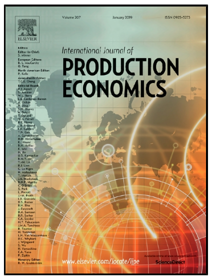
Accepted Manuscript



Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach

You Zhu, Li Zhou, Chi Xie, Gang-Jin Wang, Truong V. Nguyen

PII:	S0925-5273(19)30040-4
DOI:	10.1016/j.ijpe.2019.01.032
Reference:	PROECO 7283
To appear in:	*International Journal of Production Economics*
Received Date:	05 September 2018
Accepted Date:	25 January 2019

Please cite this article as: You Zhu, Li Zhou, Chi Xie, Gang-Jin Wang, Truong V. Nguyen, Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach, *International Journal of Production Economics* (2019), doi: 10.1016/j.ijpe. 2019.01.032

Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach

You Zhu[1], Li Zhou[2], Chi Xie[1,3,*], Gang-Jin Wang[1,3,4], Truong V Nguyen[2]

[1] College of Business Administration, Hunan University, Changsha 410082, China

[2]Business School, University of Greenwich, London, Greenwich SE10 9LS, UK

[3]Center of Finance and Investment Management, Hunan University, Changsha 410082, China

[4]Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA

[*] Corresponding author: E-Mail: xiechi@hnu.edu.cn; Tel.: +86-731-88823890

E-Mail: zy19@hnu.edu.cn (You Zhu), li.zhou@gre.ac.uk (Li Zhou), xiechi@hnu.edu.cn (Chi Xie), wanggangjin@hnu.edu.cn (Gang-Jin Wang), T.V.Nguyen@gre.ac.uk (Truong V Nguyen)

Acknowledgement

This work was supported by the [National Natural Science Foundation of China] under Grant [number 71871088, 71501066 and 71373072]; [China Scholarship Council] under Grant [number 201506135022]; and [Foundation for Innovative Research Groups of the National Natural Science Foundation of China] under Grant [number 71521061]; [China Postdoctoral Science Foundation] under Grant [2018M632960].

Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach

Abstract

In recent years, financial institutions (FIs) have tentatively utilized supply chain finance (SCF) as a means of solving the financing issues of small and medium-sized enterprises (SMEs). Thus, forecasting SMEs' credit risk in SCF has become one of the most critical issues in financing decision-making. Nevertheless, traditional credit risk forecasting models cannot meet the needs of such forecasting. Many researchers argue that machine learning (ML) approaches are good tools. Here we propose an enhanced hybrid ensemble ML approach called RS-MultiBoosting by incorporating two classic ensemble ML approaches, random subspace (RS) and MultiBoosting, to improve the accuracy of forecasting SMEs' credit risk. The experimental samples, originating from data on forty-six quoted SMEs and seven quoted core enterprises (CEs) in the Chinese securities market between 31 March 2014 and 31 December 2015, are collected to test the feasibility and effectiveness of the RS-MultiBoosting approach. The forecasting result shows that RS-MultiBoosting has good performance in dealing with a small sample size. From the SCF perspective, the results suggest that to enhance SMEs' financing ability, 'traditional' factors, such as the current and quick ratio of SMEs, remain critical. Other SCF-specific factors, for instance, the features of trade goods and the CE's profit margin, play a significant role.

Keywords: *Supply Chain Finance, Small and Medium-sized Enterprises, Credit Risk Forecasting, Machine Learning, RS-MultiBoosting, Partial Dependency Plot*

1. Introduction

The development of solutions for the financing issues of small and medium-sized enterprises (SMEs) has attracted attention from scholars and practitioners. Especially in China, the existing problems of SMEs include high financial distress, high operational risk and ambiguous financial information. Additionally, SMEs face a more uncertain competitive environment and are less equipped with the human and capital resources to withstand economic adversity than larger companies (Stiglitz and Weiss, 1981). This may result in a high non-performing loan ratio of China's SMEs. According to a report from the People's Bank of China (PBC), the non-performing loan ratio of China's large enterprises is only 1.19%, in contrast to the SMEs' ratio, which can reach 5.94%. Similarly, the Industrial and Commercial Bank of China (ICBC) points out that the non-performing loan ratio of China's SMEs in the ICBC is more than 5%. Thus, FIs generally consider China's SMEs to be of low credit quality, and they usually refuse and fear financing China's SMEs (Chen et al., 2010, Wong et al., 2016, Wang and Yang, 2014). To overcome this problem, the Chinese government has attempted to improve the current financing situation of SMEs by enforcing financial policy. 'Promoting SME Development Plan (2016-2020)', issued by the Ministry of Industry and Information Technology of the People's Republic of China, emphasizes that the Chinese government encourages qualified FIs to make loans to SMEs. The solutions to SMEs' financing issues are also proposed in this financial policy, which including four main aspects: the development of financial markets, the optimization of the financing environment, the construction of a credit guarantee system for SMEs and the construction of a credit system for SMEs. Under the guidance of financial policy, Chinese FIs have been optimizing the financing environment primarily by developing various financing solutions, which has greatly enhanced SMEs' financial capacity. According to 'The PBC 2016 Annual Report', at the end of 2016, 4.72 hundred thousand SMEs had FI loans. In 2016, the loan balance of SMEs was 10.5 trillion Yuan RMB, an increase of 5.3% year-on-year, which mostly benefited from new or improved financing solutions. One financing solution that has been widely adopted is supply chain financing (SCF), which has attracted increasing attention in academia, financial circles and industry (Yan et al., 2016, Xu et al., 2018). In the SCF solution, a high-quality core enterprise (CE) provides a credit guaranty to SMEs in the same supply chain, making it possible for SMEs to obtain low-cost financing from FIs (Klapper, 2006, Xu et al., 2018).

SCF can be defined in many ways but can primarily be classified as "financial-oriented" and "supply chain-oriented" (Gelsomino et al., 2016). For instance, More and Basu (2013) define SCF as a tool of financing to create value to stakeholders within the supply chain by planning, steering, and controlling the flow of financial resources on an inter-organizational level; Stemmler (2002) and More and Basu (2013) argue that SCF integrates cash flows into the physical supply chain, so it is an approach to supply chain management. SCF also has extensive interactions with related SMEs, which is different from the traditional financing approach (Song et al., 2018). SCF is another arrow in the quiver, encouraging the SMEs to obtain loans based on their creditworthiness (Pfohl and Gomm, 2009). Based on the above point of view, this paper defines the SCF as a financing approach in which an FI links CEs with upstream and downstream SMEs to provide flexible financial products and services, allowing SMEs to expediently receive financing through credit binding with CEs. The SCF solution extricates the members of a supply chain from short-term liquidity dilemmas and long-term financial burdens (Wuttke et al., 2013), mitigating the credit risk of the whole supply chain (Gomm, 2010, Hofmann, 2005). ~~SCF mainly includes three financing patterns, i.e., the supply chain accounts receivable financing pattern, supply chain advance account financing pattern and supply chain inventory financing pattern. In SCF solutions, one CE and some SMEs together apply for financing by specifying their financing details, such as the CE's credit rating and the SMEs' current ratio and cash ratio, which is called '1+N' SCF. Then, the FI assesses the overall credit risk of this financing project. When the whole financing project is found to have a low credit risk, the FI grants a loan to the SMEs. Otherwise, the FI refuses to finance the SMEs. SCF has the potential to increase the efficiency of SMEs' financing.~~ Moreover, FI managers hope that the CE can guarantee that their cooperating SMEs are free of the risk of fraud in the SCF. Unfortunately, this is not an easy task. First, from the supply chain point of view, risks can be spread once one of the members of the supply chain (either the CE or the SME) experiences jeopardy such as bankruptcy, disruption, and so on, which is often out of the SME's control (Hendricks and Singhal, 2005, Tang, 2006, Wuttke et al., 2013). Second, from the SME's point of view, as a supplier the SME might sometimes fall short of the capacity to meet customer demand (Richard et al., 2007). Similarly, as a buyer, the SME might encounter financial difficulty in paying for what is being procured. In both cases, SMEs' risk profile deteriorates. In China, SMEs' credit risk is still deemed the major source of risk in SCF the solution, implicating issues such as a lower credit rating, a high probability of fraud, and an undeveloped credit guarantee system (Su and Lu, 2015).

For this purpose, methods have been developed and applied to forecast SMEs' credit risk (Zhu et al., 2016, Zhu et al., 2017, Zhu et al., 2016), e.g., logistic regression analysis (LRA), an artificial neural network (ANN) and machine learning (ML). As an efficient strategy for achieving high performance forecasting, the hybrid ensemble ML approach has recently attracted increasing attention in the field of credit risk forecasting. For instance, RS-Boosting integrates boosting and random subspace (RS), and it is used to forecast corporate credit risk (Wang and Ma, 2011). RS-RAB integrates random subspace and real AdaBoost, and it is used to forecast SMEs' credit risk in SCF (Zhu et al., 2016). ~~However, these approaches struggle when dealing with a relatively small sample size.~~

In this paper, we propose a new hybrid ensemble ML approach named RS-MultiBoosting. This approach consists of two classical ensemble ML approaches, i.e., RS and MultiBoosting, and it uses a decision tree (DT) as the base classifier. The aim is to improve the forecasting performance of SMEs' credit risk in SCF when the sample size is relatively small. Additionally, to test the feasibility and precision of the RS-MultiBoosting approach, we select a real SME credit dataset from a Chinese securities market and compare RS-MultiBoosting forecasting performance against classical ML approaches, i.e., DT, RS and MultiBoosting.

The remainder of our paper is structured as follows: Section 2 reviews the literature on SMEs' credit risk and the methods of forecasting SMEs' credit risk. Section 3 explains the theory and algorithms of DT, RS and MultiBoosting. Section 4 describes the construction and theory of the proposed RS-MultiBoosting approach. Section 5 presents the case study with information about the data sources and variable definitions. It also discusses the empirical measures and design. Section 6 discusses the results of the experiment. Section 7 contains concluding remarks and future research directions.

2. Literature review

Financing difficulty is a bottleneck that restricts the development of SMEs. The current economic downturn and variations, as shaped after the 2008 global financial crisis, has further increased the financial pressure on SMEs, and thus SMEs urgently search for new ways of financing to obtain easy credit (Ali et al., 2018, Lekkakos and Serrano, 2016). In the past, SMEs would seek financing from commercial banks based on their own terms and credits (Song et al., 2016). However, a limited operating history, incomplete financial statements, insignificant performance, high levels of risk, and many other factors have constrained most SMEs to effectively receive financing through traditional methods (Song and Wang, 2013, Song et al., 2016). More important, information asymmetry has been a huge challenge for SMEs when they seek financing to develop their businesses (Gong and Cullinane, 2018).

In recent years, SCF has become an important product category of FIs because it can fulfil financing requirements and accomplish SMEs' development targets in a timely manner (Ali et al., 2018). Demica, a professional consultant firm of working capital solutions from SCF, reports that the annual growth rate of international SCF reached 30-40% between 2011 and 2013, and the growth rate will not fall below 10% before 2020 (Demica, 2014). SCF has also attracted increasing attention from academia in recent years (Xu et al., 2018). Caniato et al. (2016) propose that the benefits for a company of adopting an SCF solution can be categorized into three macro-typologies, i.e., reduced net operative working capital, increased profit and strategic benefits. Chen and Hu (2008) consider that SCF reduces the mismatch risk of supply and demand in the financial flow and creates value for the supply chain with capital constraints by integrating the FIs, the focal company and capital-constrained firms in the supply chain. Gelsomino et al. (2016) point out that the SCF provides lower debt costs, new opportunities to obtain loans and reduced working capital, especially for weak supply chain players. Gelsomino et al. (2016) and Hofmann (2005) emphasize that the FIs can improve their risk-assessment ability in estimating the probability of default, especially for SMEs using the SCF solution. Caniato et al. (2016) and Song et al. (2018) argue that SCF is an alternative method of overcoming the problem of information asymmetry, so it can control potential risks and provides easy credit to SMEs.

From an economic globalization perspective, Lekkakos and Serrano (2016) argue that the global financial crunch, credit shortages and high borrowing costs hinder SMEs from obtaining loans; nevertheless, the SCF promptly facilitates SMEs' settlement of their operations in this modern age of globalization. As set forth above, it is appreciated that SMEs are capital-constrained and weak supply chain members whose financing performance and competitiveness in supply chains are significantly improved by SCF (Song et al., 2016).

Nevertheless, the SCF solution cannot completely avoid the credit risks of SMEs (Wuttke et al., 2013, Richard et al., 2007, Hendricks and Singhal, 2005, Tang and Musa, 2011). Thus, research on SMEs' credit risk in SCF and its forecasting methods has attracted increasing interest in both academia and industry. However, the literature on SCF mainly focuses on the design and optimization of the flows of goods and information and financial flows between the members in a supply chain (Srinivasa and Mishra, 2011, Yan and Sun, 2013). Until recently, only a few studies have focused on SMEs' credit risk in SCF. In the following literature review, we consider two research areas that are the most closely related to our work: 1) SMEs' credit risk influencing factors in SCF and 2) the methods of forecasting SMEs' credit risk in SCF.

2.1 SMEs' credit risk influencing factors in SCF

The Basel Committee on Banking Supervision (BCBS) is the primary global standard setter for the prudential regulation of banks and provides a forum for regular cooperation on banking supervisory matters. Its 45 members include central banks and bank supervisors from 28 jurisdictions, such as the People's Bank of China, the Swiss National Bank, the Board of Governors of the Federal Reserve System in the US, the Bank of England in the UK and so on. In 'Principles for the Management of Credit Risk (PMCR)' (Supervision, 1999), BCBS defines credit risk as the possibility that a borrower or a lender will not fulfil his or her legal obligations according to the debt contract with the corresponding banking institution. As a member of BCBS, China agrees and follows the definition of credit risk in PMCR. In this paper, we adopt the same definition. Since SCF is different from the traditional financing approach, it has extensive interactions with related SMEs (Song et al., 2018, Stemmler, 2002). Thus, there are two main influencing factors—the "SMEs itself-oriented" factor and the "supply chain finance-oriented" factor—that may result in SMEs' inability to fulfil their legal obligations and FIs facing SMEs' credit risk in SCF (Zhu et al., 2016, Zhu et al., 2017, Zhu et al., 2016). This paper analyses SMEs' credit risk influencing factors in SCF from both sides as follows.

(1) The SMEs' credit risk influencing factors of SMEs that are self-oriented

Scholars argue that information asymmetry is the root cause of SMEs' credit risk (Altman and Sabato, 2007, Altman et al., 2010, Chen et al., 2010, Li et al., 2016, Song and Zhang, 2017, Stiglitz and Weiss, 1981). To control SMEs' credit risks, FIs attempt to overcome the problem of information asymmetry of SMEs through the SCF solution (Caniato et al., 2016, Song et al., 2018). The SCF solution has extensive interactions with related SMEs compared with traditional financing solutions, acquires complete transaction information and business credit of every member of the supply chain, and adopts specific solutions (e.g., receivable transfers, closed-loop business, relational embeddedness, and a combination of outcome control and behavioural control) that significantly reduce the information asymmetry of SMEs (Caniato et al., 2016, Martin and Hofmann, 2017, Song et al., 2018). However, FIs are not directly involved in SMEs' actual operations and production; therefore, to reduce possible the information asymmetry of SMEs, FIs take better care of SMEs' business counterparts (i.e., CEs) in the SCF solution (Song et al., 2018). Klapper (2006) argues that the credit risk becomes the default risk of the high-quality enterprise (CE) instead of the risky SMEs in the SCF solution. Thus, FIs pay more attention to information about the CE than to information about the SME. However, it is undeniable that information about SMEs remains the main influencing factor of SMEs' credit risk in the SCF solution. To guarantee that FIs will avoid the information asymmetry of SMEs, we must first understand what information about SMEs themselves is related to SMEs' credit risk in SCF.

Some scholars believe that financial information is the main factor influencing SMEs' credit risk. For instance, Altman and Sabato (2007) prove that cash, total assets, earnings before tax, interest paid, retained earnings, short-term debt and equity are the primary influencing factors of SMEs' credit risk. Chen et al. (2010) find that the asset size of an enterprise has a significant impact. Calabrese and Osmetti (2013) argue that the SMEs' credit risk is affected by SMEs' solvency ratio, return on equity, turnover per employee, added value per employee, cash flow, bank loans over turnover and total personnel costs over added. Fantazzini and Figini (2009) argue that SMEs' credit risk is affected by their liquidity ratio, debt ratio, short-term over long-term debt, provisions over sales, equity over debt and short-term debt.

However, other scholars believe that non-financial information also significantly affects SMEs' credit risk; such information includes SMEs' filing histories (Altman et al., 2010) and their existing risk information, delinquency information, historical information, historical delinquency information, credit and corresponding guarantee information, and demographic information (Derelioğlu and Gürgen, 2011). Li et al. (2016) argue that there are four main influencing factors of SMEs' credit risk: their profitability, structure, liquidity, operations and characteristics. Figini and Giudici (2011) suggest that the main influencing factors primarily include quantitative risk factors and qualitative risk factors, which are composed of financial information and non-information.

(2) SMEs' credit risk influencing factors of finance-oriented supply chain

Pfohl and Gomm (2009) and Hofmann (2005) describe SCF as a bridge of the supply chain that integrates the financing processes of all members in a supply chain for increasing the value of the supply chain, which is different from the traditional financing approach. Therefore, in SCF, SMEs' credit risk is affected not only by SMEs themselves but also by SCF factors such as the financial and non-financial state of the CE (Wuttke et al., 2013, Zhu et al., 2017, Zhu et al., 2016), the state of supply chain operations (Zhu et al., 2017, Zhu et al., 2016, Hendricks and Singhal, 2005, Tang, 2006), the object's characteristic factors of pledging (Zhu et al., 2017, Zhu et al., 2016) and so on. In particular, the CE is the guarantor of SMEs in the SCF and cooperative enterprise in a supply chain; its credit risk will be transferred to the SMEs (Wuttke et al., 2013). Song and Zhang (2017) find that most credit risks are borne by the guarantor in third-party-guaranteed loans.

2.2 Methods for forecasting SMEs' credit risk in SCF

Over the past decade, traditional statistical approaches have been applied by FIs to forecast SMEs' credit risk and make credit loan decisions based on these traditional financing channels. For example, Z-score and logistic models have been applied by most large banks in the US (e.g., Bank of America) (Altman and Sabato, 2007), the credit rating approach is widely employed by most of the banks in China (e.g., ICBC) (Chen et al., 2010) and multivariate linear discriminant analysis (MLDA) is being adopted by some large banks in Italy (e.g., UniCredit S.p.A. of Italy) (Ciampi et al., 2009). Edmister (1972) proposes a model to predict the default probability of small businesses using the MLDA approach, which is one of the earliest works in the field. Altman and Sabato (2007) develop a one-year default probability prediction model based on LRA, which is specifically intended to predict SMEs' credit risk. Since then, LRA has been widely used; for instance, Ciampi and Gordini (2009) apply LRA to predict the credit risk of SMEs in northern and central Italy. Ciampi et al. (2009) and Ciampi (2015) combine LRA with other techniques, such as linear discriminant analysis (LDA), to predict the credit risk of Italian small enterprises. Calabrese and Osmetti (2013) propose a generalized extreme value regression (GEVR) model that is suitable to predict the loan defaults of SMEs. As expected, these approaches can also be used for forecasting SMEs' credit risk in SCF. Nevertheless, traditional statistical approaches assume a certain data distribution that requires substantial historical data to classify, which makes it very challenging to collect adequate data. Often, the number of collected observations is too small to qualify for use in traditional forecasting methods (Li and Yeh, 2008, Li et al., 2012). In other words, the data gathered in SCF are often insufficient to perform reliable forecasting. Thus, new approaches are needed. ML approaches do not need to assume certain data distributions. Instead, they can extract knowledge by training the model (Wang et al., 2011). Furthermore, ML may achieve acceptable forecasting accuracy even when the dataset is small (Li et al., 2012, Li and Lin, 2008).

There are few studies focusing on credit risk prediction that are specific to SMEs by using ML. For instance, Fantazzini and Figini (2009) propose a new approach based on random survival forests (RSF) and find it performs better than traditional statistical approaches, e.g., LRA. Chen et al. (2010) develop a model based on the key mediating variable (KMV) model to forecast the credit risk of Chinese listed SMEs, and the model is robust to the change in default points in SMEs. Derelioğlu and Gürgen (2011) propose a method based on multi-layer perceptrons (MLP) to predict SMEs' credit risk in Turkey. Zhu et al. (2016) find that the performance of forecasting China's SMEs credit risk in SCF by integrating the LRA and radial basis function (RBF) is better than that of individually applying LRA or RBF. It is demonstrated that the performance of ML is generally better than that of these traditional statistical approaches, especially for classifying limited data with a non-linear distribution (Wang et al., 2011, Wang and Ma, 2012, Wang and Ma, 2011).

ML can be classified into individual ML and ensemble ML approaches. Ensemble ML approaches usually result in better forecasting performance than individual ML approaches (Chen and Huang, 2003, Nanni and Lumini, 2009, Tsai and Wu, 2008). The ensemble ML approach is a kind of ML that integrates multiple individual ML approaches for training the datasets and solving the classification problem. However, some ensemble methods are aimed at reducing the influence of noise data, i.e., the instance partitioning method, while others perform good work when there is redundant information, i.e., the attribute partitioning method. In other words, the existing ensemble methods lack diversity. To enforce the diversity of methods, numerous enhanced hybrid ensemble ML approaches have been proposed, such as RS-boosting by Wang and Ma (2011) and the random subspace-support vector machine (RSB-SVM) by Wang and Ma (2012). In particular, Zhu et al. (2017) and Zhu et al. (2016) prove that the performance of hybrid ensemble ML approaches in forecasting China's SMEs credit risk in SCF is better than that of individual ML and ensemble ML approaches.

In summary, the result of the literature review suggests that the factors influencing SMEs' credit risk in SCF are primarily sourced from "SMEs self-oriented" and "supply chain finance-oriented", and these influencing factors can be generally classified into two types: financial information and non-information. There is also evidence that the ML is becoming the prevailing approach to building the model of forecasting SMEs' credit risk; moreover, the hybrid ensemble ML is usually better than the individual ML and ensemble ML approaches.

Thus, this paper makes three contributions to SCF research. First, we develop an SME credit risk forecasting model, considering both SME-oriented and SCF-oriented influencing factors instead of only one type of factor, as was done previously. Second, we propose a new hybrid ensemble ML approach that is useful in handling relatively small datasets. Third, the result of the model provides a pragmatic guide for practitioners in terms of how to enhance SMEs' financial capability.

3. Existing machine learning approaches

To build the foundation for the enhanced hybrid ensemble ML approach developed in this paper, this section introduces each of the ML approaches that are relevant to the proposed ML model, i.e., the DT approach and two ensemble ML approaches, i.e., the RS and MultiBoosting approaches, which are widely applied to forecast credit risk.

3.1 Decision tree (DT) approach

The DT is a type of classifier that uses a tree graph to classify a sample set by starting at the root and moving through branches and notes until a leaf is encountered (Quinlan, 1993). The DT primarily consists of a decision note (the attribute of the non-classified sample), a decision branch (the different values of different decision notes) and a decision leaf note (a possible classification result). The frequently used algorithms of DT are Iterative Dichotomiser 3 (ID3) and C4.5, both proposed by Quinlan (1993). The ID3 algorithm concentrates on a multivalve attribute, which is propitious for adequately classifying the datasets. However, it is difficult to improve the accuracy of classification. In contrast to ID3, C4.5 can effectively solve the bias of multi-value attributes and improve the accuracy of classification by applying the information gain expansion of the gain ratio (Quinlan, 1993). In addition, the DT is widely used to mine non-linear data, while the C4.5 algorithm is a machine learning method that can solve credit risk forecasting problems with limited datasets. Wang et al. (2014) gain an 84.39% average prediction accuracy with 690-observation credit datasets by using the C4.5 algorithm. Similarly, Wang et al. (2011) obtain a 77.85% average accuracy with 239-observation datasets. Following the description of Quinlan (1993), the pseudo-code of C4.5 is provided in Figure 1.

3.2 Random subspace (RS) approach

Ho (1998) points out that integrating individual ML approaches might produce a highly successful approach, i.e., ensemble ML approaches. To effectively improve the forecasting accuracy and avoid the over-fitting issue of the DT approach, Ho (1998) proposes an ensemble ML approach, i.e., the RS approach. The RS approach consists of multiple trees that are constructed in randomly chosen subspaces, and it is used to improve the generalization accuracy of forecasting performance on training data. Each DT classifier of the RS model is independent. Thus, it is better adapted to learning parallel computing quickly. Nevertheless, the RS model does not lead to local optimum issues. Instead, it converges to the global minimum point. Hence, the RS approach is widely used for forecasting credit risk. In addition, RS can accurately forecast credit risk problems with limited datasets. Wang and Ma (2011) obtain average prediction accuracy rates of 81.03% and 80.68% with 239-observation and 132-observation credit datasets using the RS model. Following the description of Ho (1998), the pseudo-code of the RS approach algorithm is specified in Figure 2.

3.3 MultiBoosting approach

As an efficient strategy for improving the forecasting accuracy of ML, researchers are increasingly paying attention to ensemble ML approaches, such as boosting. Based on boosting, Freund and Schapire (1996) propose an improved version of boosting, i.e., adaptive boosting (AdaBoost), and prove that the AdaBoost is more practical and easier to implement than boosting. Bauer and Kohavi (1999) find that AdaBoost's abilities to decrease error and variance are prominent, but its ability to decrease superior variance is worse than that of other ensemble ML approaches, such as Bagging. Wagging is an improved version of Bagging that is better suited to the task of reducing superior variance than direct Bagging (Webb, 2000). In addition, AdaBoost can significantly reduce both error and variance, while Wagging has little effect on error and a greater effect on variance (Webb, 2000). Since Webb (2000) proposes a classic ensemble ML based on AdaBoost and Wagging, i.e., MultiBoosting. MultiBoosting has lower error than either AdaBoost or Wagging when using C4.5 as the base learning algorithm, which also suits parallel computing (Webb, 2000). MultiBoosting is seen as an appropriate credit risk forecasting method with limited datasets, e.g., Zhu et al. (2017) obtain an 84.08% average prediction accuracy with 377-observation credit datasets by using the MultiBoosting model. The pseudo-code of the algorithm of the MultiBoosting approach is presented in Figure 3.

4. A new hybrid ensemble machine learning approach: RS-MultiBoosting

Since individual ML methods attempt to obtain a hypothesis from the training data, the noise data and the redundant attributes will reduce accuracy (Wang et al., 2012). In addition, individual ML method is difficult to obtain robust forecasting results when the datasets are numerically limited (Li et al., 2012). In contrast to individual ML methods, the ensemble ML approaches attempt to construct a set of hypotheses and combine them to solve the same problem (Wang et al., 2012). Accordingly, ensemble ML approaches normally are more accurate than to individual ML approaches, especially for limited datasets. However, Wang and Ma (2012) argue that a good ML approach should not only improve accuracy but also enforce diversity, which means that each base learning method in ML methods makes its own contribution to the classification decision and a different form of error from each other. Some ensemble methods are aimed at reducing the influence of the noise data, i.e., the instance partitioning method, while others work well when there is redundant information, i.e., the attribute partitioning method. MultiBoosting belongs to the former while RS belongs to the latter. Individually using either RS or MultiBoosting should results in a lack of diversity. Furthermore, researchers prove that diversity can significantly improve the forecasting performance of ML methods (Wang et al., 2011, Wang and Ma, 2011, Wang and Ma, 2012, Zhu et al., 2017, Zhu et al., 2016). According to the above analysis, we propose a new hybrid ensemble ML approach, i.e., the RS-MultiBoosting approach, which integrates the RS and

MultiBoosting approaches; the DT is taken as the base learning method of RS-MultiBoosting. This new hybrid ML approach is expected to improve the accuracy and enforce the diversity of the individual and ensemble ML methods. The whole working mechanism of RS-MultiBoosting is illustrated in Figure 4.

Figure 4 shows the following: (i) The dataset is split into sub-datasets by bootstrap sampling with the replacement approach in MultiBoosting. (ii) The new sub-datasets are selected from the original sub-datasets by the RS approach. (iii) The new sub-datasets are trained by the MultiBoosting approach. (iv) The final results are aggregated by the majority vote approach. Through the above workflow, we combine the advantages of the instance partitioning approach (RS) with those of the attribute partitioning approach (MultiBoosting). Moreover, based on Opitz and Maclin (1999) and Fu et al. (2006), some researchers using C4.5 as the base learning algorithm of hybrid ensemble ML methods, such as Wang and Ma (2011), use it for RS-Boosting, Zhu et al. (2016) use it for RS-RAB, Wang et al. (2012) use it for Bagging-RS and RS-Bagging, and so on. In this paper, we also apply the C4.5 algorithm of the DT approach as the base learning algorithm of the RS-MultiBoosting approach.

Based on the above discussion, we present the pseudo-code of the algorithm of the RS-MultiBoosting approach in Figure 5.

5. Numerical example: a case study of SMEs' credit risk forecasting in China

5.1 Data sources

To compare the performance of the ML methods in forecasting SMEs' credit risk in SCF, we need a proper dataset. As a new financing solution, the SCF is not widely applied in China; it is difficult to obtain a complete dataset of SCF, especially for private corporations. Instead, SCF applications primarily involve Chinese quoted companies, such as Yonghui Superstores, a quoted company in China's Shanghai Stock Exchange; it proposes batch credit for suppliers by SCF solutions. The Amarsoft, a quoted company in China's Shenzhen Stock Exchange, constructs SCF management platforms for banks and enterprises. Thus, we focus on Chinese quoted companies when searching for a proper dataset. The sample selection criteria include three factors. First, the SMEs must be listed on the Small and Medium Enterprise Board of the Shenzhen Stock Exchange. These Chinese quoted companies are representative of SMEs, which have problems such as insignificant performance, low creditworthiness, financial pressure and so on. Second, the CEs are selected from the Shanghai Stock Exchange and the main board of the Shenzhen Stock Exchange. These Chinese quoted companies are the leading enterprises, which have some superiorities such being industry leaders, having high creditworthiness, enjoying strong financial strength and so on. Third, all quoted SMEs must have real trading relationships with one of the quoted CEs that is part of a supply chain. In other words, among the selected samples, the SME must be either a CE's supplier or its buyer. Hence, when SMEs have financing requirements, with the involvement of the CEs, SCF can take place. Based on the above sample selection criterion, we select forty-six quoted SMEs from the Small and Medium Enterprise Board of the Shenzhen Stock Exchange and seven quoted CEs from the Shanghai Stock Exchange and the main board of Shenzhen Stock Exchange during 31 March 2014 - 31 December 2015. After deleting unavailable data points, a valid quarterly 365-observation dataset remains.

5.2 Variable definitions

The forty-six quoted SMEs comprise six star-special treatments (*ST), i.e., risky SMEs (negative credit status), and forty quoted companies with normal financial indicators, i.e., non-risky SMEs (positive credit status). Based on the credit status of SMEs, we classify the dependent variables into two groups, i.e., the dependent variables are assigned the value of 0 or 1, indicating a quarterly 365-observation dataset of risky and non-risky firms. Following Zhu et al. (2016), Zhu et al. (2017) and Zhu et al. (2016), 18 original independent variables[1] are selected and can be segmented into five categories: leverage, liquidity, profitability, activity and non-financial. Table 1 defines the V_o and their categories.

Furthermore, we need to select the most important independent variables from the above variables. On the one hand, selecting appropriate independent variables for ML models is important for improving the accuracy of forecasting and reducing computation time and over-fitting (Gouvêa and Gonçalves, 2007). On the other hand, the selected independent variables can inform FI managers about which factors are important for forecasting SMEs' credit risk in SCF, SME managers about which factors are important for improving financing ability and CE managers about which factors are important for reducing the credit risk of joint liability.

5.3 Performance measures

The forecasting performances of the RS-MultiBoosting, RS, MultiBoosting and DT approaches are assessed by computing the mean values of average accuracy, the type I error, the type II error and the F-measure. These four evaluation criteria are defined as follows:

$$\text{Average Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

$$\text{Type I Error} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Type II Error} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{3}$$

[1] The "18 original independent variables" is symbolized by V_o.

$$\text{Precision Rate} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F - Measure} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{6}$$

where TP, TN, FP and FN refer to true positive, true negative, false positive and false negative, respectively. Eqs. (1), (2) and (3) imply that a good forecasting approach should have high average accuracy and a low type I error and type II error. Scholars point out that the correct discrimination of the 'positive samples' can help FIs cluster credit risk customers and non-credit risk customers (Bekhet and Eletter, 2014, Kürüm et al., 2012, West, 2000, Yap et al., 2011). Additionally, the type I error means the model incorrectly classifies the positive samples into negative samples and the type II error means the model incorrectly classifies the negative samples into positive samples. In practice, FIs incorrectly classify non-risky customers into risky customers rather than classify risky customers into non-risky customers. The former results in FIs losing potential customers, while the latter results in FIs facing risk. Thus, Wang and Ma (2011) argue that decreasing the type II error is more important to a forecasting model than decreasing the type I error. The precision rate refers to the ratio of the number of correct 'real positive' cases to the number of 'predicted positive' cases (see equation (4)). The higher the precision rate, the lower the 'false positive rate' that the ML approach obtains. In turn, the recall rate means the ratio of the number of correct predicted positive cases to the number of real positive cases (Powers, 2011) (see equation (5)). The higher the recall rate, the higher the 'true positive rate' of the ML approach (Powers, 2011). The F-measure is the arithmetical average of the precision rate (i.e., positive predictive) and the recall rate (i.e., sensitivity) (Powers, 2011) (see equation (6)). Therefore, the higher the F-measure is, the better the forecasting performance of the ML approach.

5.4 Experimental designing

To prove that the selected independent variables contribute to improving the forecasting performance of the RS-MultiBoosting model over the V_o, we construct and test models with V_o and selected independent variables. Moreover, to lower the forecasting error and the influence of the variability of the training set, we apply the 10-fold cross validation method to test the forecasting approaches based on five values (0.5, 0.6, 0.7, 0.8 and 0.9) of RS rates (i.e., the size of the subspace).

The experiments are performed on a PC with a 2.60 GHz Intel Core i7-6500U CPU and 8.00 GB RAM using the Windows 10 operating system. The data mining toolkit Waikato Environment for Knowledge Analysis (WEKA) version 3.8.2 is used for the experiment. WEKA is a popular and free suite of machine learning and data mining software written with Java and developed by the 'Machine Learning Group' at the University of Waikato in New Zealand.

To implement the experiment, we apply the 'Data Mining Processes' in the 'WEKA Knowledge-Flow Environment' that performs the experiment as follows: first read the data by 'Arff Loader' flow; second, choose the dependent variable by 'Class-Assigne' flow; third, split the valid data into desired sets by 'Cross Validation Fold Maker Customizer' flow; fourth, test and train the data set by 3 'Classifier Meta' and 1 'Classifier Tree' flows: a 'Classifier Meta' flow of RS; a 'Classifier Meta' flow of MultiBoostAB (the MultiBoosting model in WEKA); and a 'Classifier Meta' flow integrated by RS and MultiBoostAB; the 'Classifier Tree' is J48 (the C4.5 algorithm of DT model in WEKA); fifth, separately evaluate the performances of DT, RS, MultiBoosting and RS-MultiBoosting models using 4 'Classifier Performance Evaluator' flows; finally, display the evaluation results of classifiers by 'Text Viewer' flow. The above experiment steps are illustrated in Figure 6.

6. Results and discussion

6.1 Selected independent variable and discussion

The selected independent variables usually have more forecasting power than the original independent variables. Thus, they help improve forecasting performance and provide managerial insights to SME, CE and FI managers. The top 12 variables[2] (see Table 2) of the V_o are selected based on the relative importance score of the DT (Elith et al., 2008, Friedman, 2001, Ransoma et al., 2017). Table 2 indicates that the current ratio of SMEs, the features of the trade goods and the credit rating of the CE have the top 3 highest relative importance scores. Notably, six of the V_s are SME-related factors. This finding indicates that the situation of SMEs is still the main effect factor in assessing SMEs' credit risk in SCF compared with traditional financing. However, the other six variables are CE situation- or supply chain-related factors, which are newly introduced into assessing SMEs' credit risk in SCF compared with traditional financing.

To facilitate discussion, we generally divided the V_s into two categories in Table 2: traditional financing factors and supply chain financing factors. As mentioned in section 2.1, the various kinds of information regarding SMEs are the main influencing factors of SMEs' credit risk in traditional financing, thus, they are classified as traditional financing factors. The rest of the six factors are classified as supply chain financing factors.

[2] The "12 selected independent variables" are symbolized by V_s.

6.2 Model forecasting performance evaluation

The geometric and numerical results of average accuracy, the type I error, the type II error and the F-measure of the RS, MultiBoosting, DT and RS-MultiBoosting models based on the V_o and the V_s are illustrated in Table 3.

To facilitate discussion, the experimental results of the four models based on the V_o and the V_s are presented in Figures 7 and 8, respectively. The findings are discussed below.

(1) Based on the V_o, when the RS rate is set to 0.6 and 0.8, the RS and RS-MultiBoosting models gain the best forecasting performance among the five RS rates, respectively (see Figure 7).

(2) Based on the V_s, when the RS rate is set to 0.6 and 0.9, the RS and RS-MultiBoosting models gain the best forecasting performance among the five RS rates, respectively (see Figure 8).

(3) Notably, the forecasting performances of the DT and MultiBoosting models based on the V_s are worse than those of the models based on the V_o (see Table 3). However, the forecasting performances of the RS and RS-MultiBoosting models based on the V_s are improved compared with those of the models based on the V_o (see Table 3). Notably, the type II errors of the RS and RS-MultiBoosting models based on the V_s are lowered, which is important for FIs to distinguish risky SMEs from SMEs that are suitable for prospective financing.

(4) As an individual ML approach, the DT model obtains a better forecasting performance, based not only on the V_o but also on the V_s, than some ensemble ML approaches, e.g., the RS and MultiBoosting models (see Table 3). Unlike the previous literature (e.g., Wang and Ma

(2011), we find that the individual ML approach is not always weaker than the ensemble ML approach.

(5) More importantly, the results show that the proposed hybrid ensemble ML approach, i.e., RS-MultiBoosting, has the best forecasting performance among the four forecasting approaches (see Table 3).

6.3 Analysis of the partial dependency plot (PDP)

In section 6.1, we select the V_s from the V_o and list the ranking of the V_s based on the relative importance score of the DT (see Table 2). Doing so informs FI managers about which variables (i.e., independent variables) are important for managing SMEs' credit risk in SCF. In section 6.2, we find that the proposed hybrid ensemble ML approach, i.e., the RS-MultiBoosting, has better forecasting performance than the individual and ensemble ML approaches (see Table 3). Doing so helps FI managers more accurately forecast SMEs' credit risk and CE managers more confidently choose an SME as a financing partner in SCF.

In practice, however, managers usually want to know not only the most important independent variables and the performance of forecasting but also *how* these independent variables affect the predicted responses. In this way, FI managers know *how* to reduce the financing credit risk, SME managers know *how* to improve financing ability, and CE managers know *how* to reduce the credit risk of joint liability. Friedman (2001) was the first to propose the method of partial dependency plot (PDP) analysis, which is an effective way to analyse how the independent variables affect the predicted responses by using graphical visualization. The PDP visualizes the non-linear or linear relationships between the independent variables and predicts responses through a training regression model, such as the regression tree model. The

PDP can create a line plot of the predicted responses against a single feature while marginalizing over the other independent variables (Friedman, 2001). In this paper, we will adopt PDP to analyse how the V_s affect the probability of non-risky SMEs.

Because of space limitations, a full description of the partial dependence functions is not provided in this paper. The experiment is performed on a PC with a 2.60 GHz Intel Core i7-6500U CPU and 8.00 GB RAM using the Windows 10 operating system. Matrix Laboratory (MATLAB) version R2018a (9.4.0.813654) is used for the PDP experiment. In this section, we focus on how each of the variables impacts the risk assessment, as shown in Figures 9 and 10.

(1) The impact of traditional financing factors

Figure 9(a) indicates that the larger the current ratio of SMEs, the higher the probability of non-risky SMEs in the general trend. The probability of non-risky SMEs reaches the first platform (0.97) when the current ratio of SMEs is approximately 2, ranging between 1.81 and 2.25 (see the two left broken lines in the figure). The probability slightly descends from the first platform (0.97) to the second platform (0.96) when the current ratio is larger than 2.25 (see the middle broken line). Again, the probability of non-risky SMEs descends from the second platform (0.96) to the third platform (0.92) when the current ratio is higher than 3.92 (see the right broken line). This phenomenon accords with the financial feature of the current ratio. Scholars and financial managers usually consider that the short-term debt-paying ability of enterprises is very strong when the current ratio of enterprises is 2:1. Thus, the probability of non-risky SMEs is the highest when the current ratio is approximately 2. However, the enterprise owns either excessive cash or inventory when the current ratio is much higher than 2. Excessive cash means that the fund utilization efficiency of enterprises is low, and excessive inventory means that the management of enterprises is poor. Thus, the probability of non-risky SMEs decreases progressively while the current ratio increases progressively. In summary, to improve their financing ability, SMEs need to control the value of the current ratio at approximately 2:1.

Figure 9(b) indicates that a higher profit margin on the sales of SMEs signifies a higher probability of non-risky SMEs in the general trend. Normally, a higher profit margin on sales means that enterprises have stronger profitability. To improve their financing ability, SMEs need to improve their profit margin on sales.

Figure 9(c) indicates that the higher quick ratio of SMEs signifies a higher probability of non-risky SMEs in the general trend. We also note that the probability of non-risky SMEs reaches the top platform (0.97) when the quick ratio of SMEs is between 1.61 and 3.01 (see the two broken lines in the figure). In that case, the probability of non-risky SMEs descending from the top platform (0.97) to the next platform (0.93) along with the quick ratio is larger than 3.01. Scholars and financial managers usually consider that an enterprise has poor short-term debt-paying ability and high repayment risk when the quick ratio of the enterprise is less than 1. However, this does not mean that the higher the quick ratio, the better the situation of the enterprise. When the quick ratio is too high, it signifies that the production capacity of enterprises is limited. Thus, to improve their financing ability, SMEs need to control the quick ratio within reasonable limit.

Figure 9(d) indicates that a higher rate of return on total assets of SMEs signifies a higher probability of non-risky SMEs in the general trend. Additionally, the probability of non-risky SMEs reaches the top platform (0.78) when the rate of return on total assets of SMEs is above

2.38 (see the broken line in the figure). To improve their financing ability, SMEs need to improve their rate of return on total assets.

Figure 9(e) shows a non-linear behaviour. Normally, a higher total asset growth rate means a faster expansion velocity of the asset management scale of enterprises in a certain period. However, SMEs usually face the issue of blind expansion in a short period, which potentially leads to operational risk.

Figure 9(f) indicates that a higher cash ratio of SMEs signifies a higher probability of non-risky SMEs in the general trend. Normally, a higher cash ratio means a stronger ability of enterprises to pay debts that are immediately due. However, the profitability of enterprises is poor when the cash ratio is too high. Thus, this figure indicates that the probability of non-risky SMEs reaches the top platform (0.99) when the cash ratio of SMEs is between 4.16 and 5.29 (see the two broken lines in the figure). In that situation, the probability of non-risky SMEs descends from the top platform (0.99) to the next platform (0.95) when the cash ratio is higher than 5.29. To improve their financing ability, SMEs need to control the cash ratio within reasonable limits.

(2) The impact of SCF factors

Figure 10(a) indicates an interesting phenomenon, i.e., that the probability of non-risky SMEs reaches the lowest platform (0.61) when the features of trade goods are between -0.50 and 0.72 (see the two broken lines in the figure). Normally, the higher features of trade goods mean a lower debt risk in SCF. However, we appreciate that the features of the trade goods between SMEs and the CE are classified based on factors such as price rigidity, liquidation, degree of vulnerability and so on, some of which are difficult to quantify, leading to unforeseeable trends.

Figure 10(b) indicates that the higher credit rating of the CE signifies a higher probability of non-risky SMEs in the general trend. The figure indicates that there are three platforms, i.e., the left platform, medium platform and right platform, which are divided by two broken lines in this figure. Notably, the probability of non-risky SMEs is significantly improved from the left platform to the medium platform. However, the probability of non-risky SMEs is slightly improved from the medium platform to the right platform. This finding means that the effect of the CE's credit rating on the probability of non-risky SMEs constitutes diminishing marginal utility. To improve their financing ability, SMEs must cooperate with a strong CE that has a good credit rating.

Figure 10(c) indicates that the higher the profit margin on the sales of the CE, the higher the probability of non-risky SMEs in the general trend. We note that there are four platforms in this figure, which are divided by three broken lines. On a certain platform, the probability of non-risky SMEs will not be changed regardless of improving the profit margin on the sales of the CE. This finding helps SMEs to effectively select the CE for applying for SCF, and it helps FI managers make the right financing decision.

Figure 10(d) indicates the role of the industry trend. A higher value of the industry trend indicates the velocity of industry change. In general, in a relatively 'stable' industry, i.e., -0.15 to 1.5, the SME's credit risk is lower than in a rapidly changing industry, i.e., larger than 1.5. This finding reflects the reality that SMEs in conventional industry take advantage of their long-established businesses over those in rapidly changing industry, for instance, the traditional car market vs. the car-sharing business model. Thus, to decrease the financing credit risk, FI managers need to conduct research into the industry trends.

Figure 10(e) is a typical non-linear behaviour. The quick ratio reflects the short-term debt-paying ability of enterprises. The enterprise has a high repayment risk when the quick ratio of the enterprise is too low. However, this does not mean that the higher the quick ratio, the better the situation of the enterprise. When the quick ratio is too high, it signifies that the production capacity of enterprises is limited.

Figure 10(f) indicates that the probability of non-risky SMEs reaches the top platform (0.97) when the accounts receivable collection period of SMEs is between 1.06 and 1.89 (see the two broken lines in the figure). In that case, the probability of non-risky SMEs sharply declines from the top platform (0.97) to the lower platform (0.83) when the accounts receivable collection period is larger than 1.89. Normally, a shorter period of accounts receivable collection means higher efficiency in the use of the working capital of enterprises. It is widely accepted that a long accounts receivable collection period is one of the most important factors that force SMEs to apply for financing. Thus, the motivation for SMEs' financing might be called into question if the accounts receivable collection period of SMEs is short. Hence, to minimize the financing credit risk, FI managers need to pay more attention to the period of accounts receivable collection of SMEs.

It is noted that although the line PDP cannot provide a complete representation of the effect of each feature on the predicted responses, it can provide a useful reference (Friedman and Meulman, 2003).

7. Conclusion and future research

Forecasting SMEs' credit risk in SCF has become a significant task, as FIs must decide whether to finance an SME that collaborates with a CE and applies for SCF services. Increasing forecasting accuracy by even only a small percent may prevent great losses from occurring. To date, ensemble ML approaches have been widely used to refine the performance in forecasting enterprises' credit risk in traditional finance channels. However, there is little research focusing on SCF. In this paper, we **first** select the V_s from V_o extracted from the existing literature as the independent variables of the models based on the relative importance score of the DT. By comparing the relative importance scores of the V_s, we find that the current ratio of SMEs, the features of the trade goods between SMEs and the CE and the credit rating of the CE are the three most important factors related to SMEs' credit risk in SCF. **Second**, we develop a new hybrid ensemble ML approach, i.e., RS-MultiBoosting, to improve performance in forecasting the credit risk of China's SME in SCF, using a dataset consisting of forty-six SMEs and seven CEs. By comparing the forecasting performance of the DT, RS, MultiBoosting and RS-MultiBoosting approaches based on the evaluation criteria of average accuracy, the type I error, the type II error and the F-measure, we find that the RS-MultiBoosting approach obtains better results with small datasets than directly applying the individual ML approach or ensemble ML approaches. In addition, MultiBoosting and RS considerably improve the diversities of instance and feature, which significantly reduces the test error. Moreover, we prove that screening the independent variables based on the relative importance score of the DT significantly improves the accuracy of RS-MultiBoosting in forecasting SMEs' credit risk in SCF. Consequently, this paper contributes to enriching the ML approach to SMEs' credit risk forecasting in the context of SCF and provides a significant reference for assessing SMEs' credit risk in practical work, especially using small datasets. **Third**, SMEs can improve their financing ability by cooperating with a CE that has strong credit standing and financial standing in SCF. Additionally, the credit rating of the CE, the features of the trade goods between SMEs and the CE, industry trends, the accounts receivable collection period of SMEs, the profit margin on sales of the CE and the quick ratio of the CE are the new effective evaluation indicators of SMEs' credit risk in SCF compared with traditional financing. Individually fixing the financial standing of SMEs is not an effective way to improve their financing ability because FI evaluations of SMEs' credit risk are based on the whole supply chain, not the organization alone.

In view of this study, some avenues for future research on forecasting SMEs' credit risk in SCF with ML approaches also emerge. First, large datasets for experiments, particularly with primary data on SMEs in SCF, should be collected for future research. Doing so will help verify the performance of the RS-MultiBoosting method. Second, in this paper, we compare the forecasting performance of the RS-MultiBoosting approach only with that of the DT, RS and MultiBoosting ML approaches. Comparing the RS-MultiBoosting approach with other hybrid ensemble ML approaches would be worthwhile. Third, more hybrid ensemble ML approaches should be researched.

References:

Ali, Z., Gongbing, B. & Mehreen, A., 2018. Does supply chain finance improve SMEs performance? The moderating role of trade digitization. Business Process Management Journal. https://doi.org/10.1108/BPMJ-05-2018-0133.

Altman, E. I. & Sabato, G., 2007. Modeling credit risk for SMEs: evidence from the US market. A Journal of Accounting, Finance and Business Studies. 43 (2), 332-357.

Altman, E. I., Sabato, G. & Wilson, N., 2010. The value of non-financial information in small and medium-sized enterprise risk management. The Journal of Credit Risk. 6 (2), 95-127.

Bauer, E. & Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Machine Learning. 36 (1-2), 105-139.

Bekhet, H. A. & Eletter, S. F. K., 2014. Credit risk assessment model for Jordanian commercial banks: neural scoring approach. Review of Development Finance. 4 (1), 20-28.

Calabrese, R. & Osmetti, S. A., 2013. Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. Journal of Applied Statistics. 40 (6), 1172-1188.

Caniato, F., Gelsomino, L. M., Perego, A. & Ronchi, S., 2016. Does finance solve the supply chain financing problem? Supply Chain Management-AN INTERNATIONAL JOURNAL. 21 (5), 534-549.

Chen, M. C. & Huang, S. H., 2003. Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems With Applications. 24 (4), 433-441.

Chen, X. & Hu, C., 2008. The value of supply chain finance. Supply Chain Managemenet-Applications and Simulations. InTech, Rijeka, pp.111-132.

Chen, X., Wang, X. & Wu, D. D., 2010. Credit risk measurement and early warning of SMEs: An empirical study of listed SMEs in China. Decision Support Systems. 49 (3), 201-310.

Ciampi, F., 2015. Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. Journal of Business Research. 68 (5), 1012-1025.

Ciampi, F. & Gordini, N., 2009. Default prediction modeling for small enterprises: evidence from small manufacturing firms in Northern and Central Italy. Oxford Journals. 8 (1), 13-29.

Ciampi, F., Vallini, C., Gordini, N. & Benvenuti, M., 2009. Are credit scoring models able to predict small enterprise default? Statistical evidence from Italian firms. Journal of Business & Economics. 8 (1), 3-18.

Demica., 2014. A study on the growth of supply chain finance, as evidenced by SCF dedicated job titles at top European bank. Demica, London.

Derelioğlu, G. & Gürgen, F., 2011. Knowledge discovery using neural approach for SME's credit risk analysis problem in Turkey. Expert Systems with Applications. 38 (8), 9313-9318.

Edmister, R., 1972. An empirical test of financial ratio analysis for small business failure prediction. Journal of Financial and Quantitative Analysis. 7 (2), 1477-1493.

Elith, J., Leathwick, J. & Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology. 77 (4), 802-813.

Fantazzini, D. & Figini, S., 2009. Random survival forests models for SME credit risk measurement. Methodology and Computing in Applied Probability. 11 (1), 29-45.

Figini, S. & Giudici, P., 2011. Statistical merging of rating models. Journal of the Operational Research Society. 62 (6), 1067-1074.

Freund, Y. & Schapire, R. E., 1996. Experiments with a new boosting algorithm. Thirteenth International Conference on Machine Learning. 148-156.

Friedman, J. H. & Meulman, J. J., 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine. 22 (9), 1365-1381.

Friedman, J., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 29 (5), 1189-1232.

Fu, Z., Golden, B. L., Lele, S., Raghavan, S. & Wasil, E., 2006. Diversification for better classification trees. Computers & Operations Research. 33 (11), 3185-3202.

Gelsomino, L. M., Mangiaracina, R., Perego, A. & Tumino, A., 2016. Supply chain finance: a literature review. International Journal of Physical Distribution & Logistics. 46 (4), 348-366.

Gomm, M. L., 2010. Supply chain finance: applying finance theory to supply chain management to enhance finance in supply chains. International Journal of Logistics: Research and Applications. 13 (2), 133-142.

Gong, S. & Cullinane, K., 2018. Different pespectives on supply chain finance-In serach of a holistic approach. Finance and Risk Management for International Logistics and the Supply Chain. Joe Hayton, Chennai, pp.35-54.

Gouvêa, M. A. & Gonçalves, E. B., 2007. Credit risk analysis applying logistic regression, neural networks and genetic algorithms models. In: Production and Operations Management Society (POMS) 18th Annual Conference. 4-7.

Hendricks, K. B. & Singhal, V. R., 2005. Association beween supply chain glitches and operating performance. Management Science. 51 (5), 695-711.

Ho, T. K., 1998. The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence,IEEE Transactions on. 20 (8), 832-844.

Hofmann, E., 2005. Supply chain finance: some conceptual insights. Beiträge zu Beschaffung und Logistik, pp.149.

Klapper, L., 2006. The role of factoring for finacing small and medium enterprises. Journal of Banking & Finance. 30 (11), 3111-3130.

Kürüm, E., Yildirak, K. & Weber, G. W., 2012. A classification problem of credit risk rating investigated and solved by optimization of ROC curve. Central European Journal of Operations Research. 20 (3), 529-557.

Lekkakos, S. D. & Serrano, A., 2016. Supply chain finance for small and medium sized enterprises: the case of reverse factoring. International Journal of Physical Distribution & Logistics Management. 46 (4), 367-392.

Li, D. & Lin, Y., 2008. Learning management knowledge for manufacturing systems in the early stages using time series data. European Journal of Operational Research. 184 (1), 169-184.

Li, D. & Yeh, C., 2008. A non-parametric learning algorithm for small manufacturing data sets. Expert Systems with Applications. 34 (1), 391-398.

Li, D., Chang, C., Chen, C. & Chen, W., 2012. A grey-based fitting coefficient to build a hybrid forecasting model for small data sets. Applied Mathematical Modelling. 36 (10), 5101-5108.

Li, K., Niskanen, J., Kolehmainen, M. & Niskanen, M., 2016. Financial innovation: credit default hybrid model for SME lending. Expert Systems With Applications. 61 (5), 343-355.

Martin, J. & Hofmann, E., 2017. Involving financial service providers in supply chain finance practices: company needs and service requirements. Journal of Applied Accounting Researh. 18 (1), 42-62.

More, D. & Basu, P., 2013. Challenges of supply chain finance: A detailed study and a hierarchical model based on the experiences of an Indian firm. Business Process Management Journal. 19 (4), 624--647.

Nanni, L. & Lumini, A., 2009. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications. 36 (2), 3028-3033.

Opitz, D. & Maclin, R., 1999. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research. 11, 169-198.

Pfohl, H. & Gomm, M., 2009. Supply chain finance: optimizing financial flows in supply chains. Logistics Research. 1 (3-4), 149-161.

Powers, D. M. W., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies. 2 (1), 37-63.

Quinlan, J. R., 1993. Obataining the C4.5 code. C4. 5: programs for machine learning. Morgan Kaufmann Publishers, San Francisco, pp.10-11.

Ransoma, K. M., Nolan, B. T., Traum, J. A., Faunt, C. C., Bell, A. M., Gronberg, J. A. M., Wheeler, D. C., Rosecrans, C. Z., Jurgens, B., Schwarz, G. E., Belitz, K., Eberts, S. M., Kourakos, G. & Harter, T.,

2017. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley Aquifer, California, USA. Science of the Total Environment. 601-602 (5), 1160-1172.

Richard, A. B., Myers, S. C. & Allen, F., 2007. Risk Management. MCGRAW-HILL Higher Education, pp.754-815.

Song, H. & Wang, L., 2013. The impact of private and family firms' relational strength on financing performance in clusters. Asia Pacific Journal of Management. 30 (3), 735-748.

Song, H., Yu, K. & Lu, Q., 2018. Financial service providers and banks' role in helping SMEs to access finance. International Journal of Physical Distribution & Logistics Maanagement. 48 (1), 69-92.

Song, H., Yu, K., Ganguly, A. & Turson, R., 2016. Supply chain network, information sharing and SME credit quality. Industrial Management & Data Systems. 116 (4), 740-758.

Song, Z. & Zhang, X., 2017. Lending technology and credit risk under different types of loans to SMEs: evidence from China. International Review of Economics & Finance. In Press.

Srinivasa, R. N. R. & Mishra, V. K., 2011. Short-term financing in a cash-constrained supply chain. International Journal of Production Economics. 134 (2), 407-412.

Stemmler, L., 2002. The Role of Finance in Supply Chain Management. Cost Management in Supply Chains. Physica, Heidelberg, pp.165-176.

Stiglitz, J. E. & Weiss, A., 1981. Credi rartioning in markets with imperfect information. American Economic Review. 71 (3), 393-410.

Su, Y. & Lu, N., 2015. Simulation of game model for supply chain finance credit risk based on multi-agent. Open Journal of Social Sciences. 3 (1), 31-36.

Supervision, C. B., 1999. Principles for the management of credit risk. http://www.bulentsenver.com/BS_ReadingMaterials/PrinciplerForMngOfCreditRisk_bcbs75.pdf. 2016-08-28.

Tang, C. S., 2006. Perspectives in supply chain risk management. International Journal of Production Economics. 103 (2), 451-488.

Tang, O. & Musa, S. N., 2011. Identifying risk issues and research advancements in supply chain risk management. International Journal of Production Economics. 133 (1), 25-34.

Tsai, C. & Wu, J., 2008. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications. 34 (4), 2639-2649.

Wang, G. & Ma, J., 2011. Study of corporate credit risk prediction based on integrating boosting and random subspace. Expert Systems With Applications. 38 (11), 13871-13878.

Wang, G. & Ma, J., 2012. A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine. Expert Systems with Applications. 39 (5), 5325-5331.

Wang, G., Hao, J., Ma, J. & Jiang, H., 2011. A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications. 38 (1), 223-230.

Wang, G., Ma, J. & Yang, S., 2014. An improved boosting based on feature selection for corporate bankruptcy prediction. Expert Systems with Applications. 41 (5), 2353-2361.

Wang, G., Ma, J., Huang, L. & Xu, K., 2012. Two credit scoring models based on dual strategy ensemble trees. Knowledge-Based Systems. 26 (1), 61-68.

Wang, P. & Yang, Q., 2014. Analysis on financing of small and micro enterprises. Management & Engineering. 15 (1), 1838-5745.

Webb, G. I., 2000. Multiboosting: a technique for combining boosting and wagging. Machine Learning. 40 (2), 159-196.

West, D., 2000. Neural network credit scoring models. Computers & Operations Research. 27 (11), 1131-

1152.
Wong, A., Lu, W., Tjosvold, D. & Yang, J., 2016. Extending credit to small and medium size companies: relationships and conflict management. International Journal of Conflict Management. 27 (3), 331-352.
Wuttke, D. A., Blome, C. & Henke, M., 2013. Focusing the financial flow of supply chains: an empirical investigation of financial supply chain management. International Journal of Production Economics. 145 (2), 773-789.
Xu, X., Chen, X., Jia, F., Brown, S., Gong, Y. & Xu, Y., 2018. Supply chain finance: A systematic literature review and bibliometric analysis. International Journal of Production Economics. 204, 160-173.
Yan, N. & Sun, B., 2013. Coordinating loan strategies for supply chain financing with limited credit. OR Spectrum. 35 (4), 1039-1058.
Yan, N., Sun, B., Zhang, H. & Liu, C., 2016. A partial credit guarantee contract in a capital-constrained supply chain: Financing equilibrium and coordinating strategy. International Journal of Production Economics. 173, 122-133.
Yap, B. W., Ong, S. H. & Husain, N. H. M., 2011. Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications. 38 (10), 13274-13283.
Zhu, Y., Xie, C., Sun, B., Wang, G. & Yan, X., 2016. Predicting China's SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models. Sustainability. 8 (5), 433(17pages).
Zhu, Y., Xie, C., Wang, G. & Yan, X., 2016. Predicting China's SME credit risk in supply chain finance based on machine learning methods. Entropy. 18 (5), 195(8pages).
Zhu, Y., Xie, C., Wang, G. & Yan, X., 2017. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SMEs credit risk in supply chain finance. Neural Computing and Applications. 28 (1), 41-50.

Input: an attribute-valued dataset D
1. $Tree = \{\}$
2. if D is 'true' or other stopping criteria met then
3. terminate
4. end if
5. for all attribute $a \in D$ do
6. compute information-theoretic criteria if we split on a
7. end for
8. a_{best} = Best attribute according to above computed criteria
9. $Tree$ = Create a decision node that tests a_{best} in the root
10. D_v = Induced sub-datasets from D based on a_{best}
11. for all D_v do
12. $Tree_v = C4.5(D_v)$
13. Attach $Tree_v$ to the corresponding branch of $Tree$
14. end for
15. return $Tree$

Figure 1. The pseudo-code of the C4.5 algorithm of the DT approach (Quinlan, 1993)

Input: Data set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, Base classifier algorithm L ,
Number of random subspace rate k , Number of learning rounds T
1. For $t = 1, 2, ..., T$
2. Random generate a subspace sample from $D_t = RS(D, k)$
3. Train a base classifier h_t from the subspace sample
4. end

Output: $H(X) = \arg\max_{y \in Y} \sum_{t=1}^{T} 1(y = h_t(x))$;

$$\begin{cases} 1(\alpha) = 1 \text{ if } \alpha \text{ is ture} \\ 1(\alpha) = 0 \text{ otherwise} \end{cases}$$

Figure 2. The pseudo-code of the algorithm of the RS approach (Ho, 1998)

Input: D , a sequence of m labeled examples $\langle(x_1, y_1), ...(x_m, y_m)\rangle$ with labels $y_i \in Y$, Base learning algorithm L ,integer T specifying the number of iterations, vector of integers I_i specifying the iteration at which each subcommittee $i \geq 1$ should terminate
1. $D' = D$ with instance weights assigned to be 1
2. set $k = 1$
3. For $t = 1$ to $T\{$
4. If $I_k = t$ then reset D' to random weights drawn from the continuous Poisson distribution, standardize D' to sum to n , increment k
5.$C_t = L(D')$
6.$\varepsilon_t = \frac{\sum_{x_j \in D' : C_t(x_j) \neq y_j} wight(x_j)}{m}$
7. if $\varepsilon_t > 0.5$ then reset D' to random weights drawn from the continuous Poisson distribution, standardize D' to sum to n , increment k
8. go to Step 5
9. otherwise if $\varepsilon_t = 0$ then set β_t to 10^{-10} , reset D' to random weights drawn from the continuous Poisson distribution, standardize D' to sum to n , increment k
10. otherwise $\beta_t = \frac{\varepsilon_t}{(1-\varepsilon_t)}$
11. For each $x_j \in D'$
12. divide $weight(x_j)$ by $2\varepsilon_t$ if $C_t(x_j) \neq y_j$ and $2(1 - \varepsilon_t)$ otherwise
13. if $weight(x_j) < 10^{-8}$, set $weight(x_j)$ to 10^{-8}
14.$\}$
Output the final classifier:$C^*(x) = \arg\max_{y \in Y} \sum_{t:C_t(x)=y} \log \frac{1}{\beta_t}$

Figure 3. The pseudo-code of the algorithm of the MultiBoosting approach (Webb, 2000)

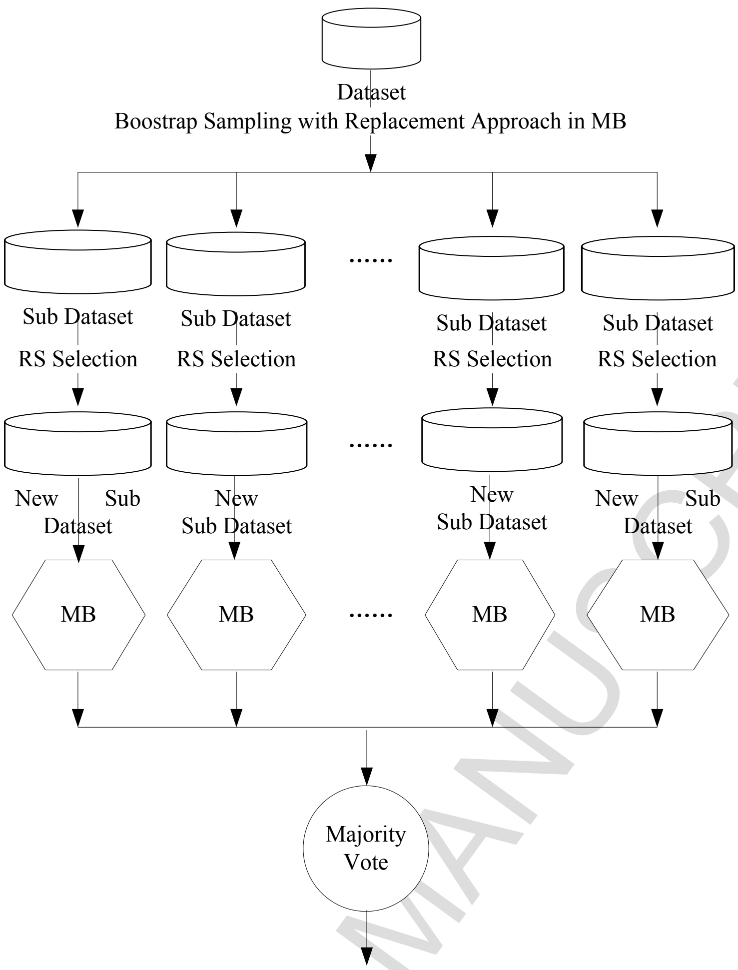


Figure 4. The working mechanism of the RS-MultiBoosting approach

Input:

Date set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ with lables $y_i \in Y$, base classifier algorithm L (Decision Tree, C4.5), number of selected features rate r, number of iterations T_{MB}, vector of integers I_i specifying the iteration T_{MB} at which each subcommittee $i \geq 1$ should terminate, number of iterations for Random Subspace T_{RS}.

Process:

D with instance weights assigned to be 1.
Set $k = 1$.
(a) For $t = 1, 2, \ldots, T_{MB}$:
(b) For $s = 1, 2, \ldots, T_{RS}$:
Random generate a subspace sample from D_t : $D_t^s = RS(D_t, r)$, Train a base classifier h_s from D using distribution D_t : $h_s = DT(D_t^s)$.
end.
$H_t^{RS}(x) = \arg\max_{y_i \in Y} \sum_{s=1}^{T_{RS}} 1(y = h_s(x))$
If $I_K = t$ then reset D_t to random weights drawn from the continuous Poisson distribution, standardize D_t to sum to n, increment k.
$\star\ h_t = DT(D_t)$.
The weighted error on the training set: $\varepsilon_t = \frac{\sum_{x_j \in D_t : h_t(x_j) \neq y_j} weight(x_j)}{m}$
If $\varepsilon_t > 0.5$ then reweight D_t, increment k.
Go to $\star$
Otherwise if $\varepsilon_t = 0$ then set β_t to 10^{-10}, reweight D_t, increment k.
Otherwise
(a) $\beta_t = \frac{\varepsilon_t}{(1-\varepsilon_t)}$,
(b) $D_{t+1} = D_t$.
For each $x_j \in D_{t+1}$, divide $weight(x_j)$ by $2\varepsilon_t$ if $h_t(x_j) \neq y_j$ and $2(1 - \varepsilon_t)$ otherwise.
If weight $(x_j) < 10^{-8}$, set $weight(x_j)$ to 10^{-8}.
end.

Output:

$H^{RS-MB}(x) = \arg\max_{y \in Y} \sum_{t : h_t(x) = y} \log \frac{1}{\beta_t}$

Figure 5. The pseudo-code of the RS-MultiBoosting algorithm

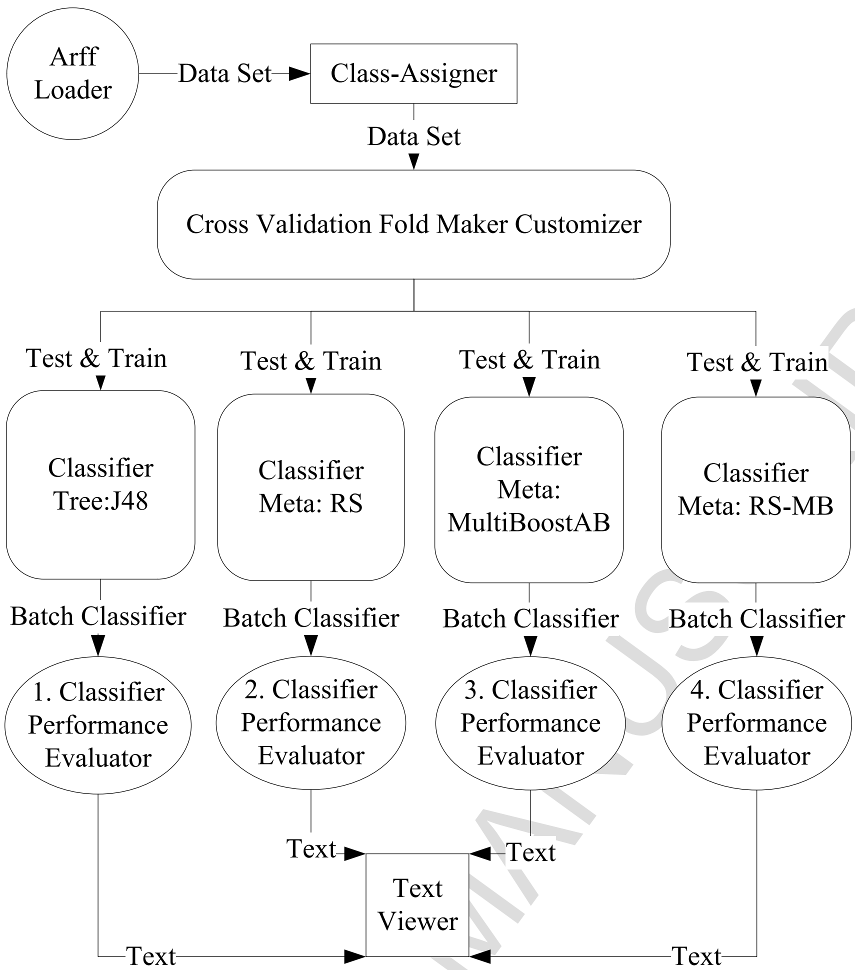


Figure 6. The Data Mining Processes in WEKA Knowledge-Flow Environment

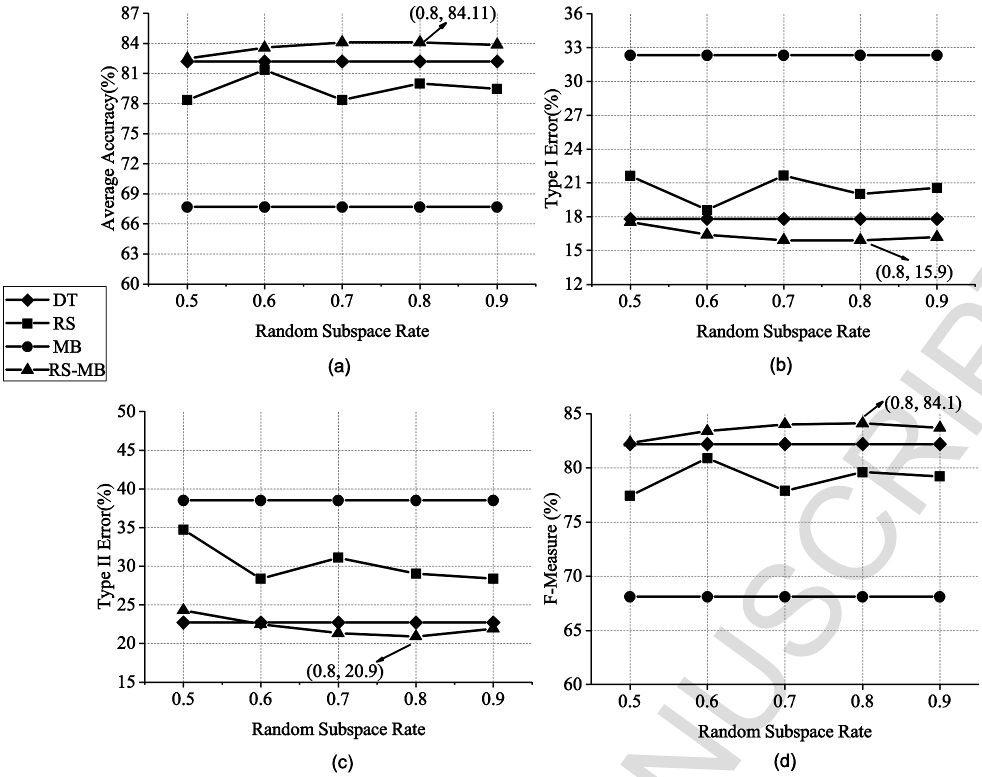


Figure 7. The geometric forecasting results of the DT, RS, MultiBoosting and RS-MultiBoosting models based on the V_o (with RS rates of 0.5, 0.6, 0.7, 0.8 and 0.9)

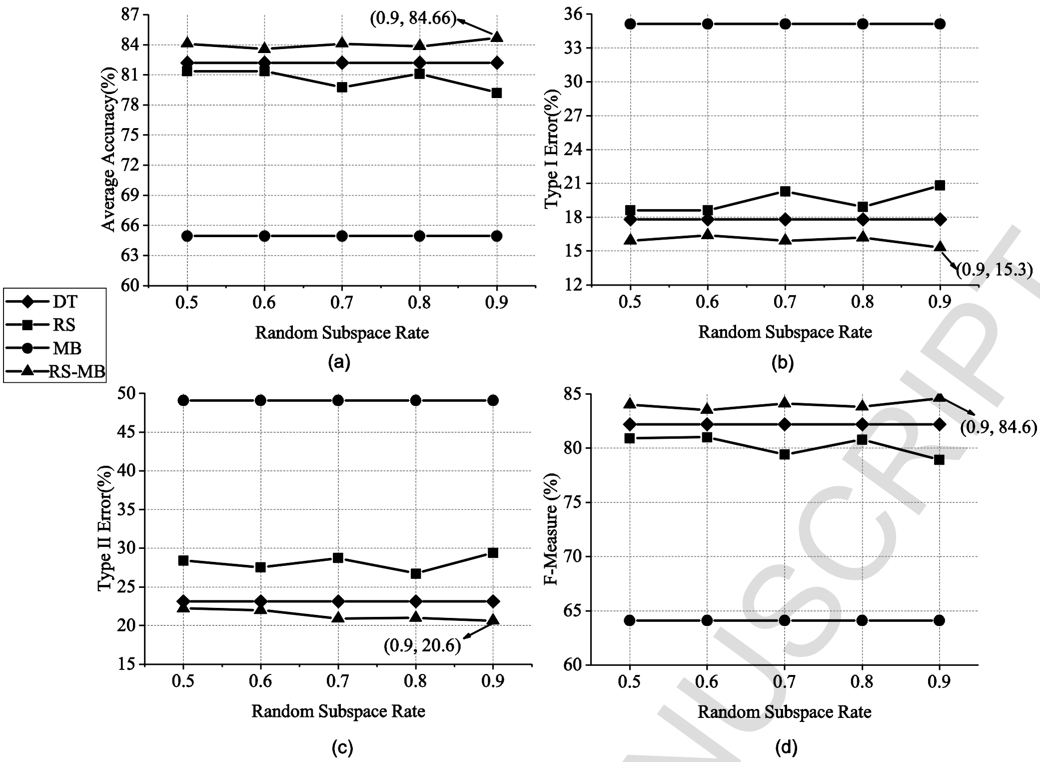


Figure 8. The geometric forecasting results of the DT, RS, MultiBoosting and RS-MultiBoosting models based on the V_s (with RS rates of 0.5, 0.6, 0.7, 0.8 and 0.9)

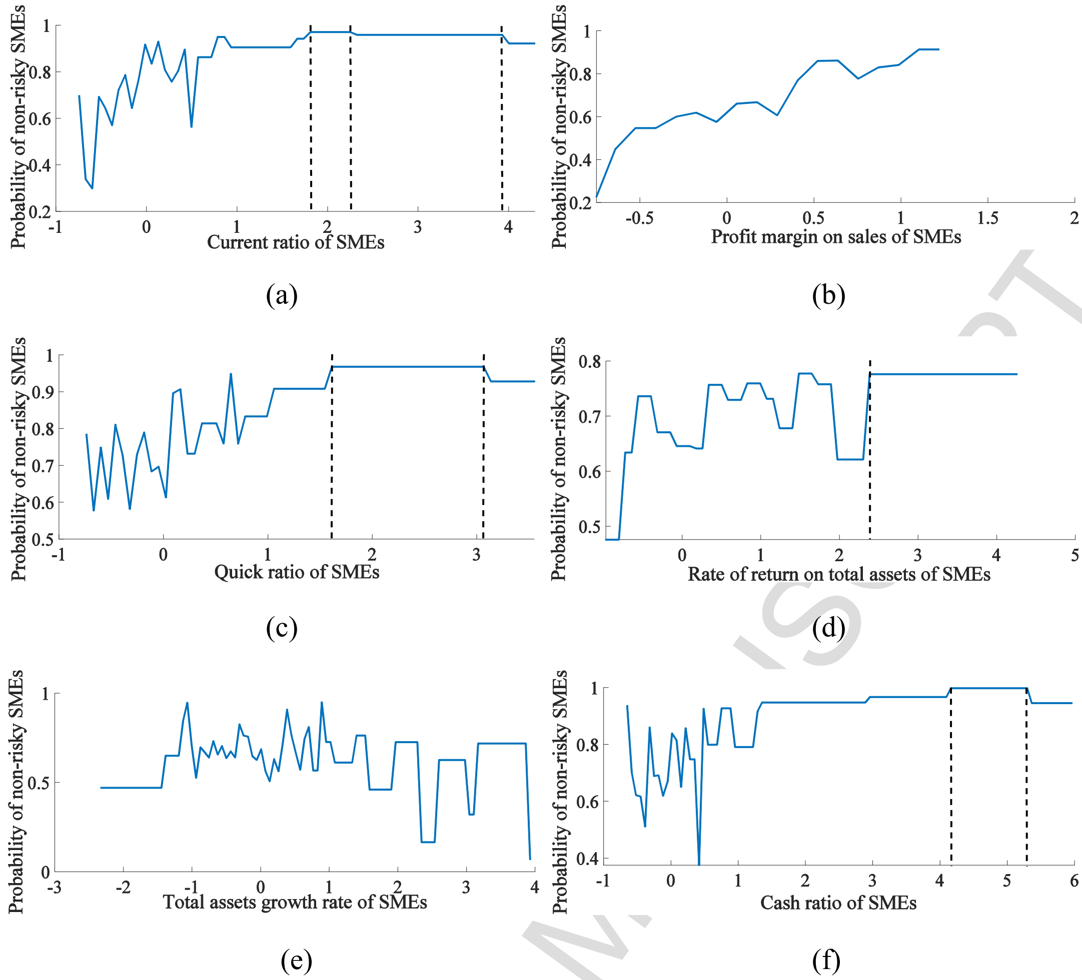


Figure 9. The line PDP of the predicted responses against each feature of the traditional financing factors

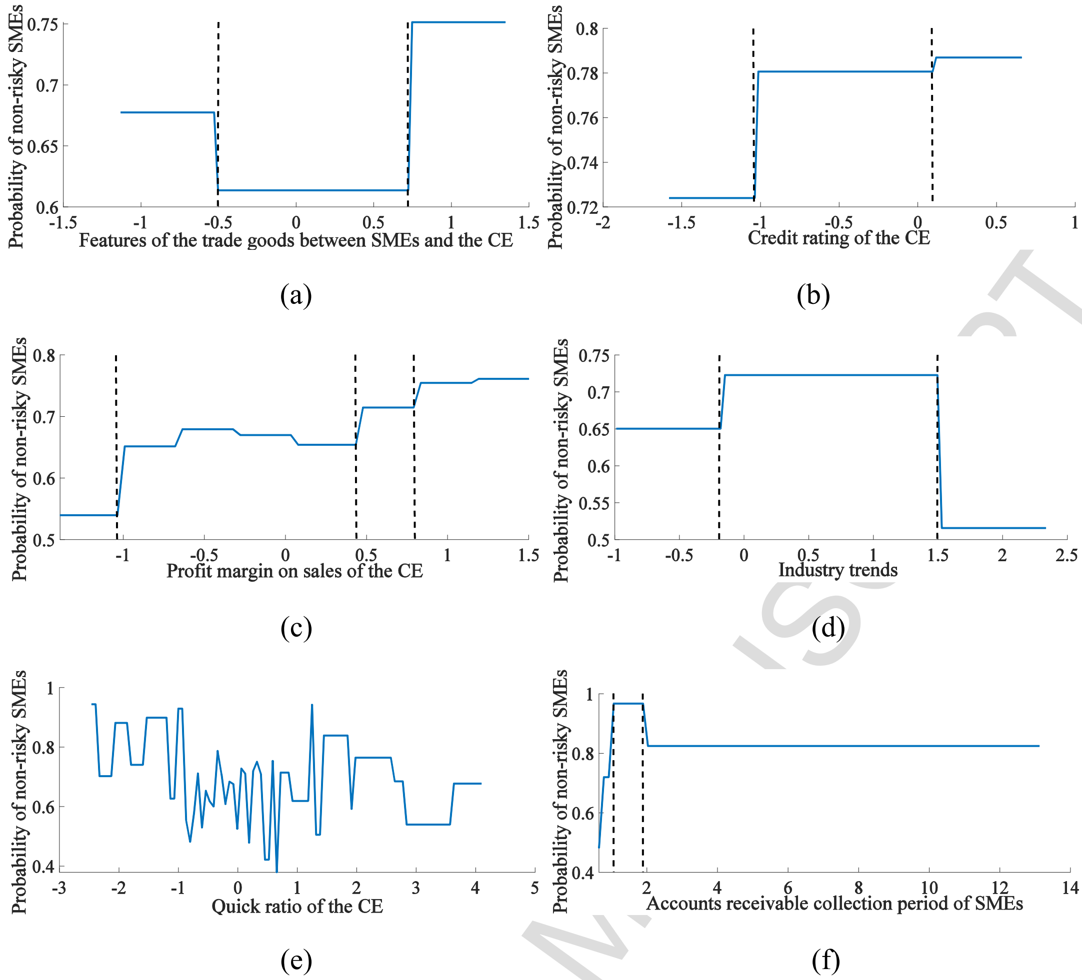


Figure 10. The line PDP of the predicted responses against each feature of SCF factors

Table 1. V_o of the forecasting models

Independent Variables	Definitions	Categories
Current ratio of SMEs	Floating assets divided by floating liabilities.	Non-financial
Quick ratio of SMEs	Floating assets minus Inventory, then divided by floating liabilities.	Liquidity
Cash ratio of SMEs	Cash and cash equivalent ending balance divided by floating liabilities.	Liquidity
Working capital turnover of SMEs	Operating receipt divided by average working capital.	Liquidity
Return on equity of SMEs	Net margin divided by the average balance of the interests of shareholders.	Liquidity
Profit margin on the sales of SMEs	Net margin divided by operating receipts.	Leverage
Rate of return on total assets of SMEs	Net margin divided by the average balance of total assets.	Profitability
Total assets growth rate of SMEs	Total assets at the end of term minus total assets at the end of last year, divided by total assets at the end of last year.	Leverage
Credit rating of the CE	An evaluation of the CE’s creditworthiness, which is divided into 7 grades in this paper.	Activity
Quick ratio of the CE	Floating assets minus inventory, divided by floating liabilities.	Non-financial

Turnover of total capital of the CE	Operating receipts divided by average general assets.	Liquidity
Profit margin on the sales of the CE	Net margin divided by operating receipts.	Liquidity
Features of the trade goods between SMEs and the CE	The degree of the features (e.g., price rigidity, liquidation, vulnerable degree and others) of trade goods, which are divided into 7 grades.	Profitability
Accounts receivable collection period of SMEs	Collection period divided by the accounts receivable turnover ratio.	Non-financial
Accounts receivable turnover ratio of SMEs	Operating receipts divided by the average accounts receivable occupation.	Leverage
Industry trends	The patterns or trends that occur within an industry, which are divided into 7 grades.	Leverage
Cooperation degree between SMEs and the CE	The transaction frequency, which is divided into 7 grades.	Non-financial
Credit rating of SMEs	An evaluation of SMEs' creditworthiness, which is divided into 7 grades.	Non-financial

Table 2. V_s of the final forecasting models

Factors	Independent Variables	Scores	Ranking
Traditional financing factors	Current ratio of SMEs	**0.0985**	1
	Profit margin on the sales of SMEs	0.0825	5
	Quick ratio of SMEs	0.0825	5
	Rate of return on total assets of SMEs	0.0816	7
	Total assets growth rate of SMEs	0.0816	7
	Cash ratio of SMEs	0.0800	10
Supply chain financing factors	Features of trade goods between SMEs and the CE	**0.0867**	2
	Credit rating of the CE	**0.0850**	3
	Profit margin on sales of the CE	0.0833	4
	Industry trends	0.0808	9
	Quick ratio of the CE	0.0800	10
	Accounts receivable collection period of SMEs	0.0774	12

Table 3. The numerical forecasting results of the RS-MultiBoosting and other ML approaches

	Evaluation Criteria	**DT**	**RS[a]**	**MultiBoosting**	**RS-MultiBoosting[b]**
Based on V_o	Average Accuracy	82.19%	81.37%	67.67%	84.11%
	Type I Error	17.80%	18.60%	32.30%	15.90%
	Type II Error	22.70%	28.40%	38.50%	20.90%
	F-Measure	82.20%	80.90%	68.10%	84.10%
Based on V_s	Evaluation Criteria	DT	RS[c]	MultiBoosting	RS-MultiBoosting[d]
	Average Accuracy	82.19%	81.37%	64.93%	84.66%
	Type I Error	17.80%	18.60%	35.10%	15.30%
	Type II Error	23.10%	27.50%	49.10%	20.60%
	F-Measure	82.20%	81.00%	64.10%	84.60%

Note: [a] the RS rate is set to 0.6

[b] the RS rate is set to 0.8

[c] the RS rate is set to 0.6

[d] the RS rate is set to 0.9