applied sciences

MDPI

Article

Research on User Default Prediction Algorithm Based on Adjusted Homogenous and Heterogeneous Ensemble Learning

Yao Lu [1,†], Kui Wang [2,†], Hui Sun [1,*], Hanwen Qu [3], Jiajia Chen [4], Wei Liu [5] and Chenjie Chang [3]

[1] School of Economics and Management, Xinjiang University, Urumqi 830046, China; luyao@stu.xju.edu.cn
[2] School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China; wangkui19@mails.ucas.ac.cn
[3] School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China; quhanwen@stu.xju.edu.cn (H.Q.); 107552104367@stu.xju.edu.cn (C.C.)
[4] Xinjiang Changji Vocational and Technical College, Changji 831100, China; en_via@163.com
[5] College of Software, Xinjiang University, Urumqi 830046, China; 107552304946@stu.xju.edu.cn
* Correspondence: shui@xju.edu.cn
† These authors contributed equally to this work.

Abstract: In the field of risk assessment, the traditional econometric models are generally used to assess credit risk. And with the introduction of the "dual-carbon" goals to promote the development of a low-carbon economy, the scale of green credit in China has rapidly expanded. But with the advent of the big data era, due to the poor interpretability of a traditional single machine learning model, it is difficult to capture nonlinear relationships, and there are shortcomings in prediction accuracy and robustness. This paper selects the adjusted ensemble learning model based on the homogeneous and heterogeneous factors for user default prediction, which can efficiently process large quantities of high-dimensional data. This article adjusts each model to adapt to the task and innovatively compares various models. In this paper, the missing value filling method, feature selection, and ensemble model are studied and discussed, and the optimal ensemble model is obtained. When comparing the predictions of single models and ensemble models, the accuracy, sensitivity, specificity, F1-Score, Kappa, and MCC of Categorical Features Gradient Boosting (CatBoost) and Random undersampling Boosting (RUSBoost) all reach 100%. The experimental results prove that the algorithm based on adjusted homogeneous and heterogeneous ensemble learning can predict the user default efficiently and accurately. This paper also provides some references for establishing a risk assessment index system.

Keywords: default prediction; ensemble learning; homogeneous ensemble; heterogeneous ensemble

check for updates

Citation: Lu, Y.; Wang, K.; Sun, H.; Qu, H.; Chen, J.; Liu, W.; Chang, C. Research on User Default Prediction Algorithm Based on Adjusted Homogenous and Heterogeneous Ensemble Learning. *Appl. Sci.* **2024**, *14*, 5711. https://doi.org/10.3390/app14135711

Academic Editors: Shahadat Uddin, Tasadduq Imam and Sisira Colombage

Received: 3 May 2024
Revised: 23 June 2024
Accepted: 27 June 2024
Published: 29 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A credit loan refers to the loan issued on the guarantee of the borrower's reputation, which is the main way for financial institutions, especially banks, to make profits. With the advent of the big data era, financial data have also ushered in explosive growth, giving rise to a variety of loan products (such as car loans, mortgage payments, credit cards, etc.) [1]. At the same time, there are also various default events. If the credit institutions can not accurately predict the user default risk, they will not be able to achieve effective risk management [2]. The aggravation and expansion of default events can affect the normal operations of banks and other financial institutions and, in serious cases, can lead to company bankruptcy or the financial crisis for the entire industry [3]. Therefore, user default prediction is the top priority in banking. In recent years, the Chinese government actively responded to the call of the international community to actively develop green finance, etc. Therefore, China has actively expanded the scale of green credit [4], which not only helps to cope with economic downturn pressure but also promotes the development of a green economy and provides solid financial support for sustainable development. And

the advent of big data and advances in machine learning have provided new opportunities for user default prediction.

Traditional methods for user default prediction include logistic regression [5,6], Decision Tree (DT) [6], and support vector machine [7], and these classification models are widely used in default prediction because of their high efficiency. Zhangjie Huang [8] used the SMOTETomek-LighTBM-LR credit default prediction model to improve the accuracy of credit default prediction, providing a new method for the credit industry to reduce the default rate and improve the efficiency of capital utilization. However, it only considers the classification imbalance of a credit dataset when processing data and has limitations in its ability to deal with complex and dynamic relationships between predictor and target variables, making it difficult to capture nonlinear relationships. Syed Nor S. H. [9] developed the prediction model for personal bankruptcies by using DT technology, which facilitates and assists financial institutions in assessing their potential borrowers. Nevertheless, the model used a single algorithm, which is still deficient in prediction accuracy and robustness. M. Z. Abedin et al. introduced a credit default prediction model that utilizes two main techniques: support vector machine (SVM) and Probabilistic Neural Network (PNN). This method aims to improve the accuracy of credit default prediction by fully utilizing the advantages of SVM and PNN algorithms [10].

Ensemble learning, as a research hotspot, aims to integrate data fusion, data modeling, and data mining into a unified framework [11]. Specifically, it is a method of combining multiple weak learners to build a machine learning method with strong performance [12]. In order to overcome the highly imbalanced default and non-default categories in small business credit risk assessment, Abedin M. Z. [13] proposed an extended integration approach based on the Weighted Synthetic Minority Oversampling Technique, called the WSMOTE-ensemble. F. N. Khan et al. [14] proposed a credit card fraud prediction and classification model based on a deep neural network and ensemble learning. The model integrates four algorithms, the Bayes classifier algorithm, logistic regression, DT, and the deep belief network, and has excellent performance in predicting a credit card customer's default, with good accuracy and interpretability. With the development of technology, more and more transactions are completed through the network, and the transaction data have characteristics such as high dimensionality and multiple sources, far exceeding the processing capacity of traditional solutions. The efficiency, accuracy, and strong robustness demonstrated by ensemble learning in multiple fields have brought new solutions and development opportunities to these problems. Ensemble learning can be divided into two types, homogeneous and heterogeneous, according to whether each base learner belongs to the same category. Homogeneous ensemble learning means that all base learners belong to the same category, which is also the most widely used. For example, Lean Y. U. [15] proposed an SVM-based ensemble learning system when studying customer risk identification and customer relationship management. The results showed that the model has multiple components and spatial selection strategies and performs best in various test samples. Heterogeneous ensemble learning refers to not all individual learners being of the same kind. He H. et al. [16] proposed a default prediction model that combines deep learning classifiers with tree-based classifiers using the ensemble learning method, and the proposed model has significantly improved the overall performance. In previous work, each model performed well in certain aspects, but there were also limitations. Through ensemble learning, we can combine the prediction results of these models, such as through voting or weighted averaging, to obtain more accurate and robust prediction results.

In this study, to ensure the accuracy of the experiments, we first used Hot-deck imputation and Mean imputation to fill the missing values in the data variables. To mitigate the impact of differences in the magnitude among the variables on the classification results, prior to model construction, we employed the z-score standardization method to adjust the numerical ranges of different features to similar levels, eliminating the influence of scale on the data. Additionally, standardization improved the distribution of existing

data, making it closer to a normal distribution, thereby enhancing the performance and interpretability of the model. Furthermore, we utilized three feature selection methods based on statistical learning, regression coefficients, and principal component analysis to select appropriate feature variables. The obtained data were then inputted into both homogeneous and heterogeneous ensemble learning models, such as CatBoost, RUSBoost, and gcForest, to predict user defaults. A comparison was made with traditional single learning models to determine the optimal model and analysis.

The experimental results of this paper show that the method based on adjusted homogeneous and heterogeneous ensemble learning can achieve good prediction of user defaults, and the accuracy of user default prediction based on the CatBoost and RUSBoost ensemble algorithms reaches 100%. This paper contributes to the development of existing user default prediction algorithms through a novel and effective approach based on homogeneous and heterogeneous ensemble learning, achieving higher performance and stronger generalization than traditional methods and state-of-the-art techniques. Our findings have important implications for practitioners and researchers in the fields of finance, banking, and insurance and lay the foundation for future research in this field.

The Section 2 describes the data and methods, the Section 3 describes the results, the Section 4 describes the discussion, and the Section 5 summarizes and comments on the paper.

2. Data and Methods

2.1. Experimental Data

The dataset selected in this experiment is the user loan dataset provided by Xinjiang Rural Credit Cooperatives. One of the main businesses of Xinjiang Rural Credit Cooperatives is to assess the default risk of borrowers and predict whether users will default based on the borrower's past fund flow, net interest rate of the main business, and other key records.

There are 117 overdue records and 117 normal non-overdue samples in the dataset. The sample has 26 characteristics, such as the return on equity, main business profit margin, and main business net interest rate. The detailed information is shown in Table 1.

Table 1. Feature details.

Feature Name	Feature Value Type	Number of Missing Values	Missing Rate (%)
Return on equity	float	2	0.85%
Main business profit margin	float	45	19.23%
Main business net interest rate	float	45	19.23%
Main business cash ratio	float	45	19.23%
Profit growth rate	float	61	26.07%
Sales growth rate	float	76	32.48%
Total asset growth rate	float	57	24.35%
Growth rate of accounts receivable	float	66	28.21%
Rate of capital accumulation	float	57	24.36%
Number of accounts receivable turnover	float	18	7.69%
Average turnover days of receivables	float	49	20.94%
Number of inventory turnover	float	20	8.55%
Average turnover days of the inventory	float	57	24.36%
Turnover of current assets	float	2	0.85%
Turnover of total capital	float	2	0.85%
Quick ratio	float	2	0.85%
Current ratio	float	2	0.85%
Interest protection multiple	float	116	49.57%
Operating cash flow liability ratio	float	4	1.71%
Asset–liability ratio	float	0	0
Long-term asset suitability rate	float	10	4.27%
Equity and liability ratio	float	0	0
Net asset	float	0	0
Net cash flow from operating activities	float	0	0
Net margin	float	0	0
Net assets and year-end loan balance ratio	float	24	10.26%

2.2. Data Preprocessing

In order to improve the performance of the model, we need to preprocess the data. Because the data variables used in this study had missing values, to fill in the missing data, we used Hot-deck imputation and Mean imputation to fill in the missing values in the dataset. In order to avoid the difference in the orders of magnitude between the data affecting the classification results, we standardized the data using the z-score normalization method before building the model to ensure that all the variables had the same scale. Splitting the data into a 7:3 ratio for the training and testing sets, we employed a five-fold cross-validation.

2.3. Feature Extraction

To identify the most important variables, we used three major types of feature selection methods, namely, statistical feature methods, regression coefficients, and principal component analysis. Using statistical methods, we calculated the information value (IV) and variance inflation factor (VIF) for each feature. The IV provides an estimate of the predictive ability of the variable, while the VIF value provides a collinearity estimate of the variable with other variables in the model [17,18]. In order to reduce the multicollinearity and improve the predictive ability of the features, we deleted the attribute VIF $\geq$ 10 and IV < 0.01 and deleted the attribute VIF $\geq$ 10 and IV < 0.001 from the data. In addition, we also used four regression models for feature selection, namely, stepwise regression based on AIC criterion, stepwise regression based on BIC criterion, Lasso regression, and Elastic Net regression. These techniques were chosen because they had been shown to be effective in previous user default prediction studies [19–21]. The third method is principal component analysis. The dimensions of the data after using different feature selection algorithms are shown in Table 2.

Table 2. Dimensions of the data after using different feature selection algorithms.

	Hot-Deck Imputation	Mean Imputation
IV < 0.001,VIF $\geq$ 10	10	9
IV < 0.01,VIF $\geq$ 10	7	6
AIC Regression	11	11
BIC Regression	9	9
Lasso Regression	9	7
Elastic Net Regression	11	7
PCA	20	20

2.4. Ensemble Learning Models

2.4.1. Random Forest

RF is a common and widely used homogeneous ensemble learning algorithm. Based on the idea of bagging [22], n training samples are extracted from the training set every time there is a return, a new training set is formed, multiple sub-models are trained, and finally the sub-model with the most votes is selected by a voting method for classification. RF consists of multiple DTs [23]. DT has the advantages of simplicity and strong interpretability and is a supervised learning algorithm. RF reduces the impact of inaccurate classification results caused by outliers in the DT. At the same time, due to the use of partial features from some samples to construct a DT in RF, the possibility of overfitting is reduced. We chose the Gini coefficient as the criterion corresponding to the CART classification tree. The RF model diagram of this experiment is shown in Figure 1.

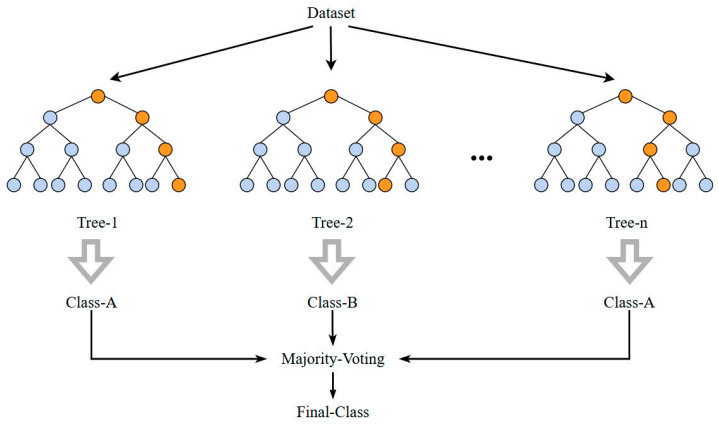


Figure 1. Structure diagram of RF model.

2.4.2. Multi-Grained Cascade Forest

Inspired by deep network learning, the Multi-Grained Cascade Forest (gcForest) algorithm is proposed. The gcForest generates a deep forest with three characteristics: layer-by-layer processing, intramodel feature transformation, and sufficient model complexity. The gcForest model divides training into two stages, Multi-Grained Scanning and Cascade Forest. Multi-Grained Scanning generates features, and Cascade Forest generates prediction results through multiple cascades of forests [24]. The gcForest has fewer parameters than deep neural networks and has considerable robustness to hyperparameter settings. In order to better adapt to the low-dimensional characteristics of credit loan data, we abandoned Multi-Grained Scanning. The structure of the gcForest model used in this experiment is shown in Figure 2. The base learners used by the model are the XGB Classifier, Random Forest Classifier, Extra-Tree Classifier, and logistic regression. Early_stopping_rounds is set to 3. The maximum number of layers is set to 100.

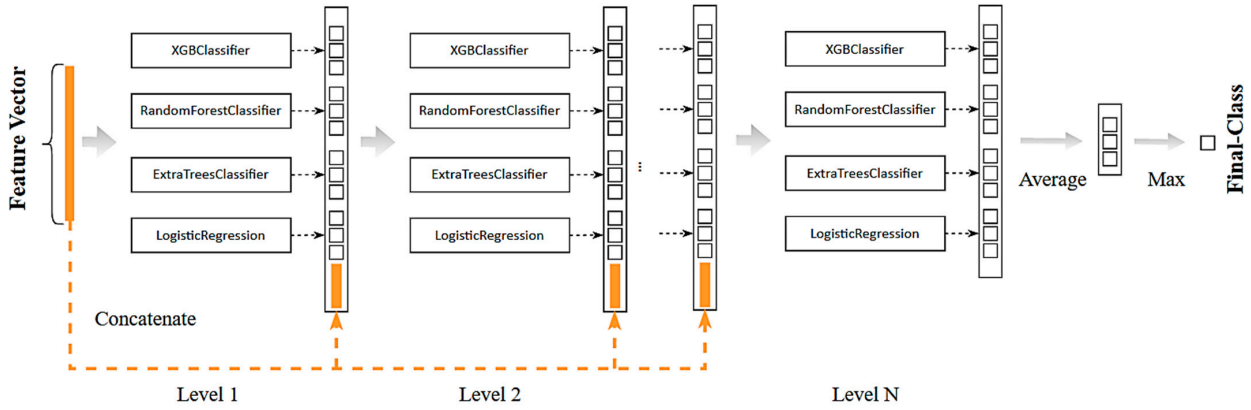


Figure 2. Structure diagram of gcForest model.

2.4.3. Categorical Features Gradient Boosting

Categorical Features Gradient Boosting (CatBoost) is a Gradient Boosting Decision Tree (GBDT) framework [25] based on a symmetric decision tree, which has fewer parameters, supports categorical variables, and has high accuracy. It can process categorical features efficiently and reasonably and binarize floating-point features, statistics values, and One-Hot Encoding features. CatBoost consists of categorical and boosting features. In addition, CatBoost solves the problems of Gradient bias and Prediction shift, thereby reducing the occurrence of overfitting and improving the accuracy and generalization ability of the algorithm. The values of the CatBoost model parameters in this paper are shown in Table 3. "Iterations" is the maximum number of trees in the model, "learning_rate" is the learning

rate of the model, "depth" is the tree depth, "grow_policy" represents the subtree growth strategy, and "depth" represents the L2 regularization parameter.

Table 3. Parameter values of CatBoost model.

Parameter	Parameter Value
iterations	1000
learning_rate	0.03
depth	6
grow_policy	'Depthwise'
12_leaf_reg	3

2.4.4. Random Undersampling Boosting

Random undersampling Boosting (RUSBoost) is a boosting ensemble method for imbalanced datasets, which extracts a certain amount of various sample types from the dataset to form a training set with balanced distribution. The boost method adopted is the Adaboost. M2 algorithm [26], which iterates the Adaboost. M2 algorithm every round. The Adaboost. M2 algorithm imposes limitations on the output form of the base classifier. The base classifier outputs a value between 0 and 1 for each class, representing the probability that the sample belongs to a certain class, and the sum of these values is equal to 1.

2.4.5. Stacking

Stacking [27] is a model fusion idea different from bagging and boosting, which improves overall performance by fusing multiple base models. The first layer contains multiple basic classifiers that provide the predicted results to the second layer. The second-layer classifiers are usually logistic regression, and the results of the first-layer classifiers are treated as feature fitting outputs. There are three base classifiers used in this experiment, namely, the K Neighbors Classifier, Random Forest Classifier, and Gaussian NB, with the parameter meta Classifier set to logistic regression. The Stacking model is shown in Figure 3.

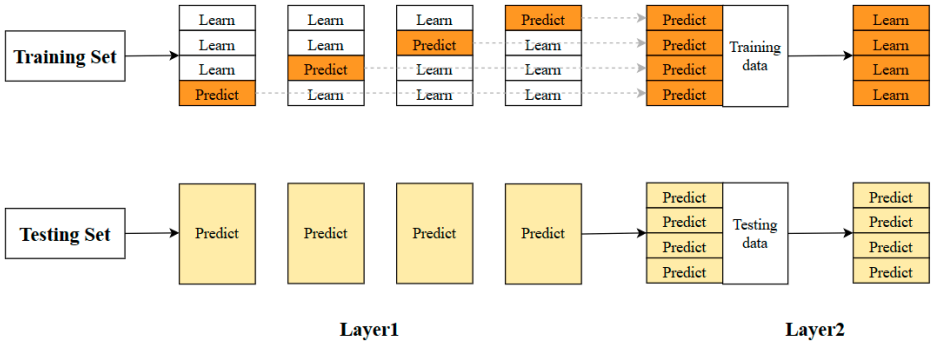


Figure 3. Structural diagram of the Stacking model.

2.5. Model Indicator

In this paper, the accuracy, specificity, sensitivity, F1-Score, Kappa, and MCC were used to evaluate the model performance. The specific evaluation methods are shown in Formulas (1)–(8). In the classification problem, the sample quantity of each category is easily imbalanced, so we introduce the kappa coefficient and MCC evaluation index.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$Specificity = \frac{TN}{FP + TN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$F(a) - Score = \frac{(a^2 + 1) \times (precision \times Sensitivity)}{a^2 \times (precision + Sensitivity)} \tag{5}$$

The commonly used index is the F1-score, which means $a = 1$, so the formula can be transformed into:

$$F1\text{-}Score = \frac{2 \times (precision \times Sensitivity)}{(precision + Sensitivity)} \tag{6}$$

The higher the value of the kappa coefficient, the higher the classification accuracy achieved by the model. The calculation of the k coefficient is based on the confusion matrix, and its expression is shown in Formula (7), where po represents the classification accuracy, that is, the accuracy pe is represented in Formula (8). R_i represents the sum of the elements in row i. C_i represents the sum of the elements in column i, and m represents the elements in the confusion matrix.

$$Kappa = \frac{po - pe}{1 - pe} \tag{7}$$

$$pe = \frac{\sum_i R_i \times C_i}{(\sum m)^2} \tag{8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9}$$

3. Results

In this paper, the effects of different ensemble learning algorithms on the prediction results of the final model were compared with those of a single classifier model. Different filling methods and feature selection algorithms also have influence on the prediction results. To prevent the loss of some important features, we also compared the impact of the data input models with and without feature selection on the results. The entire default risk prediction model process is shown in Figure 4.

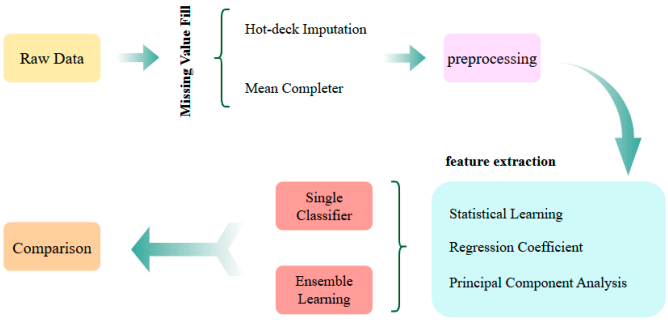


Figure 4. Experimental flow chart.

Models such as SVM, DT, KNN, LDA, RF, gcForest, CatBoost, RUSBoost, Imbalance-XGBoost, and Stacking are mainly used in the experiment. The results of the different filling methods, whether feature selection, different feature selection methods, and different models, on user default prediction are compared, respectively. Table 4 shows the impact of the different filling methods on the predicted results. We found that the result of using Mean imputation is better than that of using Hot-deck imputation, especially when using single classifiers. Then, we compared the effects of the different feature selection methods on the model performance, and the results are shown in Table 5. Using the feature selection method based on the regression algorithm, the overall effect of the model is best. In a single classifier, the feature selection using the Elastic Net regression coefficient and input

DT model have the highest accuracy, which is 98.59%. The second best single classifier model is 97.18%, and its feature selection algorithm is based on the stepwise regression of the BIC criterion. The regression coefficient can well display the relative contribution of features in the model, so features with little or no contribution can be deleted. Table 5 also shows the influence of the different feature selection methods on the results in ensemble learning. The models with better prediction results are RF and CatBoost. The results of the stepwise feature selection based on the AIC criterion were input into the RF model with the best prediction accuracy of 100.00%, and the highest accuracy of CatBoost is 98.59%. Figure 5 shows the average of various indicator values inputted into each model after four regression coefficient feature selection algorithms. The DT indicators values are higher than 95.00%. The RF, CatBoost, RUSBoost, and Stacking indicators have high values and are greater than 93.00%. In particular, the values of CatBoost are greater than 97%.

Table 4. Impact of different filling methods on prediction results, and the values in the table represent accuracy.

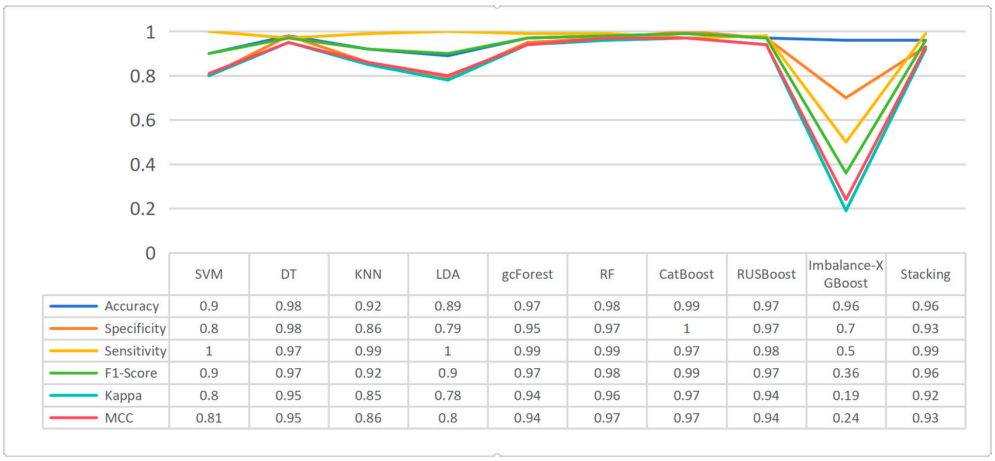
	Hot Deck	Mean
SVM	0.6761	0.8451
DT	0.7746	0.9577
KNN	0.6619	0.8732
LDA	0.7606	0.8873
gcForest	0.9718	0.9296
RF	0.9859	0.9859
CatBoost	1	0.9859
RUSBoost	1	0.9859
Imbalance-XGBoost	0.9577	0.9577
Stacking	0.9296	0.9718

Table 5. Effect of different feature selection methods on accuracy of single classifiers and ensemble models.

	IV-VIF0.01	IV-VIF0.001	AIC	BIC	Lasso	Elastic Net	PCA
SVM	0.7607	0.7746	0.8732	0.9155	0.9014	0.9014	0.8309
DT	0.8309	0.8732	0.9718	0.9718	0.9718	0.9859	0.8873
KNN	0.7887	0.8592	0.9155	0.9437	0.9155	0.9155	0.8873
LDA	0.8028	0.8028	0.8873	0.9014	0.8873	0.8873	0.8592
gcForest	0.8732	0.8591	0.9859	0.9859	0.9577	0.9577	0.9296
RF	0.9014	0.9014	1	0.9718	0.9859	0.9718	0.9155
CatBoost	0.8873	0.8169	0.9859	0.9859	0.9859	0.9859	0.9014
RUSBoost	0.8028	0.8308	0.9718	0.9859	0.9577	0.9718	0.9014
Imbalance-XGBoost	0.8028	0.8028	0.9859	0.9859	0.9437	0.9437	0.8732
Stacking	0.8591	0.9014	0.9859	0.9718	0.9437	0.9437	0.8873

Under Hot-deck imputation, we used the data without feature selection to input each ensemble learning model, and the results are shown in Figure 6, where the index values of CatBoost and RUSBoost reach 100%. We take the optimal results of each ensemble learning model, as shown in Table 6, and the results indicate that Hot-deck imputation is more suitable for ensemble learning model algorithms. Combined with Table 4, the single classifiers are more suitable for Mean imputation.

The experimental results of this paper show that the ensemble learning algorithms have better predictive ability for user defaults, and the accuracy of using the CatBoost algorithm to predict user defaults is as high as 100%, with the other indicator values also reaching 100%. The overall effect of ensemble learning is better than that of single classifiers.



	SVM	DT	KNN	LDA	gcForest	RF	CatBoost	RUSBoost	Imbalance-XGBoost	Stacking
Accuracy	0.9	0.98	0.92	0.89	0.97	0.98	0.99	0.97	0.96	0.96
Specificity	0.8	0.98	0.86	0.79	0.95	0.97	1	0.97	0.7	0.93
Sensitivity	1	0.97	0.99	1	0.99	0.99	0.97	0.98	0.5	0.99
F1-Score	0.9	0.97	0.92	0.9	0.97	0.98	0.99	0.97	0.36	0.96
Kappa	0.8	0.95	0.85	0.78	0.94	0.96	0.97	0.94	0.19	0.92
MCC	0.81	0.95	0.86	0.8	0.94	0.97	0.97	0.94	0.24	0.93

Figure 5. Average of the values of the indicators for each model using the four regression coefficient feature selection algorithms.

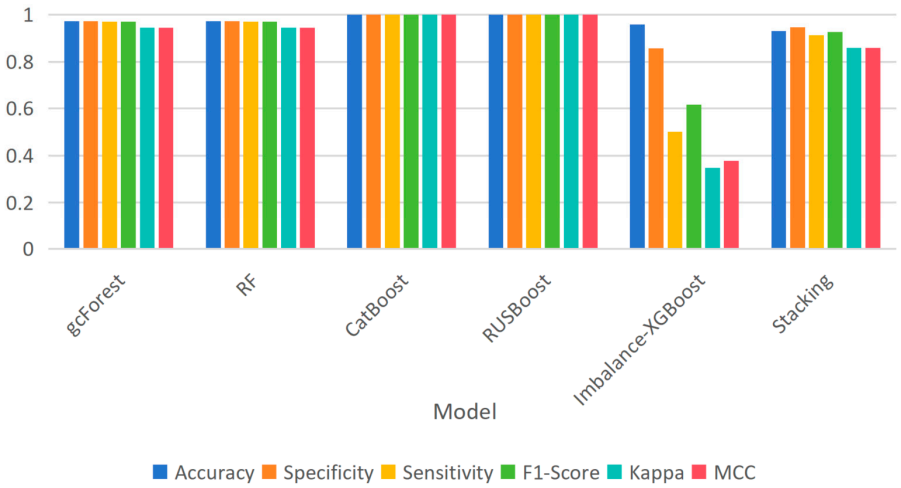


Figure 6. Results of individual ensemble learning models without feature selection.

Table 6. Comparison of optimal results of all ensemble learning models.

	Accuracy	Specificity	Sensitivity	F1-Score	Kappa	MCC
CatBoost (Hot deck)	100%	100%	100%	100%	100%	100%
RUSBoost (Hot deck)	100%	100%	100%	100%	100%	100%
gcForest (Hot deck)	97.18%	97.29%	97.06%	97.06%	94.36%	94.36%
Forest (Mean)	98.59%	97.29%	100%	98.55%	97.18%	97.22%
Imbalance (Hot deck)	95.77%	85.71%	50%	61.54%	34.78%	37.79%
Stacking (Mean)	97.18%	94.59%	100%	97.14%	94.37%	94.52%

4. Discussion

In this work, we demonstrate the effects of different adjusted ensemble learning algorithms on the final model prediction compared with a single classifier model and analyze the effects of different filling methods and feature selection algorithms on the prediction results. The results show that an ensemble learning algorithm has better predictive ability in user default prediction, especially when the CatBoost algorithm is used; its prediction accuracy is as high as 100%. Compared with a single classifier, ensemble learning has a better overall effect.

Ensemble learning algorithms combine predictions from multiple base models to reduce the bias and variance of a single model, thereby improving the overall prediction

accuracy. Additionally, due to the utilization of multiple diverse base models, ensemble learning models exhibit better robustness, maintaining good predictive performance across different data distributions or feature conditions. In our research, we found that traditional single machine learning models suffer from poor interpretability, making it difficult to capture nonlinear relationships and exhibiting shortcomings in both prediction accuracy and robustness. The heterogeneous and homogeneous ensemble learning model for user default prediction efficiently handles large volumes of high-dimensional data and enhances the model generalization capabilities. By effectively leveraging predictions from multiple models, it mitigates the risk of overfitting, thereby improving the model's generalization ability.

In addition, the results from using Mean imputation were found to be better than those from using Hot-deck imputation, especially when using single classifiers. The reason for this difference is that Hot-deck imputation requires finding all the samples without null values, but in practice, it is difficult to find a perfect Hot-deck imputation dataset. Therefore, the overall accuracy of the model using Mean imputation is better than that of Hot-deck imputation.

Future research directions could include the further optimization of ensemble learning algorithms, the exploration of more efficient feature selection methods, and deeper analysis and processing methods for unbalanced data.

5. Conclusions

In this paper, an adjusted model based on homogeneous and heterogeneous ensemble learning algorithms is proposed for user default prediction. In order to solve the missing data problem, two methods, Hot-deck imputation and Mean imputation, which are suitable for financial datasets, are adopted. The method using Mean imputation is better than Hot-deck imputation in single classifier models. In the ensemble learning model, inputting data filled with Hot-deck imputation into the Catboost and RUSboost models can achieve 100% classification accuracy, and the other five indicators (specificity, sensitivity, F1-score, kappa, and MCC) have values of 100%. The experiment proves that the ensemble learning algorithms can predict user default efficiently and accurately.

The traditional single classifier method can not adapt to complex financial scenarios, but the ensemble learning method mining the nonlinear characteristics of financial data enables it to be applied to various complex financial scenarios. And at the same time, this article also involves risk assessment and modeling methods in response to the analysis of new green credit business. The experimental exploration and results of this paper provide new ideas and prospects for risk assessment modeling.

Author Contributions: Conceptualization, C.C. and Y.L.; methodology, C.C.; software, W.L.; validation, W.L., H.Q. and K.W.; formal analysis, H.S.; investigation, J.C.; data curation, W.L.; writing—original draft preparation, H.Q.; writing—review and editing, C.C.; visualization, Y.L.; supervision, H.Q.; project administration, K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Xinjiang University's 2024 excellent doctoral student innovation project (14XJU2024BS014) and The Major project of the Ministry of Science and Technology of China (SQ2021xjkk01800).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to their containing information that could compromise the privacy of research participants.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gao, X.; Xiong, Y.; Xiong, Z.; Xiong, H. Credit Default Risk Prediction Based on Deep Learning. *Res. Sq.* **2021**. [CrossRef]
2. Çallı, B.A.; Coşkun, E. A Longitudinal Systematic Review of Credit Risk Assessment and Credit Default Predictors. *SAGE Open* **2021**, *11*, 21582440211061333. [CrossRef]
3. Kriebel, J.; Stitz, L. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *Eur. J. Oper. Res.* **2022**, *302*, 309–323. [CrossRef]
4. Hu, Y.; Jiang, H.; Zhong, Z. Impact of green credit on industrial structure in China: Theoretical mechanism and empirical analysis. *Environ. Sci. Pollut. Res.* **2020**, *27*, 10506–10519. [CrossRef]
5. Nie, G.; Rowe, W.; Zhang, L.; Tian, Y.; Shi, Y. Credit card churn forecasting by logistic regression and decision tree. *Expert Syst. Appl.* **2011**, *38*, 15273–15285. [CrossRef]
6. Padimi, V.; Sravan, V.; Ningombam, D.D. Applying Machine Learning Techniques To Maximize The Performance of Loan Default Prediction. *J. Neutrosophic Fuzzy Syst.* **2022**, *2*, 44–56. [CrossRef]
7. Ribeiro, B.; Silva, C.; Chen, N.; Vieira, A.; das Neves, J.C. Enhanced default risk models with SVM+. *Expert Syst. Appl.* **2012**, *39*, 10140–10152. [CrossRef]
8. Huang, Z. Research on Credit Default Prediction Based on Machine Learning. Master's Thesis, Chongqing Technology and Business University, Chongqing, China, 2023.
9. Syed Nor, S.H.; Ismail, S.; Yap, B.W. Personal bankruptcy prediction using decision tree model. *J. Econ. Financ. Adm. Sci.* **2019**, *24*, 157–170. [CrossRef]
10. Abedin, M.Z.; Chi, G.; Colombage, S.; Moula, F.E. Credit default prediction using a support vector machine and a probabilistic neural network. *J. Credit. Risk* **2018**, *14*, 1–27. [CrossRef]
11. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2019**, *14*, 241–258. [CrossRef]
12. Ying, C.; Qi-Guang, M.; Jia-Chen, L.; Lin, G. Advance and Prospects of AdaBoost Algorithm. *Acta Autom. Sin.* **2014**, *39*, 745–758.
13. Abedin, M.Z.; Guotai, C.; Hajek, P.; Zhang, T. Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex Intell. Syst.* **2022**, *9*, 3559–3579. [CrossRef]
14. Khan, F.N.; Khan, A.H.; Israt, L. Credit Card Fraud Prediction and Classification using Deep Neural Network and Ensemble Learning. In Proceedings of the 2020 IEEE Region 10 Symposium 2020, Dhaka, Bangladesh, 5–7 June 2020.
15. Yu, L.; Wang, S.; Lai, K.K. Developing an SVM-based ensemble learning system for customer risk identification collaborating with customer relationship management. *Front. Comput. Sci. China* **2010**, *4*, 196–203. [CrossRef]
16. He, H.; Fan, Y. A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. *Expert Syst. Appl.* **2021**, *176*, 114899. [CrossRef]
17. Murray, L.; Nguyen, H.; Lee, Y.-F.; Remmenga, M.D.; Smith, D.W. Variance Inflation Factors in Regression Models with Dummy Variables. In Proceedings of the Conference on Applied Statistics in Agriculture, Manhattan, Kansas, 29 April–1 May 2012.
18. Odhiambo Omuya, E.; Onyango Okeyo, G.; Waema Kimwele, M. Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Syst. Appl.* **2021**, *174*, 114765. [CrossRef]
19. Zizi, Y.; Jamali-Alaoui, A.; El Goumi, B.; Oudgou, M.; El Moudden, A. An Optimal Model of Financial Distress Prediction: A Comparative Study between Neural Networks and Logistic Regression. *Risks* **2021**, *9*, 200. [CrossRef]
20. Tian, S.; Yu, Y. Financial ratios and bankruptcy predictions: An international evidence. *Int. Rev. Econ. Financ.* **2017**, *51*, 510–526. [CrossRef]
21. Alonso Robisco, A.; Carbó Martínez, J.M. Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financ. Innov.* **2022**, *8*, 70. [CrossRef]
22. Lemmens, A.; Croux, C. Bagging and Boosting Classification Trees to Predict Churn. *J. Mark. Res.* **2006**, *43*, 276–286. [CrossRef]
23. Rutkowski, L.; Jaworski, M.; Pietruczuk, L.; Duda, P. The CART decision tree for mining data streams. *Inf. Sci.* **2014**, *266*, 1–15. [CrossRef]
24. Xia, M.; Wang, Z.; Han, F.; Kang, Y. Enhanced Multi-Dimensional and Multi-Grained Cascade Forest for Cloud/Snow Recognition Using Multispectral Satellite Remote Sensing Imagery. *IEEE Access* **2021**, *9*, 131072–131086. [CrossRef]
25. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *arXiv* **2017**, arXiv:1706.09516.
26. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2010**, *40*, 185–197. [CrossRef]
27. Gupta, A.; Jain, V.; Singh, A. Stacking Ensemble-Based Intelligent Machine Learning Model for Predicting Post-COVID-19 Complications. *New Gener. Comput.* **2021**, *40*, 987–1007. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.