Research Article

Personal Credit Default Discrimination Model Based on Super Learner Ensemble

Gang Li ^{1,2,3} Mengdi Shen ^{1,2} Meixuan Li ^{1,2} and Jingyi Cheng ^{1,2}

¹School of Business Administration, Northeastern University, Shenyang, Liaoning 110819, China

²School of Economics, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei 066004, China

³Institutes of science and Development, Chinese Academy of Science, Beijing 100190, China

Correspondence should be addressed to Gang Li; ligang@neuq.edu.cn

Received 6 January 2021; Revised 6 March 2021; Accepted 20 March 2021; Published 31 March 2021

Academic Editor: Shigeyuki Hamori

Copyright © 2021 Gang Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Assessing the default of customers is an essential basis for personal credit issuance. This paper considers developing a personal credit default discrimination model based on Super Learner heterogeneous ensemble to improve the accuracy and robustness of default discrimination. First, we select six kinds of single classifiers such as logistic regression, SVM, and three kinds of homogeneous ensemble classifiers such as random forest to build a base classifier candidate library for Super Learner. Then, we use the ten-fold cross-validation method to exercise the base classifier to improve the base classifier's robustness. We compute the base classifier's total loss using the difference between the predicted and actual values and establish a base classifier-weighted optimization model to solve for the optimal weight of the base classifier, which minimizes the weighted total loss of all base classifiers. Thus, we obtain the heterogeneous ensembled Super Learner classifier. Finally, we use three real credit datasets in the UCI database regarding Australia, Japanese, and German and the large credit dataset GMSC published by Kaggle platform to test the ensembled Super Learner model's effectiveness. We also employ four commonly used evaluation indicators, the accuracy rate, type I error rate, type II error rate, and AUC. Compared with the base classifier's classification results and heterogeneous models such as Stacking and Bsteking, the results show that the ensembled Super Learner model has higher discrimination accuracy and robustness.

1. Introduction

The analysis of default is a useful credit risk assessment tool. Its core notion is to utilize observable borrowers' characteristic variables to build a classification model and predict new borrowers' default probability, for loan approval and quota for banks and other financial institutions. The determination of its interest rate provides a fundamental theoretical basis. Relevant research shows that even if the default discrimination model's prediction accuracy is increased by 1%, it will significantly reduce the possible losses due to bad debts and increase financial institutions' profitability [1].

More and more scholars have attached importance to assessing personal credit defaults, and a series of studies on default discrimination models have emerged. The

representative approaches primarily rely on statistical methods and artificial intelligence. The personal credit default discrimination model started with traditional statistical methods. Typical models include the Z-score model [2], the probit analysis method [3], and the logistic analysis model [4]. However, these models may require strict assumptions. Alternatively, it is difficult to deal with the increasingly large index systems, which have limitations in practical applications. The development of artificial intelligence has resulted in machine learning techniques that can handle high-dimensional data without strict assumptions, such as artificial neural networks (ANNs) [5–7], support vector machines (SVMs) [8], and decision tree (DT) [9]. Also, these single-classifier learning algorithms can better solve nonlinear problems and improve prediction accuracy. However, the single classifier has its disadvantages and limitations when

dealing with different credit risk assessment problems. To improve the single classifier's performance, the research of machine learning technology gradually develops into the ensemble model, which overcomes the disadvantages of the single classifier. The ensemble classifier presents better performance than a single classifier [10], so it has become a research hotspot of current personal credit scoring models. Compared with the single classification models, although the ensemble models are costly, the benefits of the personal credit scoring model's accuracy improvement make up for higher operating costs. Hence, the ensemble models are more suitable for today's financial institutions [11].

In formulating credit scoring models, single classifiers such as logistic regression, decision trees, and support vector machines often serve as the base classifiers for ensembled models [12, 13]. The central combination of base classifiers falls into three types: Bagging [14], Boosting [15], and Stacking [16], where Bagging and Boosting are the typical homogeneous ensemble, and Stacking is a heterogeneous ensemble. Previous studies found that ensembling multiple differentiated single classifiers can better solve the problem of overfitting single classifiers and obtain better prediction performance and generalization on most unbalanced datasets. The heterogeneous ensemble ability, composed of multiple different base classifiers, is increasingly becoming a research hotspot [17].

Although the Stacking heterogeneous ensemble algorithm has been an amazingly effective method in the default discrimination, it has received relatively little attention [18]. Besides, existing works rarely consider the selection of base classifiers and usually directly integrates all candidate base classifiers. Therefore, some classifiers with poor prediction performance reduce the performance of the final ensembled model. The negative impact leads to the development of selective ensemble. The selective ensemble chooses some base classifiers with excellent performance and gives different weights according to the classifier's performance. Lessmann et al. focused on the selective ensemble model and proved that the selective ensemble model performs relatively well [11]. However, the selective ensemble model has not received attention in the prior default discrimination research. When constructing a default discrimination model, no model is the best because the studied problem, the data structure, and the evaluation indicators complicate the model formulation. Therefore, if an ensembled model can select the base classifier independently according to different conditions, it can optimize and integrate, thus improving the model's classification performance and robustness.

In order to make up for the shortcomings of the above research, we consider using Super Learner, the evolutionary algorithm of the Stacking algorithm, to improve the classification performance of the model and the adaptive adjustment ability of different datasets. The Super Learner ensemble algorithm is a heterogeneous ensemble algorithm based on loss minimization, which was proposed by Van der Laan et al. and proved to be optimal [19].

Super Learner integrates a variety of commonly used classification models based on cross-validation theory. It solves the optimal weighting model of the base classifiers by

minimizing classification loss. The automatic selection of the classification model is realized to ensure the classification accuracy and robustness of the ensemble model.

The Super Learner ensembled algorithm has been successfully studied extensively in medicine and social sciences [20–22]. Although Super Learner has proven to perform well in many environments, its performance of evaluating personal credit default discrimination requires further study. This paper applies the Super Learner to the personal credit default discrimination and utilizes its data adaptability to better deal with unbalanced credit datasets. The classifiers commonly used in default discrimination models are combined, including logistic regression and lasso regression, K-nearest neighbor, SVM, neural network, decision tree, random forest, GBDT, and XGBoost models, nine prediction methods to build a personal credit default discrimination model with better prediction performance robustness.

The rest of the paper is organized as follows. Section 2 reviews the application of the ensembled model in the field of credit default discrimination. Section 3 details the principle of the personal credit default discrimination model based on the Super Learner algorithm. Section 4 elaborates on the Super Learner-based heterogeneous ensemble model. Section 5 provides real cases to test the proposed model based on the Super Learner and experimental results analysis. Section 6 concludes.

2. Related Work

This section presents the related work of ensemble classification model construction.

2.1. Homogeneous Ensemble. Homogeneous ensemble methods, such as Bagging or Boosting, use a set of classifiers fitted by a classification algorithm. Although the same classification algorithm differs in the type of training datasets or input features used, making a series of classifiers finally fitted out has its characteristics. A single classifier is usually combined by majority voting or weighted voting to obtain the final classification [23]. Previously, the default discrimination research mostly established models in the form of isomorphic sets. Many practical experiences and theories have shown that the combination of models improves prediction accuracy. Paleologo et al. proposed sub-gagging ensemble for a high imbalance in credit dataset categories and missing data and conducted empirical research using samples from IBM Italian customers. The results show that using a decision tree as the most basic classifier sub-gagging ensemble improves classification performance while keeping the model simple and reasonably interpretable [24]. Yu et al. proposed an extreme value ensembled machine learning method based on a multilevel deep belief network (DBN). They conducted empirical research using the Japanese credit dataset from the UCI database, proving that the proposed method effectively improves classification accuracy [25]. Cano et al. applied RF algorithms to select features from different datasets and used them to classify the datasets. The

results revealed that the performance effect of selecting relevant variables by RF and classifying them based on them had advantages over SVM and ANN [26]. Sun et al. suggested a decision tree ensemble model for the category imbalance in enterprise credit assessment. This model combines DTE-SBD, a combined minority oversampling technology (SMOTE), and Bagging ensemble learning algorithm [27]. Hamori et al. analyzed default payment data in Taiwan and compared the prediction accuracy and classification ability of three ensemble-learning methods: Bagging, random forest, and Boosting, with those of various neural network methods using two different activation functions [28]. Plawiak et al. applied a new deep genetic cascade ensemble classifier (DGCEC) based on different support vector machine (SVM) classifiers to Statlog Australian data, determining the effectiveness of the model [29]. Alam et al. recommended a dynamic, ensemble learning algorithm for neural network ensemble design [30].

2.2. Heterogeneous Ensemble. Compared with the homogeneous ensemble method, the heterogeneous ensemble method combines a variety of different classification algorithms. The idea is that different classifiers manage the same problem in different ways, thereby complementing each other by increasing the base classifier's diversity and predictions. In addition to weighted or straightforward voting, independent ensemble learning can also use a more sophisticated method to incorporate a single classifier. The Stacking algorithm is a typical heterogeneous ensemble algorithm. Tsai et al. pointed out that heterogeneous ensemble is superior to a homogeneous combination in predicting performance [31]. Nascimento et al. believe that different classification algorithms have different representative deviations, making the base classifier have output diversity, making it easy to adapt to different datasets [32]. Li et al. designed a multiround ensemble learning model based on heterogeneous ensemble framework to predict the default risk for the uneven distribution of credit data samples in the P2P lending market, i.e., the scarcity of default sample data. The real credit data of the P2P loan market was tested and proved the model [33]. Guo et al. pointed out that credit ratings' accuracy affects financial institutions' risk control and profitability. To improve the prediction performance and adjust to different credit datasets adaptively, it introduced a technique based on the statistics and machine learning ensemble model of a multistage adaptive classifier. It validated its effectiveness using three actual datasets in UCI [34]. Papouskova et al. stated a two-stage consumer credit risk model based on heterogeneously ensemble learning, which in turn models default probability and default risk exposure, thereby modeling the overall credit risk of consumer loans based on expected losses [35]. Plawiak et al. offered a new deep genetic hierarchical network credit scoring model that integrates four primary learners of support vector machine, K-nearest neighbor, probabilistic neural network, and fuzzy system. They employed the German credit dataset in the UCI database and verified the model [36].

The previous research shows that, on the problem of default discrimination, heterogeneous ensemble models have better prediction performance and data adaptability than single classifiers and homogeneous ensemble models. In heterogeneous ensemble learning, the core means of improving prediction performance is to make the base classifier better and different, that is, to improve the accuracy and diversity of the base classifier as much as possible, thereby improving the performance of the ensemble model [37]. In existing research, however, multiple base classifiers are usually directly fitted for ensemble without considering the base classifier's selection, which makes the base learner with weak classification effects affect the ensemble model's performance. Therefore, if an ensemble model can integrate various excellent and different classifiers and independently select the appropriate classifier for the problem's data structure and optimize the ensemble, better classification performance and robustness can be obtained.

In response to the prior problems, this paper applies the ensemble Super Learner model that performs well in disease prediction to the field of personal credit default discrimination. The Super Learner algorithm is an evolution of the Stacking algorithm. Previous research results indicate that the Super Learner integrated model proposed by Van der Laan et al. can independently select the base classifier according to the data structure of the dataset and the performance of the classifier, so as to improve the classification performance and robustness of the model [38–44]. Besides, if none of the candidate models in the Super Learner's classifier algorithm library can achieve the prespecified accuracy, the performance of the Super Learner is at least as good as the best algorithm in the candidate algorithm library, or it gradually approaches the best algorithm. By including more prediction algorithms in the candidate estimation library, Super Learner will outperform any of its competitors progressively, showing better performance. This paper introduces Super Learner into the field of default discrimination, based on minimizing the classification loss, predicting the borrower's credit status, and improving the performance of personal credit default discrimination.

3. The Principle of the Personal Credit Default Discrimination Model Based on the Super Learner Algorithm

Super Learner is a heterogeneous ensemble algorithm based on loss minimization proposed by Van der Laan et al. and theoretically proves the prediction effectiveness of the Super Learner algorithm [19]. The algorithm uses ten-fold cross-validation to train a variety of differentiated base classifiers. With the goal of minimizing the weighted total loss of all base classifiers, an optimal weighted combination of base classifiers that minimizes cross-validation classification loss is obtained. Construct a Super Learner heterogeneous ensemble classifier to improve the prediction accuracy and robustness of the model. The basic steps of Super Learner are as follows.

First, build a candidate algorithm library containing multiple base classifiers. The algorithm library should include multiple algorithms commonly used to solve a problem. The algorithm in the algorithm library can be either a simple single classification model or a complex model with hyperparameter settings. As long as the algorithm can output a fitted prediction function based on the observation data, it is considered to be a classification algorithm.

Then, use the ten-fold cross-validation method to train multiple base classifiers in the candidate algorithm library to improve the robustness of the classification model, and calculate 10 sets of prediction results for each base classifier, and calculate each based on the deviation of the prediction results from the true value. For 10 sets of loss of a base classifier, the total loss of the base class is calculated according to the 10 sets of losses of each base classifier. The smaller the total loss of the base class, the better the classification performance. The ensemble process should give the base classifier a larger loss. In order to ensure the classification performance of the ensemble classifier, the greater the total loss, the worse the discrimination performance, the smaller the role of the classifier in the ensemble classifier, even the classifier is not considered in the ensemble process.

Finally, with the objective of minimizing the total weighted loss of all base classifiers, the optimal weighting model of base classifiers is established with the constraint that the weight of each base classifier is nonnegative and the sum of the weights is 1. Fit the selected base classifier in the complete dataset and combine the optimal weights to construct the Super Learner ensemble classifier.

The Super Learner ensemble algorithm has been extensively studied in the fields of medicine and social sciences [20–22]. Existing research usually uses Super Learner to integrate a variety of commonly used learning algorithms in this field to improve patient mortality prediction or predict personal behavior, etc. and has proved through empirical evidence that its prediction performance is good. In view of its excellent predictive performance, this study applies the Super Learner ensemble algorithm to the field of personal credit, thereby constructing a personal credit default discrimination model with better classification performance and robustness.

In the personal credit problem, assume that the data structure of the current credit dataset is $O_i = (X_i, Y_i) \sim P_0$, $i = 1, 2, \dots, n$, where O_i represents the observation data of the i -th borrower and P_0 its probability distribution. $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ represents the p observation index of the i -th borrower, and Y_i represents the default state of the borrower, which is a binary classification result variable.

In forecasting research, researchers usually use a variety of forecasting algorithms to estimate the probability distribution and find the algorithm with the best forecasting performance. It can be found from previous research that for a particular data distribution, the prediction performance of a particular algorithm is superior to other prediction algorithms. However, in personal credit research, researchers cannot know in advance which classification algorithm is

most suitable for a given credit dataset. Researchers usually evaluate the performance of the prediction algorithm through the loss function. A prediction algorithm has the best expected performance for the loss function used, and it is regarded as the best prediction algorithm. Common loss functions include absolute error loss function, squared error loss function, and negative log loss function suitable for binary dependent variables. Super Learner refers to the value of the loss function as risk, and the process of minimizing risk is the process of minimizing classification losses in personal credit problems and then constructs the optimal weighted combination of classifiers.

The estimated function of choosing the best default discriminant model is $Q_0(X) = E_0(Y|X)$, and then the objective function is expressed as the minimum value of the expected loss:

$$Q_0(X) = \arg \min E[L(X, Q(X))]. \quad (1)$$

The specific principles of Super Learner are as follows:

- (1) Define the default discrimination candidate classifier library K and express the number of algorithms in the library as $k(n)$ and the base classifier as $\Psi_k(X)$, $k = 1, 2, \dots, k(n)$.
- (2) Divide the complete credit dataset O into a training set and a verification set. The role of the training set is to fit the classification model, and the validation set is used to evaluate the prediction ability of each model. Cross-validation aims to use multiple training/validation partitions to generate multiple models and do multiple sets of different training/validation to cope with the problem of too one-sided model and insufficient data. In V -fold cross-validation, the ordered sample data are divided into V parts of equal size, $T(v)$ and $V(v)$ are used to denote the training set and the verification set, and the v th part is used as the verification set $V(v)$; then, the remaining $V-1$ copies are used as the training set $T(v)$, where $v = 1, \dots, V$.
- (3) For the V -fold, on the training set $T(v)$, use various algorithms in the candidate classifier library to fit a variety of base classifiers, and use the data in the verification set $V(v)$ to test the performance of the classifier and save its prediction result as $\Psi_{kT(v)}(X_i)$, $O_i \in V(v)$, $v = 1, \dots, V$.
- (4) Repeat the above steps V times, and put the prediction results obtained by each algorithm into an $n \times K$ matrix $Z = \{\Psi_{kT(v)}(X_{V(v)}), v = 1, \dots, V \& k = 1, \dots, K\}$, where $X_{V(v)} = (X_i; O_i \in V(v))$ represents the observation data in the verification set $V(v)$.
- (5) Estimate the regression $E(Y|Z)$ from Y to Z according to the dataset. If the model $m(Z_i|\alpha)$ is a linear regression model, then $m(Z|\alpha) = aZ$; if the dependent variable is a binary categorical variable, the model can use the logistic regression model $m(Z|\alpha) = 1/(1+\exp(-\alpha Z))$. Then, use least squares or maximum likelihood estimation to estimate the weight vector. You can also use constrained least squares regression, such as lasso regression. And restrict the sum of

nonnegative weights to 1 to achieve constraints. On this basis, the weighted combination family of candidate classifiers indexed by the weight vector can be expressed as

$$m(z/\alpha) = \sum_{k=1}^k \alpha_k \Psi_{k,T(v)}(X_{V(v)}), \alpha_k \geq 0 \forall k, \sum_{k=1}^k \alpha_k = 1. \quad (2)$$

- (6) On all admissible α combinations, select the vector value α that minimizes the cross-validation risk of all candidate estimators $\sum_{k=1}^K \alpha_k \Psi_k$, and obtain the optimal weight vector $\hat{\alpha}$:

$$\hat{\alpha} = \arg \min \sum_{i=1}^n (Y_i - m(z_i/\alpha))^2. \quad (3)$$

- (7) According to the classifier selected in the previous step, fit the base classifier $\Psi_k(X)$ on the complete credit dataset, $k = 1, \dots, K$, and combine with the optimal weight vector $\hat{\alpha}$ obtained in the previous step to construct the Super Learner ensemble model:

$$\Psi_{SL}(X) = \sum_{k=1}^K \hat{\alpha}_k \Psi_k(X). \quad (4)$$

The construction flowchart of the Super Learner is shown in Figure 1.

4. Establishment of Personal Credit Default Discrimination Model Based on Super Learner Heterogeneous Ensemble

4.1. Constructing Candidate Base Classifier Library and Determining Its Parameters. When building an ensemble model, the base classifier's choice is an essential part of the ensemble model. The diversity and difference of the base classifier are crucial to the final performance generalization ability. Generally speaking, it is done manually selected. For example, Wang et al. chose LR, DT, ANN, and SVM as the primary classifier [18]. Ala'raj et al. chose RF, DT, NB, NN, and SVM [45]. Li et al. used XGBoost, DNN, and LR as the base classifier [33]. Guo et al. used nine classifiers such as KNN, LDA, LR, DT, SVM, MLP, RF, AdaBoost, and GBDT as candidate classifiers [34]. Papouskova et al. chose LR, DT, SVM, and RF as the base classifier [35]. He et al. used KNN, LR, LDA, SVM, DT, RF, GBDT, AdaBoost, and XGBoost, as the base classifier [46].

The candidate base classifier library constructed based on sorting out relevant literature uses first single classifiers and adds an ensemble classifier with excellent performance to further enhance the prediction performance and robustness of the model. The classification algorithm library set in this article includes logistic regression (LR), lasso regression (Lasso), K-nearest neighbor (KNN), support vector machine (SVM), neural network (NN), decision tree (DT), random forest (RF), GBDT, and XGBoost algorithms. Among them, logistic regression is the most commonly used single

classifier in credit scoring and has a good prediction performance. At the same time, random forest, GBDT, and XGBoost are the more typical applications in Bagging ensemble and Boosting ensemble, respectively, and the prediction performance as compared to traditional single classification. The classifier is excellent, so these nine classifiers are used as the base classifier to get better prediction performance and robustness.

For the comparison of all datasets, this paper keeps the base classifier candidate library unchanged. In order to further improve the performance of each classifier, this paper uses the enumeration method to optimize the parameters of each classifier [47]. That is to say, the possible values of each parameter are arranged and combined, and each possibility is tried through loop traversal, and the best parameter is regarded as the final parameter combination. The brief introduction and parameter setting of each classifier are shown in Table 1.

4.2. Calculate the Total Classification Loss of the Candidate Base Classifier. Cross-validation divides the original dataset into a training set and a test set. The training data train the classifiers, while the test data evaluate these classifiers' performance. Using cross-validation can enhance the robustness of the classification model. It divides the complete dataset into ten equal-sized subsets and uses nine of them as the training set. Then, the cross-validation trains the base classifier in the candidate classifier library, builds a prediction model, and tests the prediction performance through the test data. After storing the prediction results, the procedure repeats ten times until each becomes a verification set, thus obtaining ten sets of prediction results. In the following, the cross-validation calculates the base classifier's classification loss by using the true attributes and prediction results of each set of data to obtain the total loss of the base classifier. The above process terminates until the total loss of each base classifier in the candidate classifier library is obtained. The smaller the total loss of the base classification, the better the classification performance. The ensemble process gives the base classifier a more significant weight to ensure the ensemble classifier's classification performance. The more significant the total loss, the worse the discrimination performance, and the smaller role the classifier played in the ensemble classifier. The classifier is not even considered in the ensemble process to obtain better prediction performance of the ensemble model.

4.3. Solve the Optimal Weight of Each Base Classifier and Build the Super Learner Ensemble Model. The choice of the base classifier is an integral part of the ensemble model. In applying the model, if fixing the base classifier, the model cannot adaptively select a more suitable base classifier for the data structure of different datasets, failing to obtain a better ensemble model. In the existing research, most ensemble credit scoring models usually directly integrate all the base classifiers constructed and rarely consider base classifiers' choice. However, since each model's weight is equal, the base classifier with poor prediction performance will affect the

FIGURE 1: Construction flow diagram of Super Learner.

TABLE 1: Base classifier and its main parameters.

Classifier	Characteristic	Parameters
LR	Explainability, robustness, and generalization	—
Lasso	Sparseness and reduced multicollinearity	—
KNN	Simple, can deal with nonlinearity and no assumptions about data	—
SVM	Processing small samples and high-dimensional data with high classification accuracy	—
NN	Self-learning and sophisticated nonlinearity handling	decay, maxit
DT	Simple, can handle type variables, and easy to overfit	max_depth, cp
RF	Handling of high-dimensional features, interpretability, and simple implementation	num.trees, mtry
GBDT	Iterative decision tree and strong generalization ability	num.trees, mtry
XGBoost	High efficiency and scalability	ntrees, max_depth, shrinkage

Note: LR=logistic regression; Lasso=lasso regression; KNN=K-nearest neighbor classifier; SVM=support vector machine; NN=neural network; DT=decision tree; RF=random forest; GBDT=gradient boosting decision tree; XGBoost=extreme gradient boosting.

performance of the final ensemble model. The emergence of selective ensemble overcomes this shortcoming. It selects the best-performing base classifier for ensemble or gives different base classifiers different weights. It has become a research hotspot in the field of ensemble learning.

The Super Learner theoretical framework aims to build an algorithm library containing weighted combinations of multiple prediction models and expects one of these weighted combinations to perform better than each prediction algorithm. Based on this, a candidate algorithm library can be given in advance for the problem to be solved. Hence, we can construct an infinite set of weighted candidate combination families and select the weighted optimal

combination by minimizing cross-validation. We have obtained the total loss of each base classifier in the candidate base classifier library in the previous step. The goal is to minimize the weighted total loss of all base classifiers, and the weight of each base classifier is nonnegative. The sum is one as the constraint condition, and it establishes the optimal weighting model of the base classifier, selects the weight vector value that minimizes the total weighted loss of all base classifiers among all allowed weight combinations, and finds the optimal value of each base classifier weights. Different datasets have different data structures. It may even happen that some classifiers are not considered during the ensemble process to make the prediction performance of the

ensembled model better; that is, the weight value of the classifier is 0.

According to the credit dataset's data structure, the most suitable base classifier in the candidate base classifier library is chosen, and the optimal weight of each base classifier is obtained. Each algorithm is refitted in the complete dataset to generate the final base classifier and weighted combination with each classifier's weight to generate the Super Learner ensemble model.

5. Empirical Research

5.1. Experimental Dataset. In the empirical study, four real credit datasets are used to evaluate the performance of the model. It includes three credit datasets of Australia, German, and Japanese in UCI database [48], and a large real credit dataset of Give Me Some Credit (GMS C) provided by Kaggle platform. The details of the datasets are shown in Table 2.

There are 690 samples in the Australian dataset, of which 307 are nondefault samples, and 383 are default samples. Each observation datum contains 6 categorical features, 8 numerical features, and a result variable (accept or reject).

There are 1000 samples in the German dataset, of which 700 are nondefault samples, and 300 are default samples. Each observation datum contains 13 categorical features, 7 numerical features, and a result variable (accept or reject).

There are 690 samples in the Japanese dataset, of which 307 are nondefault samples, and 383 are default samples. Each observation datum contains 11 categorical features, four numerical features, and a result variable (accept or reject). The Japanese dataset is not a complete dataset, which contains specific missing values.

There are 150000 samples in the GMS C dataset, of which 139975 are nondefault samples, and 10025 are default samples. Each observation datum contains 10 numerical features and a result variable (accept or reject). The GMS C dataset is not a complete dataset, which contains some missing values.

5.2. Data Preprocessing. In practice, missing data and outliers inevitably exist in credit data, which is not conducive to the construction and application of the model. Data preprocessing can make the data complete and more normative, which is an indispensable step in the modeling process. In this study, the data preprocessing includes three steps: missing value filling, qualitative index virtual coding, and data standardization. After preprocessing the original data through these steps, new data are obtained.

In multistep data preprocessing, the first step is to fill in missing values in the data. According to the type of missing data in the original dataset, fill them separately, create a new category for the category feature to replace the missing value in the feature, and use the mean value to replace the numerical feature's missing value. Then, the second step of virtual encoding is performed. Because of the incomparability between unordered feature values and multicategory feature values, the coding of virtual variables is used to quantify the categorical variables according to the feature

category. Generally, a feature with a category k is encoded as a set of $k-1$ derivative dummy variables, which can effectively avoid multiple collinearities and can also represent all categories within the feature (that is, the benchmark comparison category is set to 0 when the $k-1$ dummy variables are all 0, and it is the base class). Finally, the third step of data normalization is performed to eliminate the numerical difference between features. In classifiers, support vector machines and other classification models based on distance metrics are extremely sensitive to the difference in the order of magnitude between data. A vast difference in the order of magnitude between data will cause a significant classification error. To avoid the influence of the magnitude difference between the data on the classification results, before building the model, the dataset should be standardized through standardization.

This paper uses the Z-score standardization method to standardize the data [39, 41]. The Z-score standardization method is shown in

$$x' = \frac{x - \bar{x}}{s}, \quad (5)$$

where x' represents the processed value, x is the original value, $\bar{x}$ denotes the mean of the feature, and s stipulates the standard deviation of the feature.

In this paper, four credit datasets are used in the empirical study, which are processed according to the multistep data preprocessing method described above. First, fill in the missing values in the data. For the datasets including missing values, Japanese datasets, and GMS C datasets, the missing values of classification indicators are filled according to the new categories, and the missing values of numerical characteristics are filled according to the average values of corresponding indicators, and the complete dataset is obtained. The Australian dataset and the German dataset are already complete datasets and do not need to be processed. Then, the classification variables are processed. The Japanese dataset contains classification indexes and needs dummy variable processing. The Australian dataset and the German dataset have processed the classification indicators, while the GMS C dataset has numerical indicators, which do not need to be processed separately. Finally, based on the complete dataset obtained after two-step data preprocessing, the Z-score of numerical indicators is standardized to get the complete dataset.

After multistep data preprocessing, the dataset is divided into the total training set and the test set according to the ratio of 8:2. It means 80% of the data is used to train the model, and 20% of the data is used to verify the effectiveness of the model. To further improve the performance of each model, the total training set containing 80% data is divided into two parts according to the same proportion, 80% is used as the training set and 20% as the verification set, and the enumeration method is used to adjust the model parameters.

5.3. Evaluation Indicators. There are many indicators to evaluate the classification performance of the model. Commonly used indicators are accuracy rate, recall rate, F value, and AUC value, among others. Besides, four

TABLE 2: Description of the four datasets used in the study.

Dataset	Samples	Good	Bad	Category features	Numerical features	Features
Australian	690	307	383	6	8	14
German	1000	700	300	13	7	20
Japanese	690	307	383	11	4	15
GMSC	150000	139975	10025	0	10	10

evaluation indicators are used in this paper to estimate the performance of the model, accuracy rate, AUC, type I error rate, and type II error rate, which are obtained based on the confusion matrix shown in Table 3. According to Table 3, true positive (TP) means that if the actual values are zero, then the prediction is zero. However, false negative (FN) means that the actual value of zero may result in the prediction value of one. Also, false positive (FP) gives a prediction value of zero, although the real value is one, while the true negative (TN) presents the same results in any cases.

Based on these factors, the accuracy, the type I error rate, and the second type error rate are expressed as follows:

$$\begin{aligned} \text{accuracy} &= \frac{TP + TN}{TP + FN + FP + TN}, \\ \text{type I error} &= \frac{FN}{TP + FN}, \\ \text{type II error} &= \frac{FP}{TP + FP}. \end{aligned} \quad (6)$$

The ROC curve is a curve describing the binary classifier system's performance when its recognition threshold changes. The curve is created by plotting the true positive rate and false positive rate under various threshold settings. AUC is defined as the probability that the prediction model ranks randomly selected positive instances higher than randomly selected negative instances. Hence, AUC can be computed as the area under the ROC curve. For AUC, the larger the value, the better the classifier performance. In the literature on credit risk, AUC is a suitable measure for performance evaluation due to its robustness to unbalanced data. The calculation method of AUC is as follows:

$$AUC = \frac{\sum_{ins_i \in \text{pos}} \text{rank}_{ins_i} - M \times (M + 1)/2}{M \times N}, \quad (7)$$

where rank_{ins_i} represents the index number of the i -th sample; that is, when ranking the probability scores in increasing order, rank_{ins_i} is in the rank's position. M and N are the number of positive samples and negative samples, respectively. $\sum_{ins_i \in \text{pos}}$ means only to add up the index numbers of positive samples.

In the above evaluation indicators, the type II error rate refers to the misjudgment of borrowers with a high probability of default as being able to borrow, which will cause more significant losses for banks and financial institutions and will focus on this article. Besides, although the indexes currently used to evaluate the classification performance of the model are accuracy rate, among others, for unbalanced samples, AUC can better

TABLE 3: Confusion matrix for credit scoring.

		Predicted	
Real	Positive (nonrisk)	Positive (nonrisk) True positives (TPs)	Negative (risk) False positives (FNs)
	Negative (risk)	False negatives (FPs)	True negatives (TNs)

evaluate the model's performance, so AUC will also act as a critical evaluation index.

5.4. Experimental Results and Analysis. Using the three real credit datasets of Australia, German, and Japanese in the UCI database, GMSC datasets of Kaggle platform, and the constructed Super Learner ensemble model, the selection of base classifiers in each dataset is obtained. The optimal weight coefficients of each base classifier are calculated. In solving the optimal weights, the smaller the total loss of the base classifier, the better the classification performance. The ensemble process should give the base classifier a more considerable weight to ensure the final classification performance of the ensembled classifier; the more the total loss, the worse the discrimination performance, and the smaller the role of the classifier in the ensembled classifier, and if the weight coefficient of the classifier is 0, it means that the final ensembled model has better discrimination accuracy. The classifier is not considered in the ensemble process. The experiment uses R3.5.5 for empirical analysis and uses the Super Learner package to fit the relevant models.

In the Australian dataset, the final ensemble model uses only three basic classifiers of LR, KNN, and GBDT, and other classifiers with weighting coefficients of 0 did not participate in the final ensemble. Among the three classifiers used, GBDT has the most significant weight coefficient of 0.7457, indicating that GBDT plays the most crucial role in the Super Learner ensembled model built on the Australian dataset. In the German dataset, four classifiers of LR, SVM, RF, and GBDT constitute the final Super Learner ensembled model, of which RF plays the most extensive role. Based on the Japanese dataset, the Super Learner ensemble uses seven base classifiers, including LR, KNN, SVM, NN, RF, GBDT, and XGBoost process, and XGBoost plays the most critical role. In the GMSC dataset, the final integration model uses LR, KNN, DT, RF, GBDT, and XGBoost, among which GBDT plays the most critical role.

The above results show that Super Learner integrates different base classifiers for credit datasets with different data structures. Also, according to each classifier's role in the

ensemble process, Super Learner gives different weights to different classifiers. Even, to ensure the classification performance of the ensembled classifier, the proposed Super Learner does not consider specific classifiers with little effect in the ensemble process. This design also reveals that the Super Learner ensembled model has good data adaptability and can choose the classifier independently according to the dataset's situation to build a Super Learner with better predictive performance. The ensembled model is selected instead of integrating all candidate classifications such as traditional ensemble learning, so that some base classifiers with inadequate classification affect the ensembled model's accuracy.

Table 4 shows the classifiers' adaptive selection results on each dataset, that is, the classifier used to construct the final model and the weight coefficients of each classifier.

The base classifier and its corresponding weight coefficients are determined, combined with the base classifier constructed in the complete training set, thus constructing the final Super Learner model. The performance of various comparison models and the proposed Super Learner model on different datasets is shown in Table 5.

Table 5 shows the running results of the proposed model and 10 comparison models on the same datasets and compares them with four evaluation indicators.

The top three classifiers on each evaluation indicator are highlighted in bold. Because different datasets have different data structures and evaluation indicators, no classifier is optimal. Each classifier has a different performance on each dataset, and each dataset has its own suitable classifier. Super Learner and neural networks show excellent overall discriminative performance on the Australian dataset, where the discriminatory accuracy of the classifier is 0.8913. On the German dataset, the accuracy of heterogeneous integration model composed of majority voting is 0.785, and that of super learner is 0.78. Super Learner is slightly inferior to GBDT on the Japanese dataset with an accuracy rate of 0.913, but it still maintains high overall accuracy. Another vital aspect of the classification model is that the model can maintain excellent performance on multiple datasets. On the Australian dataset, Super Learner and XGBoost performed better overall; on the German dataset, the heterogeneous integration model composed of Super Learner and majority voting performs better; on the Japanese dataset, the top two classifiers overall performed were GBDT and Super Learner; on the GMSC dataset, the accuracy of Super Learner is the highest, which is 0.9369, and GBDT and XGBoost are the second.

Thus, in these four credit datasets, the overall discriminant performance of the Super Learner we constructed maintains the best or the second best. Besides, Super Learner has better robustness than other models and can better adapt to different datasets.

The accuracy of discrimination of ensembled algorithms such as random forest and XGBoost is also high, but it is slightly inferior to the Super Learner algorithm used in this paper from the perspective of stability.

On the one hand, the above results prove the accuracy and robustness of the constructed Super Learner model; on

the other hand, random forest, GBDT, and XGBoost are the mainstream models in Bagging and Boosting ensemble algorithms. The heterogeneous ensemble model composed of majority voting is also the most common heterogeneous ensemble model. Its good performance further shows that the ensemble model is better than a single classifier in most cases.

Figures 2–9 intuitively compare each classifier's performance on the accuracy rate, AUC, type I error rate, and type II error rate of each evaluation index on the datasets.

Figures 2 and 3 show the performance of ten comparative models and Super Learner ensemble classifier on the Australian dataset.

The accuracy and AUC of each classifier show the same trend. When the two metrics of a classifier are larger, the classification loss will be smaller. On the contrary, the type I error rate and the type II error rate are on the contrary. The smaller the two index values of the classifier, the lower the error rate of the classifier. Figure 2 shows the comparison results on accuracy and AUC of each classifier. Super Learner has the highest accuracy, while AUC is not the highest but remains in the top three. It can be seen from Figure 3 that the type I error rate and the type II error rate show an opposite trend. When the type I error rate of the classifier is small, the type II error rate is relatively large, while our Super Learner has good performance in both types of error rates, both of which remain at the best or the second best.

From Figures 4 and 5, we can clearly see that the Super Learner ensemble classifier has the highest accuracy and AUC compared with other classifiers on the German dataset, indicating its excellent discrimination performance. Although Super Learner's performance in the type I error rate and the type II error rate is not the best, it remains in the top three.

Figures 6 and 7 show the performance of the ten comparison models and Super Learner ensemble classifiers on the Japanese dataset.

As can be seen from Figure 6, Super Learner is slightly inferior to GBDT in accuracy and AUC. Figure 7 shows the comparison of the type I error rate and the type II error rate of each classifier. Although Super Learner's performance in the first and second type error rates is not the best, it still keeps in the top three.

Figures 8 and 9 show the performance of ten comparison models and Super Learner ensemble classifiers on the GMSC dataset. Expressed by Figure 8, Super Learner is the best in accuracy and AUC. Because the GMSC dataset is an extremely unbalanced dataset, the ability of each classification model to identify a few default samples is poor. Figure 9 shows the comparison results of each classifier on the type I error rate and the type II error rate. From the type I error rate, the performance of single classification is better, while the performance of the type II error rate is worse. However, the ensemble model performs better in the type II error rate, which means its ability to identify default samples is stronger. Although the performance of the Super Learner model is not the best in the second kind of error rate, it remains in the second place.

TABLE 4: Selection of classifiers and corresponding coefficients on each dataset.

Dataset	Classifier								
	LR	Lasso	KNN	SVM	NN	DT	RF	GBDT	XGBoost
Australian	0.1026	0	0.1517	0	0	0	0	0.7457	0
German	0.3134	0	0	0.2490	0	0	0.3620	0.0755	0
Japanese	0.1206	0	0.0364	0.0547	0.0254	0	0.1350	0.2141	0.4135
GMSC	0.0323	0	0.0367	0	0	0.0005	0.1858	0.5191	0.2256

TABLE 5: Evaluation results of the proposed model and various comparison models on different datasets.

Dataset	Classifier	Evaluation metrics			
		Accuracy	AUC	Type I error	Type II error
Australian	LR	0.8768	0.9334	0.0615	0.1781
	Lasso	0.8623	0.9313	0.0615	0.2055
	KNN	0.8551	0.9344	0.0923	0.1918
	SVM	0.8623	0.9336	0.0615	0.2055
	NN	0.8913	0.9045	0.0923	0.1233
	DT	0.8768	0.9231	0.0462	0.1918
	RF	0.8768	0.9534	0.0769	0.1644
	GBDT	0.8841	0.9498	0.0462	0.1781
	XGBoost	0.8841	0.9523	0.0615	0.1644
	Fused with MV	0.8841	0.8879	0.0462	0.1781
	Super Learner	0.8913	0.9511	0.0462	0.1644
German	LR	0.7700	0.7661	0.1538	0.4211
	Lasso	0.7600	0.7623	0.1329	0.5088
	KNN	0.7550	0.7113	0.0559	0.7193
	SVM	0.7750	0.7616	0.0979	0.5439
	NN	0.7350	0.6900	0.1399	0.5789
	DT	0.6950	0.6784	0.1818	0.6140
	RF	0.7650	0.7711	0.0909	0.5965
	GBDT	0.7700	0.7651	0.1119	0.5263
	XGBoost	0.7400	0.7518	0.1329	0.5789
	Fused with MV	0.7850	0.6914	0.0909	0.5263
	Super Learner	0.7800	0.7723	0.0979	0.5263
Japanese	LR	0.8623	0.9308	0.1311	0.1429
	Lasso	0.8986	0.9474	0.0656	0.1299
	KNN	0.8841	0.9241	0.1967	0.0519
	SVM	0.8623	0.9591	0.1148	0.1558
	NN	0.8913	0.9232	0.2131	0.0260
	DT	0.8913	0.9457	0.1475	0.0779
	RF	0.9058	0.9645	0.1148	0.0779
	GBDT	0.9203	0.9729	0.1148	0.0519
	XGBoost	0.8986	0.9453	0.1311	0.0779
	Fused with MV	0.9130	0.9119	0.0984	0.0779
	Super Learner	0.9130	0.9650	0.1148	0.0649
GMSC	LR	0.9331	0.6997	0.0024	0.9572
	Lasso	0.9332	0.7000	0.0022	0.9592
	KNN	0.9334	0.7186	0.0041	0.9303
	SVM	0.9349	0.6443	0.0037	0.9118
	NN	0.9364	0.8603	0.0083	0.8276
	DT	0.9344	0.6574	0.0099	0.8336
	RF	0.9365	0.8504	0.0076	0.8349
	GBDT	0.9368	0.8617	0.0094	0.8053
	XGBoost	0.9368	0.8627	0.0079	0.8260
	Fused with MV	0.9361	0.8560	0.0045	0.8836
	Super Learner	0.9369	0.8649	0.0081	0.8223

FIGURE 2: Accuracy and AUC of eleven classification models on the Australian dataset.

FIGURE 3: Type I error rate and type II error rate of eleven classification models on the Australian dataset.

FIGURE 4: Accuracy and AUC of eleven classification models on the German dataset.

FIGURE 5: Type I error rate and type II error rate of eleven classification models on the German dataset.

FIGURE 6: Accuracy and AUC of eleven classification models on the Japanese dataset.

FIGURE 7: Type I error rate and type II error rate of eleven classification models on the Japanese dataset.

FIGURE 8: Accuracy and AUC of eleven classification models on the GMSC dataset.

FIGURE 9: Type I error rate and type II error rate of eleven classification models on the GMSC dataset.

In most cases, although the implementation of a single classifier is relatively simple, its classification performance is inferior to its ensembled structure. At the same time, Super Learner heterogeneous ensemble shows the potential to improve the default discrimination model's performance. It achieved the best or remained in the top three among the four evaluation indicators of all datasets, compared with other models. Based on maintaining a high discriminating ability, the robustness of the model is guaranteed, and it is more suitable as a useful tool for distinguishing potential default borrowers from banks and other financial institutions in reality.

5.5. Comparison of Other Studies. According to the principle in Section 3 of this paper, the construction principle of the proposed Super Learner heterogeneous ensemble model is to consider the robustness of the base classifier

first and the accuracy of the base classifier second. Therefore, the Super Learner heterogeneous ensemble model may lose part of its accuracy in ensuring the robustness of default judgment.

This paper compares the results of other researchers on the same credit dataset, further verifies the characteristics of the Super Learner heterogeneous ensemble model, and proves the applicability of the model in the field of credit. The specific results are shown in Table 6. From the fifth and sixth columns of Table 6, it can be clearly found that on the Japanese and GMSC datasets, the performance of the Super Learner model in each indicator is better than other comparative models. Columns 3 and 4 in Table 6 show the performance of the Super Learning model and the comparison models on the Australian and German datasets. Although the Super Learning model is not the best, it remains the second best. From the above analysis, it can be seen that the Super Learner heterogeneous ensemble model

TABLE 6: Performance comparison with other default discrimination models.

Model	Evaluation metrics	Australian	German	Japanese	GMSC
Super Learner model	Accuracy	0.8913	0.7800	0.9130	0.9369
	AUC	0.9511	0.7723	0.9650	0.8649
	Type I error	0.0461	0.0979	0.1147	0.0081
	Type II error	0.1643	0.5263	0.0649	0.8223
Stacking model (2011) [18]	Accuracy	0.8657	0.7597	—	—
	AUC	—	—	—	—
	Type I error	0.1291	0.1076	—	—
	Type II error	0.1368	0.5500	—	—
Bsteking model (2018) [49]	Accuracy	0.8828	0.7866	—	—
	AUC	0.9280	0.7948	—	—
	Type I error	—	—	—	—
	Type II error	—	—	—	—
GCSSE model (2020) [50]	Accuracy	0.8689	0.7376	—	—
	AUC	0.9128	0.7561	—	—
	Type I error	—	—	—	—
	Type II error	—	—	—	—
GSCI model (2020) [51]	Accuracy	0.9116	0.7775	0.8913	—
	AUC	0.9143	0.7042	0.8948	—
	Type I error	—	—	—	—
	Type II error	—	—	—	—
Boughaci et al.'s work (2020) [52]	Accuracy	0.8570	0.7750	0.8480	0.9240
	AUC	0.9540	0.8060	0.9180	0.8550
	Type I error	—	—	—	—
	Type II error	—	—	—	—
Zhang et al.'s work (2019) [53]	Accuracy	0.8754	0.7682	0.8720	0.9352
	AUC	0.9370	0.8029	0.9387	0.8577
	Type I error	—	—	—	—
	Type II error	—	—	—	—

shows good robustness and maintains the top three ranking in all indicator performance. In addition, it also has good performance in accuracy, which is only lower than the GSCI model in the Australian dataset and lower than the Bsteking model in the German dataset. This shows that the Super Learning heterogeneous ensemble model constructed in this paper has excellent robustness and good accuracy.

6. Conclusion

Establishing a borrower's default discrimination model is an essential task for banks and other financial institutions to make loan decisions. Therefore, the discriminative performance and robustness of the default discriminant model are crucial to financial institutions' profitability, such as banks. In this study, we utilized the heterogeneous ensemble default discriminant model. The ensembled Super Learner model that determines the optimal combination of multiple base classifiers using cross-validation performs well in disease prediction in the medical field. This paper considers introducing the Super Learner algorithm into personal credit default evaluation research to build a default discrimination model with heterogeneous ensemble for better default discrimination accuracy and robustness.

First, we construct a base classifier candidate library containing a single classifier with better prediction performance such as logistic regression and SVM and a homogeneous isolator with better performance, such as a random

forest. Second, we calculate the total classification loss of each base classifier. Ten-fold cross-validation is used to train multiple base classifiers in the candidate library separately, calculate ten sets of prediction results for each base classifier, and calculate ten sets of losses for each base classifier according to the deviation of the prediction results from the actual values. The ten sets of loss of a base classifier calculate the total loss of the base class. The smaller the total loss of the base class, the better the classification performance. The ensemble process should give the base classifier a more substantial weight to ensure the ensembled classifier's classification performance. The higher the total loss, the worse the discrimination performance, the smaller the classifier's role in the ensembled classifier, even the classifier is not considered in the ensemble process. Then, we created the base classifier's optimal weighting model to solve the optimal weight of each base classifier. The model aims to minimize the weighted total loss of all base classifiers, with each base classifier's weight being nonnegative and the sum of weights being one as a constraint. Thus, we can find the optimal combination of the classifiers with the smallest cross-validation classification loss. Finally, in the empirical research, we use four commonly used indicators of accuracy, AUC, type I error rate, and type II error rate as evaluation indicators to verify the effectiveness and reliability of the Super Learner integrated model on UCI (Australian dataset, German dataset, and Japanese dataset) and GMSC datasets. In order to prove the superiority of the Super Learner

ensemble model, we compare it with the single classifier such as logistic regression, SVM, homogeneous ensemble model such as random forest and GBDT, and heterogeneous ensemble model combined by majority voting. Simultaneously, we also compare the results with the latest research results of other researchers on the basis of using the same datasets. The results show that the Super Learner heterogeneous ensemble model constructed in this paper has better discrimination accuracy and robustness than other single classifiers and ensembled classifiers.

Data Availability

The data used to support the findings of this study are from a previous study [48].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The study was supported by the General Project of the Natural Science Foundation of Hebei Province (no.G2019501105), the Top Young Talents Program of Higher Learning Institutions of Hebei Province in 2019 (no. BJ2019213), the Fundamental Research Funds for the Central Universities (no. N182314006), and the China Postdoctoral Science Foundation (nos. 2016M590136 and 2017T100111).

References

- [1] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 160, no. 3, pp. 523–541, 1997.
- [2] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *Journal of Finance*, vol. 23, no. 9, pp. 589–609, 1968.
- [3] W. J. Boyes, D. L. Hoffman, and S. A. Low, "An econometric analysis of the bank credit scoring problem," *Journal of Econometrics*, vol. 40, no. 1, pp. 3–14, 1989.
- [4] D. Martin, "Early warning of bank failure," *Journal of Banking & Finance*, vol. 1, no. 3, pp. 249–276, 1977.
- [5] I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.
- [6] A. Khashman, "Neural networks for credit risk evaluation: investigation of different neural models and learning schemes," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6233–6239, 2010.
- [7] S. Oreski, D. Oreski, and G. Oreski, "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12605–12617, 2012.
- [8] S. R. Sain, "The nature of statistical learning theory," *Technometrics*, vol. 38, no. 4, p. 409, 1996.
- [9] D. Feldman and S. Gross, "Mortgage default: classification trees analysis," *The Journal of Real Estate Finance and Economics*, vol. 30, no. 4, pp. 369–396, 2005.
- [10] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis," *Expert Systems with Applications*, vol. 37, no. 1, pp. 534–545, 2010.
- [11] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.
- [12] B. Wang, L. Ning, and Y. Kong, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications*, vol. 128, pp. 301–315, 2019.
- [13] J. Abellán and C. J. Mantas, "Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3825–3830, 2014.
- [14] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [15] R. E. Schapire, "The strength of weak learnability," *Proceedings of the Second Annual Workshop on Computational Learning Theory*, vol. 5, no. 2, pp. 197–227, 1989.
- [16] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
- [17] E. Mayhua-López, V. Gómez-Verdejo, and A. R. Figueiras-Vidal, "A new boosting design of Support Vector Machine classifiers," *Information Fusion*, vol. 25, pp. 63–71, 2015.
- [18] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Systems with Applications*, vol. 38, no. 1, pp. 223–230, 2011.
- [19] M. J. Van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner," *Statistical Applications in Genetics & Molecular Biology*, vol. 6, no. 1, pp. 1–23, 2007.
- [20] R. Wyss, S. Schneeweiss, M. Van der Laan, S. D. Lendle, C. Ju, and J. M. Franklin, "Using Super Learner prediction modeling to improve high-dimensional propensity score estimation," *Epidemiology*, vol. 29, no. 1, pp. 96–106, 2018.
- [21] B. Valerio and H. Edward, "Principled machine learning using the super learner: an application to predicting prison violence," *Sociological Methods & Research*, vol. 48, no. 3, pp. 698–721, 2019.
- [22] H. Lin, W. Zou, T. Li et al., "A super-learner model for tumor motion prediction and management in radiation therapy: development and feasibility evaluation," *Scientific Reports*, vol. 9, 2019.
- [23] C.-F. Tsai and C. Hung, "Modeling credit scoring using neural network ensembles," *Kybernetes*, vol. 43, no. 7, pp. 1114–1123, 2014.
- [24] G. Paleologo, A. Elisseeff, and G. Antonini, "Subagging for credit scoring models," *European Journal of Operational Research*, vol. 201, no. 2, pp. 490–499, 2010.
- [25] L. Yu, Z. Yang, and L. Tang, "A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment," *Flexible Services and Manufacturing Journal*, vol. 28, no. 4, pp. 576–592, 2016.
- [26] G. Cano, J. Garcia-Rodriguez, A. Garcia-Garcia et al., "Automatic selection of molecular descriptors using random forest: application to drug discovery," *Expert Systems with Applications*, vol. 72, pp. 151–159, 2017.
- [27] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," *Information Sciences*, vol. 425, pp. 76–91, 2018.
- [28] S. Hamori, M. Kawai, T. Kume et al., "Ensemble learning or deep learning? Application to default risk analysis," *Journal of Risk and Financial Management*, vol. 11, no. 2, pp. 1–14, 2018.

- [29] P. Plawiak, M. Abdar, and U. R. Acharya, "Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring," *Applied Soft Computing*, vol. 84, 2019.
- [30] K. M. R. Alam, N. Siddique, and H. Adeli, "A dynamic ensemble learning algorithm for neural networks," *Neural Computing and Applications*, vol. 32, no. 12, pp. 8675–8690, 2020.
- [31] C.-F. Tsai, Y.-F. Hsu, and D. C. Yen, "A comparative study of classifier ensembles for bankruptcy prediction," *Applied Soft Computing*, vol. 24, pp. 977–984, 2014.
- [32] D. S. C. Nascimento, A. L. V. Coelho, and A. M. P. Canuto, "Integrating complementary techniques for promoting diversity in classifier ensembles: a systematic study," *Neurocomputing*, vol. 138, pp. 347–357, 2014.
- [33] W. Li, S. Ding, Y. Chen, and S. Yang, "Heterogeneous ensemble for default prediction of Peer-to-Peer lending in China," *IEEE Access*, vol. 6, pp. 54396–54406, 2018.
- [34] S. Guo, H. He, and X. Huang, "A multi-stage self-adaptive classifier ensemble model with application in credit scoring," *IEEE Access*, vol. 7, pp. 78549–78559, 2019.
- [35] M. Papouskova and P. Hajek, "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," *Decision Support Systems*, vol. 118, no. 3, pp. 33–45, 2019.
- [36] P. Plawiak, M. Abdar, J. Plawiak et al., "A new deep genetic hierarchical network of learners for prediction of credit scoring," *Information Sciences*, vol. 516, pp. 401–418, 2020.
- [37] L. Nanni and A. Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3028–3033, 2009.
- [38] A. W. Van der Vaart, S. Dudoit, and M. J. Van der Laan, "Oracle inequalities for multi-fold cross validation," *Stats & Decisions*, vol. 24, no. 3, pp. 351–371, 2006.
- [39] S. Rose, "Mortality risk score prediction in an elderly population using machine learning," *American Journal of Epidemiology*, vol. 177, no. 5, pp. 443–452, 2013.
- [40] R. Pirracchio and M. Carone, "The balance Super Learner: a robust adaptation of the Super Learner to improve estimation of the average treatment effect in the treated based on propensity score matching," *Statistical Methods in Medical Research*, vol. 27, no. 8, pp. 2504–2518, 2018.
- [41] A. I. Naimi and L. B. Balzer, "Stacked generalization: an introduction to super learning," *European Journal of Epidemiology*, vol. 33, no. 5, pp. 459–464, 2018.
- [42] C. Ju, M. Combs, S. D. Lendle et al., "Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods," *Journal of Applied Statistics*, vol. 46, no. 12, pp. 2216–2236, 2019.
- [43] M. A. Hedeya, A. H. Eid, and R. F. Abdel-Kader, "A Super-Learner ensemble of deep networks for Vehicle-Type classification," *IEEE Access*, vol. 8, pp. 98266–98280, 2020.
- [44] M. Cherifa, A. Blet, A. Chambaz, E. Gayat, M. Resche-Rigon, and R. Pirracchio, "Prediction of an acute hypotensive episode during an ICU hospitalization with a Super Learner machine-learning algorithm," *Anesthesia & Analgesia*, vol. 130, no. 5, pp. 1157–1166, 2020.
- [45] M. Ala'raj and M. F. Abbod, "Classifiers consensus system approach for credit scoring," *Knowledge-Based Systems*, vol. 104, no. 6, pp. 89–105, 2016.
- [46] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: adaption of different imbalance ratios," *Expert Systems with Applications*, vol. 98, no. 5, pp. 105–117, 2018.
- [47] Z. Han, H. Zhang, D. Wu et al., "Performance optimization for a novel combined cooling, heating and power-organic Rankine cycle system with improved following electric load strategy based on different objectives," *Energy Conversion and Management*, vol. 221, Article ID 113294, 2020.
- [48] A. Asuncion and D. J. Newman, *UCI Machine Learning Repository*, School of Information and Computer Science, University of California, Irvine, CA, USA <http://www.ics.uci.edu/?mlearn/MLRepository.html>.
- [49] Y. Xia, C. Liu, B. Da, and F. Xie, "A novel heterogeneous ensemble credit scoring model based on bstacking approach," *Expert Systems with Applications*, vol. 93, no. 3, pp. 182–199, 2018.
- [50] J. Xiao, X. Zhou, Y. Zhong et al., "Cost-sensitive semi-supervised selective ensemble model for customer credit scoring," *Knowledge Based Systems*, vol. 189, p. 2020.
- [51] X. Chen, S. Li, X. Xu et al., "A novel GSCI-based ensemble approach for credit scoring," *IEEE Access*, vol. 8, 2020.
- [52] D. Boughaci and A. A. K. Alkhalwaldeh, "Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: a comparative study," *Risk and Decision Analysis*, vol. 8, no. 1-2, pp. 15–24, 2020.
- [53] W. Zhang, H. He, and S. Zhang, "A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: an application in credit scoring," *Expert Systems with Applications*, vol. 121, pp. 221–232, 2019.