IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 3 August 2024, accepted 15 August 2024, date of publication 19 August 2024, date of current version 28 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445499

RESEARCH ARTICLE

Feature Enhanced Ensemble Modeling With Voting Optimization for Credit Risk Assessment

DONGQI YANG AND BINQING XIAO
School of Management and Engineering, Nanjing University, Nanjing 210008, China

Corresponding author: Binqing Xiao (bengking@nju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 72342024 and Grant 72071102, and in part by the National Social Science Foundation of China under Grant 23&ZD175.

ABSTRACT Machine learning methods have gained widespread utilization in small and micro enterprise credit risk assessment. However, the practical application of these methods encounters a conundrum involving accuracy and interpretability. In this study, a multi-stage ensemble model is proposed to enhance the model's interpretability. To strengthen predictive portraits, a multi-feature enhancement method is proposed to integrate non-financial behavioral information and soft information on credit rating into the annual loan ledger data, thereby bolstering the explanatory capacity of the features. To rectify the issue of data imbalance and avoid information loss, a new bagging-based oversampling method is proposed to oversample the minority class samples in multiple parallelized subsets divided by the bagging strategy. To unleash the performance potential of base classifiers, a new voting-weight optimization method is proposed to optimize the soft voting weights of the candidate base classifiers. The experiment results of an annual loan ledger dataset of a commercial bank in China (with an accuracy of 97.9%, an area under the curve of 0.97, a logistic loss of 0.07, a Brier score of 0.01, and a Kolmogorov-Smirnov statistic of 0.38) and the other five public datasets indicating excellent model fit. By focusing on the widespread soft information and data structures characteristic of SME loan risk assessment data, an additional SHAP model explanation method enhances interpretability. This method reveals that the enhanced 'debt-to-income ratio,' along with non-financial behavioral information and features derived from soft information, are essential for predicting loan defaults. Such enhancements help to alleviate the issue of information asymmetry in SME loan risk assessment.

INDEX TERMS Credit risk, ensemble modeling, feature enhancement, model interpretability, voting optimization.

I. INTRODUCTION

For Small and Medium-sized Enterprises (SMEs), the challenge in credit risk assessment lies in the identification of default characteristics. SMEs often lack effective financial information and have weaker risk resistance capabilities, which adds to the complexity of the assessment. Traditional banking and financial features may lack specificity when assessing the credit risk of SMEs [1]. Levine et al. highlighted that the primary assessment for SME loans is based on unquantifiable "soft information" provided by account managers [2]. Therefore, the credit risk assessment for SMEs should transcend traditional financial information and take into account the financial status of the business owners as well as external data sources (non-financial and soft information) to provide a comprehensive reflection of their creditworthiness.

Furthermore, with the increasing demand for SME loans, the development of new credit characteristics and rating methods has become urgent. Given the significant similarity between corporate bankruptcy prediction and credit scoring, incorporating features from credit scoring approaches [3], Features from credit scoring approaches, such as the debt-to-income ratio that is effective in credit scoring, can enhance the identification of default characteristics in the credit risk assessment for SMEs.

Another issue arises due to the data imbalance in the corporate loan dataset, where non-defaulted SMEs (majority

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

class samples) far outnumber defaulted SMEs (minority class samples), and the dataset size is relatively small [4]. Consequently, this results in a scarcity of default samples in the corporate loan dataset and a highly imbalanced data distribution. Applying undersampling methods to address data imbalances would result in the removal of a substantial amount of information, aggravating the scarcity of an already limited sample size [5]. Therefore, one of the main objectives of this study is to develop an oversampling method that can effectively generate minority class samples to alleviate the data imbalance issue.

Traditional linear models may exhibit limitations in accuracy when dealing with highly complex credit risk assessment problems [3]. Ensemble learning algorithms have been demonstrated to provide higher accuracy compared to traditional models across various datasets [4]. In the era of big data, credit risk assessment models are required to handle vast amounts of data with complex features, as is the case with the experimental dataset adopted in this study. Ensemble methods can combine the predictive results of multiple models, enhancing the understanding of data complexity and capturing nonlinear relationships within the data. Furthermore, ensemble learning algorithms generally possess better generalization capabilities. By combining multiple weak learners, they can mitigate the risk of overfitting and offer more accurate predictions on unseen data [6]. These models leverage their robust predictive capabilities, which are typically constructed using ensemble methods such as classifiers adopted in this study, i.e., gradient boosting decision tree (GBDT) [7], random forest (RF) [8], adaptive boosting (AdaBoost) [9], extreme gradient boosting (XGBoost) [10], extremely randomized trees (ExtraTree) [11], bootstrap aggregating (Bagging) [12], and light gradient-boosting machine (LightGBM) [13]. However, ensemble-based models have been criticized for their lack of interpretability compared to traditional linear models. Simply providing accurate predictions is inadequate to support practical applications. Therefore, this study aims to unveil and improve the transparency of machine learning models by deconstructing their ''black box'' nature.

This study presents a novel multi-stage ensemble model for assessing the risk of SME loans. First, a multi-feature enhancement method is proposed to incorporate external data sources (e.g., external legal risk features) and soft information (e.g., expert credit evaluations) about SME loans. This integration aims to enhance the explanatory capability of the features and generate a feature-enhanced training set. Subsequently, a new bagging-based oversampling method is proposed to partition the feature-enhanced training set into multiple parallelized subsets through the bagging strategy. This method overcomes imbalances in the minority class samples through the use of Synthetic Minority Oversampling Technique (SMOTE) [14], leading to a balanced training set. Furthermore, a new voting-weight optimization method is proposed for optimizing the soft voting weights of the candidate base classifiers in the classifier ensemble, utilizing the L-BFGS-B algorithm [15] for constructing the stacking-based model. Finally, the SHapley Additive exPlanations (SHAP) explanation method is employed to evaluate the default features in the proposed model from the perspective of machine learning interpretability. Empirical findings demonstrate [16], [17] the exceptional performance of the SHAP explanation method in assigning feature importance. The SHAP explanation method indicates that, beyond loan information, financial information, and non-financial basic information included in the annual loan ledger data, the enhanced features (e.g., the enhanced ''debt-to-income ratio,'' non-financial behavioral information, and soft information on credit rating) are effective features of SMEs credit risk assessment.

The main contributions of this paper are highlighted below:

(1) The external data sources and soft information are integrated to enhance credit risk models for SMEs. By incorporating these elements, the model captures a comprehensive view of SME creditworthiness, addressing limitations in traditional financial metrics.

(2) The soft voting weights of the classifier ensemble are optimized adaptively by the L-BFGS-B algorithm, which enhances the performance of the ensemble model, ensuring more accurate and reliable credit risk predictions.

(3) The SHAP method is employed to evaluate feature importance and enhance the interpretability of the model, which provides insights into key features affecting credit risk, such as the debt-to-income ratio and non-financial behavioral information, making the model more transparent and trustworthy for practical applications.

The remainder of this study is organized as follows: Section II reviews relevant literature on the proposed model. Section III provides detailed explanations of the model. Section IV introduces the experimental settings. Section V analyzes the experimental results. Section VI concludes this study and discusses future research directions.

II. RELATED WORK

A. FEATURE ENHANCEMENT

The rapid advancement of big data and artificial intelligence technology in recent years has prompted a growing interest in using external data sources as supplementary and expansive resources to traditional data within the financial industry and academic community. For instance, researchers have investigated the utilization of social media data, geographical data, and online behavior data for evaluating borrowers' credit risks [18]. Moreover, studies have examined the practical applications of mobile phone and email data [19]. These data sources provide innovative insights and perspectives for credit risk assessment, empowering financial institutions to conduct thorough and precise credit evaluations and risk control [20].

However, in the credit risk assessment of SMEs, risks from legal proceedings would greatly affect business operations [21]. Moreover, the business model of commercial banks assigns a certain level of importance to expert credit evaluations of SMEs [22]. Consequently, credit risk assessment for SMEs should encompass not only their loan information and financial data but also take into account external data sources such as non-financial basic information, non-financial behavioral information, and soft information on credit rating. The previous study demonstrated that derived features from outlier algorithms could improve the interpretability of the dataset [23]. In this study, a multi-feature enhancement method was proposed to extend the work [23] by integrating external data sources and soft information in the credit risk assessment of SMEs. To incorporate the legal and compliance aspects of SME business operations, the relevant legal case information was included in the dataset. Additionally, to account for the influence of soft information, the ''account manager's evaluation of customer credit'' data was enhanced in the dataset. These feature enhancements play a crucial role in constructing a more comprehensive and informative training set.

While there has been interest in the use of external data sources for credit risk assessment within the financial industry and academic community, the systematic integration of these sources into a multi-feature enhancement method tailored for SMEs has been lacking. The multi-feature enhancement method goes beyond traditional financial data, which is often limited in scope, particularly for SMEs that may lack comprehensive financial records. These enhancements are expected to result in a more comprehensive and informative training set, thereby improving the accuracy and reliability of credit risk assessments for SMEs. By proposing the multi-feature enhancement method that integrates these diverse data sources, including soft information and legal case data, the study seeks to fill the existing gap and provide a more holistic approach to credit risk assessment for SMEs.

B. BALANCED SAMPLING

Assessing the credit risk of SMEs is a vital undertaking for financial institutions, as it enables them to determine the creditworthiness and potential default risks associated with providing loans to these enterprises. There has been considerable focus on using oversampling methods as a viable approach to handling the imbalanced nature of corporate loan datasets, where the number of default samples is often significantly smaller than that of non-default samples.

In many practical applications, the minority class may contain critical information. Oversampling methods aim to mitigate the issue of data imbalance by artificially increasing the representation of samples from the minority class [24]. These techniques encompass synthetic oversampling methods, like the SMOTE, which generate synthetic instances by interpolating between neighboring samples of the minority class, thereby aiding the model in better learning the characteristics of the minority class [14].

By balancing the dataset, SMOTE helps to enhance the model's recognition capability for the minority class, potentially improving overall classification accuracy, recall, and F1 scores [25]. Besides, SMOTE does not require a large number of minority class samples, which is particularly useful in this study when the cost of obtaining additional samples is high or impractical.

Applying SMOTE methods to SME credit risk assessment can help address the imbalanced distribution of default and non-default samples, leading to improved model performance in predicting loan defaults. Nevertheless, it is essential to consider the potential drawbacks and challenges associated with oversampling methods, including the heightened computational complexity. Therefore, in this study, a new bagging-based oversampling method is proposed to partition the feature-enhanced training set into multiple parallelized subsets through the bagging strategy. The subsets are balanced using the SMOTE and subsequently merged into a balanced training set. This method plays a pivotal role in enhancing the learning capability of classifiers and reducing time overhead, ultimately leading to a more effective assessment of credit risk for SMEs.

Building upon the related work, this study identifies a research gap where traditional oversampling methods, despite their benefits, may introduce computational complexity. To bridge this gap, the study introduces an innovative bagging-based oversampling method. This approach segments the feature-enhanced training set into multiple subsets, which are then independently balanced using SMOTE and merged to form a balanced dataset. This methodology is designed to bolster the classifiers' learning capabilities and to curtail computational overhead, thereby facilitating a more precise and efficient credit risk assessment for SMEs. The integration of this method into the existing body of knowledge advances the field by offering a practical solution to the prevalent issue of data imbalance, enhancing model performance in predicting loan defaults for SMEs.

C. ENSEMBLE MODELING & INTERPRETABILITY

Ensemble modeling has replaced logistic regression as the dominant approach in credit risk assessment due to its superior performance and adaptability to large-scale data, with stacking [26] being a representative method. Stacking is based on the premise that combining diverse models can outperform a single model by capturing various aspects of the underlying data distribution through the hard and soft voting mechanisms, supporting multiple aggregation choices for single-classifier predictions [25].

However, the enhanced accuracy and performance achieved through ensemble modeling often come at the cost of interpretability due to its significantly more complex structure compared to logistic regression. Interpretability pertains to the capacity to comprehend and elucidate a model's decision-making process, specifically regarding the contribution and significance of each feature. The ensemble models can make it challenging to differentiate individual feature contributions and comprehend the underlying rationale behind predictions.

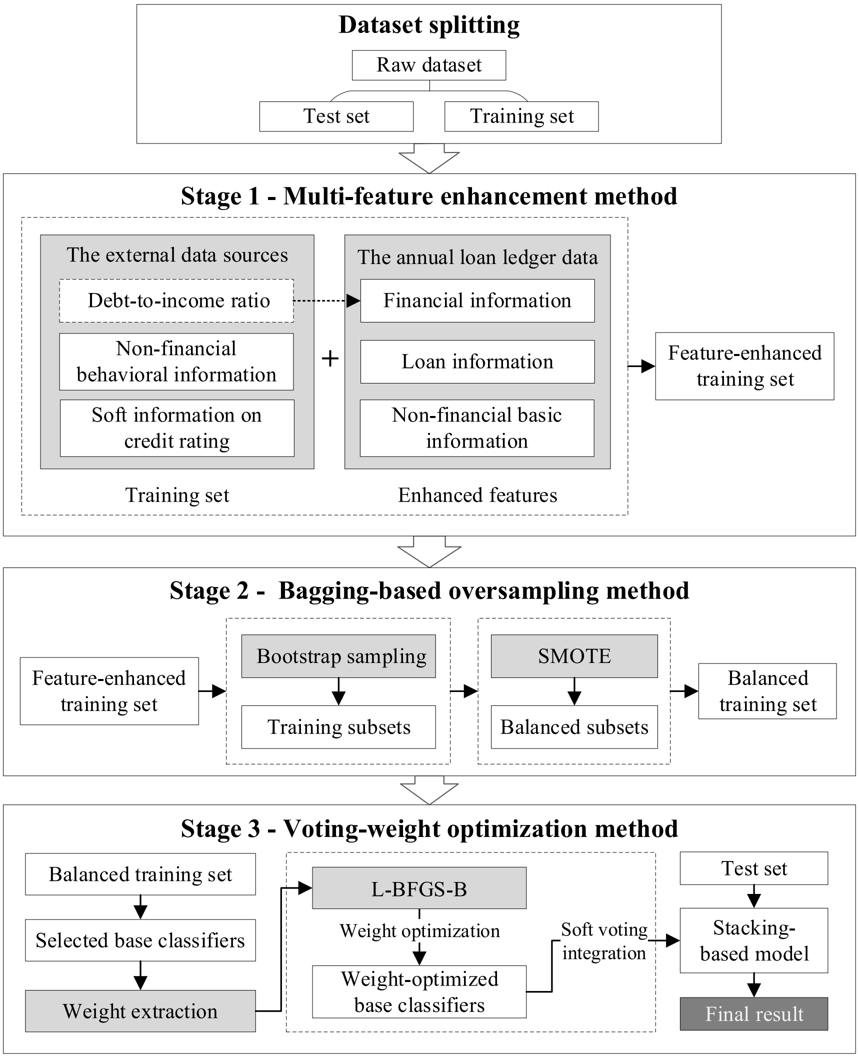

FIGURE 1. Framework of the proposed model.

Researchers have dedicated considerable efforts to resolving the trade-off between interpretability and accuracy in ensemble modeling. Methods such as SHapley Additive exPlanations (SHAP) [27] and feature importance measures have been proposed to quantify the influence of features in ensemble models. However, the pursuit of both high accuracy and interpretability remains an ongoing research challenge.

This study extends prior research by aiming to improve the performance of the ensemble model through the optimization of the soft voting decision weight for the classifier ensemble. The principle is as follows: the soft voting mechanism assumes equal decision weights predicted by different base classifiers, implying that the prediction bias of base classifiers uniformly influences the overall model performance. Therefore, a more effective approach is to adaptively adjust the soft voting weights of the base classifiers in a classifier ensemble based on different data sources and prediction performance. This adaptation reduces the decision weight of the base classifier with poor performance, thereby mitigating excessive degradation in model performance. In this study, a new voting-weight optimization method was proposed to optimize the soft voting weights of base classifiers in the classifier ensemble using the L-BFGS-B algorithm for adaptive optimization. Furthermore, a stacking-based heterogeneous ensemble model was constructed with optimized weights for the classifiers.

This study advances the field by extending prior research through the optimization of soft voting decision weights in ensemble models. The traditional soft voting mechanism assumes equal decision weights from base classifiers, which implies that the prediction bias of each base classifier uniformly affects the overall performance. Challenging this assumption, the study proposes an adaptive adjustment of soft voting weights based on the performance of each base classifier. This optimization aims to diminish the impact of poorly performing base classifiers, thus preventing a significant decline in overall model performance. By addressing the need for performance optimization and interpretability in ensemble models, this study contributes to resolving the persistent challenges in the field and offers a pathway toward more effective and transparent credit risk assessment models.

III. PROPOSED MODEL

This study proposes a novel multi-stage ensemble model that assesses the risk of SME loans. The framework is illustrated in Figure 1. The proposed model comprises three main stages: multi-feature enhancement, bagging-based oversampling, and voting-weight optimization. Details of these three stages are presented in the following subsections.

A. MULTI-FEATURE ENHANCEMENT METHOD

The loan default status of small and medium-sized enterprises (SMEs) is not solely reliant on their financial situation [28]. External data sources [29], [30] and soft information [31], [32] have demonstrated their effectiveness in providing predictive information. Therefore, a multi-feature enhancement method was proposed to integrate various external data sources, including non-financial behavioral information and soft information on credit rating, with the financial information, loan information, and non-financial basic information found in the annual loan ledger data, as illustrated in Figure 2. This integration aims to enhance the explanatory power of the features. Firstly, the debt-to-income ratio was enhanced as a financial information feature due to its high similarity to corporate bankruptcy prediction and credit scoring. Next, relevant legal case information (the number of plaintiff/accused cases), which includes external legal risk features that influence daily business operations, was enhanced as non-financial behavioral information in the dataset. Additionally, the expert's credit evaluations were enhanced with soft information on credit ratings provided by account managers who deeply understand their clients. To incorporate qualitative soft information, a discretization approach was employed to transform soft information into multiple graded labels, enabling its quantitative integration into the model. These feature enhancements are merged into the annual loan ledger data of SMEs to create the feature-enhanced training set.

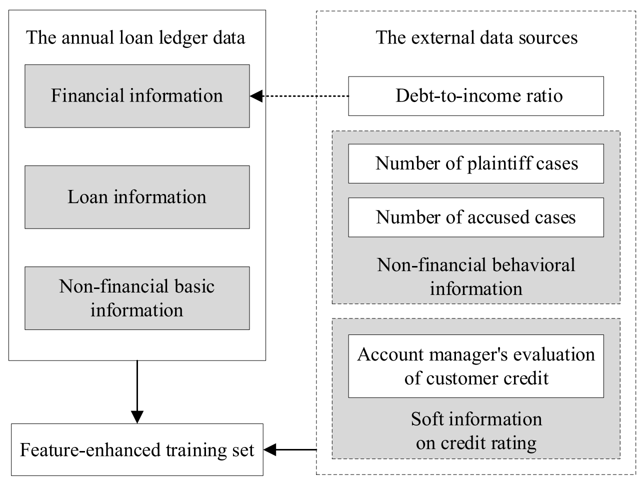

FIGURE 2. Schematic of the multi-feature enhancement method.

B. BAGGING-BASED OVERSAMPLING METHOD

In the scenario of assessing credit risk for SMEs, special attention is given to defaulted SMEs that provide critical and valuable default information, which are considered minority class samples in the dataset [33]. Since the number of defaulted SMEs is relatively small in real business, the dataset is severely imbalanced. Therefore, employing an oversampling method is crucial to effectively expanding and extracting intrinsic information while minimizing the risk of losing important data. To address the issue of high time overhead associated with the oversampling method, a bagging-based oversampling technique is proposed, leveraging the time-saving characteristic of the bagging strategy. As illustrated in Figure 3, the method involves dividing the feature-enhanced training set into multiple parallelized subsets using the bagging strategy. To ensure balance, the minority class samples are oversampled using the SMOTE technique until their size matches that of the majority class samples. Finally, the balanced subsets are combined into a balanced training set.

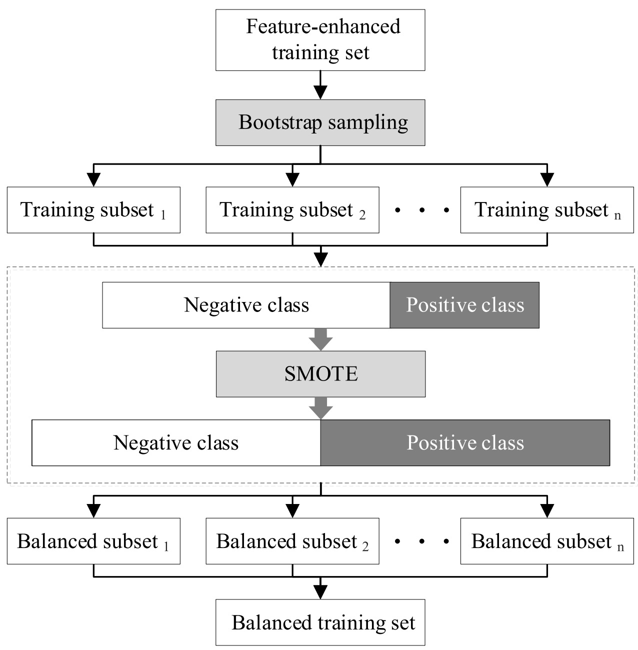

FIGURE 3. Schematic of the bagging-based oversampling method.

C. VOTING-WEIGHT OPTIMIZATION METHOD

The balanced training set was used to train the base classifiers, and their performance was assessed by calculating the area under the receiver operating characteristic curve (AUC) [34]. As depicted in Figure 4, the selected base classifiers (i.e., $Clf\ 1, Clf\ 2, \ldots, Clf\ n$) were selected to constitute the candidate classifier pool for subsequent ensemble modeling using the soft voting mechanism. To optimize the capabilities of the selected base classifiers, the soft voting weights (i.e., $W_{Clf\ 1}, W_{Clf\ 2}, \ldots, W_{Clf\ n}$) of the candidate base classifiers were fine-tuned utilizing the L-BFGS-B algorithm.

Consequently, the weight-optimized base classifiers (i.e., $Oclf\ 1, Oclf\ 2, \ldots, Oclf\ n$) were obtained to create the optimized classifier pool for the stacking integration process. This formed the basis for constructing the stacking model, which was further applied to predict the test set and derive the final result.

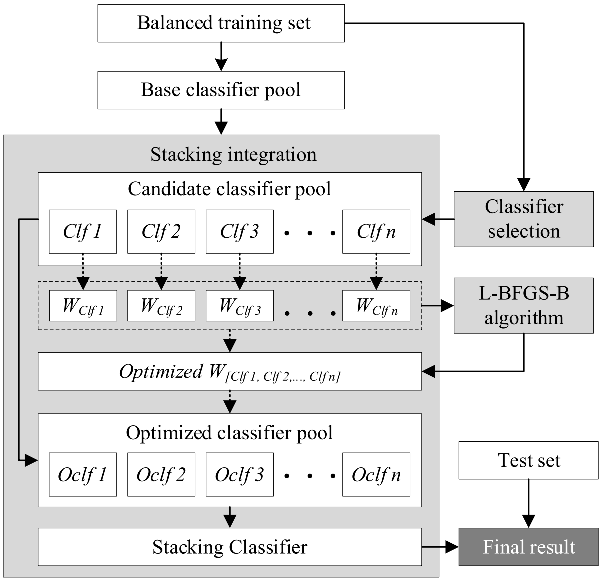

FIGURE 4. **Schematic of the voting-weight optimization method.**

IV. MATERIALS AND METHODS

A. DATASET DESCRIPTION AND DATA PREPROCESSING

This study utilizes the annual loan ledger data obtained from a commercial bank located in Jiangsu Province, China, which is called ChinaZJB and is presented in Table 1. The ChinaZJB dataset consists of 1,329 valid samples of SMEs after merging the non-financial behavioral information and soft information on credit rating with the financial information, loan information, and non-financial basic information found in the annual loan ledger data. Among them, 108 SMEs have default records, while 1,221 SMEs have no default records, resulting in an imbalanced ratio of approximately 1:11.

To check the robustness of the proposed model, five datasets from the UC Irvine (UCI) machine-learning repository, that is, the Polish 1, Polish 2, Polish 3 [35], Australian, and Taiwan credit datasets [36], were used for robustness checks in this study. The details of these datasets are presented in Table 2.

In this study, basic data preprocessing approaches, including standardization and normalization, were applied to process the raw datasets. Binary variables are represented as 0-1 variables for data processing. For categorical variables with multiple categories, one-hot encoding is utilized to convert each category into a binary feature. This approach ensures that the distances between categories are equal, thereby enhancing the characterization of variable relationships. The dataset contains a few missing values, which are addressed through a combination of case-by-case verification and mean imputation. To test for multicollinearity, the variance inflation factor (VIF) method is employed. The VIF serves as an index for assessing the issue of multicollinearity within a regression model. Multicollinearity refers to a high degree of linear association among the independent variables of the model, which may lead to imprecise estimation of regression coefficients. Consequently, the standard errors may be underestimated, thereby affecting the statistical inference of the model. For the i th independent variable X_i, the calculation formula for the VIF is given by Equation (1).

$$VIF_i = \frac{1}{1 - R_i^2}. \tag{1}$$

Here, R_i^2 represents the coefficient of determination (R-squared) obtained from a linear regression model that includes X_i as the dependent variable and all other independent variables as predictors. In essence, a regression of X_i on the other independent variables yields R_i^2, which is then utilized in the aforementioned formula to compute the VIF. It indicates that the VIF values, which are less than 5, demonstrate weak multicollinearity among the enhanced features.

B. EVALUATION METRICS

In this study, five evaluation metrics were used: accuracy (ACC) [37], area under the curve (AUC), logistic loss (Loss) [38], Brier score [39], and Kolmogorov-Smirnov (KS) [40]. These metrics were determined based on the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Predictive accuracy is defined by Equation (2).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

The AUC is a widely used metric in binary classification tasks. It represents the area under the ROC curve and the coordinate axis. The value for this area was less than one. A higher AUC indicates a better classification ability compared to a lower AUC value.

Logistic loss is a measure of loss for the classification model. As is depicted in Equation (3), i represents the sequence number of the predicted sample and $i \in \{1, 2, \ldots, n\}$, n represents the number of samples, y_i and p_i represent the real value and probability prediction, respectively, and $y_i \in \{0, 1\}$. The performance of the model is better when the log loss is lower.

$$L_{logistic} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \times Log\,(p_i) + (1 - y_i) \times Log\left(1 - p_i\right)) \tag{3}$$

The Brier score quantifies the mean squared difference between the predicted probability and the actual label. It can be considered as a loss function. A lower Brier score indicates better model performance.

The KS rate is used to measure the ability of a binary classification model to classify positive and negative samples. A higher KS value corresponds to better model performance.

TABLE 1. Data description.

Feature category	Name	Explanation	Numerical type
Default status	Classification label	1= an overdue loan, and 0= a normal loan	Category variable
Loan information	Number of loans	Current number of loans	Continuous variable
	Total borrowing	Total loan amount	Continuous variable
	Repayment rate	For the single loan is the annualized interest rate of the loan, and the average annualized interest rate of each loan is calculated for the multiple loan	Continuous variable
	Mortgage	1= mortgage, 0= other	Category variable
	Acceptance bill	1= acceptance bill loan, 0= no acceptance bill	Category variable
	Invest in category	Refers to the industry classification in which the borrower invests	Category variable
	Autonomous payment	1= autonomous payment, and 0= entrusted payment	Category variable
	Business purpose	1= the business use, and 0= the use of fixed assets	Category variable
Financial information	Total assets	Current total assets	Continuous variable
	Operating income	Operating income this year	Continuous variable
	Registered capital	Registered capital	Continuous variable
	Debt-to-income ratio	Operating income divided by total borrowing	Continuous variable
Non-financial basic information	Business duration	The time from the establishment time to the loan time	Continuous variable
	Industry category	Refers to the industry classification in which the borrower is located	Category variable
	Business type	Refers to the organizational form to which the borrower belongs	Category variable
	Holding subject	1= state-owned or collective holding, and 0= private holding	Category variable
	Nature of holding	1= absolute holding, and 0= non-absolute holding	Category variable
	Whether it is an urban SME	1= urban SME, and 0= rural SME	Category variable
	Number of employees	Total number of current employees	Continuous variable
	Credit rating	Credit rating of SMEs (AAA, AA, A, BBB, BB, B, C)	Category variable
Non-financial behavioral information	Number of plaintiff cases	The number of times the SME is the defendant in legal proceedings	Continuous variable
	Number of accused cases	The number of times the SME is the plaintiff in legal proceedings	Continuous variable
Soft information on credit rating	Account manager's evaluation of customer credit	The expert's credit evaluations of the account manager of the borrowing SME, including excellent, good, general, poor, deterioration, and shutdown	Category variable

TABLE 2. Description of the datasets.

Dataset	Sample size	Positive samples	Negative samples	Dimension of the input features
Polish 1	7027	271	6756	65
Polish 2	10173	400	9773	65
Polish 3	10503	495	10008	65
Australian	690	307	383	15
Taiwan	30000	6636	23364	24

C. ALGORITHMS

In this study, seven base classifiers (XGBoost, GBDT, Adaboost, RF, LightGBM, Bagging, and ExtraTree) were evaluated, and five evaluation metrics (ACC, AUC, Loss, Brier score, and KS) were adopted to evaluate the performances of the base classifiers and ensemble models. The raw datasets were divided into a training set and a test set using five-fold cross-validation [41], repeated 50 times to calculate the mean performance. Five-fold cross-validation is a technique for evaluating a model's generalization capability, dividing the dataset into five parts, using four for training and one for validation in each iteration. Unlike setting aside a separate test set, cross-validation ensures all data is used for both training and validation, which is especially important when data is limited. By training and validating the model on multiple subsets of data, five-fold cross-validation helps detect if the model is overfitting to specific data subsets.

V. EXPERIMENT

A. EXPERIMENT SETUP

Data preprocessing approaches and the base classifiers GBDT, Adaboost, RF, Bagging, and ExtraTree were implemented using the Python module "sklearn." The SMOTE algorithm was imported from the Python module "imblearn." The base classifier XGBoost was imported from the Python module "xgboost," and the base classifier LightGBM was imported from the Python module "lightgbm." The L-BFGS-B algorithm was imported from the Python module "SciPy." The approaches and algorithms mentioned above are imported with default parameters. All experiments were executed on a laptop running Python Version 3.9 with a 12th Gen Intel Core i7-12700H processor, 16 GB of RAM, and the Microsoft Windows 11 operating system. The software and hardware requirements are presented in Table 3.

B. BASELINE RESULTS

The performance of the proposed model was verified by evaluating the baseline results using five metrics. As listed in Table 4, seven base classifiers were applied.

C. PERFORMANCE EVALUATION OF THE MULTI-FEATURE ENHANCEMENT METHOD

The multi-feature enhancement method integrates various external data sources, including non-financial behavioral information and soft information on credit rating, into the ChinaZJB dataset. This did not apply to the other five UCI datasets (Polish 1, Polish 2, Polish 3, Australian, and Taiwan). So the effectiveness of the multi-feature enhancement method was evaluated only on the ChinaZJB dataset using five metrics, as outlined in Table 5. The values of the evaluation metrics are highlighted in bold if the base classifiers performed the same or better after applying the multi-feature enhancement method. This observation

TABLE 3. Software and hardware requirements.

Category	Description	Version/Parameters
Software	Data preprocessing approaches Base classifiers: GBDT, Adaboost, RF, Bagging, ExtraTree	Python module "sklearn"
	SMOTE algorithm	Python module "imblearn"
	Base classifier XGBoost	Python module "xgboost"
	Base classifier LightGBM	Python module "lightgbm"
	L-BFGS-B algorithm	Python module "SciPy"
Hardware	Operating System	Microsoft Windows 11
	Processor	12th Gen Intel Core i7-12700H
	RAM	16 GB
Python Version	Python Version	3.9

TABLE 4. Baseline results.

Datasets	Base classifiers	ACC	AUC	Loss	Brier	KS
ChinaZJB	XGBoost	0.92481	0.79001	0.29797	0.06267	0.27830
	GBDT	0.92481	0.80288	0.25920	0.06751	0.22667
	AdaBoost	0.91353	0.79924	0.62357	0.21535	0.24241
	RF	0.91729	0.79079	0.25905	0.06962	0.29635
	LightGBM	0.92105	0.80211	0.40636	0.07399	0.22667
	Bagging	0.90977	0.68387	1.57143	0.08481	0.22667
	ExtraTree	0.90977	0.75391	0.39275	0.07702	0.24241
Polish 1	XGBoost	0.89502	0.92326	0.23824	0.07397	0.29004
	GBDT	0.88094	0.92908	0.28637	0.08575	0.30887
	AdaBoost	0.89787	0.92788	0.63944	0.22321	0.28850
	RF	0.82703	0.86422	0.43177	0.13531	0.23992
	LightGBM	0.90014	0.92904	0.22365	0.07008	0.27370
	Bagging	0.85462	0.90724	0.36822	0.11284	0.31259
	ExtraTree	0.83812	0.79480	0.43279	0.13364	0.37615
Polish 2	XGBoost	0.88737	0.89265	0.26464	0.08120	0.23390
	GBDT	0.84963	0.87745	0.36510	0.11116	0.24071
	AdaBoost	0.85779	0.87325	0.65906	0.23298	0.21442
	RF	0.79410	0.80560	0.46598	0.14897	0.18503
	LightGBM	0.88501	0.89572	0.25886	0.08017	0.23425
	Bagging	0.82801	0.84940	0.42239	0.13259	0.22252
	ExtraTree	0.83400	0.75393	0.45094	0.14041	0.29510
Polish 3	XGBoost	0.88425	0.91213	0.28139	0.08662	0.28153
	GBDT	0.83760	0.85414	0.36727	0.11372	0.32517
	AdaBoost	0.85654	0.90324	0.65742	0.23216	0.30136
	RF	0.77125	0.83301	0.45588	0.14824	0.22574
	LightGBM	0.87949	0.90453	0.27396	0.08507	0.28503
	Bagging	0.80152	0.83433	0.42577	0.13743	0.26160
	ExtraTree	0.80371	0.72363	0.43726	0.13819	0.29130
Australian	XGBoost	0.85507	0.93855	0.37538	0.10669	0.38760
	GBDT	0.85217	0.94363	0.30563	0.09746	0.38611
	AdaBoost	0.83188	0.91705	0.63736	0.22222	0.35409
	RF	0.87391	0.94626	0.31073	0.09312	0.38695
	LightGBM	0.84928	0.93572	0.36844	0.10943	0.36749
	Bagging	0.85942	0.93229	0.71701	0.09972	0.37195
	ExtraTree	0.85652	0.92567	0.34236	0.10387	0.38750
Taiwan	XGBoost	0.77033	0.77315	0.51439	0.16740	0.07966
	GBDT	0.76583	0.77807	0.54230	0.17829	0.11518
	AdaBoost	0.76750	0.77381	0.65347	0.23024	0.11855
	RF	0.78433	0.76681	0.51215	0.16694	0.06171
	LightGBM	0.76967	0.77593	0.52693	0.17259	0.03953
	Bagging	0.77600	0.76743	0.52223	0.17035	0.10621
	ExtraTree	0.79050	0.76628	0.49711	0.16039	0.06881

indicates an improvement in the performance of most evaluation metrics. In particular, the KS indicator has been greatly improved, which shows that the multi-feature enhancement method would help the proposed model perform better in dealing with data in real business scenarios, providing evidence for the effectiveness of the integrated various external data sources.

D. PERFORMANCE EVALUATION OF THE BAGGING-BASED OVERSAMPLING METHOD

The effectiveness of the proposed bagging-based oversampling method is outlined in Table 6. In Table 6, the performance evaluation of the ChinaZJB dataset was compared to Table 5, and the performance evaluation of five UCI datasets (Polish 1, Polish 2, Polish 3, Australian, and Taiwan) was

TABLE 5. Performance evaluation of the multi-feature enhancement method.

Dataset	Base classifiers	ACC	AUC	Loss	Brier	KS
	XGBoost	0.92105	**0.80713**	**0.28466**	**0.06193**	0.25983
	GBDT	**0.93358**	0.79607	**0.21606**	**0.05647**	**0.32687**
	AdaBoost	**0.92233**	**0.81473**	**0.62109**	**0.21410**	**0.32891**
ChinaZJB	RF	**0.93609**	**0.80149**	**0.24839**	**0.05556**	**0.34906**
	LightGBM	**0.93233**	**0.81484**	**0.30618**	**0.05843**	**0.34669**
	Bagging	**0.92231**	**0.71097**	**1.09258**	**0.07036**	**0.34669**
	ExtraTree	**0.93484**	**0.80607**	**0.27794**	**0.05284**	**0.30876**

Note: significant values were boldfaced.

TABLE 6. Performance evaluation of the bagging-based oversampling method.

Datasets	Base classifiers	ACC	AUC	Loss	Brier	KS
	XGBoost	**0.92732**	**0.84199**	**0.25158**	**0.05853**	**0.35791**
	GBDT	0.92857	**0.81999**	**0.21292**	0.06081	**0.32933**
	AdaBoost	**0.92353**	**0.82838**	**0.24337**	**0.06893**	**0.35199**
ChinaZJB	RF	0.92331	**0.81103**	**0.23118**	**0.05236**	0.30724
	LightGBM	0.92982	**0.82759**	**0.23826**	**0.05613**	**0.35193**
	Bagging	**0.92880**	**0.72473**	**1.08962**	0.07364	**0.37895**
	ExtraTree	**0.93857**	**0.81692**	**0.22603**	0.05936	**0.33438**
	XGBoost	**0.92639**	**0.94234**	**0.08040**	**0.01949**	**0.37995**
	GBDT	**0.92851**	0.92439	**0.20644**	**0.05198**	0.15825
	AdaBoost	**0.91511**	**0.93248**	**0.61460**	**0.21085**	0.16419
Polish 1	RF	**0.89145**	**0.88898**	**0.17241**	**0.04057**	**0.33333**
	LightGBM	**0.92482**	**0.94493**	**0.08116**	**0.02082**	**0.39267**
	Bagging	**0.88624**	0.84605	0.37833	**0.02532**	0.28848
	ExtraTree	**0.86690**	0.75123	**0.23674**	**0.04920**	**0.38136**
	XGBoost	**0.91248**	**0.90426**	**0.18082**	**0.02340**	0.15854
	GBDT	**0.92541**	0.84458	**0.27817**	**0.07473**	**0.28749**
	AdaBoost	**0.90835**	**0.87680**	**0.62883**	**0.21791**	**0.26823**
Polish 2	RF	**0.87705**	**0.82333**	**0.19708**	**0.04651**	**0.26713**
	LightGBM	**0.92002**	0.89543	**0.16160**	**0.02707**	0.16147
	Bagging	**0.86147**	0.79670	0.45141	**0.03225**	**0.28035**
	ExtraTree	**0.88430**	0.70938	**0.28118**	**0.05380**	**0.30683**
	XGBoost	**0.92411**	**0.91434**	**0.16389**	**0.02939**	0.15271
	GBDT	**0.88729**	**0.90542**	**0.30705**	**0.08911**	0.15750
	AdaBoost	**0.90249**	0.89900	**0.63103**	**0.21901**	0.22321
Polish 3	RF	**0.87755**	0.82143	**0.20909**	**0.05351**	**0.28914**
	LightGBM	**0.93859**	**0.91259**	**0.18210**	**0.03447**	0.19997
	Bagging	**0.85355**	**0.85510**	**0.40203**	**0.03599**	0.25542
	ExtraTree	**0.84679**	**0.78190**	**0.27707**	**0.04657**	**0.38992**
	XGBoost	**0.85652**	**0.94730**	**0.31335**	**0.09793**	**0.39410**
	GBDT	0.84638	**0.94497**	**0.30491**	0.09753	**0.39710**
	AdaBoost	**0.84203**	**0.93352**	**0.63508**	**0.22107**	**0.36843**
Australian	RF	0.86957	**0.95023**	**0.30911**	**0.09194**	**0.40496**
	LightGBM	**0.85362**	**0.94424**	**0.33022**	**0.10262**	**0.38093**
	Bagging	0.85217	**0.94275**	**0.40246**	**0.09709**	**0.38609**
	ExtraTree	**0.86377**	**0.93640**	**0.32523**	**0.09878**	**0.39578**
	XGBoost	**0.81633**	0.76573	**0.44630**	**0.14096**	**0.12191**
	GBDT	**0.79283**	**0.78085**	**0.50230**	**0.16204**	**0.16043**
	AdaBoost	0.75300	**0.77598**	0.65490	0.23094	**0.17053**
Taiwan	RF	**0.80683**	**0.76913**	**0.46652**	**0.14856**	**0.11892**
	LightGBM	**0.80233**	**0.78170**	**0.46223**	**0.14728**	**0.14809**
	Bagging	**0.81300**	0.72279	**0.45882**	**0.14135**	**0.12079**
	ExtraTree	**0.80883**	0.75934	**0.45957**	**0.14593**	**0.08115**

Note: significant values were boldfaced.

compared to Table 4. The values of the evaluation metrics are highlighted in bold if the base classifiers achieved equal or improved performance after applying the bagging-based oversampling method. This observation indicates an overall improvement in the performance of most evaluation metrics on six datasets, which indicates that the oversampled samples generated by the proposed bagging-based oversampling method help increase the sample information in the data and improve the training effect of the classifier.

E. PERFORMANCE EVALUATION OF THE VOTING-WEIGHT OPTIMIZATION METHOD

After multi-feature enhancement and bagging-based oversampling, base classifiers were selected to form a candidate

TABLE 7. Composition of the best-performing classifier ensemble corresponding to the six datasets.

Dataset	The proposed model	XGBoost	GBDT	AdaBoost	RF	LightGBM	Bagging	ExtraTree
ChinaZJB	*Eclf* 1	●		●		●		
Polish 1	*Eclf* 2	●		●		●		
Polish 2	*Eclf* 3	●		●		●		
Polish 3	*Eclf* 4	●	●			●		
Australian	*Eclf* 5	●	●		●			
Taiwan	*Eclf* 6		●	●		●		

TABLE 8. Final performance of the proposed model.

Dataset	The proposed model	ACC	AUC	Loss	Brier	KS
ChinaZJB	*Eclf* 1	**0.97995**	**0.97966**	**0.07828**	**0.01833**	**0.38300**
Polish 1	*Eclf* 2	**0.94317**	**0.95890**	0.18188	**0.01585**	**0.40187**
Polish 2	*Eclf* 3	**0.94017**	**0.92293**	**0.13205**	0.03744	**0.32185**
Polish 3	*Eclf* 4	**0.94067**	**0.93464**	**0.11469**	**0.01998**	0.33160
Australian	*Eclf* 5	**0.88101**	**0.96278**	**0.22078**	**0.06934**	**0.43428**
Taiwan	*Eclf* 6	**0.83350**	**0.79076**	**0.42609**	**0.13261**	0.10396

classifier pool. Subsequently, the soft voting weights of the candidate classifiers were optimized using the L-BFGS-B algorithm and then used for classifier permutations and combinations to form classifier ensembles. Classifier ensembles that performed better on the AUC indicator for each dataset were selected as the best-performing classifier ensembles. As the performance of Bagging and ExtraTree are relatively weak in the experiment, they are not contained in the best-performing classifier ensemble. As listed in Table 7, the best-performing classifier ensembles corresponding to the six datasets are *Eclf* 1, *Eclf* 2, *Eclf* 3, *Eclf* 4, *Eclf* 5, and *Eclf* 6, respectively. For example, *Eclf* 1 is an ensemble of XGBoost, AdaBoost, and LightGBM.

The final performance of the proposed ensemble model is presented in Table 8. In Table 8, the values of the evaluation metrics are highlighted in bold if the proposed ensemble model outperforms or achieves equal performance compared to the base classifiers after both the multi-feature enhancement and bagging-based oversampling methods. The evaluation indicators show that the performance of the proposed ensemble model has been improved to varying degrees in most indicators after the voting-weight optimization method is used, indicating that the modeling idea of the voting-weight optimization method, i.e., the optimization of soft voting decision weights in ensemble models aims to diminish the impact of poorly performing base classifiers, thus preventing a significant decline in overall model performance is effective.

F. PERFORMANCE EVALUATION OF THE SHAP EXPLANATION METHOD

An essential challenge encountered in the practical application of machine learning models in business domains is the lack of transparency compared to linear regression [42]. This lack of transparency makes it difficult for operators to understand the crucial features at play, leading to a potential lack of trustworthiness in the process, despite its reliable results. To address this issue in model interpretation, the study introduces the SHAP explanation method to explore the ChinaZJB dataset.

The core idea of SHAP explanations is to compute the Shapley values of each feature over all possible combinations of feature values. These values can be used to explain individual prediction results or the overall behavior of the model on a dataset. For a feature j, its SHAP value ϕ_j can be calculated using the Equation (4):

$$\phi_j = \sum_{S\subseteq\{1,\ldots,M\}\backslash\{j\}} \frac{|S|!(M-|S|-1)!}{M!} \times [f(x_S \cup \backslash\{j\}) - f(x_S)] \tag{4}$$

where M is the total number of features, S is a subset of features not containing feature j, x_S represents the feature values for the subset S, $f(x_S \cup j)$ is the model output for the feature set including j, $f(x_S)$ is the model output for the feature set not including j.

Figure 5 is the feature importance bar graph to illustrate the global significance of the top 20 features in the credit risk assessment of SMEs. This significance is based on the average of absolute SHAP values per feature across the dataset. The features are ranked in descending order, and a red color indicates a positive correlation with failure prediction, while a blue color indicates a negative correlation. Figure 6 is the beeswarm plot to display an information-dense summary of how the top 20 features impact the model's output. The features are ranked in descending order. On the x-axis, the signs of the effect on business failure prediction are depicted, and the color of the dots reflects the magnitude of their effect. Blue indicates a low effect, while red indicates a high effect.

Apart from loan information and financial data, which have consistently demonstrated their significance, it has been discovered that non-financial basic information plays a pivotal role in assessing the likelihood of default among SMEs. Among these features of the ChinaZJB dataset, "business duration" ranks first, surpassing other features in identifying defaults. The longer the lifespan of an SME, the more experienced its managers become, enabling them to better resist external risks and increasing the likelihood of stable operations. Consequently, this strengthens the SME's ability to repay debts. Non-financial behavioral information also serves as valuable explanatory features. Ranking third, the "number of accused cases" reflects whether an SME has violated laws or regulations. This highlights the substantial impact of an SME's adherence to legal and compliant business behavior during the loan period on loan quality. Additionally, this feature also provides insight into an SME's credit risk management and operational capabilities.

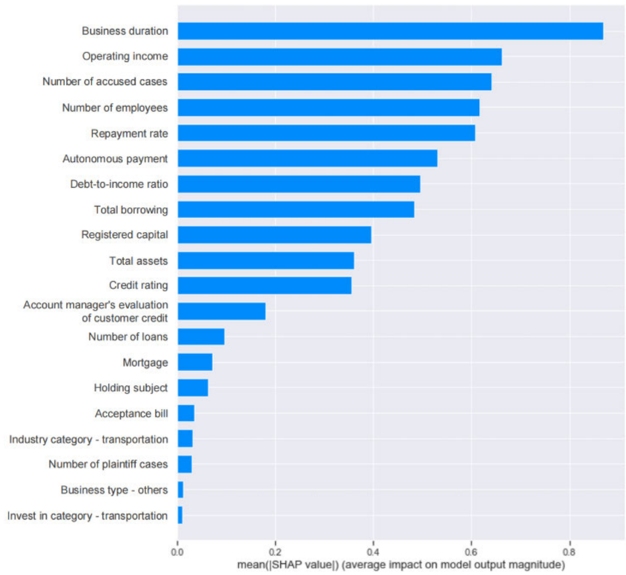

FIGURE 5. **Feature importance bar graph for global interpretability of the ChinaZJB dataset.**

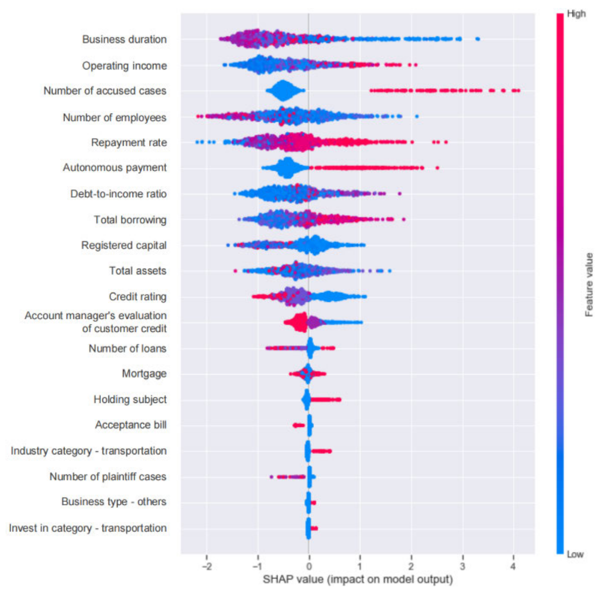

FIGURE 6. **Beeswarm plot for global interpretability of the ChinaZJB dataset.**

The "debt-to-income ratio" derived from credit scoring, has also proven to be an effective feature. This feature represents the extent to which the SME's "operating income" covers its "total borrowing" with higher ratios indicating lower default risks. The key to credit risk assessment is effectively measuring the SME's capacity to repay, willingness to repay, and total borrowing (typically consisting of loans from various financial institutions). An SME's funds and loans are different in terms of capacity to repay and willingness to repay. Therefore, "total borrowing" is important because it revises estimates of SMEs' capacity and willingness to repay so that the "debt-to-income ratio" is important as well.

The integration of the SHAP explanation method within the context of this study offers a transformative perspective for financial institutions and practitioners engaged in credit risk assessment for SMEs. By demystifying the complexities of machine learning models, this approach provides financial institutions and practitioners with a transparent and actionable framework to dissect the contributions of various features impacting creditworthiness. The practical upshot of this transparency is a more nuanced understanding of the risk profiles of SME borrowers, which can lead to the refinement of lending policies that are not only robust but also responsive to the multifaceted nature of business operations.

The identification of "business duration" as a pivotal feature points to the value of longevity and experience in business management, suggesting that longer-established SMEs may present lower credit risks due to their managers' enhanced ability to navigate external challenges. This insight could prompt financial institutions to reassess their lending criteria, potentially extending favorable terms to businesses with proven track records of stability.

Moreover, the significance of the "number of accused cases" as a feature underscores the imperative of legal compliance in financial health. For practitioners, this indicates that SMEs with a cleaner legal slate may be more reliable borrowers, influencing the development of risk assessment models that factor in legal and regulatory adherence.

The study's findings also highlight the "debt-to-income ratio" as a critical quantitative indicator, providing a straightforward yet effective measure for financial institutions to gauge an SME's capacity to service debt. This metric can be instrumental in formulating lending decisions that are underpinned by a clear assessment of the borrower's financial capacity.

Lastly, the study's exploration of "soft information" challenges the traditional dichotomy between relationship and transaction lending, demonstrating that qualitative assessments such as credit rating perceptions hold substantial sway even in the transactional sphere. This realization can encourage financial institutions to broaden their evaluative scope to include non-traditional data sources, thereby enriching their risk management strategies.

VI. CONCLUSION AND FUTURE WORK

The imperative to reconcile the precision of predictions with the clarity of explanations constitutes an essential research frontier, pivotal for instilling trust, transparency, and accountability in the deployment of ensemble models within practical domains. In this study, we introduce a new multi-stage ensemble model designed to amplify the interpretability of features, thereby addressing the quintessential challenge of balancing accuracy with comprehensibility. The efficacy of our model is rigorously appraised through a quintet of evaluative metrics: accuracy (ACC), area under the curve (AUC), loss, Brier score, and Kolmogorov-Smirnov (KS) statistic. The empirical evidence garnered from our experiments substantiates the model's superior performance, underscoring its efficacy.

Harnessing annual loan ledger data from a commercial bank, we engage the SHAP explanation method to distill and elucidate the significance of pivotal features, unveiling the mechanisms at play. This approach not only underscores the SHAP method's preeminence in bestowing interpretability upon machine learning models but also accentuates the indispensable role of non-traditional data in risk assessment. In particular, we demonstrate that non-financial attributes, such as legal compliance records, the longevity of enterprises, and the seasoned insights of account managers, are paramount in identifying SME default risk.

This study transcends the conventional boundaries by offering financial institutions a novel lens through which to assess borrowers' credit risks and repayment capacities in credit loans. It equips banks with the means to achieve precise outcomes using contemporary machine learning models and adeptly navigates the potential pitfalls of over-reliance on data-driven algorithms. The insights and methodologies proffered by this study are poised to invigorate the financial sector's approach to credit risk evaluation, fostering a more discerning and nuanced appraisal of borrower profiles.

Despite the innovative approach presented in this study through a multi-stage ensemble model for assessing SME loan risk, several limitations remain. The model's effectiveness depends on the availability and quality of external data sources, such as external legal risk features and expert credit evaluations. Variability in these data sources could adversely impact the model's performance and generalizability. Although the SHAP explanation method offers valuable insights into feature importance, it relies on the completeness and relevance of the selected features. Misrepresentation or omission of key variables could lead to biased interpretations of feature significance.

Future research could expand upon these findings in several ways. For example, conducting rigorous external validation across diverse SME loan datasets from different geographic and economic contexts is essential to enhance the generalizability of the proposed model and to ensure its robustness beyond specific datasets. This effort might involve integrating real-time data and continuously updating the model to maintain accuracy and relevance. Moreover, exploring advanced interpretability techniques beyond SHAP, such as model-agnostic methods or visualization tools, could provide deeper insights into decision-making processes, thereby improving stakeholder and regulatory trust and facilitating broader adoption of the model in practical applications.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY AND ACCESS

The raw datasets for this study called ChinaZJB and the UCI datasets Polish 1, Polish 2, Polish 3, Australian, and Taiwan credit datasets were uploaded to IEEE DataPort (https://dx.doi.org/10.21227/n048-7q52).

REFERENCES

[1] M. Stevenson, C. Mues, and C. Bravo, "The value of text for small business default prediction: A deep learning approach," *Eur. J. Oper. Res.*, vol. 295, no. 2, pp. 758–771, Dec. 2021.

[2] R. Levine, C. Lin, Q. Peng, and W. Xie, "Communication within banking organizations and small business lending," *Rev. Financial Stud.*, vol. 33, no. 12, pp. 5750–5783, Dec. 2020.

[3] R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu, "Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending," *Appl. Econ.*, vol. 47, no. 1, pp. 54–70, Jan. 2015.

[4] G. Yao, X. Hu, and G. Wang, "A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain," *Expert Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 117002.

[5] S. J. Yen and Y. S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation* (Lecture Notes in Control and Information Sciences), vol. 344, D. S. Huang, K. Li, and G. W. Irwin, Eds., Berlin, Germany: Springer, 2006, pp. 731–740.

[6] M. Papouskova and P. Hajek, "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," *Decis. Support Syst.*, vol. 118, pp. 33–45, Mar. 2019.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[9] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn.*, Bari, Italy, 1996, pp. 148–156.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[11] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[12] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, 2017, pp. 3146–3154.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[15] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.

[16] R. I. Hamilton and P. N. Papadopoulos, "Using SHAP values and machine learning to understand trends in the transient stability limit," *IEEE Trans. Power Syst.*, vol. 39, no. 1, pp. 1384–1397, Feb. 2023.

[17] R. Kitani and S. Iwata, "Verification of interpretability of phase-resolved partial discharge using a CNN with SHAP," *IEEE Access*, vol. 11, pp. 4752–4762, 2023.

[18] L. Roa, A. Correa-Bahnsen, G. Suarez, F. Cortés-Tejada, M. A. Luque, and C. Bravo, "Super-app behavioral patterns in credit risk models: Financial, statistical and regulatory implications," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114486.

[19] M. Óskarsdóttir, C. Bravo, C. Sarraute, J. Vanthienen, and B. Baesens, "The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics," *Appl. Soft Comput.*, vol. 74, pp. 26–39, Jan. 2019.

[20] N. Kozodoi, S. Lessmann, K. Papakonstantinou, Y. Gatsoulis, and B. Baesens, "A multi-objective approach for profit-driven feature selection in credit scoring," *Decis. Support Syst.*, vol. 120, pp. 106–117, May 2019.

[21] C. Yin, C. Jiang, H. K. Jain, and Z. Wang, "Evaluating the credit risk of SMEs using legal judgments," *Decis. Support Syst.*, vol. 136, Sep. 2020, Art. no. 113364.

[22] J. H. Dahooie, S. H. R. Hajiagha, S. Farazmehr, E. K. Zavadskas, and J. Antucheviciene, "A novel dynamic credit risk evaluation method using data envelopment analysis with common weights and combination of multi-attribute decision-making methods," *Comput. Oper. Res.*, vol. 129, May 2021, Art. no. 105223.

[23] W. Zhang, D. Yang, and S. Zhang, "A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring," *Expert Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114744.

[24] W. Zhang, F. He, and S. Zhang, "A novel fairness-aware ensemble model based on hybrid sampling and modified two-layer stacking for fair classification," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 11, pp. 3883–3896, Nov. 2023.

[25] W. Zhang, D. Yang, S. Zhang, J. H. Ablanedo-Rosas, X. Wu, and Y. Lou, "A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113872.

[26] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.

[27] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 4768–4777.

[28] D. Barber, W. O. Peake, and M. L. Harris, "Can playing defense yield gains? Examining the relationships among regulatory focus, innovation, and SME performance," *J. Small Bus. Manage.*, vol. 62, no. 3, pp. 1469–1497, May 2024.

[29] F. Guo, R. Krishnan, and J. Polak, "The influence of alternative data smoothing prediction techniques on the performance of a two-stage short-term urban travel time prediction framework," *J. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 214–226, May 2017.

[30] V. B. Djeundje, J. Crook, R. Calabrese, and M. Hamid, "Enhancing credit scoring with alternative data," *Expert Syst. Appl.*, vol. 163, Jan. 2021, Art. no. 113766.

[31] S. Agarwal, V. Mikhed, and B. Scholnick, "Peers' income and financial distress: Evidence from lottery winners and neighboring bankruptcies," *Rev. Financial Stud.*, vol. 33, no. 1, pp. 433–472, 2020.

[32] Y. Xu, A. Saunders, B. Xiao, and X. Li, "Bank relationship loss: The moderating effect of information opacity," *J. Banking Finance*, vol. 118, Sep. 2020, Art. no. 105872.

[33] F. Sigrist and C. Hirnschall, "Grabit: Gradient tree-boosted Tobit models for default prediction," *J. Banking Finance*, vol. 102, pp. 177–192, May 2019.

[34] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[35] M. Zięba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, Oct. 2016.

[36] A. Asuncion and D. Newman, *UCI Machine Learning Repository*. Irvine, CA, USA: University of California, 2007.

[37] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sens. Environ.*, vol. 62, no. 1, pp. 77–89, Oct. 1997.

[38] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*, 2nd ed., London, U.K.: Springer, 2009.

[39] H. M. Fletcher, "The structure of bone," in *Fundamentals of Osteology*, 3rd ed., New York, NY, USA: Springer, 1998, pp. 12–35.

[40] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[41] M. A. Newton and A. E. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 56, no. 1, pp. 3–26, Jan. 1994.

[42] S. Wei, D. Yang, W. Zhang, and S. Zhang, "A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning," *IEEE Access*, vol. 7, pp. 99217–99230, 2019.

DONGQI YANG received the B.S. and M.S. degrees from Zhejiang University of Finance and Economics, China. He is currently pursuing the Ph.D. degree with the School of Management and Engineering, Nanjing University, China. His research interests include small and micro enterprise credit rating and machine learning.

BINQING XIAO received the B.S., M.S., and Ph.D. degrees from Nanjing University, China. He is currently a full-time Professor with the School of Management and Engineering, Nanjing University. In the past five years, he has published over ten articles in international journals, including *Production and Operations Management*, *Information Systems Research*, and *Journal of Banking and Finance*. His research interests include financial risk management (credit risk management, small and micro enterprise credit rating, and operational risk management), microfinance and rural finance, and financial service operations.