment, and they are also a driving force for innovation.

Consequently, this study will employ commonly used machine learning credit models to conduct an in-depth investigation into the creditworthiness of Chinese small and medium-sized enterprises (SMEs). It aims to construct a scientific credit risk evaluation indicator system and identify the key features that significantly influence corporate credit. The findings are expected to provide valuable insights and assistance to financial institutions and enterprises in the realm of credit management.

2 Method

2.1 Dataset

The dataset utilized in the thesis is sourced from China Chengxin International Credit Rating Co., Ltd., a prominent Chinese credit rating agency, and is publicly available as open-source data. The dataset comprises 1,958 enterprise records and 13 features. Each sample corresponds to a small or medium-sized enterprise (SME) in China. The features include both financial and non-financial indicators, such as credit score, registered capital, enterprise scale score, enterprise performance score (financial data features), as well as organizational form of capital, business registration date, and industry type (non-financial factors).

In the credit risk evaluation of SMEs, the construction of a scientific and rational evaluation indicator system is of paramount importance. This system must ensure the accuracy of the evaluation results while also guaranteeing the objectivity, universality, and standardization of the evaluation framework. Therefore, in the selection of features, the principles of independence and availability must be strictly adhered to Chen [10].

Historically, investigators have leveraged financial records and textual information from financial statements to uncover instances of corporate fraud. However, the study by Wei Dong et al. (2018) demonstrates that financial social media data offers greater value in fraud detection, outperforming baseline methods that use only financial ratios or language-based features [11].

Considering the characteristics of SMEs, this study focuses specifically on the fundamental information that reflects the current business conditions and future development trends of enterprises, as well as their scale and performance, to provide a more comprehensive perspective for credit risk assessment. The dataset is divided into four categories of features: enterprise basic information, enterprise scale and efficiency, employee welfare and stability, and corporate honors and certifications. These features reflect the business scope and operating scale, profitability, stability of enterprise development, and innovation capability and social reputation of the enterprise.

2.2 Model

2.2.1. Logistic regression

Logistic regression, a variant of Generalized Linear Models (GLMs), is commonly utilized for binary classification tasks and can also be adapted for multi-class scenarios. Its fundamental concept lies in applying the Sigmoid function to the linear combination of feature variables, transforming it into the range (0, 1). This process enables the prediction of the probability of the target variable falling into a specific class. This is a supervised learning algorithm, meaning it uses labeled training data to learn a model that can classify new, unseen data. The derivation and computation of logistic regression are similar to those of linear regression.

2.2 2. Decision tree

The decision tree classifier, a hierarchical model utilized in supervised learning, is applicable to both classification and regression analyses. Commencing at the root node, it evaluates specific attributes of a sample and directs it to a corresponding child node in accordance with the evaluation outcome. This process is recursively repeated, selecting the optimal feature for splitting until a leaf node is reached, ultimately forming a tree-like structure. The advantage of decision trees is their ease of understanding and interpretability, with an intuitive decision-making process. They also make fewer distributional assumptions about the data, making them suitable for various types of data. However, decision tree classifiers are prone to overfitting, resulting in poor model generalization. Additionally, they are sensitive to noise and outliers in the data.

2.2 3 Random forest

Random forest is a committee-based learning technique that integrates numerous decision trees for classification and regression purposes. It builds diverse decision trees by bootstrapping the training data and randomly choosing feature subsets for each tree's training. For predictions, it uses majority voting in classification or averaging in regression. This ensemble approach boosts the model's robustness and precision. Random forest can process many input variables, making it ideal for high-dimensional datasets. Additionally, it's resilient to outliers and data noise. However, it demands significant training time and memory, particularly with a large number of trees. Unlike a single decision tree, random forest is less interpretable.

2.2.4 LightGBM

LightGBM is a high-performance gradient boosting algorithm. It uses a histogram method, converting continuous feature values into K discrete integers. This, combined with its leaf-wise growth strategy, minimizes splits and boosts training speed. LightGBM is fast in training and low in memory consumption, making it suitable for large-scale data. It supports parallel training, further enhancing training efficiency. It also offers various regularization methods to effectively prevent overfitting. However, LightGBM is sensitive to hyperparameters, requiring meticulous tuning.

2.2.5 CatBoost

CatBoost is a machine learning algorithm based on gradient boosting decision trees. It employs an Ordered Boosting algorithm that can directly handle categorical features without transformation, while incorporating various optimization techniques such as symmetric binary tree layout, histogram-based acceleration, and data parallel computation. CatBoost provides automatic feature scaling, simplifying data preprocessing steps. It performs well with large-scale data, offering high training efficiency and accuracy.

2.2.6 Multilayer perceptron

The Multilayer Perceptron (MLP) is a feedforward neural network model composed of one or more hidden layers and an output layer. Each layer in the model consists of multiple neuron nodes, each connected to all nodes in the previous layer with trainable weights and bias values. The weights and bias values are updated through the backpropagation algorithm (BP algorithm) to minimize the error. It can handle nonlinear problems and has strong generalization capabilities. However, during model training, it is prone to getting stuck in local optima.

3 Results

3.1 Data preprocessing

To reduce the influence of differing numerical feature magnitudes on the model, normalization of continuous variables is necessary. In this case, min-max normalization is applied to scale feature values to between 0 and 1. The features subjected to normalization include registered capital, enterprise-scale score, enterprise performance score, social security stability score, business registration duration, and housing provident fund stability score. This approach ensures that large differences in magnitudes do not distort the importance of features, thereby maintaining the stability of model predictions.

One-Hot encoding is a method for transforming categorical variables into numerical representations, widely used in machine learning and data analysis. Its core concept involves representing each category of a categorical variable as an independent binary column, where each column contains only 0 or 1. This encoding method preserves the independence among different categories of categorical variables, thereby preventing ordinal relationships among categories from being mistakenly interpreted as numerical magnitudes. This is a universal method applicable to most machine learning algorithms.

To enable the model to recognize categorical variables, it applies One-Hot encoding to categorical features such as organizational form of capital, primary industry category, strategic emerging industry category, high-tech enterprise status, leading enterprise at the municipal level, and model employer for labor protection and integrity.

For the credit score feature, following the method of Huang et al. (2024), it classifies enterprises with a credit score of 50 or below as high-risk, those with scores between 50 and 80 as medium-risk, and those with scores above 80 as low-risk [12]. This categorization transforms the credit score into a multi-class classification problem suitable for machine learning modeling. The model is then trained based on the credit score feature.

3.2 Experimental results

For multi-class classification tasks, accuracy, precision, recall, and F1 score are commonly used as the primary metrics for evaluating model performance. Higher values of these metrics indicate better model performance, although their calculation methods differ. The preprocessed data were split into training and testing sets in an 8:2 ratio. The experimental results of the test set were used as the final criteria for model evaluation. The performance metrics are shown in Table 1.

Table 1. Each model evaluates the results in the test set

Model	Accuracy/%	Precision/%	Recall/%	F1 Score/%
Logistic Regression	90.82	90.37	90.82	90.12
Decision Tree	91.58	91.21	91.58	91.29
Random Forest	94.64	94.54	94.64	94.45
LightGBM	97.19	97.21	97.19	97.15
CatBoost	95.15	95.04	95.15	94.92
MLP	91.33	91.07	91.33	91.18

In the assessment of various machine learning models on the test set, it has been observed that the Random Forest, LightGBM, and CatBoost models exhibit superior F1 Scores in predicting the credit ratings of SMEs. Among these, the LightGBM model demonstrates the most outstanding performance. LightGBM employs a leaf-based algorithm and histogram technique, which enables it to efficiently process large-scale datasets and effectively address class imbalance issues. These features provide it with a significant advantage in credit risk classification tasks, indicating that the model possesses excellent generalization capabilities and is most beneficial for practical applications.

In the research on credit risk evaluation of SMEs, the feature importance analysis conducted by the Random Forest model has revealed the relative weighting relationships among various features. This approach provides a reference for financial institutions and businesses, assisting them in gaining a deeper understanding of the credit conditions of the enterprises being assessed. The feature importance of the Random Forest model is illustrated in Fig. 1.

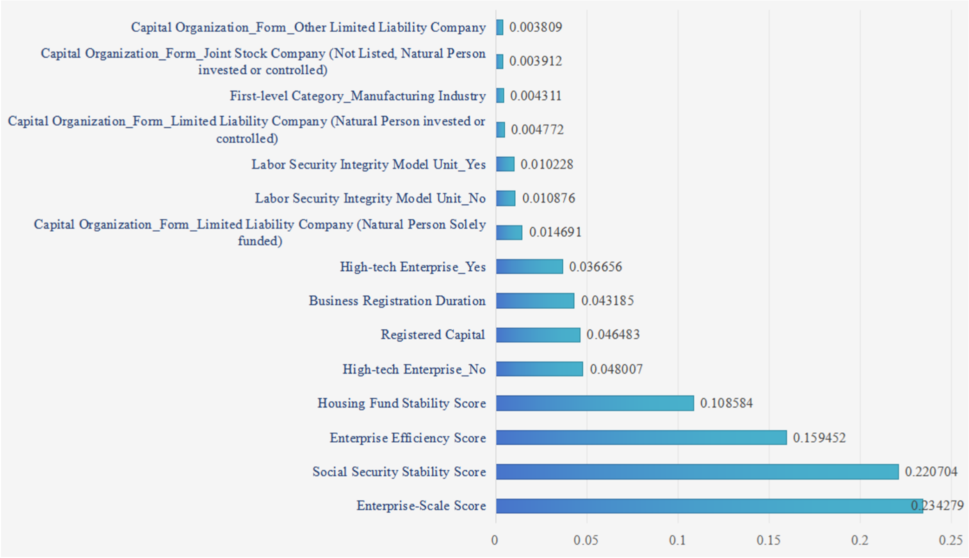

Fig. 1. Random forest classifier feature importance top (Photo/Picture credit: Original).

4 Discussion

Research has indicated that corporate governance factors are associated with financial risk, which is one of the reasons leading to increased credit risk. However, few studies have examined the impact of corporate governance on credit risk. In terms of evaluation models, individual models may suffer from overfitting, while the Random Forest model can effectively improve this shortcoming, offering good predictive performance and robustness, making it suitable for addressing credit risk assessment issues. The experimental results of this paper also reflect that the Random Forest further enhances classification performance, significantly outperforming the single decision tree in all metrics. This occurs because of its random choice of feature subsets and sample subsets, lowering the likelihood of overfitting and strengthening the model's stability and generalization abilities.

CatBoost has unique advantages in handling categorical features, and through the gradient boosting framework and innovative feature combination methods, it can fully mine information from data。 However, its performance is slightly lower than that of LightGBM. It performs quite well in handling large-scale data, with high training efficiency and accuracy.

Qiu W constructed a three-level credit rating model based on XGBoost, classifying customers into stable good customers, unstable good customers, and bad customers, which has good performance and strong generalization ability [13]. In the results of this paper, XGBoost also shows good model performance.

LightGBM achieved the highest scores in terms of accuracy, precision, recall, and F1 score, indicating its optimal comprehensive performance and generalization capability in the classification task of credit risk for small and medium-sized enterprises. This superior performance is attributed to its efficient algorithmic design and robust handling of imbalanced datasets. LightGBM reduces computational load and runs at impressive speeds. LightGBM is particularly convenient. It can handle categorical variables directly, with no encoding needed. It's also an accurate ensemble algorithm.

5 Conclusion

This study involves training machine learning models to predict corporate credit ratings using a public dataset, concluding that machine learning models exhibit robust predictive capabilities for the credit of SMEs. Among various models, the LightGBM model demonstrates the most outstanding performance and generalization ability due to its efficient algorithm and adaptability.

Overall, company size and performance, along with the stability of employee benefits, are identified as the two most critical features in credit risk assessment for SMEs. These features reflect the economic strength and operational and welfare stability of the enterprise. Larger company size corresponds to a higher enterprise risk rating, while stable employee benefits indicate a willingness of the enterprise to pay higher remuneration to its staff. Following these are the basic information of the company, including the form of capital organization, whether it is a high-tech enterprise, and whether it belongs to the manufacturing sector.

Future research may further optimize the hyperparameters of the LightGBM model to enhance its performance even more. Additionally, by integrating various data sources and feature engineering techniques, such as incorporating more non-financial features, there is potential to further improve the accuracy of credit risk classification. Moreover, considering the importance of model interpretability in credit risk management, it is worth exploring how to enhance the interpretability of complex models while maintaining their performance, to better meet practical business needs.

References

1. E. J. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Finance **23**, 589–609 (1968)
2. S. M. Lin, J. Ansell, G. Andreeva, Predicting default of a small business using different definitions of financial distress. J. Oper. Res. Soc. **63**, 539–548 (2012)
3. J. Gupta, The value of operating cash flow in modelling credit risk for SMEs. Appl. Financ. Econ. **24**, 649–660 (2014)
4. L. Cultrera, X. Brédart, Bankruptcy prediction: the case of Belgian SMEs. Rev. Account. Financ. **15**, 101–119 (2016)
5. M. Karan, Ö. Arslan, Credit risks and internationalization of SMEs. J. Bus. Econ. Manag. **10**, 361–368 (2009)
6. S. Locke, W. N. Hewa, Factors affecting the probability of SME bankruptcy: a case study on New Zealand unlisted firms. Bus. J. Entrepreneurs (2012)
7. L. Zhou, J. Lu, J. Ma, A study on the identification of early warning features in enterprise credit information based on random forest algorithm. China Inform. **6**, 57–60 (2023)
8. H. Zhong, C. Miao, Z. Shen, Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. Neurocomputing **128**, 285–295 (2014)
9. Z. Qiu, B. He, A study on the construction of credit risk assessment system based on machine learning algorithms: an analysis of individual credit risk evaluation based on China UnionPay data. Price: Theory Pract. **10**, 89–92, 194 (2021)
10. R. Chen, Research on credit risk assessment model for technology-based small and medium-sized enterprises. Bengbu: Univ. Anhui Financ. Econ. (2022)
11. D. Wei, S. Liao, Z. Zhang, Leveraging financial social media data for corporate fraud detection. J. Manag. Inf. Syst. **35**, 461–487 (2018)

12. F. Huang, T. Ma, X. Wang, Research on credit risk assessment of SMEs based on random forest model. Modern Bus. **23**, 96–99 (2024)
13. W. Qiu, Credit risk prediction in an imbalanced social lending environment based on XGBoost. In: 2019 5th Int. Conf. Big Data Inf. Anal. (BigDIA) (2019)