International Review of Financial Analysis 89 (2023) 102755

Contents lists available at ScienceDirect

International Review of Financial Analysis

journal homepage: www.elsevier.com/locate/irfa

# A two-stage credit scoring model based on random forest: Evidence from Chinese small firms

Ying Zhou [a], Long Shen [a], Laura Ballester [b,*]

[a] School of Economics and Management, Dalian University of Technology, Dalian City, Liaoning Province, China
[b] Faculty of Economics, University of Valencia, Avenida de los Naranjos s/n, 46022 Valencia, Spain

ARTICLE INFO

*JEL classification:*
C10
C61
G15
G32

*Keywords:*
Credit scoring
Small firms
Expert system
Dominance analysis

ABSTRACT

Small firms are major contributors to most economies, often supported by government policies. However, the credit scoring of small firms is complicated and costly, making it a challenging field of research. Using loan data from 3045 small firms in China, we design a two-stage expert system for default prediction that quantifies the variables and thresholds that have a key impact. Firstly, we use SMOTE to deal with the imbalanced data and secondly, we employ random forest to build predictive credit features. Dominance analysis shows that, when making default assessments on Chinese small firms, it is important to consider not only financial factors, but also non-financial and macroeconomic factors. In particular, the net cash profit, the firm's legal disputes and the per capita disposable income of urban residents are key factors in credit scoring. Robustness tests show that our proposed methodology performs better than other machine learning models, and this result is robust with observations from other countries.

## 1. Introduction

Small firms are an important part of the global economy. In China, small firms are the foundation of economic growth and have become major drivers of social progress (Kou et al., 2021). The Chinese government has taken various measures to support the growth of private companies, including increasing credit and cutting taxes and fees. Nevertheless, these companies must make major efforts to carry out some basic financial tasks, such as obtaining access to financing directly related to the evaluation of their credit risk. For large firms, certified audited financial statements are used to evaluate credit risk and support financial decision-making, such as granting loans (Tian, Yu, & Guo, 2015). For small firms, however, due to a lack of reliable data and other factors, credit risk assessment is complicated and costly for banks. On the one hand, banks use relationship lending to accumulate soft information over time to deal with the lack of credit data; on the other hand, small firms often face high costs when accessing financing due to the opacity of information and the high risk of failure (Altman & Sabato, 2007). In 2018, the Financial System Survey Report of Chinese Small Firms, issued by the People's Bank of China, noted that small businesses face various macroeconomic constraints, such as changes in the domestic economic structure and international trade friction, which have also made it more difficult for small firms to access financing. Therefore, the study of credit scoring for small firms is of great significance to alleviate the financial inequality facing such entities.

In the early days, loan institutions generally adopted methods based on the subjective experience of credit experts. With the development of statistical techniques, methods such as linear discriminant analysis, survival analysis, logit and probit regression and comprehensive evaluation are today applied to credit scoring (Altman, 1968; Bellotti & Crook, 2009a; Thomas, 2000). Due to its high interpretability, logistic regression is still the most used credit scoring model in academia and business (Audrino, Kostrov, & Ortega, 2019). However, this method has several drawbacks. It cannot capture interaction and handle high-dimensional variables, it has requirements on the distribution of the variables, and they are limited by multicollinearity. This results in an inability to identify complex patterns within millions of data points, and thus make accurate inferences and predictions. Additionally, while conventional methods may operate well during tranquil periods, they are not particularly sensitive to financial datasets characterized by high volatility and financial distress. The performance of conventional methods depends not only on the mathematical algorithms, but also on the dataset quality. Our study strives to fill this gap. We seek to develop a feature transformation approach that leverages advanced machine

* Corresponding author.
*E-mail address:* laura.ballester@uv.es (L. Ballester).

https://doi.org/10.1016/j.irfa.2023.102755
Received 31 October 2022; Received in revised form 24 April 2023; Accepted 27 June 2023
Available online 3 July 2023

learning techniques to improve the quality of the data, thereby compensating for the inadequacies of logistic regression predictions that lack precision and cannot accommodate high-dimensional variables, with a view to enhancing the effectiveness of default discrimination. We use random forests (Breiman, 2001; Dumitrescu, Hué, Hurlin, & Tokpavi, 2022) to construct credit scoring. Specifically, this transformation is based on traditional decision tree methods and included in the latest machine learning techniques that have already been widely adopted (Alonso-Robisco & Carbó, 2022). Our data transformation methodology involves the use of random forests, which construct predictive credit features by combining decision rules along decision pathways from root nodes to terminal nodes.

The objective of the paper is to design a two-stage expert system for small firm credit scoring in China using data from 3045 small business loans and 81 variables including financial, non-financial and macroeconomic factors. The first stage involves calculating the default probability. We use the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic defaulted samples to deal with the imbalanced data from small firms and reduce the misevaluation of defaulted firms. In the second stage, we utilize the random forest to establish predictive credit features that serve as novel explanatory variables in logistic regression. This methodology enables us to capture potential non-linear effects in credit scoring data while preserving the intrinsic interpretability of logistic regression. Finally, dominance analysis allows us to identify the relevant variables that predict small business default in China.

Our contributions to the existing literature are as follows. First, we contribute to the most recent credit scoring literature (Liu et al., 2022; Yu, Yao, Zhang, Yin, & Liu, 2020; Zhu, Zhang, Wu, Li, & Li, 2022) in terms of methodology. We employ advanced machine learning techniques to construct sparse credit features with default discrimination ability based on a large number of variables, which allows the logistic regression to accommodate high dimensional variables. In addition, the method can capture non-linear effects that can enhance the credit scoring data to improve model performance. We have also expanded the sample of defaulted firms to improve the accuracy. Second, this study contributes to the growing literature on factors impacting firm default status (Cathcart, Dufour, Rossi, & Varotto, 2020). Most credit scoring studies use a limited set of potential explanatory variables, typically financial ratios, market price variables, or a combination of both. We consider seven types of firm characteristics that include financial, non-financial and macroeconomic factors, increasing the number of variables commonly used in the credit scoring literature. Finally, this research extends the literature on the credit scoring of small firms in emerging markets (Kou et al., 2021). Emerging market financial systems, financing guarantee systems and small and medium-sized enterprises (hereafter referred to as SMEs) are very different from those in developed countries. Existing theories neither explain nor generalize the emerging market scenario well. We explore an appropriate credit scoring system for small firms in China based on the theory of information asymmetry, which helps to alleviate the financial inequality facing small firms.

From the training sample data obtained by SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) in the first stage, random forest constructs a total of 771 credit features based on 81 variables that correspond to seven types of firm characteristics including financial, non-financial and macroeconomic factors in the second stage. The magnitude of the absolute value of the logistic regression coefficients identifies the credit features that have a critical impact on the default status of Chinese small firms. We conclude that the most important variables are the cash component of net profits (financial factor), the firm's legal disputes (non-financial factor) and the per capita disposable income of urban residents (macroeconomic factor). Our findings reinforce the existing literature, which has found significant impact variables on corporate credit risk in other countries (Carling, Jacobson, Lindé, & Roszbach, 2007; Henisz & McGlinch, 2019; Tobback, Bellotti, Moeyersoms, Stankova, & Martens, 2017). Therefore, when making default judgments about Chinese small firms, it is important to consider not only financial factors but also non-financial and macro factors that are found to be relevant. The robustness results show that our proposed methodology outperforms other machine learning models and that this result is robust with samples from other countries.

The remainder of this paper is organized as follows. In the next section we present the literature review. Section 3 shows the data employed in the analysis. Section 4 contains the methodology applied. Section 5 presents the empirical results for small firms in China, while Section 6 summarizes the robustness tests. Specifically, we apply our proposed method to other economies and compare it to popular machine learning methods. Finally, Section 7 provides the conclusions.

## 2. Literature review

In this section, we present an overview of the literature most related to this study.

### 2.1. Credit scoring literature for small firms

According to Calabrese, Marra, and Angela Osmetti (2016), the efficiency of the credit system would increase if banks were better able to predict whether small firms will default on their loans. Therefore, a thorough credit risk assessment is essential to address the issue of small firms' access to financing. The two main approaches to assessing the credit risk of small firms are to identify the variables that influence credit risk and to estimate the likelihood of default. Since financial variables usually reflect the financial performance and repayment capacity of small and medium-sized companies, many researchers have focused on the role of financial characteristics in assessing SME credit risk. Behr and Guettler (2007) develop a scoring model using logit analysis based on a unique data set of loans made to German SMEs between 1992 and 2002, which allows SMEs to monitor their bank's pricing behavior and reduces information asymmetries between lenders and borrowers. Altman and Sabato (2007) develop a distress prediction model specifically for the SME sector and analyze its effectiveness compared to a generic corporate model using a logit regression technique for US firms over the period 1994–2002. They find that five financial ratios perform best in aggregate in predicting SME default.

The informational opacity of SMEs has been identified as a key challenge in earlier research, despite the fact that financial ratios offer a meaningful and objective mathematical representation of SME performance (Voulgaris, Doumpos, & Zopounidis, 2000). The financial statements of many SMEs may be partial or unaudited because they are not listed on financial exchanges. As a consequence of the restricted volume and scope of hard information, numerous studies have highlighted the substantial impact of non-financial information on the estimation of SME credit risk. Keasey, Pindado, and Rodrigues (2015) study the determinants of the expected costs of financial distress for a sample of European manufacturing SMEs for the period 1999–2006. They find that a firm's *ex ante* financial distress expenses depend not only on the risk of financial distress, but also on factors affecting the duration and cost of insolvency proceedings.

Financial data and firm-specific information are combined with legal decisions by Yin, Jiang, Jain, and Wang (2020) to assess SME credit risk in China. They conclude that characteristics extracted from persuasive court rulings greatly improve the performance of default discrimination. Tobback et al. (2017) gather data from Belgian and UK firms between 2011 and 2014 to build a bankruptcy prediction ensemble model that combines output scores from the relational model with structured data. They find that the relational model gives improved predictions over a simple financial model in detecting the riskiest firms. Cathcart et al. (2020) analyze a sample of firm-year data from large companies and SMEs in six European nations between 2005 and 2015 to determine the effect of leverage and various funding sources on default risk. They

observe that financial leverage affects the likelihood of default for SMEs more than for large enterprises. Ciampi, Cillo, and Fiano (2020) combine corporate prior payment behavior and Kohonen maps to predict small firm default using a sample of Italian small firms using logistic regression and discrete-time hazard models. They find that financial factors are not the only determinants of business insolvency. In a case where financial (accounting) data are not needed, Kou et al. (2021) propose a bankruptcy prediction model for SMEs that combines transactional data and payment network-based factors. Test results, both online and offline, support the economic value and predictive power of transactional data-based factors.

### 2.2. Imbalanced data processing literature

In credit scoring, imbalanced data is common due to the small proportion of "poor firms" in default compared to the total number of firms. Imbalanced data can lead to inaccurate default discrimination. Many methods are currently used to solve the imbalanced data problem, including sampling methods (Nasir, Simsek, Cornelsen, Ragothaman, & Dag, 2021), cost-sensitive learning models (Audrino et al., 2019; Liu et al., 2022) and ensemble learning models (Shahabadi, Tabrizchi, Rafsanjani, Gupta, & Palmieri, 2021). Among these, sampling methods are more frequently used. Sampling methods are based on data-level imbalance processing methods, which mainly include oversampling and undersampling.[1] This method saves computational effort but leads to data information loss and training model inaccuracy. Therefore, as a first stage, this paper adopts the Synthetic Minority Oversampling Technique (SMOTE) based on virtual sample synthesis, which has been widely adopted (Chawla et al., 2002; Haixiang et al., 2017).

### 2.3. The generation of decision rules literature

In recent years, researchers have also used machine learning methods such as neural networks (Baesens, Setiono, Mues, & Vanthienen, 2003), support vector machines (Bellotti & Crook, 2009b; Yu et al., 2020) and random forests (Dumitrescu et al., 2022; Zhu et al., 2022) to construct credit scoring. The decision rule generation methods used in existing research mostly employ nonlinear classification models to transform the original features into credit features. When Baesens et al. (2003) conducted credit risk assessment, they used the fully connected neural network to extract decision rules from the original variables and visualized the rules as decision tables using three public credit datasets. These explanatory rules will explain the learned knowledge embedded in the neural network, which can help credit risk managers explain why particular loan applicants are classified as good or bad. Mues, Baesens, Files, and Vanthienen (2004) considered that the decision tree may become too large to be understood by experts due to the inherent replication of the isomorphic subtree. The neural network is used to transform original variables into a series of decision rules, and the generated rules are simplified, which improves the intuitiveness and availability of knowledge using three public credit datasets.

Wu and Hsu (2012) proposed an enhanced decision support model (EDSM) that combines decision trees with the relevance vector machine using 3069 public electronic companies in Taiwan from 2005 to 2009. Decision trees are used to overcome the opacity of relevance vector machines, generate knowledge as logical statements, and help to interpret the predicted results. Setiono, Baesens, and Mues (2011) removed the network connection from the input unit to the hidden unit through a simple pruning process using a German credit dataset. They not only found a neural network with good prediction performance, but also clearly explained the prediction results obtained by the neural network. Huang, Tsaih, and Yu (2014) proposed a hierarchical growth self-organizing map to discover the topological pattern of fraudulent financial reporting in China from 1992 to 2008 and carried out the effective detection of fraudulent financial reporting and feature extraction. The above scholars have conducted a deep study on the generation of decision rules and made considerable progress, but they are directly using the generated rules to make decisions.

The random forest method, that is, the randomized version of bagged decision trees (Breiman, 2001), significantly outperforms logistic regression and has gradually evolved into one of the main models in the credit scoring sector (Liu et al., 2022). Lessmann, Baesens, Seow, and Thomas (2015) compared 41 algorithms using various evaluation criteria and different credit rating datasets and found that the random forest method was superior in default discrimination capability among the multitude of default prediction methods and has gradually developed into one of the main methods in the field of credit scoring. Butaru et al. (2016) used a large credit card delinquency dataset collected by a US financial regulatory agency over a six-year period for credit scoring. They found that random forest yielded a 6% gain in recall compared to logit. Moreover, random forest has greater interpretability compared to other advanced machine learning algorithms (Uddin, Chi, Al Janabi, & Habib, 2022). However, although random forest has been found to outperform logistic regression and has greater interpretability compared to other advanced machine learning algorithms, it is still important to address the issue of explainability and interpretability in machine learning approaches used in the credit scoring industry. Many of these algorithms, particularly ensemble approaches (including random forest), are often viewed by customers and regulators as "black boxes" because it is difficult to understand how the corresponding scorecards and credit approval process work. This concern aligns with the current focus of financial authorities on controlling artificial intelligence and the need for interpretability, particularly in the credit rating sector.

Our study fills this gap. We propose a hybrid credit scoring technique to maintain the inherent interpretability of the scoring model while enhancing the predictive performance of the logistic regression model by data pre-processing and feature engineering. The difference between this paper and the logistic regression approach to credit scoring is that the variables used in this study are the optimal credit features obtained by generalizing the decision rules from the root node to the leaf nodes of the random forest.

## 3. Data

### 3.1. Data collection and preparation

The variables used in this paper come from the credit management system of small firms of China Regional Commercial Bank (Uddin et al., 2022).[2] Specifically, 3045 small firms from the last 20 years were

[1] The undersampling method draws some firms from a sample of non-defaulting firms (the majority class), and the sampled firms and the defaulting firms (the minority class) together form a balanced data set. The main idea of the undersampling method is to remove some of the non-defaulting firms so that the sample size of non-defaulting firms in the data set is closer to the sample size of defaulting firms.

[2] The bank is a large state-owned commercial bank headquartered in Dalian, China. With a strong branch network and >20,000 employees, the bank has established a comprehensive service network that covers the entire Liaoning Province. As of the end of June 2022, its total assets were 475.5 billion yuan, the balance of various loans was 255.3 billion yuan, and the balance of various deposits was 299.3 billion yuan. In July 2022, *The Banker* magazine ranked the bank 313th among the top 1000 largest banks in the world. Given the bank's size and extensive branch network, as well as its state-owned status, it clearly plays a significant role in the local economy and financial sector. Therefore, our data is highly representative of Chinese small firms.

selected.[3] This sample includes 2995 non-defaulted firms and 50 defaulted firms, and includes construction, wholesale and retail trade, and 12 other industries. We selected the variables based on a thorough review of the literature and expert opinions covering all aspects of the firm, including its management, consumers, the economic sector and the public sector. We considered seven types of firm characteristics, including financial, non-financial and macroeconomic factors. Specifically, (1) internal financial factors, such as the ratio of liabilities to assets, the operating cash flow ratio, and the growth rate of retained earnings; (2) internal non-financial factors, such as the years of employment in related industries, auditing, and the status of the branded products held by a firm owner; (3) the basic situation of the firm's legal representative, such as university degree, age, loan default records and monthly family income; (4) the basic credit situation of the firm, including the firm's registered fund category, corporate credit in last three years; (5) the firm's reputation, such as firm tax records, firm legal disputes, and the compliance status of the firm's operations; (6) macro conditions, such as the industry prosperity index, per capita savings of urban and rural residents, and GDP growth rate; and (7) pledge guarantee factors, including the pledge score. The data for macro indicators were obtained from the official website of the National Bureau of Statistics of the People's Republic of China and the CNKI database of China's economic and social development statistics and was manually merged by us.

The firm characteristics are listed in Table 1, and the 82nd row shows the actual default state identification of small firms, in which the defaulted state identification is 1 and the non-defaulted state identification is 0. For qualitative variables, this paper scores the different states of qualitative variables based on existing research (Kou et al., 2021) and expert opinions. The qualitative variables are quantified between [0,1]. The better the credit status, the higher the score given to the variable state. The scoring criteria for 26 observable qualitative variables are presented in Table 2. The first 81 rows in Column *e* of Table 1 are the original data for the quantitative variables and the numerical data obtained according to the qualitative variable scoring criteria in Table 2.[4] Table S1 in the Annex contains all the 81 variables with their corresponding definitions.

The 81 variables included in Column *e* of Table 1 are randomly divided into training samples and test samples according to the proportion of defaulted and non-defaulted firms. The proportion of training samples was 60%, and the total number was $3045 \times 60\% = 1827$, including 1797 non-defaulted firms and 30 defaulted firms. The proportion of test samples was 40%, with a total of $3045 \times 40\% = 1218$, of which 1198 were non-defaulted firms and 20 were defaulted firms.

### 3.2. Imbalanced sample processing based on SMOTE

The number of defaulted firms in small firm samples is far less than that of non-defaulted firms. In the first stage, we use the Synthetic Minority Oversampling Technique (SMOTE) to process the imbalanced training samples (Chawla et al., 2002). The advantage of SMOTE is that a new data vector is generated by a random linear interpolation between two defaulted firms with a close Euclidean distance.[5] Thus, several defaulted firms are generated to achieve the balance of the overall training samples and ensure that the generated defaulted firm samples are similar to the existing defaulted firm samples, which guarantees the accuracy of the default prediction model for defaulted firm discrimination, reducing the type II error.[6] The virtual defaulted firm sample (1769) that is equal to the training non-defaulted firm sample (1797) minus the training defaulted firm sample (30) is synthesized by the SMOTE algorithm, and a new training sample is formed with the original training sample. The unbalanced sample is processed by SMOTE to equalize the number of defaulted and non-defaulted firms in the sample. The defaulted firm sample consists of the original defaulted firm sample and the newly generated dummy defaulted firm samples.

## 4. Methodology

### 4.1. Credit feature construction based on Random Forest

The second stage of the estimation process is to apply random forest. Random forest is an extension of the bagging algorithm based on constructing a bagging ensemble with decision trees, which further introduces random sample selection and random variable selection in the training process of decision trees (Breiman, 2001). For a decision tree, given $N$ firms, we use bootstrap sampling to randomly select a sample of firms in the sampling set, and after sampling $N$ times, the training sample $D$ of the decision tree is obtained. The decision tree learning algorithm recursively selects the optimal variable and divides the training samples according to the variable, so that each sub-sample has the best classification process. In this paper, the decision tree uses the Gini Impurity to select the optimal feature and determine the critical point of the optimal binary segmentation of the feature.[7]

There are three criteria for determining leaf nodes. First, the node cannot continue to split, that is, only one sample is included in the node. Second, all samples in the node belong to the same category, and the Gini Impurity of the node sample is 0. Third, to ensure that the credit feature constructed by random forest can be interpreted robustly, the decision rules constructed from root node to leaf node must not be too complex. This study uses a pre-pruning strategy to set the maximum depth of the decision trees. By analogy, a total of $n$ decision trees are constructed to obtain a random forest, $n = 100$ in this study (default setting in python's scikit-learn package). Due to the different training samples used in the construction of all decision trees, the variables used are different, which ensures the diversity of the decision trees. The decision rules constructed between each decision tree must be very different.

Based on the idea of constructing credit features in the random forest, a decision tree divides a sample according to the decision rules and falls on a leaf node. The decision rule from the root node to the leaf node represents the variable state obtained by the optimal binary segmentation point. The sample satisfies different states of all the variables on the decision path, and the various states of the different variables together constitute a new credit feature. The number of leaf nodes in the random forest is the number of decision paths and credit features. Credit features extracted from the random forest model are to replace the firm samples one by one in the random forest model trained. The interaction term of the specific state of the variable on the decision path from the root node to the leaf node of the decision tree in the random forest model is regarded as a credit feature. The details are as follows.

Suppose that the set of decision trees in the random forest is $T = \{T_1, T_2, T_3, \ldots, T_n\}$, and $n$ is the number of decision trees in the random forest

[3] According to the small and medium-sized firm standard specification of the Chinese Ministry of Industry and Information Technology, the National Bureau of Statistics of China, the National Development and Reform Commission, and the Ministry of Finance.

[4] Since random forests are not sensitive to the distribution of firm samples and the dimension between different variables, all variables are not normalized and the extreme values are processed.

[5] Detailed information regarding SMOTE is presented in section 1 of the Annex.

[6] Type II error indicates the percentage of defaulted firms misclassified as non-defaulted firms by dividing the number of defaulted firms misclassified by the total number of defaulted firms. A bank can lose much more by accepting a "bad firm" than by rejecting a "good firm".

[7] Detailed information regarding the Gini Impurity for the random forest method is presented in section 2 of the Annex.

**Table 1**
Original data after scoring small firm qualitative variables.

| (*a*) No. | (*b*) Categories | (*c*) Variables | (*d*) Variable types | (*e*) Original Data $v_{ij}$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | (1) Firm 1 | (2) Firm 2 | ... | (3044) Firm 3044 | (3045) Firm 3045 |
| 1 | $C_1$ Internal financial factors | $X_1$ Liabilities to assets ratio | Negative | 0.640 | 0.624 | ... | 0.182 | 0.346 |
| 2 | | $X_2$ Operating cash flow ratio | Positive | 1.125 | 0.000 | ... | 0.198 | 0.349 |
| ... | | ... | ... | ... | ... | ... | ... | ... |
| 19 | | $X_{19}$ Net cash flow ratio | Positive | 1.124 | 0.000 | ... | 0.198 | 0.349 |
| 20 | | $X_{20}$ EBITDA-to-total liability ratio | Positive | 0.031 | 0.000 | ... | 0.017 | 0.631 |
| 21 | | $X_{21}$ Return on equity | Positive | 0.055 | 0.000 | ... | 0.003 | 0.224 |
| 22 | | $X_{22}$ Net present value of sales | Positive | 0.916 | 0.000 | ... | 0.032 | 0.054 |
| ... | | ... | ... | ... | ... | ... | ... | ... |
| 32 | | $X_{32}$ Net cash flows from operations | Positive | 34,074,356.977 | 0.000 | ... | 131,991.370 | 2,545,297.450 |
| 33 | | $X_{33}$ Subtotal of cash inflows from operations | Positive | 62,816,920.073 | 10.000 | ... | 4,075,102.400 | 53,245,183.750 |
| 34 | | $X_{34}$ Accounts receivable turnover ratio | Positive | 1.937 | 0.048 | ... | 5.070 | 3.000 |
| 35 | | $X_{35}$ Inventory turnover rate | Positive | 3.000 | 0.004 | ... | 20.330 | 3.000 |
| ... | | ... | ... | ... | ... | ... | ... | ... |
| 42 | | $X_{42}$ Accounts payable turnover ratio | Negative | 7.483 | 1,026.371 | ... | 1,026.371 | 85.402 |
| 43 | | $X_{43}$ Cash conversion cycle | Positive | −2.546 | −1,412.309 | ... | −1,412.309 | −79.402 |
| 44 | | $X_{44}$ Growth rate of business revenue | Positive | 0.000 | 0.000 | ... | 0.000 | 4.085 |
| 45 | | $X_{45}$ Profit growth rate | Positive | 0 | 0 | .. | 0 | 5 |
| 46 | | $X_{46}$ Total asset growth rate | Positive | 0.041 | 0.000 | ... | −0.638 | −0.036 |
| 47 | | $X_{47}$ Capital accumulation rate | Positive | 0 | 0 | ... | 1.15 | 0 |
| 48 | | $X_{48}$ Retained earnings growth rate | Positive | 1.809 | 0.034 | ... | −1.041 | 0.888 |
| 49 | $C_2$ Internal non-financial factors | $X_{49}$ Years of employment of the firm's legal representative | Qualitative | 1 | 1 | ... | 1 | 0 |
| 50 | | $X_{50}$ Audit status | Qualitative | 0 | 0 | ... | 0 | 0 |
| ... | | ... | ... | ... | ... | ... | ... | ... |
| 56 | | $X_{56}$ Status of the company's branded products. | Qualitative | 0 | 0 | . | 0 | 0 |
| 57 | | $X_{57}$ Proportion of the total amount of money returned by firms through the bank | Positive | 0.000 | 0.000 | ... | 0.018 | 0.597 |
| 58 | $C_3$ Basic situation of the firm's legal | $X_{58}$ Degree | Qualitative | 0 | 0 | ... | 0.4 | 0.7 |
| 59 | representative | $X_{59}$ Legal representative loan default records | Qualitative | 0 | 0 | ... | 0.7 | 0.7 |
| ... | | ... | ... | ... | ... | ... | ... | ... |
| 67 | | $X_{67}$ Family monthly income | Positive | 0 | 0 | ... | 5000 | 2000 |
| 68 | | $X_{68}$ Time of holding position | Qualitative | 0 | 0 | ... | 0 | 0 |
| 69 | $C_4$ Basic credit of firm | $X_{69}$ Firm Registered Fund Category | Qualitative | 0 | 0 | ... | 1 | 1 |
| 70 | | $X_{70}$ Corporate Credit in last 3 Years | Qualitative | 0 | 0 | ... | 1 | 1 |
| 71 | $C_5$ Reputation of firm | $X_{71}$ Firm tax records | Qualitative | 0 | 0 | ... | 1 | 0.25 |
| 72 | | $X_{72}$ Firm's legal disputes | Qualitative | 0.2 | 0.2 | ... | 1 | 1 |
| 73 | | $X_{73}$ Compliance status of the firm's operations | Qualitative | 0 | 0 | ... | 0.5 | 0.5 |
| 74 | | $X_{74}$ Number of contract breaches involving the firm | Qualitative | 0 | 0 | ... | 1 | 1 |
| 75 | $C_6$ Macro conditions | $X_{75}$ Industry prosperity index | Positive | 127.960 | 151.810 | ... | 134.750 | 134.750 |
| 76 | | $X_{76}$ Urban and rural residents' per capita savings at the end of the year | Positive | 23,388.050 | 23,388.050 | ... | 23,388.050 | 23,388.050 |
| 77 | | $X_{77}$ GDP growth rate | Positive | 15.200 | 15.200 | ... | 15.200 | 15.200 |
| 78 | | $X_{78}$ Consumer Price Index | Interval [101,105] | 100.600 | 100.600 | ... | 100.600 | 100.600 |
| 79 | | $X_{79}$ Per capita disposable income of urban residents | Positive | 9,101.350 | 9,101.350 | ... | 9,101.350 | 9,101.350 |
| 80 | | $X_{80}$ Engel coefficient | Negative | 39.400 | 39.400 | ... | 39.400 | 39.400 |
| 81 | $C_7$ Pledge guarantee factors | $X_{81}$ Pledge Score | Qualitative | 0.100 | 0.190 | ... | 0.610 | 0.350 |
| – | – | State of default | – | 1 | 1 | ... | 0 | 0 |

This table presents the original data after scoring small firms for qualitative variables. Column *a, b, c, d* and *e* show the serial number, the variable category, the variable name, the variable type and the original data of the firm, respectively. Due to space limitations, some variables are omitted.

**Table 2**
Scoring criteria for qualitative variables.

| (*a*) No. | (*b*) Category | (*c*)Variables | (*d*) Score standard | (*e*) Score values |
|---|---|---|---|---|
| 1 | Financial information | Employment years of relevant industries | 1. Working years ≥8 years | 1.00 |
| | | | 2. 5 years ≤Working years<8 years | 0.70 |
| | | | 3. 2 years ≤Working years<5 years | 0.40 |
| | | | 4. 0 ≤ Working years<2 years, or missing data | 0.00 |
| ... | | ... | ... | ... |
| 10 | Basic situation of leaders | Time of holding position | 1. Time of loan origination - Time of holding position≥8 years | 1.00 |
| | | | 2. 5 years ≤Time of loan origination-Time of holding position<8 years | 0.70 |
| | | | 3. 2 years ≤Time of loan origination-Time of holding position<5 years | 0.40 |
| | | | 4. 0 ≤ Time of loan origination-Time of holding position<2 years or missing data | 0.00 |
| ... | | ... | ... | ... |
| 21 | Credit records | Registered funds category | 1. Fund registered funds | 1.00 |
| | | | 2. Funds and physical registered funds | 0.80 |
| | | | 3. Physical registered funds | 0.60 |
| | | | 4. Other, or missing data | 0.00 |
| ... | | ... | ... | ... |
| 23 | Firm Reputation | Tax records | 1. Tax history of 3 years and more without arrears | 0.00 |
| | | | 2. Tax records for <3 years without arrears | 1.00 |
| | | | 3. No tax records | 0.75 |
| | | | 4. Two or more tax arrears or missing data | 0.25 |
| ... | | ... | ... | ... |
| 26 | | Number of contract breaches involving the firm | 1. zero times | 1.00 |
| | | | 2. one time | 0.60 |
| | | | 3. two times | 0.30 |
| | | | 4. three times | 0.00 |

This table presents the scoring criteria for the 26 qualitative variables. The better the credit status corresponding to the firm state described by the variable, the higher the corresponding score. Columns *a*, *b*, *c*, *d* and *e* represent the ordinal number of the variable, the category to which the variable belongs, the variable name, the scoring criteria for the variable and the score, respectively. Due to space limitations, some variables are omitted.

model. The value of sample $x_i$ of the $i$th firm falling into the leaf node of the decision tree $T_r$ is $K_r$, $K_r = \{K_{r1}, K_{r2}, K_{r3}, \ldots, K_{r,mr}\}$, $K_{rs} \in \{0,1\}$, and $m_r$ is the number of leaf nodes in $T_r$. When the sample $x_i$ is input into the decision tree $T_r$ and falls on the sth leaf node, $K_{rs} = 1$, otherwise $K_{rs} = 0$. The number of decision paths in random forests is the number of all leaf nodes, and the number of new credit features. Therefore, the instance $x_i$ of the ith firm is divided according to the trained random forest, and the new sample data $X_i^* = \{X_{i,1}^*, X_{i,2}^*, X_{i,3}^*, \ldots, X_{i,q}^*\}$ of the $i$th firm can be obtained, where $q$ is the number of new credit features.

Variables with significant distinguishing power between default and non-default are used by the minimum priority of the Gini Impurity to make the decision rule. The credit rating variable system consisting of the interaction term of the variable state is established, which makes up for the fact that the interaction between variables cannot be captured using logistic regression alone.[8]

**Table 3**
Confusion matrix for credit scoring.

| Actual status | Predicted status | | |
|---|---|---|---|
| | Positive (Default) | Negative (Non-default) | Total |
| Positive (Default) | TP (True Positive): The number of positive instances correctly classified as positive ones | FN (False Negative): The number of positive instances incorrectly classified as negative ones | Total number of actual positive instances (TP + FN) |
| Negative (Non-default) | FP (False Positive): The number of negative instances incorrectly classified as positive ones | TN (True Negative): The number of negative instances correctly classified as negative ones | Total number of actual negative instances (FP + TN) |
| Total | Total number of predicted positive instances (TP + FP) | Total number of predicted negative instances (FN + TN) | Total number of instances |

This table presents the corresponding confusion matrix with the defaulted or non-defaulted repayment status of small firms. TP indicates a true positive, where the actual status is non-defaulted and the predicted status is non-defaulted. FP stands for a false positive, where the actual status is non-defaulted but the predicted status is defaulted. TN means a true negative, where the actual status is defaulted and the predicted status is defaulted. FN denotes a false negative, where the actual status is defaulted but the predicted status is non-defaulted.

### 4.2. Credit scoring equation based on logistic regression

Let $p_i$ be the default probability of the $i$th firm; $y_i$ is the default state of the $i$th firm, where default is 1 and non-default is 0; $\boldsymbol{x}$ are the firm variables, $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{im})$; $b$ is a constant term; $m$ is number of variables of firms; $w_j$ is the coefficient of the $j$th variable; and $x_{ij}$ is the variable $j$ of firm $i$. The logistic regression model for credit scoring (Audrino et al., 2019) characterizes the nonlinear relationship between the default probability $p_i$ of the $i$th firm and variable data $\boldsymbol{x_i}$.

$$p_i = p(y_i = 1|x) = \left[1 + exp\left(-\left(b + \sum_{j=1}^{m} w_j x_{ij}\right)\right)\right]^{-1} \quad (1)$$

By substituting the data $x_{ij}$ of the $i$th firm ($i = 1, 2, \ldots, n$) and the $j$th variable ($j = 1, 2, \ldots, m$) into Eq. (1), the theoretical value $p_i$ of each group of variables $\boldsymbol{x_i}$ for the judgment of the firm's default state can be calculated. When $p_i$ is greater than the critical point of credit scoring 0.5, the firm is judged to be defaulted; if this is not the case, it is considered non-defaulted, as shown in Eq. (2).

$$\hat{y}_i \begin{cases} 1 \; p_i \geq 0.5 \\ 0 \; p_i < 0.5 \end{cases} \quad (2)$$

Similarly, the discriminant values $\hat{y}$ of the logistic regression model for the default status of all firms are obtained. By comparing the predicted value $\hat{y}$ with the actual default state $y$ of the firm, the confusion

[8] The economic significance of using random forest to transform features is as follows. The newly constructed credit features are composed of all decision rules on a decision path in random forest, which is the combination of different feature states of the original features. If a firm sample satisfies the conditions constrained by the decision rules on the decision path, the credit feature of the sample is valued as 1, otherwise it is valued as 0, which is convenient for further studying the interaction between features in the construction of the credit scoring model.

matrix of credit scoring can be obtained as shown in Table 3. We will use parameters TP, FN, FP and TN[9] to calculate some performance measures.

### 4.3. Variable importance analysis

We use dominance analysis (DA)[10] to identify the relative importance of the original variables (OV) (Luchman, 2014). DA calculates the relative variable importance (RVI) based on the contribution of each OV to the prediction using a series of outcomes that deconstruct and evaluate each contribution to an overall default discrimination accuracy (AUC). RVI indicates the importance of the variable; the larger the RVI, the stronger the default discriminatory power of the variable.

### 4.4. Model performance

To check the precision of our approach, we use some performance measures:

(1) *Precision* represents the ratio of defaulted firms to be judged correctly.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

(2) *Accuracy* represents the ratio that correctly identifies all defaulted and non-defaulted firms.

$$Accuracy = \frac{TP + TN}{TN + FP + TP + FN} \tag{4}$$

(3) *G-mean* represents the geometric mean of the ratio of the non-defaulted sample judgment rate to the correct ratio of the defaulted sample judgment rate.

$$G - mean = \sqrt{\frac{TN}{TN + FP} \bullet \frac{TP}{TP + FN}} \tag{5}$$

(4) *Mathew correlation coefficient (MCC)* measures the correlation of the true classes $y$ with the predicted labels $\hat{y}$. *MCC* is a robust metric that summarizes the classifier performance in a single value, if defaulted and non-defaulted cases are of equal importance.

$$MCC = \frac{cov(y, \hat{y})}{\sigma_y \bullet \sigma_{\hat{y}}} = \frac{(TP \bullet TN - FP \bullet FN)}{\sqrt{(TP + FP) \bullet (TP + FN) \bullet (TN + FP) \bullet (TN + FN)}} \tag{6}$$

(5) *The Kolmogorov-Smirnov value* (*KS*) represents the extent to which the model can distinguish between defaulted and non-defaulted firms.

$$KS = MAX\left(\frac{TP}{TP + FN} - \frac{FP}{FP + TN}\right) \tag{7}$$

(6) *Bookmaker Informedness (BM)* and (7) *Markedness (MK)* are comprehensive evaluation measures. *BM* denotes the sum of the percentage correct for defaulting customers and the percentage correct for non-defaulting customers minus 1. *MK* denotes the percentage correct for the sample predicted to be defaulting customers versus the percentage correct for the sample predicted to be non-defaulting customers minus 1. These two metrics treat defaulted and non-defaulted customers as equally important.

$$BM = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \tag{8}$$

$$MK = \frac{TP}{TP + FP} + \frac{TN}{TN + FN} - 1 \tag{9}$$

Next, we compute (8) the *AUC,* the Area Under the ROC (receiver operating characteristic) Curve, which is a widely used evaluation criterion when considering imbalanced data.[11]

$$AUC = \int_0^1 \frac{TP}{TP + FN} \bullet \frac{FP}{TN + FP} \tag{10}$$

We then compute the average performance ranking ($AR_j$) of the $j$th credit scoring model, where $D$ represents the number of datasets and $r_i^j$ represents the ranking of the $j$th credit scoring model on the $i$th data set.

$$AR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j \tag{11}$$

Finally, to test the difference between the performance measures of the model, we compute the Friedman test (Friedman, 1937) and also the post-hoc Bonferroni-Dunn test (Dunn, 1964) to demonstrate the superiority of the proposed model. The Friedman test contrasts the null hypothesis that there are no differences in the rankings of the performance measures across different credit scoring models against the alternative that the performance measure rankings across different credit scoring models are different.

$\chi_F^2$ represents the Friedman test statistics, where $D$ represents the number of used datasets and $K$ represents the number of credit scoring models used.

$$\chi_F^2 = \frac{12D}{K(K+1)}\left[\sum_{j=1}^{K} AR_j^2 - \frac{K(K+1)^2}{4}\right] \tag{12}$$

The post-hoc Bonferroni-Dunn test contrasts the null hypothesis that there are no differences in the ranking of performance measures between two models against the alternative that there are differences in the ranking of performance measures between two models. $Z_{ij}$ represents the post-hoc Bonferroni-Dunn test statistics, where $N$ is the total number of observations across all groups, $T_S$ is the number of observations tied at the sth specific tied value, and $n_i$ is the total number of observations in the $i$th group.

$$Z_{ij} = \frac{AR_i - AR_j}{\sqrt{\frac{N(N+1)}{12} - \left[\sum T_S^3 - \frac{T_S}{12(N-1)}\Big/\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right]}} \tag{13}$$

[9] TP indicates a true positive, where the actual status is non-defaulted and the predicted status is non-defaulted. FP stands for a false positive, where the actual status is non-defaulted but the predicted status is defaulted. TN denotes a true negative, where the actual status is defaulted and the predicted status is defaulted. FN means a false negative, where the actual status is defaulted but the predicted status is non-defaulted.

[10] DA is a conceptual expansion of Shapley value decomposition from Collaborative Game Theory, which aims to develop ways to split "payoffs" among "players" based on their relative contributions. The payoff is normally taken to be a single value. The DA implementation of Shapley value decomposition works using a brute force method. The data are used to estimate sub-models that reflect all conceivable combinations of included or excluded OVs, and the AUC associated with each sub-model is gathered. If the pre-selected model has $p$ OVs, obtaining all feasible combinations of the OVs yields $2^p$ sub-models to be evaluated.

[11] The AUC is a simple and widely used metric for judging the discriminatory power of credit scoring. The AUC ranges from 50% (purely random prediction) to 100% (perfect prediction). Following Iyer, Khwaja, Luttmer, and Shue (2016), an AUC of 60% is generally considered desirable in information-scarce environments, whereas AUCs of 70% or greater are the goal in information-rich environments.

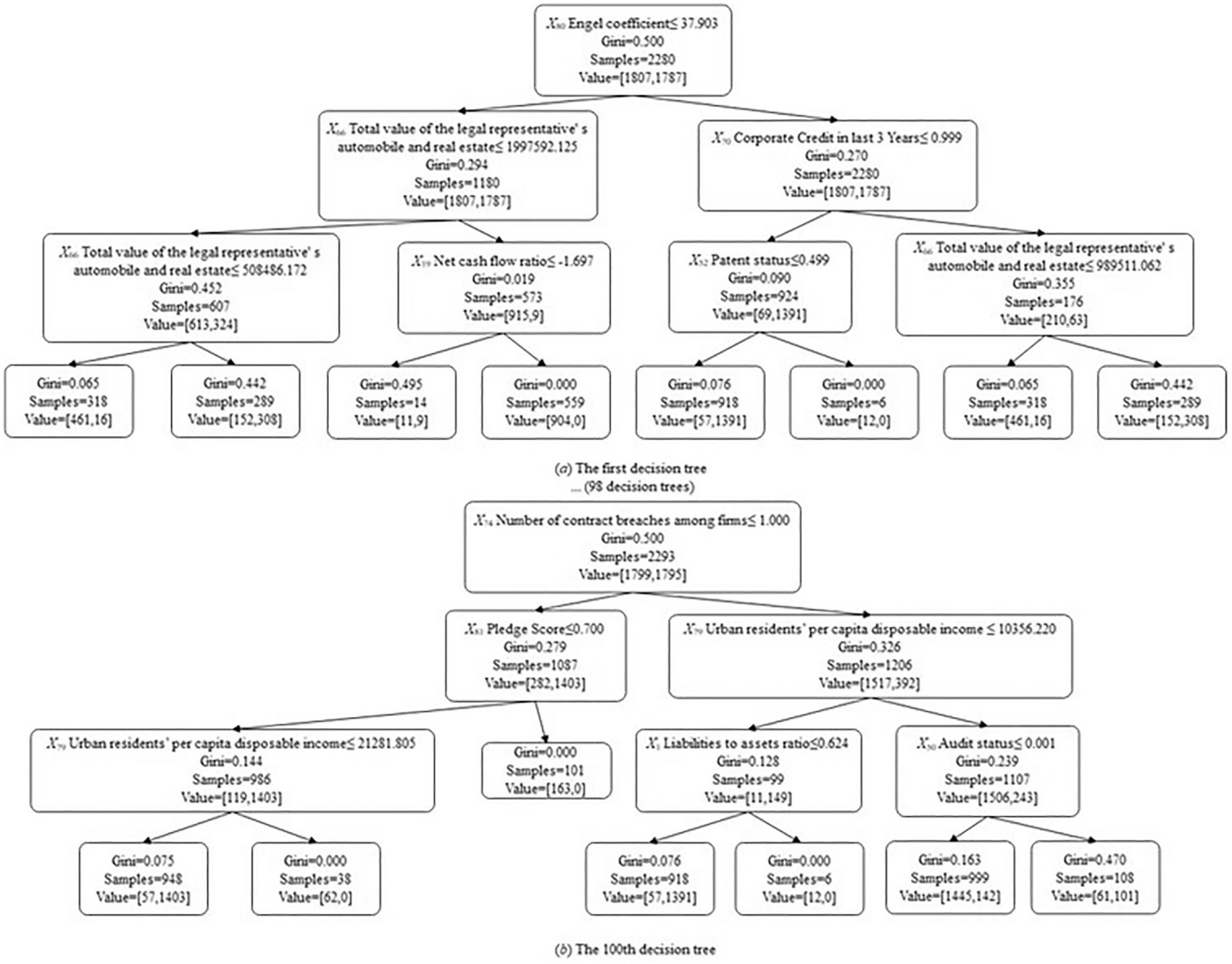


**Fig. 1.** Random forest constructed by small firm samples.
This figure illustrates the random forest constructed by small firm samples. There are 100 decision trees in the random forest. Panel (*a*) represents the first decision tree in the random forest and Panel (*b*) represents the 100th decision tree in the random forest. The first line of the node indicates the variable name of the node and the threshold for the variable interval division, except for the box for the position of the leaf node in the last line. The second row indicates the Gini index of the node's sample set. The third row indicates the number of non-duplicate firms in the node's sample set; as it is a put-back sample, there are bound to be duplicates of firms. The first value in the fourth row indicates the number of non-defaulted firms in the node sample set, and the second value in the fourth row indicates the number of defaulted firms in the node sample set. Since the third row excludes the number of duplicate firms, the fourth row does not exclude the number of duplicate firms, so the sum of the third and fourth rows is not equal. Due to space limitations, some information is omitted.

## 5. Empirical results

Based on the training sample data obtained by SMOTE, the random forest is constructed as shown in Fig. 1[12] and the credit feature data for non-defaulted and defaulted firms are shown in Table 4. As can be seen from column *d*, we construct a total of 771 credit features based on 81 variables. For a firm in column *e*, if the conditions corresponding to the variables in column *d* are satisfied, the variable for that firm takes the value of 1, otherwise it takes the value of 0. As a result, we obtained a numerical matrix $x_{ij}$ of credit features consisting of 0/1 for 3594 firms.[13]

The next step is to introduce into Eq. (1) the matrix $x_{ij}$ and the corresponding default state $y_j$ corresponding to the data sample of 3594 firms with 771 credit features. Table 5 lists the credit features in descending order according to the coefficients of the logistic regression. The coefficient $\beta_i$ and the intercept term $\alpha$ of the logistic regression are obtained by maximum likelihood estimation. These estimated coefficients are listed in Table 5 (column *e*), and the equation obtained is shown in Eq. (14).[14]

$$p_i = p(y_i = 1|x^*) = \left[1 + exp\left(-\left(b + \sum_{j=1}^{81} w_j x_{ij}\right)\right)\right]^{-1} = \left[1 + exp\left(-\left(-1.705 - 0.090x_1^* + 0.144x_2^* + 0.048x_3^* + \ldots - 0.013x_{769}^* + 0.211x_{770}^* + 0.169x_{771}^*\right)\right)\right] \quad (14)$$

[12] The feature is $X_{80}$ Engel coefficient, and the optimal cut-off critical point of the feature is 37.9.

[13] As can be seen, each variable is an indicator with a specific economic meaning, which is one of the points that distinguishes our study from Dumitrescu et al. (2022). Our computer program can easily transform the variables and output the rules and values corresponding to each of them.

[14] The *p*-value of the *F*-test of the default discriminant equation of Eq. (14) is $5.163 \times 10^{-4}$.

**Table 4**
Credit features extracted based on training samples.

| (*a*) No. | (*b*) No. of decision tree | (*c*) No. of leaf node | (*d*) Credit Feature | (*e*) Credit Feature Data | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Non-defaulted firms | | | Defaulted firms | | |
| | | | | (1) Firm 1 | ... | (1797) Firms 1797 | (1798) Firms 1798 | ... | (3594) Firms 3594 |
| 1 | Decision tree 1 | Leaf 1 | $X_1^*$:$X_{80} \leq 37.9$, $X_{66} \leq 1{,}997{,}592.12 \times {}_{66} \leq 508{,}486.17$ | 0 | ... | 1 | 0 | ... | 0 |
| 2 | | Leaf 2 | $X_2^*$:$X_{80} \leq 37.9$, $X_{66} \leq 1{,}997{,}592.12$, $X_{66}>508{,}486.17$ | 0 | ... | 0 | 0 | ... | 1 |
| 3 | | Leaf 3 | $X_3^*$:$X_{80} \leq 37.9$, $X_{66} \leq 1{,}997{,}592.12$, $X_{19} \leq -1.7$ | 0 | ... | 0 | 0 | ... | 0 |
| 4 | | Leaf 4 | $X_4^*$:$X_{80} \leq 37.9$, $X_{66} \leq 1{,}997{,}592.12$, $X_{19}>-1.7$ | 1 | ... | 0 | 0 | ... | 0 |
| 5 | | Leaf 5 | $X_5^*$: $X_{80} \leq 37.9$, $X_{70} \leq 1.0$, $X_{52} \leq 0.5$ | 0 | ... | 0 | 1 | ... | 0 |
| 6 | | Leaf 6 | $X_6^*$: $X_{80} \leq 37.9$, $X_{70} \leq 1.0$, $X_{52}>0.5$ | 0 | ... | 0 | 0 | ... | 0 |
| 7 | | Leaf 7 | $X_7^*$: $X_{80} \leq 37.9$, $X_{70} \leq 1.0$, $X_{66} \leq 989{,}511.06$ | 0 | ... | 0 | 0 | ... | 0 |
| 8 | | Leaf 8 | $X_8^*$: $X_{80} \leq 37.9$, $X_{70} \leq 1.0$, $X_{66}>989{,}511.06$ | 0 | ... | 0 | 0 | ... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 765 | Decision tree 100 | Leaf 765 | $X_{765}^*$: $X_{74} \leq 1.0$, $X_{81} \leq 0.7$ $X_{79} \leq 21{,}281.805$ | 0 | ... | 0 | 0 | | 1 |
| 766 | | Leaf 766 | $X_{766}^*$: $X_{74} \leq 1.0$, $X_{81} \leq 0.7$ $X_{79} > 21{,}281.805$ | 0 | ... | 0 | 0 | | 0 |
| 767 | | Leaf 767 | $X_{767}^*$: $X_{74} \leq 1.0$, $X_{81} > 0.7$ | 0 | ... | 0 | 0 | | 0 |
| 768 | | Leaf 768 | $X_{768}^*$: $X_{74} \leq 1.0$, $X_{79} \leq 10{,}356.22$, $X_1 \leq 0.624$ | 0 | | 0 | 0 | | 0 |
| 769 | | Leaf 769 | $X_{769}^*$: $X_{74} \leq 1.0$, $X_{79} \leq 10{,}356.22$, $X_1>0.624$ | 0 | | 0 | 0 | | 0 |
| 770 | | Leaf 770 | $X_{770}^*$: $X_{74} \leq 1.0$, $X_{79} \leq 10{,}356.22$, $X_{50} \leq 0.001$ | 1 | | 1 | 1 | | 0 |
| 771 | | Leaf 771 | $X_{771}^*$: $X_{74} \leq 1.0$, $X_{79} \leq 10{,}356.22$, $X_{50}>0.001$ | 0 | | 0 | 0 | | 0 |

This table presents the credit features extracted based on training samples. Columns *a, b, c, d* and *e* show the ordinal number of the credit feature, the ordinal number of the decision tree in the random forest, the ordinal number of the leaf nodes, the constructed credit feature and the corresponding decision rule, and the ordinal number of the original training data transformed by the random forest, respectively. We construct a total of 771 credit features based on 81 variables. The values of the credit features are all 0 or 1. Due to space limitations, some information is omitted.

**Table 5**
Logistic regression results.

| (*a*) No. | (*b*) No. of decision tree | (*c*) No. of leaf node | (*d*) Credit feature | (*e*) Coefficient |
|---|---|---|---|---|
| 1 | Leaf 531 | 69–7 | $X^*_{531}$: $X_{73}$ Compliance status of the firm's operations ≤0.5, $X_2$ Operating cash flow ratio ≤ −0.298, $X_{78}$ Consumer price index>100.161 | −0.286 |
| 2 | Leaf 24 | 4–1 | $X^*_{24}$: $X_{62}$ Residential status of the firm's legal representative ≤0.998, $X_{63}$ Duration of residence of the firm's legal representative ≤1.0, $X_{68}$ Time of holding position ≤0.696 | 0.271 |
| 3 | Leaf 655 | 85–5 | $X^*_{655}$: $X_{79}$ Per capita disposable income of urban residents ≤13,348.194, $X_{13}$ Capital immobilization ratio ≤ 1.513, $X_{11}$ Net cash profit ≤5.009 | −0.255 |
| 4 | Leaf 167 | 22–3 | $X^*_{167}$: $X_{73}$ Compliance status of the firm's operations ≤0.499, $X_{16}$ Unpaid loans to net assets ratio ≤ 0.0, $X_{15}$ Long-term asset suitability rate ≤ 20.436 | 0.254 |
| 5 | Leaf 243 | 32–3 | $X^*_{243}$: $X_{63}$ Duration of residence of the firm's legal representative ≤1.0 $X_{16}$ Unpaid loans to net assets ratio ≤ 0.0, $X_{35}$ Inventory turnover rate ≤ 4.344 | 0.254 |
| 6 | Leaf 60 | 8–5 | $X^*_{60}$: $X_{80}$ Engel coefficient ≤ 37.902, $X_{58}$ Education ≤0.9, $X_{62}$ Residential status of the firm's legal representative ≤0.986 | 0.250 |
| 7 | Leaf 622 | 81–4 | $X^*_{622}$: $X_{63}$ Duration of residence of the firm's legal representative ≤1.0 $X_{17}$ Unpaid loans to total assets ratio ≤ 0.0, $X_{13}$ Capital immobilization ratio > 0.051 | 0.238 |
| 8 | Leaf 321 | 42–7 | $X^*_{321}$: $X_{49}$ Employment years of the corporate legal representative ≤1.0 $X_{76}$ Urban and rural residents' per capita savings at the end of the year ≤25,849.673, $X_{80}$ Engel coefficient ≤ 40.7 | −0.235 |
| 9 | Leaf 454 | 59–5 | $X^*_{454}$: $X_{80}$ Engel coefficient ≤ 37.902, $X_{68}$ Time of holding position ≤0.69, $X_{62}$ Residential status of the firm's legal representative ≤0.998 | 0.235 |
| 10 | Leaf 97 | 13–2 | $X^*_{97}$: $X_{68}$ Time of service ≤0.4, $X_{73}$ Compliance status of the firm's operations ≤0.5, $X_{13}$ Capital immobilization ratio > 0.048 | 0.234 |
| ... | ... | ... | ... | ... |
| 769 | Leaf 769 | 52–1 | $X^*_{769}$: $X_{80}$ Engel coefficient ≤ 37.902, $X_{33}$ Subtotal cash inflows from operations ≤782,155,680.0, $X_{79}$ Per capita disposable income of urban residents ≤18,988.607 | 0.000 |
| 770 | Leaf 770 | 17–4 | $X^*_{770}$: $X_{70}$ Corporate credit in last 3 years ≤1.0, $X_{35}$ Inventory turnover rate ≤ 4.109, $X_{11}$ Net cash profit> 9.16 | 0.000 |
| 771 | Leaf 771 | 47–6 | $X^*_{771}$: $X_{69}$ Types of registered capital for firms ≤1.0, $X_2$ Operating cash flow ratio ≤ −0.564, $X_{37}$ Current asset turnover rate > 0.192 | 0.000 |
| | | | Intercept | −1.705 |

This table presents the credit features and the corresponding logistic regression coefficient values. We rank the coefficients of credit features in column *e* according to their absolute values of credit feature coefficients in descending order, so the ordinal numbers in column *b* are not in order. The first number in column *c* indicates the ordinal number of the decision tree and the second number indicates the ordinal number of the leaf node of the corresponding decision tree. Due to space limitations, we have detailed a list of the top 10 credit characteristics of importance.

Based on the magnitude of the absolute value of the logistic regression coefficients, we identify the credit features that have a critical impact on the default status of Chinese small firms. The larger the absolute value of the coefficient, the greater the importance of the variable. We now have an intuitive understanding of the credit features. There are three credit features that have the most critical impact on default prediction (Table 5). First, there is feature $X^*_{531}$, whose coefficient is −0.286. If the firm satisfies the condition of the constraint, $X^*_{531} = 1$, the default probability calculated by Eq. (14) is smaller, i.e., the credit status of the firm will be better. In $X^*_{531}$, "compliance status of the firm's operations" ($X_{73}$) is a positive variable, the "operating cash flow ratio" ($X_2$) is a positive variable, and the "consumer price index" ($X_{78}$) is an interval variable.[15] It can be seen that if the firm has not been commended or sanctioned by Chinese government authorities in the current year in terms of legal compliance, and if the firm was sanctioned in the previous year, then even if the operating cash flow ratio of the firm is poor, as long as the "consumer price index" in the external macro conditions is higher than 100.161 (100 in the previous year), the firm is not likely to default. This is because improvements in the external environment can increase the firm's revenues and cash flow, making up for the lack of internal cash flow and allowing the firm to continue to meet its debt obligations. In reality, small firms are small in scale, irregular in their management, and vulnerable to the macroeconomic environment. Therefore, when judging the default risk of small firms, the "consumer price index" is very important. When the "consumer price index" is slightly higher than that of the previous year, other variables of small firms can be considered to determine whether to grant credit.

Second, there is feature $X^*_{24}$, whose coefficient is 0.271, which shows that the firm is more likely to default when the CEO does not own a house, the CEO has lived in the locality for less than one year, and the CEO's tenure is <0.696 years. Therefore, the CEO's residency status ($X_{62}$), the CEO's length of residency ($X_{63}$), and the CEO's tenure ($X_{68}$) will directly affect the likelihood of loan repayment. The CEO's personal situation is crucial.

[15] The "compliance status of the firm's operations" refers to the absence of possible sanctions against the company in the present year by state authorities; the "operating cash flow ratio" is the cash generated by the company's normal business operations; and the "consumer price index" is the average change in the prices of goods and services consumed by households, which indicate the level of inflation in the region where the company operates.

**Table 6**
Analysis of the default discrimination ability of key credit features.

| (*a*) No. | (*b*) Credit feature | (*c*) RVI | (*d*) S·D of RVI |
|---|---|---|---|
| 1 | $X^*_{531}$: $X_{73}$ Compliance status of the firm's operations ≤0.5, $X_2$ Operating cash flow ratio ≤ −0.298, $X_{78}$ Consumer price index>100.161 | 0.012727 | 0.000045 |
| 2 | $X^*_{24}$: $X_{62}$ Residential status of the firm's legal representative ≤0.998, $X_{63}$ Duration of residence of the firm's legal representative ≤1.0, $X_{68}$ Time of holding position ≤0.696 | 0.011973 | 0.000002 |
| 3 | $X^*_{655}$: $X_{79}$ Per capita disposable income of urban residents ≤13,348.194, $X_{13}$ Capital immobilization ratio ≤ 1.513, $X_{11}$ Net cash profit ≤5.009 | 0.002692 | 0.000499 |
| 4 | $X^*_{167}$: $X_{73}$ Compliance status of the firm's operations ≤0.499, $X_{16}$ Unpaid loans to net assets ratio ≤ 0.0, $X_{15}$ Long-term asset suitability rate ≤ 20.436 | 0.002503 | 0.000661 |
| 5 | $X^*_{243}$: $X_{63}$ Duration of residence of the firm's legal representative ≤1.0 $X_{16}$ Unpaid loans to net assets ratio ≤ 0.0, $X_{35}$ Inventory turnover rate ≤ 4.344 | 0.001388 | 0.000257 |
| 6 | $X^*_{60}$: $X_{80}$ Engel coefficient ≤ 37.902, $X_{58}$ Education ≤0.9, $X_{62}$ Residential status of the firm's legal representative ≤0.986 | 0.000624 | 0.000073 |
| 7 | $X^*_{622}$: $X_{63}$ Duration of residence of the firm's legal representative ≤1.0 $X_{17}$ Unpaid loans to total assets ratio ≤ 0.0, $X_{13}$ Capital immobilization ratio > 0.051 | 0.000582 | 0.000091 |
| 8 | $X^*_{321}$: $X_{49}$ Employment years of the corporate legal representative ≤1.0 $X_{76}$ Urban and rural residents' per capita savings at the end of the year ≤25,849.673, $X_{80}$ Engel coefficient ≤ 40.7 | 0.000546 | 0.000066 |
| 9 | $X^*_{454}$: $X_{80}$ Engel coefficient ≤ 37.902, $X_{68}$ Time of holding position ≤0.69, $X_{62}$ Residential status of the firm's legal representative ≤0.998 | 0.000504 | 0.000077 |
| 10 | $X^*_{97}$: $X_{68}$ Time of service ≤0.4, $X_{73}$ Compliance status of the firm's operations ≤0.5, $X_{13}$ Capital immobilization ratio > 0.048 | 0.000447 | 0.000075 |

This table illustrates the default discrimination ability of key credit features (relative variable importance, RVI) of different categories. We use dominance analysis (DA) to identify the relative importance of credit features (CF). DA calculates the RVI based on each CFs contribution to prediction using a series of outcomes that deconstruct and evaluate each CF's contribution to an overall default discrimination accuracy (AUC). RVI indicates the importance of credit features, and the larger the RVI, the stronger the default discriminatory power of the credit feature. We only list the top ten credit features in the RVI ranking.

Finally, there is feature $X^*_{655}$, whose coefficient is −0.255. It includes "per capita disposable income of urban residents", the "net cash profit" and the "capital immobilization ratio".[16] It can be seen that even if the "per capita disposable income of urban residents" where the firm is located ($X_{79}$) is lower than 13,348.194 and the "net cash profit" ($X_{11}$) is lower than 5.009, the firm is unlikely to default as long as the "capital immobilization ratio" ($X_{13}$) is lower than 1.513. It is easy to understand

[16] The "per capita disposable income of urban residents" in the region in which the company operates indicates the purchasing power of consumers in the region. The "net cash profit" is calculated by dividing net cash flow from operating activities by net profit. A company's "net cash flow from operating activities" indicates whether any additional cash has been brought into or taken out of the business. It includes any changes to net income as well as any adjustments made to non-cash items. The "capital immobilization ratio", calculated as net fixed assets divided by total capital, reflects the proportion of a company's capital investment that is in long-term fixed assets.

**Table 7**
Results for default discrimination on test samples.

| (*a*) Actual status | (*b*) Predicted status | | (*c*) Total | (*d*) *U* test |
|---|---|---|---|---|
| | Positive (Default) | Negative (Non-default) | | |
| Positive (Default) | 14 | 6 | 20 | 555,661.000*** |
| Negative (Non-default) | 76 | 1,122 | 1,198 | |
| Total | 90 | 1,128 | 1,218 | |

This table presents the default prediction results for the test sample. The default prediction model constructed in this study has a recognition rate of 14/(14 + 6) = 70.000% for defaulted small firms, 1122/(76 + 1122) = 93.656% for non-defaulted small firms, (18 + 1032)/1218 = 93.268% for all firms, and a *G*-mean of (0.879 × 0.913)1/2 = 80.967% for all firms. We use the *U* test to test the robustness of the results. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively.

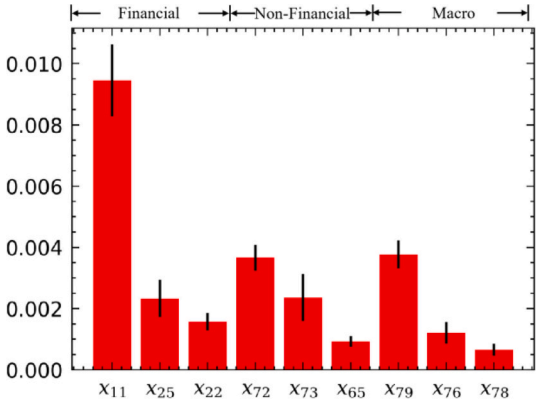


**Fig. 2.** Comparison of the default discrimination ability of key indicators of different categories.
This figure illustrates the default discrimination ability of the key indicators (relative variable importance, RVI) of different categories. We use dominance analysis (DA) to identify the relative importance of the original variables (OV). DA calculates the RVI based on each OV's contribution to the prediction using a series of outcomes that deconstruct and evaluate each OV's contribution to an overall default discrimination accuracy (AUC). RVI indicates the importance of the variable, and the greater the RVI, the stronger the default discriminatory power of the variable. We only list the top three variables in the RVI ranking for different categories.

the significance of a high rate of capital immobilization, as it implies that the firm's investment in fixed asset exceeds its own capacity, that the firm has used part of its liquidity to borrow, and that the daily working capital is funded by loans due to excessive shortages of capital, resulting in a deterioration in the firm's financial situation. Therefore, the "capital immobilization ratio" should be taken into account in the credit scoring of small firms.

By looking at the top ten credit features in terms of importance, we find that the "Engel coefficient" appeared three times, in $X^*_{60}$, $X^*_{321}$ and $X^*_{454}$. The "Engel coefficient' refers to the proportion of household food expenditure in its total consumption expenditure, which measures the standard of living of a population in a country or region (Su & Hoshino, 2016). The greater the "Engel coefficient", the higher the population's standard of living and the better the macroeconomic situation. Small firms are most vulnerable to macroeconomic conditions, especially to the impact of living standards. Therefore, when judging the default risk of small firms, the Engel coefficient is vital.

As we do not have direct access to the degree of importance of the original variables, we compare the credit features with the original variables in terms of their default discriminatory power. We use dominance analysis to calculate the degree of importance of the credit

**Table 8**
Analysis of the default discrimination ability of key variables.

| (*a*) No. | (b) Categories | (c)Variables | (d) RVI | (e) S·D of RVI |
|---|---|---|---|---|
| 1 | Financial Factors | $X_{11}$ Net cash profit | 0.009452 | 0.001171 |
| 2 | | $X_{25}$ Operating profit margin ratio | 0.002334 | 0.000612 |
| 3 | | $X_{22}$ Net present value of sales | 0.001574 | 0.000289 |
| 4 | Non-Financial Factors | $X_{72}$ Firm's legal disputes | 0.003661 | 0.000422 |
| 5 | | $X_{73}$ Compliance status of the firm's operations | 0.002363 | 0.000768 |
| 6 | | $X_{65}$ Age of legal representative | 0.000928 | 0.000171 |
| 7 | Macro Factors | $X_{79}$ Per capita disposable income of urban residents | 0.003769 | 0.000453 |
| 8 | | $X_{76}$ Urban and rural residents' per capita savings at the end of the year | 0.001214 | 0.000348 |
| 9 | | $X_{78}$ Consumer price index of residents | 0.000655 | 0.000195 |

This table illustrates the default discrimination ability of key variables (relative variable importance, RVI) of different categories. We use dominance analysis (DA) to identify the relative importance of original variables (OV). DA calculates the RVI based on each OV's contribution to prediction using a series of outcomes that deconstruct and evaluate each OV's contribution to an overall default discrimination accuracy (AUC). RVI indicates the importance of the variable, and the larger the RVI, the stronger the default discriminatory power of the variable. We only list the top three variables in the RVI ranking for different categories.

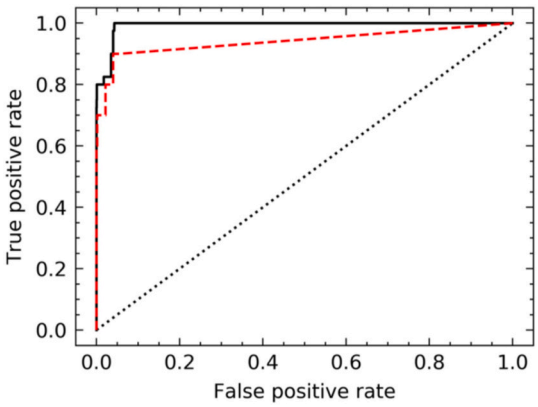


**Fig. 3.** ROC curve of the default discriminant model.
This figure illustrates the ROC (receiver operating characteristic) curve of the default discriminant model on training samples and testing samples. The ROC curve is used to calculate the AUC. The AUC is a fundamental and commonly used statistic for assessing the discriminatory power of credit scores. The AUC ranges from 50% (utterly random prediction) to 100% (perfect accuracy).

features and the original variables separately. The results of the dominance analysis of the original variables are shown in Table 6 (which are the same as those in Table 5).

We need to validate the default prediction capability of the proposed method to verify its validity. By substituting the test sample into Eq. (14) to get the default probability $p_j$ of the test firms and substituting the default probability into Eq. (2) to get the theoretical default state $\hat{y}_j$ of the test firms, and then comparing the discriminant value $\hat{y}_j$ with the actual default state $y_j$ of the firms, we can obtain the default discrimination confusion matrix of the test sample, as shown in Table 7. The default prediction model constructed in this study has a recognition rate of 70.000% for defaulted small firms, 93.656% for non-defaulted small firms, 93.268% for all firms, and a *G*-mean of 80.967% for all firms, which can accurately identify the default status of Chinese small firms and ensure the identification of the default status of "bad" firms. Column *d* in Table 6 shows the *U* test of the default prediction results, which further illustrates that the model constructed in this paper can significantly distinguish the default status of defaulted and non-defaulted small firms. We also present the receiver operating characteristic (ROC) curve in Fig. 3, used to calculate the AUC. The AUC values in the training and test samples reached 99.181% and 95.102%, respectively. This shows that it is reasonable to use credit features as variables for prediction, and the predicted scores are significantly different in the default and non-default samples.

The results of the dominance analysis of the original variables of different categories are shown in Fig. 2 and Table 8. We conclude that the most important variables among the financial factors are net cash profit, operating profit margin ratio and net present value of sales.[17] This result, which is similar to the financial variables used by Altman and Sabato (2007) in their credit scoring model for small and medium-sized firms in the US, Behr and Guettler (2007) for their German counterparts, and Tobback et al. (2017) for those in the UK, corroborates their findings.

In addition, among the non-financial factors, the most important variables are the firm's legal disputes, the compliance status of the firm's operations and the age of the legal representative.[18] Boyer and Blazy (2014) looked at French data and found that gender, age, belonging to a minority group, professional experience, and financial resources of managers of micro-enterprises can affect the future survival of a business. Bruno, Cornaggia, and Cornaggia (2016) found a significant correlation between the operating conditions of American firms and their credit risk. Research by Henisz and McGlinch (2019) on global firms and Ko, Lee, and Anandarajan (2019) on Taiwanese firms found that a legal dispute can have a negative impact on a firm's credit rating, potentially affecting its ability to secure loans or other forms of financing. Our research on non-financial variables impacting the credit risk of small businesses in China attests to these findings in other regions.

Finally, among the macro factors the relevant variables are the per capita disposable income of urban residents, the per capita year-end balance of savings of urban and rural residents and the consumer price index of residents.[19] Salas and Saurina (2002) studied the problematic loans of Spanish banks, incorporating macroeconomic factors such as GDP growth rate and the indebtedness of firms and families. Carling et al. (2007) found that macroeconomic factors such as the output gap, the yield curve, and household expectations have a significant impact on the credit risk of Swedish corporations. Our conclusion is similar, further advancing the study of macroeconomic factors affecting the credit risk of small firms. Given these results, it is important to consider not only financial factors, but also non-financial and macro factors in the credit evaluation of Chinese small firms. As can be seen from the RVIs in Tables 6 and 8, the credit features obtained are much better at identifying defaults compared to the original variables. Therefore, the way we construct the credit characteristics is close to what we are looking for and our technique makes sense.

We can conclude that net cash profit (financial factor) has the strongest default discriminatory power among all the indicators, and the RVI is much higher than that of the other indicators. This variable reflects the quality and profitability of the firm, and the higher the indicator, the better. The larger the size of the firm, the greater the contribution of net profit from operating operations to cash flow, the

[17] The "operating profit margin ratio" is calculated by subtracting the cost of goods sold and selling, general and administrative expenses (also called operating expenses or SG&A) from net sales and dividing by net sales. The "net present value of sales" is calculated by dividing the net cash flow from operating activities by sales revenue.

[18] The "firm's legal disputes" refers to any legal disputes involving the company, indicating its legal risks and liabilities.

[19] The per capita savings of urban and rural residents at the end of the year in the region in which the company operates, indicating the economic conditions of the region.

greater its capacity to recover sales, the lower its reliance on inventories and accounts receivable, and the higher its operating turnover. Small firms do not have an advantage over large firms in terms of financing. It is difficult to raise funds at low cost without the policy support of special industries. The source of funds is seed capital investment and operating cash inflow. Therefore, the primary focus should be on the net cash profit of the firm when making default judgments about Chinese small businesses.

The RVI of the firm's legal disputes (non-financial factor) ranks third among all indicators in terms of default identification. While the transformation of China's economic structure is accelerating and market competition is becoming increasingly fierce, small firms are prone to legal disputes due to imperfect internal management systems, the lack of democratic procedures for major decisions, and poor legal awareness of business operators (among other things) and their ability to prevent and control legal risks is relatively weak. Once their legal risks have evolved from being hidden dangers to causing real damage, they will weaken the strength and market competitiveness of firms and even jeopardize their survival. Therefore, when identifying the default risk of Chinese small firms, we should also focus on the legal disputes of the firms.

The per capita disposable income of urban residents (macro factor) ranks second among all indicators in terms of default discriminatory power. A high urban disposable income per capita indicates that the city in which the firm is located has a large share of tertiary industry; a large number of high-quality jobs, which can form a virtuous circle of attracting talent; increased tax revenue, which further supports and promotes industrial upgrading; and a good business environment for the development of small firms. Therefore, when making default judgments regarding Chinese small firms, we should pay attention to the per capita disposable income of the city in which the small firms are located.

Within the category of non-financial factors, the variable age of the legal representative ranks third among the non-financial indicators in terms of identifying default risk and the sixth in importance for all the categories, which is well-researched and significant. This variable can reflect the future development potential of small firms. If the legal representative is older, the development strategy of the firm may be conservative and less innovative, but the capital and customers and other resources will be relatively greater. If the legal representative is younger, the development strategy of the firm may be more aggressive, but the capital and customers and other resources will be relatively smaller. Therefore, the age of the legal representative should be considered when making default assessments on Chinese small firms.

## 6. Additional analysis

We verify the default identification ability of the proposed model on four other public credit datasets. They are Australian, Japanese, Taiwanese and Polish credit datasets from the Center for Machine Learning and Intelligent Systems (UCI Machine Learning Archive). The basic information from these four datasets is described in Table 9. These credit datasets are widely used in the literature (Sariannidis, Papadakis, Garefalakis, Lemonakis, & Kyriaki-Argyro, 2020; Smiti & Soui, 2020). The datasets for Australia and Japan are balanced and the datasets for Taiwan and Poland are unbalanced, but our proposed methodology is applicable to all. We use N × 2-fold cross-validation to evaluate the performance of the model based on test samples to ensure the robustness of the results.[20] In addition to including other country samples, we compare the results of our model (*RFLR*, Random Forest Logistic Regression) with six typical machine learning models in order to verify the superiority of our proposed approach compared to popular machine learning methods. Specifically, we select logistic regression (*LR*), the support vector machine (*SVM*), Naive Bayes (*NB*), decision tree (*DT*), K-nearest neighbor (*KNN*) and the linear discriminant model (*LDA*). For each of the above models, we compute the performance measures detailed in Section 4.4.

**Table 9**
Description of datasets for comparative analysis.

| (*a*) Dataset | (*b*) Number of samples | (*c*) Number of variables | (*d*) Non-defaulted firms | (*e*) Defaulted firms |
|---|---|---|---|---|
| Australian | 690 | 14 | 383 | 307 |
| Japanese | 690 | 15 | 383 | 307 |
| Taiwanese | 30,000 | 23 | 23,364 | 6,636 |
| Polish | 10,503 | 64 | 10,008 | 495 |

This table presents the description of datasets for comparative analysis. Column *a, b, c, d* and *e* are the dataset name, the number of samples, the number of variables (financial factors), the number of non-defaulted firms and the number of defaulted firms in the dataset, respectively.

Table 10 shows the results of the model performance comparison, where each value is the average of $100 \times 2 = 200$ groups of results. The average ranking (column *d*) shows that our model (*RFLR*) performs better than the other machine learning models, as the average rank is always the lowest compared to the other models. This result holds not also for Chinese small firms, but also for all the different country samples used. We can affirm that, on average, our model performs better than other alternative models, and this result is robust with other country samples. Regarding the individual performance measures (column *c*), we can state that for small firms in China, six of the eight performance measures point to the model estimated in this paper (*RFLR*) as the most appropriate in terms of the selected performance measures. We are certain that this model is the most appropriate given our particular sample. Only the *G-mean* and *BM* performance measures select another model (*KNN*) as the most appropriate. Looking at the results using other countries samples, the results are mixed. For Japan, Australia and Poland it is clear that our model performs better than others; however, for Taiwan the robustness is not guaranteed. We know that no model will be suitable for all scenarios, but the most important measures of *G-mean* and AUC, and the average ranking, allow us to maintain confidence in our proposed methodology.

In addition, Table 11 shows the results for the Friedman test and the post-hoc Bonferroni-Dunn test that compare the difference between the performance measures of the proposed model. We observe that the null hypothesis is rejected for all the individual performance measures except for the precision. Specifically, we obtain seven *p*-values $<0.01$ in eight performance measures. This result indicates that there are significant differences in the ranking of performance criteria, except for precision.[21] The results of the Bonferroni test show that the performance measure of the proposed model is significantly different from other traditional models, except for *LDA*.[22]

The above results illustrate that our proposed methodology has good applicability and robustness for the default prediction of small firms in China and outperforms popular machine learning algorithms. In general, when we apply our default prediction model to other economies, the results are robust.

[20] The whole sample is divided into two parts, sample A and sample B, without changing the proportion of defaulted and non-defaulted firms. Sample A is used to determine the coefficient of the logistic regression, and the determined coefficient is used to discriminate the default status of sample B. Sample B is used to determine the coefficient of the logistic regression, and the determined coefficient is used to discriminate the default status of sample A. This process is repeated $N = 100$ times.

[21] But this does not mean that our proposed methodology is not the best in terms of measure precision, only that we cannot prove it statistically. In terms of average ranking, our proposed method is still the best in terms of measure precision.

[22] But this does not make our proposed methodology worse than *LDA*, only that we could not prove it statistically. In terms of average ranking, our proposed methodology is still better than *LDA*.

**Table 10**
Comparison of model performance.

| (*a*) *Dataset* | (*b*) *Model* | (*c*) *Performance* | | | | | | | | (*d*) Average ranking |
|---|---|---|---|---|---|---|---|---|---|---|
| | | (1) *Precision* | (2) *Accuracy* | (3) *G-mean* | (4) *MCC* | (5) *KS* | (6) *BM* | (7) *MK* | (8) *AUC* | |
| *China Small firms* | *RFLR* | **0.680** (0.091) | **0.989** (0.002) | 0.764 (0.066) | **0.625** (0.076) | **0.863** (0.052) | 0.586 (0.101) | **0.673** (0.091) | **0.965** (0.027) | **2.3** |
| | *LR* | 0.246 (0.039) | 0.960 (0.007) | 0.809 (0.066) | 0.393 (0.053) | 0.759 (0.068) | 0.647 (0.107) | 0.241 (0.039) | 0.911 (0.044) | 4.5 |
| | *SVM* | 0.383 (0.075) | 0.976 (0.006) | 0.788 (0.068) | 0.478 (0.063) | 0.77(0.073) | 0.619 (0.107) | 0.377 (0.075) | 0.918(0.04) | 3.5 |
| | *NB* | 0.062 (0.014) | 0.814 (0.049) | 0.750 (0.051) | 0.169 (0.032) | 0.585 (0.091) | 0.511 (0.091) | 0.056 (0.014) | 0.797 (0.064) | 6.9 |
| | *DT* | 0.519 (0.099) | 0.984 (0.004) | 0.767 (0.073) | 0.547 (0.087) | 0.590 (0.111) | 0.590 (0.111) | 0.512 (0.100) | 0.795 (0.056) | 3.9 |
| | *KNN* | 0.354 (0.042) | 0.973 (0.005) | **0.872** (0.055) | 0.514 (0.049) | 0.805 (0.087) | **0.758** (0.095) | 0.351 (0.042) | 0.91(0.046) | 3.1 |
| | *LDA* | 0.222 (0.036) | 0.951 (0.009) | 0.854(0.06) | 0.395 (0.047) | 0.807 (0.067) | 0.722 (0.103) | 0.218 (0.036) | 0.933 (0.037) | 3.9 |
| *Australia* | *RFLR* | 0.847 (0.027) | **0.863** (0.014) | 0.861 (0.014) | 0.724 (0.029) | 0.747 (0.027) | 0.723 (0.029) | **0.725**(0.03) | **0.933** (0.011) | **1.8** |
| | *LR* | 0.796 (0.021) | 0.857 (0.014) | 0.861 (0.014) | 0.721 (0.027) | 0.747 (0.026) | 0.725 (0.027) | 0.717 (0.026) | 0.920 (0.012) | 2.9 |
| | *SVM* | 0.787 (0.017) | 0.855 (0.013) | 0.859 (0.013) | 0.720 (0.025) | 0.742 (0.026) | 0.723 (0.026) | 0.716 (0.025) | 0.921 (0.013) | 4.1 |
| | *NB* | **0.851** (0.026) | 0.813 (0.018) | 0.795 (0.022) | 0.623 (0.036) | 0.719 (0.027) | 0.604 (0.039) | 0.643 (0.035) | 0.897 (0.017) | 5.5 |
| | *DT* | 0.788 (0.027) | 0.813 (0.019) | 0.810 (0.020) | 0.622 (0.039) | 0.622 (0.040) | 0.622 (0.040) | 0.623 (0.039) | 0.811(0.02) | 6.5 |
| | *KNN* | 0.806 (0.026) | 0.846 (0.016) | 0.847 (0.015) | 0.692 (0.031) | 0.700 (0.026) | 0.695 (0.031) | 0.69(0.031) | 0.898 (0.012) | 4.9 |
| | *LDA* | 0.795 (0.021) | 0.859 (0.014) | **0.863** (0.014) | **0.726** (0.026) | **0.753** (0.027) | **0.730** (0.027) | 0.722 (0.026) | 0.926 (0.011) | 1.9 |
| *Japan* | *RFLR* | 0.841 (0.026) | **0.862** (0.015) | **0.861** (0.015) | **0.723** (0.029) | 0.744 (0.026) | 0.723 (0.030) | **0.723**(0.03) | **0.930** (0.011) | **1.4** |
| | *LR* | 0.796 (0.021) | 0.856 (0.014) | 0.859 (0.014) | 0.719 (0.027) | 0.742 (0.027) | 0.722 (0.027) | 0.715 (0.026) | 0.908 (0.014) | 3.6 |
| | *SVM* | 0.787 (0.018) | 0.855 (0.014) | 0.859 (0.014) | 0.72(0.028) | 0.736 (0.027) | 0.723 (0.028) | 0.716 (0.027) | 0.910 (0.016) | 3.5 |
| | *NB* | **0.853** (0.021) | 0.814 (0.016) | 0.796 (0.020) | 0.624 (0.033) | 0.699 (0.027) | 0.605 (0.036) | 0.644 (0.031) | 0.887 (0.017) | 5.6 |
| | *DT* | 0.781 (0.030) | 0.808 (0.021) | 0.806 (0.021) | 0.613 (0.042) | 0.613 (0.042) | 0.613 (0.042) | 0.614 (0.042) | 0.807 (0.021) | 6.8 |
| | *KNN* | 0.823 (0.024) | 0.854 (0.014) | 0.854 (0.015) | 0.707 (0.028) | 0.711 (0.028) | 0.709 (0.029) | 0.706 (0.028) | 0.889 (0.015) | 4.8 |
| | *LDA* | 0.790 (0.019) | 0.857 (0.014) | 0.861 (0.013) | **0.723** (0.025) | **0.747** (0.026) | **0.727** (0.026) | 0.72(0.025) | 0.912 (0.013) | 1.9 |
| *Taiwan* | *RFLR* | 0.459 (0.013) | 0.757 (0.008) | **0.669** (0.005) | 0.344(0.01) | 0.370 (0.008) | 0.365 (0.009) | 0.324 (0.013) | **0.746** (0.005) | **1.9** |
| | *LR* | 0.378 (0.005) | 0.687 (0.006) | 0.669 (0.004) | 0.292 (0.006) | 0.370 (0.007) | 0.340 (0.007) | 0.25(0.006) | 0.722 (0.004) | 3.1 |
| | *SVM* | **0.47**(0.013) | **0.763** (0.008) | 0.668 (0.005) | **0.350** (0.009) | **0.371** (0.007) | **0.367** (0.008) | **0.333** (0.013) | 0.718 (0.005) | 2.0 |
| | *NB* | 0.253 (0.004) | 0.391 (0.014) | 0.471 (0.017) | 0.144 (0.009) | 0.385 (0.008) | 0.143 (0.013) | 0.145 (0.007) | 0.726 (0.005) | 5.6 |
| | *DT* | 0.356 (0.007) | 0.697 (0.006) | 0.592 (0.006) | 0.205 (0.009) | 0.223(0.01) | 0.223 (0.010) | 0.189 (0.009) | 0.612 (0.005) | 5.8 |
| | *KNN* | 0.352 (0.004) | 0.667 (0.004) | 0.642 (0.004) | 0.246 (0.007) | 0.287 (0.008) | 0.287 (0.008) | 0.211 (0.006) | 0.686 (0.004) | 5.5 |
| | *LDA* | 0.378 (0.006) | 0.688 (0.006) | 0.669 (0.004) | 0.292 (0.006) | 0.37(0.007) | 0.340 (0.007) | 0.251 (0.006) | 0.722 (0.004) | 2.9 |
| *Poland* | *RFLR* | **0.250** (0.019) | 0.903 (0.008) | **0.698** (0.021) | **0.318** (0.022) | **0.540** (0.025) | **0.450** (0.031) | **0.226**(0.02) | **0.845** (0.012) | **1.1** |
| | *LR* | 0.095 (0.003) | 0.680 (0.013) | 0.678 (0.012) | 0.160(0.01) | 0.373 (0.021) | 0.357 (0.023) | 0.072 (0.004) | 0.718 (0.012) | 3.4 |
| | *SVM* | 0.089 (0.004) | 0.632 (0.023) | 0.677 (0.011) | 0.156(0.01) | 0.375(0.02) | 0.359 (0.022) | 0.068 (0.004) | 0.721 (0.012) | 3.9 |
| | *NB* | 0.049 (0.001) | 0.171 (0.014) | 0.347 (0.017) | 0.019 (0.013) | 0.079 (0.058) | 0.031 (0.021) | 0.012 (0.008) | 0.537 (0.033) | 7.0 |
| | *DT* | 0.243 (0.021) | **0.907** (0.007) | 0.650 (0.023) | 0.287 (0.026) | 0.385 (0.033) | 0.385 (0.033) | 0.215 (0.022) | 0.692 (0.016) | 2.6 |
| | *KNN* | 0.08(0.005) | 0.761 (0.007) | 0.547(0.02) | 0.083 (0.014) | 0.200 (0.027) | 0.164 (0.028) | 0.042 (0.007) | 0.612 (0.015) | 5.6 |
| | *LDA* | 0.094 (0.004) | 0.685 (0.017) | 0.672 (0.014) | 0.156 (0.012) | 0.362 (0.024) | 0.345 (0.027) | 0.07(0.005) | 0.712 (0.014) | 4.3 |

This table presents the comparison of model performance. Columns *a*, *b*, *c* and *d* show the dataset, the default prediction method, the model accuracy and the average ranking of the corresponding method on a single dataset for the eight performance measures, respectively. We compare six typical machine learning models with our model (*RFLR, Random Forest Logistic Regression*), including logistic regression (*LR*), support vector machine (*SVM*), Naive Bayes (*NB*), decision tree (*DT*), K-nearest

neighbor (*KNN*) and linear discriminant model (*LDA*). We use N × 2-fold cross-validation (Berg et al., 2020) to evaluate the performance of the model based on test samples. The numbers in parentheses are the standard deviations of performance. The best performance for each evaluation measure is highlighted in bold.

**Table 11**
The average rankings across datasets for different performance measures.

| (*a*) *Model* | (*b*) *Performance* | | | | | | | | (*c*) Average ranking |
|---|---|---|---|---|---|---|---|---|---|
| | (1) Precision | (2) Accuracy | (3) *G*-Mean | (4) MCC | (5) KS | (6) BM | (7) MK | (8) AUC | |
| *RFLR* | **1.6(/)** | **1.4(/)** | 2.6(/) | **1.5(/)** | **2.1(/)** | 3(/) | **1.2(/)** | **1(/)** | **1.8(/)** |
| *LR* | 3.9** | 4.2** | 2.6 | 3.9** | 3.7 | 3.3 | 3.8** | 3.7** | 3.6(/) |
| *SVM* | 4.4** | 3.6* | 3.7 | 3.3* | 3.4 | 2.8 | 3.2* | 3.2* | 3.5(/) |
| *NB* | 4.6** | 6.7*** | 7*** | 6.6*** | 5.2** | 7*** | 6.6*** | 5.4*** | 6.1(/) |
| *DT* | 4.4** | 3.9** | 5.6** | 4.8** | 5.8*** | 5* | 4.8*** | 6.6*** | 5.1(/) |
| *KNN* | 4.4** | 4.6** | 4.4* | 4.8** | 5.2** | 4.4 | 5*** | 5.4*** | 4.8(/) |
| *LDA* | 4.7** | 3.6* | **2.1** | 3.1 | 2.6 | **2.5** | 3.4* | 2.7 | 3.1(/) |
| $\chi_F^2$ | 7.607 | 15.836*** | 20.721*** | 16.714*** | 13.007** | 16.436*** | 18.514*** | 23.679*** | (/) |

This table presents the average rankings across datasets for different performance measures. Columns *a, b* and *c* indicate the default prediction method used, the model performance ranking and the average precision ranking, respectively. *rflr* denotes the method used in this paper. The best performance for each performance measure is highlighted in bold. We use the Friedman test to test the differences between the performance measures of the model. We use the post-hoc Bonferroni-Dunn test to prove the superiority of the proposed model. *, **, and *** indicate significance at the 10%, 5% and 1% levels, respectively. Slanted lines indicate statistically insignificant.

## 7. Conclusions

In this study, we propose a new credit scoring model for small firms in China that is based on random forest decision rules as independent variables for logistic regression. We use data from 3045 small business loans and select seven variable categories that include financial, non-financial and macroeconomic factors (81 variables in total) to construct the credit scoring model. As a first stage, we use SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic default samples to deal with the imbalanced data of small firms. Secondly, we adopt random forest to build predictive credit features that are taken as a new explanatory variable in the logistic regression.

Using dominance analysis, we conclude that the most important variables among the financial factors are the net cash profit, the operating profit margin ratio, and the net present value of sales. Regarding the non-financial category, the firm's legal disputes, the legal compliance of the firm and the age of the legal representative are the most representative. The most important variables among the macroeconomic factors are the per capita disposable income of urban residents, the per capita year-end balance of savings of urban and rural residents and the consumer price index of residents. These results indicate that the role of non-financial and macro factors in the credit evaluation of Chinese small businesses cannot be ignored. The applied methodology is robustly tested against other popular machine learning algorithms and using other economies.

Our results have three main implications. First, they reinforce an idea that emerged in the literature a few years ago that model accuracy depends not only on data mining techniques, but also on how the data are presented during the modelling process. Our results invite the scientific community interested in this topic to renew the conceptual basis used to design models. Second, our findings inform financial institutions of the true value of our model by showing them the limitations of their own models and providing them with a reliable modelling framework that can meet their needs. Finally, the conclusions can help lending institutions, business managers and government agencies in their decision making.

## Declaration of Competing Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

## Acknowledgements

Ying Zhou and Long Shen acknowledge financial support from the General Programs of National Natural Science Foundation of China (Grant 72071026), the Key Programs of National Natural Science Foundation of China (Grant 71731003), the General Programs of National Natural Science Foundation of China (Grant Numbers 72173096, 71971051, 71971034, 71873103, and 72271040), the Youth Programs of National Natural Science Foundation of China (Grant Numbers 71901055, 71903019, and 72201098), the National Natural Science Foundation of China for Regional Science Fund Project (Grant Number 72161033), and the Major Projects of National Social Science Foundation of China (Grant Number 18ZDA095). Laura Ballester acknowledges financial support from Fundación Ramón Areces and the Spanish Ministry of Science and Innovation (Grant PGC2018-095072-B-I00 and Grant PGC2018-093645-B-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF. A way to make Europe"). We all thank the reviewers and the editor for invaluable comments and suggestions.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.irfa.2023.102755.

## References

Alonso-Robisco, A., & Carbó, J. M. (2022). Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio. *International Review of Financial Analysis, 84*, Article 102372.
Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609.
Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus, 43*(3), 332–357.
Audrino, F., Kostrov, A., & Ortega, J. P. (2019). Predicting US bank failures with MIDAS logit models. *Journal of Financial and Quantitative Analysis, 54*(6), 2575–2603.
Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science, 49*(3), 312–329.
Behr, P., & Guettler, A. (2007). Credit risk assessment and relationship lending: An empirical analysis of German small and medium-sized enterprises. *Journal of Small Business Management, 45*(2), 194–213.
Bellotti, T., & Crook, J. (2009a). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society, 60*, 1699–1707.
Bellotti, T., & Crook, J. (2009b). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*, 3302–3308.
Boyer, T., & Blazy, R. (2014). Born to be alive? The survival of innovative and non-innovative French micro-start-ups. *Small Business Economics, 42*, 669–683.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Bruno, V., Cornaggia, J., & Cornaggia, K. J. (2016). Does regulatory certification affect the information content of credit ratings? *Management Science, 62*(6), 1578–1597.
Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 72*, 218–239.
Calabrese, R., Marra, G., & Angela Osmetti, S. (2016). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *Journal of the Operational Research Society, 67*(4), 604–615.
Carling, K., Jacobson, T., Lindé, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking & Finance, 31*(3), 845–868.
Cathcart, L., Dufour, A., Rossi, L., & Varotto, S. (2020). The differential impact of leverage on the default risk of small and large firms. *Journal of Corporate Finance, 60*, Article 101541.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research, 16*, 321–357.
Ciampi, F., Cillo, V., & Fiano, F. (2020). Combining Kohonen maps and prior payment behavior for small enterprise default prediction. *Small Business Economics, 54*(4), 1007–1039.
Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research, 297*, 1178–1192.
Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics, 6*(3), 241–252.
Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239.
Henisz, W. J., & McGlinch, J. (2019). ESG, material credit events, and credit risk. *Journal of Applied Corporate Finance, 31*(2), 105–117.
Huang, S. Y., Tsaih, R. H., & Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert Systems with Applications, 41*(9), 4360–4372.
Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science, 62*(6), 1554–1577.
Keasey, K., Pindado, J., & Rodrigues, L. (2015). The determinants of the costs of financial distress in SMEs. *International Small Business Journal, 33*(8), 862–881.
Ko, C., Lee, P., & Anandarajan, A. (2019). The impact of operational risk incidents and moderating influence of corporate governance on credit risk and firm performance. *International Journal of Accounting & Information Management, 27*(1), 96–110.
Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., & Kou, S. (2021). Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decision Support Systems, 140*, Article 113429.
Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.
Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., & Li, A. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. *International Review of Financial Analysis, 79*, Article 101971.
Luchman, J. N. (2014). Relative importance analysis with multicategory dependent variables: An extension and review of best practices. *Organizational Research Methods, 17*(4), 452–471.
Mues, C., Baesens, B., Files, C. M., & Vanthienen, J. (2004). Decision diagrams in machine learning: An empirical study on real-life credit-risk data. *Expert Systems with Applications, 27*(2), 257–264.
Nasir, M., Simsek, S., Cornelsen, E., Ragothaman, S., & Dag, A. (2021). Developing a decision support system to detect material weaknesses in internal control. *Decision Support Systems, 151*, Article 113631.
Salas, V., & Saurina, J. (2002). Credit risk in two institutional regimes: Spanish commercial and savings banks. *Journal of Financial Services Research, 22*(3), 203.
Sariannidis, N., Papadakis, S., Garefalakis, A., Lemonakis, C., & Kyriaki-Argyro, T. (2020). Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: Decision making based on machine learning (ML) techniques. *Annals of Operations Research, 294*(1), 715–739.
Setiono, R., Baesens, B., & Mues, C. (2011). Rule extraction from minimal neural networks for credit card screening. *International Journal of Neural Systems, 21*(04), 265–276.
Shahabadi, M. S. E., Tabrizchi, H., Rafsanjani, M. K., Gupta, B. B., & Palmieri, F. (2021). A combination of clustering-based under-sampling with ensemble methods for solving imbalanced class problem in intelligent systems. *Technological Forecasting and Social Change, 169*, Article 120796.
Smiti, S., & Soui, M. (2020). Bankruptcy prediction using deep learning approach based on borderline SMOTE. *Information Systems Frontiers, 22*(5), 1067–1083.
Su, L., & Hoshino, T. (2016). Sieve instrumental variable quantile regression estimation of functional coefficient models. *Journal of Econometrics, 191*(1), 231–254.
Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149–172.
Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance, 52*, 89–100.
Tobback, E., Bellotti, T., Moeyersoms, J., Stankova, M., & Martens, D. (2017). Bankruptcy prediction for SMEs using relational data. *Decision Support Systems, 102*, 69–81.
Uddin, M. S., Chi, G., Al Janabi, M. A., & Habib, T. (2022). Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *International Journal of Finance & Economics, 27*(3), 3713–3729.
Voulgaris, F., Doumpos, M., & Zopounidis, C. (2000). On the evaluation of Greek industrial SME's performance via multicriteria analysis of financial ratios. *Small Business Economics, 15*(2), 127–136.
Wu, T. C., & Hsu, M. F. (2012). Credit risk assessment and decision making by a fusion approach. *Knowledge-Based Systems, 35*, 102–110.
Yin, C., Jiang, C., Jain, H. K., & Wang, Z. (2020). Evaluating the credit risk of SMEs using legal judgments. *Decision Support Systems, 136*, Article 113364.
Yu, L., Yao, X., Zhang, X., Yin, H., & Liu, J. (2020). A novel dual-weighted fuzzy proximal support vector machine with application to credit risk analysis. *International Review of Financial Analysis, 71*, Article 101577.
Zhu, W., Zhang, T., Wu, Y., Li, S., & Li, Z. (2022). Research on optimization of an enterprise financial risk early warning method based on the DS-RF model. *International Review of Financial Analysis, 81*, Article 102140.