Ensemble Learning with Feature Optimization for Credit Risk Assessment

Guanghui Zeng
Northeastern University
Weixin Su
Xiamen University of Technology
Chaoqun Hong
cqhong@xmut.edu.cn

Xiamen University of Technology

Research Article

Keywords: Credit risk assessment, Machine learning, Ensemble learning, Feature Optimization

Posted Date: July 29th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4665987/v1

Additional Declarations: No competing interests reported.

Ensemble Learning with Feature Optimization for Credit Risk Assessment

Guanghui Zeng[1], Weixin Su[2], Chaoqun Hong[2*]

[1]School of Business Administration, Northeastern University, China.
[2] School of Computer and Information Engineering, Xiamen University of Technology, China.

*Corresponding author(s). E-mail(s): cqhong@xmut.edu.cn ;
Contributing authors: zenggh@gx-credit.com; wxsu@s.xmut.edu.cn;

Abstract

Credit risk assessment stands as a cornerstone in financial decision-making, with significant implications for economic stability and growth. This paper highlights the transformative advantages of credit big data over traditional methods, particularly in enhancing the creditworthiness evaluation of small and medium-sized enterprises (SMEs). We delineate the distinctive features of the big data financial innovation model across six economic dimensions, showcasing its potential to reshape financial practices. To address the inefficiencies of traditional expert-driven approaches, we introduce an innovative 'Feature Selector-classifier Optimization Framework' that streamlines the credit risk prediction process. This framework not only refines the accuracy and efficiency of predictions but also integrates seamlessly with economic analysis, offering a robust tool for financial decision-makers. Our ensemble classifier delivers remarkable performance, exemplified by its high accuracy and AUC scores across multiple datasets, thereby validating the framework's efficacy in enhancing predictive power while ensuring operational efficiency.

Keywords: Credit risk assessment, Machine learning, Ensemble learning, Feature Optimization

1 Introduction

Credit risk assessment is an indispensable component of financial decision-making, essential for evaluating the creditworthiness of individuals and businesses seeking loans

or credit facilities. Small and medium-sized enterprises (SMEs) play a crucial role in economic development by supporting growth, employment, and innovation. They constitute over 90% of market entities, contribute to 80% of employment, hold about 70% of patents, and account for more than 60% of GDP and over 50% of tax revenue. Despite their significance, SMEs face substantial challenges in securing financing. The primary issue lies in the incomplete and asymmetrical credit information, which hinders financial institutions' ability to accurately assess and manage credit risk before, during, and after lending. This financing difficulty has significant economic implications, as it affects the stability and growth potential of these enterprises. Traditional credit risk assessment methods have primarily relied on expert judgment and manual feature engineering. These approaches, while valuable, are often limited in their ability to capture the complex dynamics of financial data, especially for SMEs. They tend to be time-consuming and fail to leverage the strengths of diverse data sources and advanced analytical techniques. Consequently, there is a pressing need for more efficient and accurate methods to assess credit risk, particularly for SMEs that lack extensive financial histories or substantial collateral.

The advent of big data and the rapid development of machine learning technologies offer new opportunities to enhance credit risk assessment. Big data finance integrates vast amounts of structured and unstructured data from various sources, providing a more comprehensive view of an enterprise's creditworthiness. This approach allows for the identification of patterns and trends that traditional methods might overlook, thereby improving the accuracy and reliability of credit risk predictions. This shift towards data-driven approaches is particularly crucial for small and medium-sized enterprises (SMEs), which often lack the resources to employ sophisticated risk assessment tools. The application of machine learning techniques has been shown to improve the predictive accuracy and reliability of credit risk assessments, offering a more nuanced understanding of the financial health of enterprises (Chen, Ribeiro, and Chen (2016)).

The global push for financial inclusion has further emphasized the importance of accessible and reliable credit risk assessments, particularly for underserved populations. By leveraging machine learning, financial institutions can refine their risk assessment strategies, extending credit services to a broader demographic while maintaining prudent risk management(Mhlanga (2021)).

With the continuous improvement of the social credit system and the rapid development of big data technology, credit big data provides a new method and path for enterprise credit risk control. Credit big data integrates data from diverse sources, including financial records, government databases, and social media, enabling a more comprehensive and accurate assessment of credit risk. This paper starts from the concept of credit big data, summarizes its advantages over traditional credit data, and elaborates on the characteristics of big data financial innovation models. Taking pure credit loans of engineering construction enterprises as the research object, this paper analyzes the construction method of big data financial innovation models, with empirical analysis demonstrating the effectiveness of these models.

In response to these challenges, this study introduces an innovative "Feature Selector-classifier Optimization Framework" for credit risk assessment, integrating big

data finance principles with advanced machine learning techniques. This framework automates the feature engineering and model selection processes, leveraging ensemble learning to enhance predictive performance. By systematically exploring combinations of feature selectors and classifiers, the framework aims to optimize the feature space, reduce computational costs, and improve the overall accuracy and efficiency of credit risk predictions. The main contributions of this study are as follows:

- This study introduces an innovative credit risk assessment model grounded in economic principles, specifically focused on the evaluation of credit risk for Small and Medium-sized Enterprises (SMEs). The model provides a comprehensive framework by delving into macroeconomic factors, industry-specific indicators, and non-traditional data sources, offering financial decision-makers a holistic assessment approach.
- Through the process of automating feature engineering and model selection, this study significantly enhances the precision of credit risk forecasting. An automated "Feature Selector-Classifier Optimization Framework" is implemented, systematically exploring various combinations of feature selectors and classifiers. This process, coupled with the framework's adept handling of complex financial data, leads to a substantial improvement in predictive capabilities.
- Empirical analysis across multiple datasets, especially with imbalanced and large-scale data, validates the superior performance of our ensemble learning strategy in credit risk assessment. By offering clear credit risk evaluation results, it aids decision-makers in understanding the rationale behind assessments, leading to more transparent and justified financial decisions.

The remainder of this paper is organized as follows: Section 2 reviews the literature on credit risk assessment, focusing on the transition from traditional methods to machine learning and ensemble techniques. Section 4 introduces the 'Feature Selector-classifier Optimization Framework,' detailing its components and rationale. Section 5 outlines the experimental setup, including the datasets used, the preprocessing steps, and the benchmark models. Section 6 presents the performance results of the individual classifiers and the ensemble classifier, along with feature importance analysis. Finally, Section 7 discusses the implications of the results, concludes the paper, and suggests directions for future research.

2 Related Work

2.1 Credit Big Data and Economic Analysis

The burgeoning field of credit big data has become increasingly significant in economic analysis and credit risk assessment. Credit big data encompasses a vast array of information related to credit generated by individuals, enterprises, government departments, and social organizations. In China, this data is primarily sourced from financial institutions, government departments, other public institutions, and the internet (Y.T. Liu, Tang, and Liang (2016)). The aggregation of data from these varied sources forms the credit big data ecosystem, which is instrumental in supporting the development of social credit systems in the era of big data (Hu (2022)).

This paper explores the risk assessment of SMEs through the lens of credit big data, examining the transformation between big data financial innovation models and traditional financial models across various aspects such as credit subjects, risk models, risk management, and loan acquisition difficulties (Xiong (2022)). The construction method of the big data financial innovation model is detailed, with "Engineering Credit" serving as an empirical analysis case to validate the model's effectiveness.

2.2 Integration of Machine Learning in Credit Risk Assessment

The advent of machine learning has marked a significant leap in the precision and efficiency of credit risk assessment. Galindo and Tamayo (2000) set the stage with their pioneering work on statistical and machine learning methodologies for credit risk modeling. In a comprehensive review, Chen et al. (2016) highlighted the industry's shift towards machine learning, underscoring its role in enhancing predictive accuracy and reliability.

AI's innovative applications in credit risk, such as Guanghui (2021) exploration of "Xinyidai" for SME financing, have opened new avenues for financial inclusion and credit evaluation. C. Liu, Xie, Zhao, Xie, and Liu (2019) introduced an evolutionary multi-objective soft subspace clustering algorithm, demonstrating AI's potential in handling complex datasets for credit risk assessment. The application of neural networks has further advanced the field, with Di Lorenzo, Piscopo, and Sibillo (2024) utilizing a neural network algorithm to estimate Conditional Value at Risk (CVaR) in reverse mortgages, showcasing the suitability of neural networks for fitting nonlinear functions. S. Liu and Vicente (2022) tackled the issue of fairness in machine learning models used in credit scoring, introducing a stochastic multi-objective optimization approach that effectively manages the trade-offs between accuracy and fairness, ensuring more equitable outcomes.

Ensemble learning techniques have been instrumental in refining credit risk prediction. Ying and Lihua (2020) improved SVM ensembles with noise elimination, enhancing model robustness. Comparative analyses by Bhatore, Mohan, and Reddy (2020) and Wang, Liu, and Qi (2022) have shed light on the strengths of various machine learning models and the benefits of ensemble methods.

Systematic literature reviews, such as the one conducted by Bussmann, Giudici, Marinelli, and Papenbrock (2021), have emphasized the importance of explainable machine learning in credit risk management. The field has since evolved, with Li, Paraschiv, and Sermpinis (2022) contributing an explainable CBR approach, enhancing model interpretability—a critical factor for stakeholder trust and decision-making. These reviews, along with comparative studies, offer valuable benchmarks for model development and the integration of diverse algorithms.

2.3 Advances in Credit Risk Modeling Techniques

Machine learning techniques have been instrumental in advancing credit risk modeling. Sousa, Gama, and Brandão (2016) introduced a dynamic modeling framework, addressing the need for adaptability in credit risk assessments, especially with high-dimensional data. Additionally, the challenge of imbalanced datasets in credit fraud

detection has been addressed by Zhao, Li, Lyu, Ma, and Zhu (2023) through a cost-sensitive ensemble deep forest model. Gärtner, Kaniovski, and Kaniovski (2021) proposed a heuristic algorithm for solving maximization problems of combinatorial complexity, providing a probability distribution approach for financial risk analysis. Corazza, De March, and Di Tollo (2021) proposed adaptive Elman networks, demonstrating neural networks' robustness in credit risk assessment. Further advancements include the use of Bayesian methods to integrate individual ratings and credit performance, as demonstrated by Bu, Guo, and Li (2022), which provides a robust tool for portfolio credit risk analysis. Le, Ku, and Jun (2021) applied sequence-based clustering, providing insights into the temporal aspects of credit risk.

The development of rule-based and evolutionary models, such as the one proposed by Soui, Gasmi, Smiti, and Ghédira (2019), and the application of deep learning by Gicić, Đonko, and Subasi (2023), have significantly improved the interpretability and accuracy of credit risk assessments. Mahajan et al. (2022) contributed to the field by employing a Gaussian process-based approach for uncertainty handling in credit risk predictions.

Data-driven approaches, as explored by Huang et al. (2020), and hybrid models that integrate different machine learning techniques, such as the work by Yu, Zhang, and Yin (2022), have become crucial for extracting meaningful patterns from complex datasets and addressing data scarcity issues.

Despite the advancements, the field faces challenges in balancing model complexity with interpretability. Bussmann et al. (2021) and Xu, Ding, and Pan (2018) have highlighted the need for models that are not only accurate but also transparent in their decision-making processes. Su and Ren (2019) proposed a multilayer fusion network to better understand the interplay between social life and credit activities, emphasizing the importance of considering various factors in model development.

2.4 Challenges in Credit Risk Assessment Research

In summary, existing studies lay a foundation for credit risk assessment, emphasizing machine learning, data-driven approaches, and ensemble learning. Additionally, existing studies tend to focus on individual methodologies, overlooking the potential of comprehensive exploration of feature selectors and model combinations. In the wave of the data economy, credit big data is experiencing explosive growth. Therefore, it is crucial to extract potential huge value from a large amount of credit data. Our proposed framework is built upon these methods, providing a comprehensive and automated solution that integrates ensemble learning for enhanced performance.

3 Big Data Financial Innovation Model

3.1 Credit Big Data

As the sources of credit big data continue to expand, the content of credit data becomes more diverse, and the application scenarios and service fields are also more extensive. The advantages of credit big data over traditional credit data are shown in Table 1. Credit big data mainly has the following five advantages.

Table 1 Advantages of credit big data over traditional credit data

	Traditional credit data	Credit big data
Data source	A single source primarily	Multi source data, sea source data
Data structure	Structured data primarily	Equal emphasis on structured and unstructured data
Data feature	Financial data primarily	Equal emphasis on financial and non-financial data
Data storage	Centralized data primarily	Centralized and decentralized data
Data security	Low safety protection capability	Blockchain, privacy computing, sandbox technology

- High dimensional data sources. Traditional data usually comes from data from credit institutions and mainly relies on credit data from the People's Bank of China. And credit big data has the characteristic of "full data", breaking through traditional limitations in its collection scope. It integrates public credit information such as water, electricity, gas, and market credit information from financial data of banking institutions and relevant data generated by government departments, including internet big data such as warehousing, logistics, social networking, payment, e-commerce, and transportation, making the dimensions and scale of credit big data richer and larger.
- The data content is unstructured. Most of the traditional credit data are structured data, mainly standardized text; and credit big data presents heterogeneity, with structured, semi structured, and unstructured data characteristics, including XML data files, text documents, images, videos, audio, and so on.
- Data features are non-financial. Traditional credit data is mainly based on financial data for risk assessment, which includes indicators such as profitability, asset management ability, and debt repayment ability. Non-financial indicators are ignored for comprehensive credit assessment of small and medium-sized enterprises. Credit big data integrates the evaluation of non-financial data, including comprehensive information such as the company's potential development capabilities, industry status, research and development investments and their results, market share, technical goals, corporate and executive characteristics.
- Decentralized data storage. The traditional data storage method uses centralized servers to store data, while in the era of credit big data, data storage has shifted from centralized to decentralized. Data is stored in social institutions, third-party service institutions, platform enterprises, etc., and is interconnected in data applications.
- Intelligent data security. The traditional model has low data security protection capabilities, while in the era of big data, various technologies are adopted to ensure data security, such as using blockchain encryption technology to solve security issues in the data circulation and sharing process using asymmetric encryption algorithms, and privacy algorithms to ensure that potential attackers cannot reverse deduce accurate sensitive information, Fully isolated and lightweight virtualization technology using sandbox technology automatically identifies specific high-risk software isolation operations.

3.2 Big data financial innovation model

The financial industry has also entered the era of "big data finance", providing more possibilities for solving "financing difficulties" by capturing and analyzing the usage traces of enterprises online or on mobile terminals. For financing, big data finance refers to the collection of massive unstructured data, the real-time analysis of enterprise data through big data, the Internet, cloud computing and other information methods, the provision of comprehensive information for enterprises, the mastery of enterprise information through mining data, the improvement of financial service platform efficiency and the reduction of credit risk. Big data financial risk assessment has unparalleled advantages in data and models compared to traditional finance. These advantages help to obtain enterprise information and achieve innovation in enterprise

Table 2 Comparison of big data financial innovation model and traditional financial service model

	Traditional financial services	Big data financial innovation model
Credit reporting subject	Passive solicitation	Proactive push
Risk model	Relatively fixed model	Intelligent algorithms, continuous iteration
Risk management	Post loan risk disposal	Full process data risk monitoring and risk warning for pre loan, during loan, and post loan
Risk preference	Heavy mortgage and heavy guarantee	Heavy mortgage and heavy guarantee
Difficulty in obtaining loans	Difficult initial loan and easy increase in credit	Easy down loan and difficult credit enhancement
Price system	Low interest rates for large enterprises and high interest rates for small enterprises	Differentiated pricing

financial management, mainly reflected in six aspects: credit reporting subjects, risk models, risk management, risk preferences, difficulty in obtaining loans, and price mechanisms, as shown in Table 2.

- Credit reporting subject. In terms of information collection, traditional financial services mainly collect information passively through manual investigation and active collection, which can easily lead to moral hazard and adverse selection due to information asymmetry. However, in the era of big data finance, this information collection method is inefficient. The information collected by the big data financial innovation model has gradually expanded from financial data to government platform data and external credit data. At the same time, enterprises are encouraged to fill in data independently, integrate these multi-source basic data sources to build an industry database, realize active push, and explore applications in the financial field, eliminate information asymmetry between banks and enterprises during the lending process, and improve bank lending efficiency .
  When providing credit services for SMEs, the big data financial innovation model can provide two advantages from the aspect of credit subjects: first, the introduction of cross domain data can establish credit files for enterprises without credit records, improve the availability of financing for SMEs, guide the capital flow of SMEs, and promote the development of inclusive finance. The second is to ensure the authenticity of data through the use of financial technology. In the context of credit big data, according to the relevant provisions of the National Public Credit Information Basic Catalogue, enterprise entities can independently declare data, and financial institutions can obtain multiple data sources and verify the authenticity of the data through multiple data sources.
- Risk model. In traditional financial services, although traditional risk control models set evaluation indicators based on industry classification, the models are relatively fixed, and their computational power is not accurate enough, which limits their ability to solve risk assessments.
  In the big data financial innovation model, the combination of rule engine and machine learning algorithm model in the credit field can realize the intelligent algorithm of continuous iteration. The main function of risk control models is to more accurately predict future operating conditions and changes in repayment ability, while also more effectively identifying loan applicants who are reliable and have a high willingness to repay. Financial technology can greatly improve the speed and scale of processing, thereby significantly reducing the operating costs of credit, and monitoring changes in credit risk to minimize the possibility of default by borrowing enterprises, Conduct more accurate customer screening and risk control.
- Risk management. The risk management of traditional financial services mainly involves dealing with risks after loans, with poor risk control and insufficient timeliness of risk control. Based on the model of big data financial innovation, the whole process of data risk monitoring and risk early warning before, during and after the loan is realized. Through the double verification of the enterprise credit access model, after excluding the unqualified enterprises, the credit situation of the admitted enterprises is quantitatively evaluated using the enterprise credit evaluation model, and the whole process of tracking and monitoring the enterprise risk data,

and according to the risk impact, Output different risk level prompts to financial institutions.

The big data financial innovation model uses scientific methods to digitize the risk model, provide objective risk quantification, and reduce subjective judgments, so as to improve the efficiency of risk management, give timely warning to enterprise risks, and take timely countermeasures against risks.

- Risk appetite. The traditional bank credit model relies on collateral and guarantees, without further management of the purpose of the borrowed funds. Moreover, some companies have short survival cycles and poor risk resistance, making it difficult to quantitatively evaluate their repayment ability. Using financial technology and big data financial innovation model to build a full process online credit operation process, manage the loan purpose, form a closed-loop management of capital flow, information flow, logistics, and order flow, break away from the traditional credit strategy of "heavy mortgage and heavy guarantee", and gradually develop a pure credit model of "no mortgage and no guarantee".

  The financial service model of pure credit loans mainly involves conducting intelligent risk control cooperation online, promoting the migration of banks' risk control models to the online market. With the help of new technologies such as big data, artificial intelligence and blockchain, it solves the pain point of technology enterprises lacking mortgage guarantees and enhances their survival and development capabilities.
- Getting a loan is difficult and easy. The so-called first loan, also known as the first loan, refers to the first time an enterprise without a loan record in the credit report of the People's Bank of China obtains a loan from a banking financial institution. In traditional financial services, small and medium-sized enterprises often face significant information asymmetry issues between banks and enterprises due to non-standard information, resulting in pain points such as lack of credit information, insufficient collateral, and weak guarantees, making it difficult to obtain down loans. And some companies that have obtained loans will have higher credit and more advantages.

  The big data financial innovation model strengthens the sharing of credit information, reduces the information asymmetry between banks and enterprises, collects more comprehensive credit data, focuses on the analysis and judgment of loan use, improves the efficiency of credit approval, effectively solves the problem of small and medium-sized enterprises' first loan difficulty, and also prevents excessive credit granting to enterprises with good credit.
- Price mechanism. Compared to large enterprises, small and medium-sized enterprises have a higher cost of obtaining loans and financing. They not only cannot obtain preferential interest rates, but also have to pay more floating interest. In traditional financial services, due to the fact that small and medium-sized enterprises often use mortgage or guarantee methods for loans, not only the procedures are complex, but also more guarantee fees and mortgage asset evaluation fees need to be paid, resulting in low interest rates for large enterprises and high interest rates for small enterprises.

In the context of big data finance, big data is used to collect enterprise data and evaluate enterprise risks from the data. Based on the assessed enterprise risks, more accurate and fair pricing is made, and different enterprises have different pricing, achieving differentiated pricing.

4 Feature Selector-classifier Optimization Framework

4.1 Framework Overview

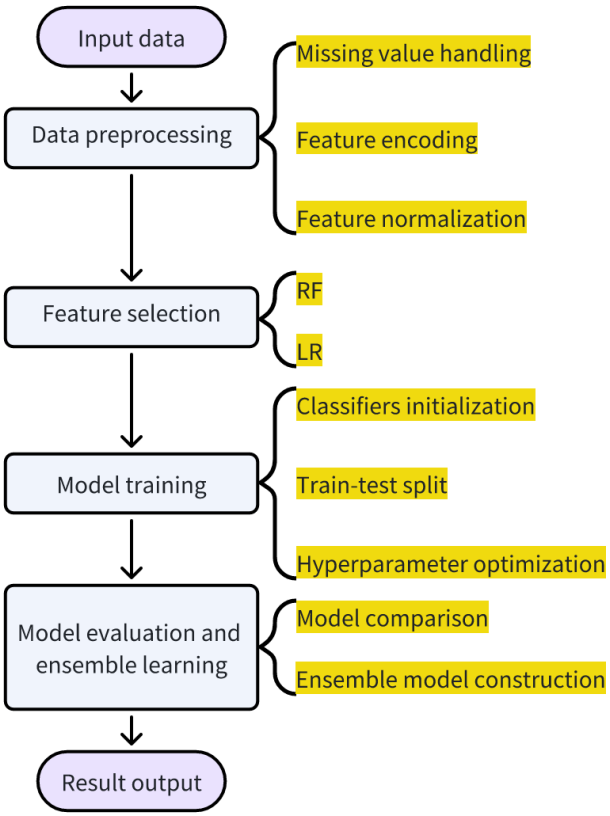

Fig. 1 The flowchart of the automated feature engineering and model selection process in the Feature Selector-classifier Optimization Framework for credit risk assessment

Conventional credit assessment methodologies heavily rely on expert experience for laborious feature engineering and often confine themselves to a single classification model. These approaches are not only time-consuming but also fail to leverage the strengths of different models, posing significant challenges for financial institutions in effectively identifying and controlling credit risk. To address these issues, an innovative

"Feature Selector-classifier Optimization Framework" is introduced, automating the feature engineering and model selection process, as depicted in Figure 1 and algorithm 1.

During the data preprocessing phase, features with more than 90% missing values are initially removed to minimize noise. For the remaining features with fewer missing values, imputation is performed using mean or median filling techniques to maintain data integrity. Subsequently, discrete features are encoded using one-hot encoding to convert them into numerical data that can be directly processed by the model. Numerical features are then normalized using min-max scaling to a range of [0,1], which accelerates model convergence and enhances stability. The feature selection and classification modeling phase involves initializing a set of feature selectors and classifiers, traversing the feature selectors to select the most relevant feature subsets, and iterating through classifiers for training and optimization. For each classifier, the data is split into training and testing sets, with oversampling applied to the training set to balance the sample distribution. Hyperparameter optimization for each classifier is achieved through grid search combined with cross-validation to determine the optimal model parameters. After training all feature Selector-classifier combinations, the performance of each model is evaluated on the test set. Finally, several single-model classifiers with superior performance are selected to construct an ensemble classifier, further enhancing the predictive performance.

The proposed framework automates the optimization process by systematically exploring combinations of feature engineering and model selection and leverages ensemble learning to improve performance, providing a more accurate method for credit risk prediction.

4.2 Ensemble Learning Based on Voting

The key to our framework's predictive power lies in its ensemble learning approach, which synergistically combines the strengths of multiple classification models. Ensemble learning has been widely recognized for its ability to improve the accuracy and robustness of predictions by reducing both bias and variance.

4.2.1 Voting Classifier Mechanism

Ensemble learning is a powerful approach in machine learning that enhances the predictive accuracy and robustness by leveraging the collective strength of multiple classification models. This technique has demonstrated stability in model performance when dealing with complex and dynamic datasets by minimizing prediction bias and variance. The essence of ensemble learning lies in its ability to:

- Reduce Bias: Integrating a variety of models helps to capture a broader range of data patterns that a single model might miss, thus lowering the overall bias of the predictions.
- Reduce Variance: Sensitivity of individual models to minor fluctuations in the training data can lead to high variance in predictions. Ensemble learning addresses this issue by averaging the predictions from different models, leading to a more stable and reliable outcome.

Algorithm 1 Feature Selector-classifier Optimization Framework

Input: Credit dataset $D = \{X, y\}$
Preprocess dataset D to handle missing values and normalize data
Encode categorical features using one-hot encoding
Initialize feature selector set $S = \{\text{Random Forest, Logistic Regression}\}$
Initialize classifier set $C = \{\text{RF, LR, GBDT, DT, NN, K-NN, XGB, LightGBM}\}$
Initialize ensemble models set E
Initialize performance metrics set M
for each feature selector s_i in S **do**
    $X_{selected} \leftarrow$ Apply s_i on X to select features
    for each classifier c_j in C **do**
        Split D into D_{train} and D_{test}
        Apply oversampling on D_{train} to balance classes
        Optimize c_j using grid search and cross-validation on D_{train}
        Evaluate optimized c_j on D_{test}, compute metrics m
        $M \leftarrow M \cup \{m\}$
        $E \leftarrow E \cup \{c_j\}$ based on performance
    end for
end for
Select top-performing models from E for ensemble
Construct voting ensemble classifier from selected models
Evaluate ensemble model on D_{test}, compute metrics $m_{ensemble}$
$M \leftarrow M \cup \{m_{ensemble}\}$
Output: Performance metrics M

The implementation of voting classifiers within our ensemble learning framework mirrors a democratic decision-making process. Each model's prediction is equivalent to a vote, and the final prediction is an aggregation of all these votes. In the soft voting scheme, we not only consider the categorical predictions from each classifier but also the associated probability scores. This approach allows for a more refined and precise final prediction, which can be encapsulated by the following formula:

$$P_{\text{ensemble}} = \arg\max_i \sum_{j=1}^{n} w_j \cdot p_{ij} \tag{1}$$

In this equation, P_{ensemble} represents the ensemble's final prediction, w_j denotes the weight assigned to the j-th classifier, and p_{ij} is the probability score assigned by the j-th classifier for class i. The sum totals over all n classifiers within the ensemble.

4.2.2 Construction and Optimization of Ensemble Models

The construction of our ensemble model is meticulously designed to harness the collective strengths of a diverse array of base learners, thereby enhancing the precision and robustness of credit risk predictions. Our approach strategically amalgamates a spectrum of models, each selected for its distinctive capacity to elucidate the intricate patterns embedded in credit risk data.

At the core of our ensemble lies a judicious blend of linear and nonlinear models. Linear models, exemplified by logistic regression, lay the foundation of our ensemble, offering a transparent and interpretable framework for pinpointing pivotal risk indicators and delineating baseline data relationships. These models excel in scenarios where the data manifests perceptible linear trends, thereby providing a dependable initial point for risk assessments.

To capture the nuances and complexities that linear models may overlook, we incorporate a suite of nonlinear models, including Random Forest, Gradient Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). These models are instrumental in discerning intricate interactions and non-linear relationships among variables, which is essential for navigating the dynamic and multifaceted nature of credit risk.

The optimization of each base learner within our ensemble is conducted through a hyperparameter tuning process using cross-validation. For a given base learner, let Θ_i represent the set of candidate hyperparameters, and let D_{train} and D_{test} denote the training and test datasets, respectively. The goal is to find the optimal hyperparameter configuration θ_i^* that maximizes the AUC-ROC score on the test set:

$$\theta_i^* = \arg\max_{\theta \in \Theta_i} \text{AUC-ROC}(\theta, D_{\text{test}}) \tag{2}$$

This process is repeated for each base learner, and the resulting models are combined into an ensemble. The integration of both linear and nonlinear models within our ensemble is a deliberate and strategic choice, aimed at providing a comprehensive understanding of credit risk that is adaptable to various financial contexts and data scenarios.

In summary, our ensemble model's construction and optimization are guided by a strategic synthesis of diverse machine learning algorithms, each contributing to a holistic and nuanced perspective on credit risk. This methodical approach is crafted to yield an analytical tool that is not only accurate and reliable but also capable of delivering profound insights for informed financial decision-making.

4.3 Economic Decision Support

Our proposed framework not only enhances the credit risk assessment process but also serves as a robust decision support system for economic decision-making. It plays a pivotal role in risk management and asset allocation, enabling financial institutions to make more informed decisions.

4.3.1 Risk Management

In terms of risk management, the framework provides precise credit risk evaluations that assist financial institutions in identifying and quantifying potential loan default risks. The ensemble learning approach, which combines predictions from multiple models, enhances the ability to capture risk dynamics. This allows institutions to formulate more effective risk mitigation strategies and pricing decisions.

4.3.2 Asset Allocation

For asset allocation, the framework's insights allow portfolio managers to make more informed asset distribution based on the credit status of enterprises and market trends. By optimizing the feature space and reducing computational costs, financial institutions can predict the returns and risks of different investments more accurately, leading to better asset combinations and higher investment returns.

4.3.3 Decision Transparency

Furthermore, the framework's decision support capabilities offer transparency by providing clear credit risk assessment results. This helps decision-makers understand the logic behind the assessments, leading to more transparent and justifiable decisions. Such transparency is crucial for meeting regulatory requirements, enhancing investor confidence, and establishing a financial institution's market reputation.

Overall, the framework, through its advanced machine learning techniques and ensemble methods, offers financial institutions a new perspective for making comprehensive and refined economic decisions, particularly in the key areas of risk management and asset allocation.

5 Experimental Settings

5.1 Dataset Description

The experimental evaluation of our Feature Selector-classifier Optimization Framework was conducted using two credit risk datasets from Kaggle: the 30S-CR Small and Medium Enterprise Credit Data and the Credit Risk dataset. The 30S-CR SME Credit Data comprises 1707 samples with 87 features, representing MSMEs that have received bank financing, labeled as 0 (non-default) or 1 (default). The Credit Risk dataset includes 32581 samples with 12 features, representing individual borrower loan records with similar default labels. Both datasets were preprocessed to handle missing values, and categorical features were encoded using one-hot encoding. Features were normalized to ensure consistency and improve model convergence.

Table 3 presents basic statistical information for the two credit risk datasets used in this study.

Table 3 Dataset Overview

Dataset Name	Sample Size	Feature Count
30S-CR SME Credit Data	1707	87
Credit Risk Dataset	32581	12

5.2 Economic Context of the Datasets

The economic context of the datasets used in our experiments is of paramount importance as it provides insights into how credit risk evolves across different economic

conditions. Our datasets, which encompass a wide range of financial records from both small and medium-sized enterprises (SMEs) and individual borrowers, offer a unique opportunity to analyze the behavior of credit risk during various economic cycles.

5.2.1 Economic Cycles and Credit Risk

The datasets include historical data that spans over different phases of the economic cycle, including periods of growth, stability, and recession. By examining the credit performance of loans originated during these distinct phases, we can discern patterns and trends that are indicative of the sensitivity of credit risk to macroeconomic factors. For instance, during periods of economic downturn, we observe an increase in default rates, which is a critical factor for financial institutions to consider when assessing and managing credit risk.

5.2.2 Sector-Specific Economic Influences

Additionally, the datasets contain information on loans across various industry sectors, allowing us to analyze the impact of sector-specific economic influences on credit risk. Some sectors may be more cyclically sensitive, exhibiting higher credit risk during economic contractions compared to more stable sectors. Understanding these sector-specific dynamics is vital for making informed credit decisions and for developing tailored risk management strategies.

5.2.3 Temporal Variation in Credit Behavior

The temporal variation in credit behavior is another key aspect captured by our datasets. We analyze the evolution of creditworthiness over time, considering factors such as changes in borrower income, employment status, and market conditions. This temporal analysis is crucial for financial institutions to adapt their credit policies in response to shifting economic landscapes and borrower circumstances.

By incorporating these economic dimensions into our experimental settings, we ensure that our credit risk assessment framework is grounded in real-world economic contexts, thereby enhancing its relevance and applicability to actual financial decision-making processes.

5.3 Benchmark Models and Evaluation Metrics

In our quest to evaluate and enhance the performance of credit risk assessment models, we have established a set of benchmark models that are widely recognized for their effectiveness in classification tasks. The models include Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Decision Tree (DT), Neural Network (NN), K-Nearest Neighbors (KNN), XGBoost (XGB), and LightGBM (LGBM). These models serve as a diverse foundation for our ensemble learning approach, each bringing unique strengths to the collective prediction process. The choice of base learners is based on their individual strengths and their potential to contribute to the ensemble's predictive power. The models serve as a benchmark against which the performance of

the ensemble classifier is compared. This comparison not only validates the effectiveness of the ensemble approach but also highlights the improvements gained through the integration of diverse models.

Our evaluation of the credit risk assessment models is based on a set of established benchmarks and metrics that objectively measure the performance of each model. The key metrics used in this study are Accuracy and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC).

5.3.1 Confusion Matrix Analysis

The confusion matrix is a critical tool for evaluating the performance of classification models. It provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, which are essential for calculating various performance metrics. The number of instances where defaulting samples are correctly identified as defaults is termed **True Positives (TP)**; the count of defaulting samples incorrectly classified as non-defaults is labeled **False Negatives (FN)**; the quantity of non-defaulting samples accurately determined as non-defaults is known as **True Negatives (TN)**; and the number of non-defaulting samples mistakenly categorized as defaults is called **False Positives (FP)**, as illustrated in Table 4.

Table 4 Confusion Matrix

	Predicted	
Actual	Negative	Positive
Negative	TN	FP
Positive	FN	TP

5.3.2 Accuracy

Accuracy is a fundamental metric that calculates the proportion of correctly classified instances out of the total number of instances. It is given by the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP (True Positives) is the number of positive instances correctly classified, TN (True Negatives) is the number of negative instances correctly classified, FP (False Positives) is the number of negative instances incorrectly classified as positive, and FN (False Negatives) is the number of positive instances incorrectly classified as negative.

5.3.3 Area Under the ROC Curve (AUC-ROC)

The AUC-ROC metric is a graphical representation of a model's diagnostic ability. It measures the probability that the model ranks a randomly chosen positive instance

higher than a randomly chosen negative instance. The AUC is calculated by integrating the True Positive Rate (TPR) over the False Positive Rate (FPR) for all possible cut-off points. The formula for AUC is given by:

$$\text{AUC} = \int_0^1 TPR(x)\, dFPR(x) \tag{4}$$

where $TPR(x) = \frac{TP}{TP+FN}$ is the True Positive Rate and $FPR(x) = \frac{FP}{FP+TN}$ is the False Positive Rate at a given threshold x.

These metrics provide a comprehensive view of the models' effectiveness in credit risk assessment, with Accuracy offering a quick overview of overall correctness and AUC-ROC providing a measure of the model's ability to discriminate between the two classes across all classification thresholds.

5.4 Hyperparameter Tuning and Model Optimization

A critical component in refining predictive models is the careful adjustment of hyperparameters. Our approach involves a thorough procedure for hyperparameter optimization, which employs a systematic search strategy across a predefined range of values, coupled with a validation technique to ascertain the most effective configuration for each model. The goal of this process is to maximize the Area Under the ROC Curve (AUC-ROC), a key performance metric in binary classification tasks.

The hyperparameter tuning process involved defining a grid of potential values for each hyperparameter and conducting cross-validation for each combination to evaluate the model's performance. The optimal hyperparameters were selected based on the highest AUC-ROC score achieved during the cross-validation process.

The optimal hyperparameters for the benchmark models on the 30S-CR SME Credit Data and the Credit Risk dataset are presented in Tables 5 and 6, respectively.

Table 5 Optimal Hyperparameters for Benchmark Models on the 30S-CR SME Credit Data

Classifier	Best Params
Random Forest	$\{'max\_depth' : \text{None}, 'n\_estimators' : 300\}$
Logistic Regression	$\{'C' : 10\}$
Gradient Boosting	$\{'learning\_rate' : 0.5, 'n\_estimators' : 300\}$
Decision Tree	$\{'max\_depth' : \text{None}\}$
Neural Network	$\{'activation' : 'relu', 'hidden\_layer\_sizes' : (256,)\}$
K-Nearest Neighbors	$\{'n\_neighbors' : 7\}$
XGBoost	$\{'learning\_rate' : 0.1, 'n\_estimators' : 200\}$
LightGBM	$\{'learning\_rate' : 0.1, 'n\_estimators' : 300\}$

Table 6 Optimal Hyperparameters for Benchmark Models on the Credit Risk Dataset

Classifier	Best Params
Random Forest	$\{'max\_depth' : \text{None},' n\_estimators' : 300\}$
Logistic Regression	$\{'C' : 0.1\}$
Gradient Boosting	$\{'learning\_rate' : 0.5,' n\_estimators' : 300\}$
Decision Tree	$\{'max\_depth' : 10\}$
Neural Network	$\{'activation' :' relu',' hidden\_layer\_sizes' : (256,)\}$
K-Nearest Neighbors	$\{'n\_neighbors' : 5\}$
XGBoost	$\{'learning\_rate' : 0.5,' n\_estimators' : 300\}$
LightGBM	$\{'learning\_rate' : 0.1,' n\_estimators' : 300\}$

Table 7 Performance on the 30S-CR MSME Credit Data

Feature Selector	Classifier	Accuracy	AUC
Random Forest	RF	0.927	0.727
	LR	0.681	0.694
	GBDT	0.915	0.614
	DT	0.857	0.525
	NN	0.918	0.618
	K-NN	0.746	0.583
	XGB	**0.934**	**0.729**
	LightGBM	0.933	0.653
	Voting Classifier	**0.941**	**0.749**
Logistic Regression	RF	0.939	0.694
	LR	0.719	0.702
	GBDT	0.921	0.686
	DT	0.836	0.538
	NN	0.898	0.486
	K-NN	0.795	0.642
	XGB	0.939	0.675
	LightGBM	0.939	0.643
	Voting Classifier	0.939	0.738

6 Results

6.1 MSME Credit Data by 30S-CR Results

In the context of the MSME Credit Data provided by 30S-CR, our Feature Selector-classifier Optimization Framework demonstrated a significant enhancement in the predictive modeling of credit risk. The strategic integration of the Random Forest feature selector with the XGBoost classifier led to an impressive accuracy of 0.934 and an AUC score of 0.729, as indicated in Table 7. These results are not merely numerical accomplishments but represent a substantial advancement in the realm of credit risk assessment for SMEs, which often grapple with limited resources and complex financial landscapes.

The ensemble classifier, comprising the most efficacious base learners—Random Forest, XGBoost, and LightGBM—achieved an accuracy of 0.941 and an AUC of 0.749. The ensemble's performance is a testament to the power of collective intelligence in machine learning. By amalgamating the predictions of multiple models, the

Table 8 Performance on the Credit Risk Dataset

Feature Selector	Classifier	Accuracy	AUC
Random Forest	RF	0.897	0.904
	LR	0.785	0.839
	GBDT	0.904	0.910
	DT	0.867	0.864
	NN	0.846	0.875
	K-NN	0.796	0.832
	XGB	**0.905**	**0.907**
	LightGBM	**0.908**	0.906
	Voting Classifier	**0.903**	**0.913**
Logistic Regression	RF	0.856	0.873
	LR	0.799	0.855
	GBDT	0.878	0.888
	DT	0.857	0.823
	NN	0.874	0.899
	K-NN	0.857	0.866
	XGB	0.869	0.883
	LightGBM	0.865	0.881
	Voting Classifier	0.883	0.899

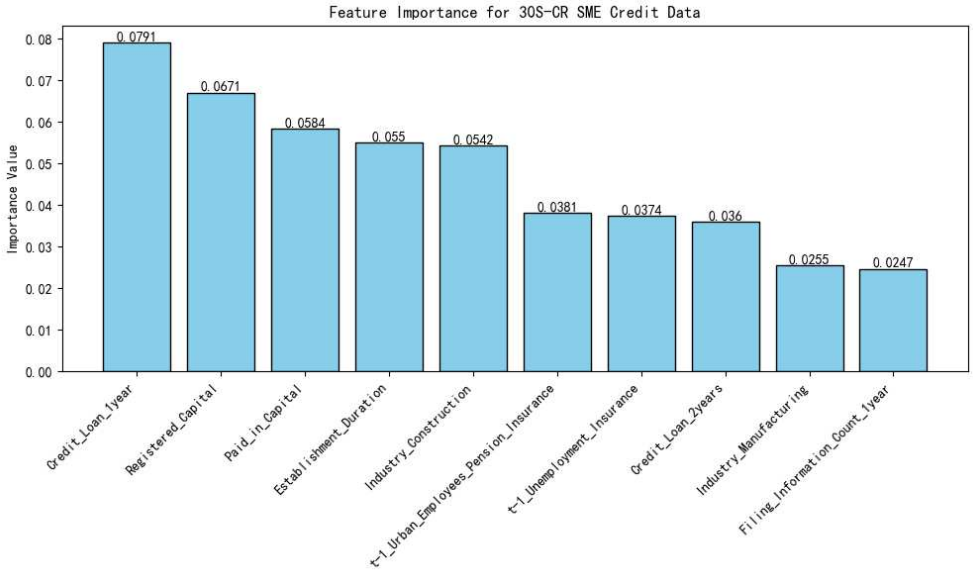

Fig. 2 Feature Importance Analysis for the 30S-CR SME Credit Data

ensemble classifier mitigates the vulnerabilities inherent in individual models, thereby providing a more nuanced and stable prediction of credit risk.

The feature importance analysis, as determined by the Random Forest feature selector, is depicted in Figure 2. This analysis reveals the critical features that substantially enhance the model's predictive accuracy.. The feature 'Credit Loan 1year'

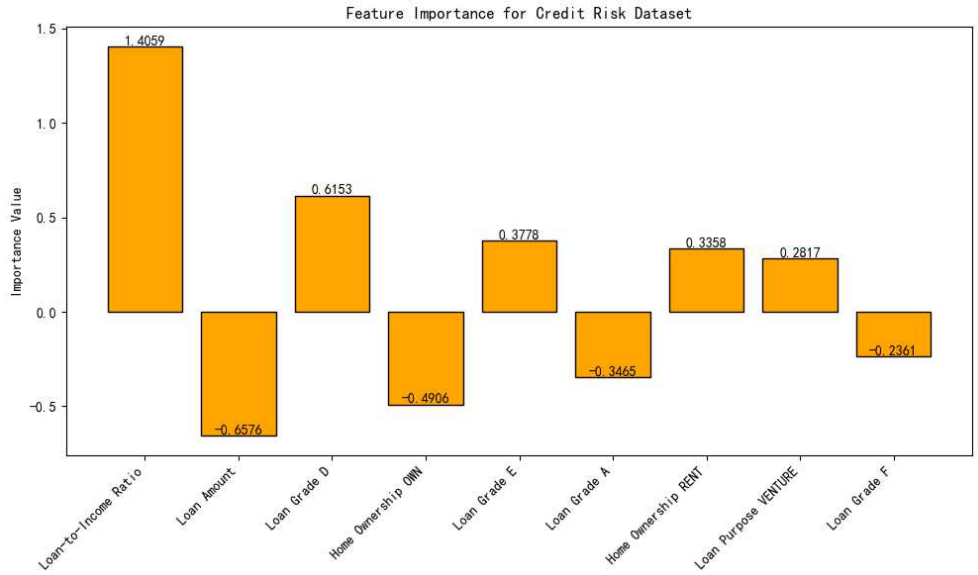

Fig. 3 Feature Importance Analysis for the Credit Risk Dataset

garnered the highest importance score, highlighting its predictive relevance in the SME credit risk context. This feature's prominence underscores the criticality of recent credit history in evaluating the creditworthiness of SMEs. The 'Registered Capital' and 'Paid in Capital' also emerged as significant determinants, reflecting the financial stability and commitment of the business entities.

Moreover, features such as 'Establishment Duration,' 'Industry Construction,' and 't-1 Urban Employees Pension Insurance' were found to contribute significantly to the predictive efficacy of the model. These insights are invaluable for financial institutions and regulators as they seek to craft credit risk assessment models that are both comprehensive and attuned to the unique characteristics of SMEs.

6.2 Credit Risk Dataset Results

The empirical analysis utilizing the Credit Risk dataset was instrumental in validating the efficacy of our Feature Selector-classifier Optimization Framework in the domain of credit risk assessment. The synergistic application of the Random Forest feature selector with the XGBoost classifier yielded a notable accuracy of 0.905 and an AUC score of 0.907, as delineated in Table 8. These metrics underscore the framework's capability to discern complex patterns and relationships within credit data, thereby enhancing the predictive accuracy of risk assessments.

The ensemble classifier, which integrates the top-performing base learners—Random Forest, XGBoost, and LightGBM—exhibited a competitive performance with an accuracy of 0.903 and an AUC of 0.913. This outcome is particularly significant as it demonstrates the ensemble's ability to leverage the strengths of diverse

models, leading to a more robust and generalizable prediction model. The ensemble's superior performance is attributed to its capacity to reduce the impact of individual model biases and to improve the overall stability of predictions.

The feature importance analysis, as elucidated by the Random Forest feature selector, is presented in Figure 3. The feature 'loan percent income' emerged as the most influential with a significant importance score of 0.217474. This finding is aligned with financial intuition, as the proportion of loan amount relative to the borrower's income is a pivotal indicator of repayment capacity. The close correlation between this feature and credit risk underscores the necessity for lenders to meticulously assess borrowers' income stability and debt servicing ability.

Following 'loan percent income,' the features 'loan amnt' and 'loan grade (D)' were identified with importance scores of -0.657615 and 0.61531, respectively. The negative score associated with 'loan amnt' suggests that larger loan amounts may be associated with a higher risk profile, which is a valuable insight for risk management strategies. Concurrently, 'loan grade (D)' exhibited a positive importance score, indicating that this grading system is effective in distinguishing creditworthiness levels and should be considered in credit risk modeling.

The feature importance analysis not only provides actionable insights for financial institutions but also contributes to the theoretical understanding of credit risk factors. By identifying key predictors, our framework enables lenders to refine their risk assessment strategies and to develop more targeted underwriting criteria. Moreover, the emphasis on data-driven feature selection and model integration offers a methodological advancement over traditional expert-driven approaches, which may be subject to individual biases and limitations.

6.3 Consistency with Economic Credit Data

The results obtained from our framework are consistent with the principles outlined in the study of credit big data and its application to SME financing innovation. The integration of multi-source data, including financial records and industry-specific metrics, corroborates the effectiveness of our approach in providing a comprehensive credit risk assessment. This aligns with the empirical analysis presented in the aforementioned study, which demonstrated the advantages of credit big data in evaluating the credit risk of engineering construction enterprises.

In conclusion, the results section showcases the predictive prowess of our framework and its ability to identify key credit risk indicators. The integration of diverse data sources and the application of advanced machine learning techniques have proven to be a formidable combination in the field of credit risk assessment.

7 Conclusion

The present study introduces the Feature Selector-classifier Optimization Framework, an innovative approach designed to enhance the accuracy and efficiency of credit risk assessment. Our framework integrates ensemble learning and feature optimization techniques, demonstrating its effectiveness in managing the complexities inherent in

credit risk prediction. This conclusion synthesizes our key findings and discusses the broader implications of our research.

Our empirical investigation has demonstrated that the ensemble classifier consistently surpasses its individual counterparts in terms of accuracy and AUC scores across both the 30S-CR SME Credit Data and the Credit Risk dataset. This remarkable performance not only highlights the robustness of our framework but also its adeptness at capturing the intricate and nuanced dynamics of credit risk. The framework's exceptional handling of imbalanced and large-scale datasets is particularly noteworthy, showcasing its capacity to navigate the complexities of real-world financial data. Such success positions our framework as a potentially transformative tool in the financial industry's ongoing quest for innovation in credit risk management.

While our study presents promising results, it is not without limitations. The framework's performance, as evaluated on existing datasets, must be tested under dynamic real-world economic conditions. Future research should explore the framework's applicability in various economic contexts and investigate its integration with emerging financial technologies such as blockchain and alternative data sources.

In conclusion, the Feature Selector-classifier Optimization Framework represents a significant advancement in the field of credit risk assessment. It provides a data-driven, automated approach that enhances predictive accuracy and informs decision-making in financial institutions. The framework's success on diverse datasets and its alignment with economic credit data principles highlight its potential as an invaluable instrument in the financial sector's ongoing quest for innovation in credit risk management.

References

Bhatore, S., Mohan, L., Reddy, Y.R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *4*, 111–138,

Bu, S., Guo, N., Li, L. (2022). Rating frailty, bayesian updates, and portfolio credit risk analysis. *Quantitative Finance*, *22*(4), 777–797,

Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*, 203–216,

Chen, N., Ribeiro, B., Chen, A. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, *45*, 1–23,

Corazza, M., De March, D., Di Tollo, G. (2021). Design of adaptive elman networks for credit risk assessment. *Quantitative Finance*, *21*(2), 323–340,

Di Lorenzo, E., Piscopo, G., Sibillo, M. (2024). Addressing the economic and demographic complexity via a neural network approach: risk measures for reverse mortgages. *Computational Management Science*, *21*(1), 11,

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational economics*, *15*, 107–143,

Gärtner, T., Kaniovski, S., Kaniovski, Y. (2021). Numerical estimates of risk factors contingent on credit ratings. *Computational Management Science*, *18*(4), 563–589,

Gicić, A., Đonko, D., Subasi, A. (2023). Intelligent credit scoring using deep learning methods. *Concurrency and Computation: Practice and Experience*, *35*(9), e7637,

Guanghui, Z. (2021). Promoting financing for small and medium-sized enterprises through "xinyidai". *Macroeconomic Management*, ,

Hu, J.T. (2022). The rule of credit data opening and protection. *Credit*, *40*(11), 7-13, https://doi.org/10.15896/j.xjtuskxb.202206011

Huang, Y., Zhang, L., Li, Z., Qiu, H., Sun, T., Wang, X. (2020). Fintech credit risk assessment for smes: Evidence from china.

Le, R., Ku, H., Jun, D. (2021). Sequence-based clustering applied to long-term credit risk assessment. *Expert Systems with Applications*, *165*, 113940,

Li, W., Paraschiv, F., Sermpinis, G. (2022). A data-driven explainable case-based reasoning approach for financial risk detection. *Quantitative Finance*, *22*(12), 2257–2274,

Liu, C., Xie, J., Zhao, Q., Xie, Q., Liu, C. (2019). Novel evolutionary multi-objective soft subspace clustering algorithm for credit risk assessment. *Expert Systems with Applications*, *138*, 112827,

Liu, S., & Vicente, L.N. (2022). Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, *19*(3), 513–537,

Liu, Y.T., Tang, S.S., Liang, M. (2016). The formation, application value and enhancement strategy of credit big data. *South China Finance*, *11*, 47-53, https://doi.org/10.1109/ACCESS.2017.2789283

Mahajan, S., Nayyar, A., Raina, A., Singh, S.J., Vashishtha, A., Pandit, A.K. (2022). A gaussian process-based approach toward credit risk modeling using stationary activations. *Concurrency and Computation: Practice and Experience*, *34*(5), e6692,

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International journal of financial studies*, *9*(3), 39,

Soui, M., Gasmi, I., Smiti, S., Ghédira, K. (2019). Rule-based credit risk assessment model using multi-objective evolutionary algorithms. *Expert systems with applications*, *126*, 144–157,

Sousa, M.R., Gama, J., Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications*, *45*, 341–351,

Su, W., & Ren, J. (2019). Risk propagation model based on social life and credit activities multilayers fusion network. *Concurrency and Computation: Practice and Experience*, *31*(10), e4732,

Wang, T., Liu, R., Qi, G. (2022). Multi-classification assessment of bank personal credit risk based on multi-source information fusion. *Expert systems with applications*, *191*, 116236,

Xiong, Z.D. (2022). Data and credit: The hegemony of capitalist credit in the digital age and its critique. *Journal of Xi'an Jiaotong University (Social Sciences)*, *42*(06), 104-111, https://doi.org/10.15896/j.xjtuskxb.202206011

Xu, P., Ding, Z., Pan, M. (2018). A hybrid interpretable credit card users default prediction model based on ripper. *Concurrency and Computation: Practice and*

Experience, *30*(23), e4445,

Ying, L., & Lihua, H. (2020). Supply chain finance credit risk assessment using support vector machine - based ensemble improved with noise elimination. *International Journal of Distributed Sensor Networks*, *16*(1), 1550147720903631,

Yu, L., Zhang, X., Yin, H. (2022). An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity. *Expert Systems with Applications*, *202*, 117363,

Zhao, F., Li, G., Lyu, Y., Ma, H., Zhu, X. (2023). A cost-sensitive ensemble deep forest approach for extremely imbalanced credit fraud detection. *Quantitative Finance*, *23*(10), 1397–1409,

Declarations

Funding

This work was supported in part by the National Natural Science Foundation of China, Major Research Program Integration Project (No. 92146005), National Natural Science Foundation of China (No. 62173285) and the Fujian Provincial Natural Science Foundation of China (No. 2021J011181).

Conflicts of interest/Competing interests

No.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Not applicable.

Code availability

Not available.