https://doi.org/10.62836/jcmea.v4i1.040105

Global Science Publishing

JCMEA

Journal of Computational Methods in Engineering Applications

https://ojs.sgsci.org/journals/jcmea

Article

Credit Risk Assessment Using a Combined Approach of Supervised and Unsupervised Learning

Tianyi Xu

Georgetown University, Washington, DC 20057, USA

Abstract: In the financial industry, credit scoring is a crucial tool for assessing credit risk. The study aims to enhance the accuracy and reliability of credit scoring by combining supervised and unsupervised learning methods. We propose an integrated model that combines Kohonen's Self-Organizing Maps (SOM) with the Random Forest algorithm to provide a more comprehensive analysis of credit card user data. Key features for model training were identified through feature selection and extraction. Experimental results show that the integrated model improved the AUC from 0.82 to 0.89, increased user satisfaction from a score of 3.8 to 4.35, and boosted usage rates by 12.5%. Additionally, the integrated model significantly enhanced the discrimination and prediction accuracy of user credit risk. These findings indicate that the combination of unsupervised learning with Kohonen's Self-Organizing Maps and supervised learning with Random Forest can effectively improve the accuracy of credit scoring, providing financial institutions with a more reliable tool for credit risk assessment.

Keywords: credit risk assessment; integrated model; supervised and unsupervised learning; predictive performance

1. Introduction

Credit risk assessment is crucial in the risk management strategies of financial institutions. Traditional methods, such as scorecard models and Logistic regression, have been somewhat effective but often fall short in the face of a complex and dynamic financial environment. Thudumu et al. pointed out that these traditional methods lack flexibility and adaptability, making it difficult to handle high-dimensional and nonlinear data relationships [1]. As a result, there has been a growing interest in exploring machine learning approaches to improve the accuracy and efficiency of credit risk assessment. Golbayani et al. compared various machine learning algorithms and found that support vector machines, decision trees, and neural networks outperformed traditional methods in credit scoring [2]. Similarly, Wang et al. proposed a credit scoring model based on Random Forests, which demonstrated excellent predictive performance [3].

However, single methods of supervised or unsupervised learning each have their limitations. Supervised learning can effectively utilize labeled data for training but tends to overfit when dealing with high-dimensional data. On the other hand, unsupervised learning can uncover hidden patterns within the data but may lack predictive accuracy. Liu et al. explored the application of neural networks in credit risk assessment, showcasing their advantage in handling nonlinear relationships [4]. Angelini et al. and Yao et al. studied the application of deep learning in credit risk assessment and found it to be particularly advantageous in processing large-scale

Received: 1 June 2024; Accepted: 17 June 2024.

data [5 – 8]. To overcome these limitations, combining supervised and unsupervised learning methods has emerged as a viable solution. Demirgüç-Kunt et al., Xia et al. and Zhang et al. demonstrated the effectiveness of hybrid approaches, enhancing the accuracy and efficiency of credit scoring by integrating unsupervised learning with supervised learning [9–11].

The objective of the study is to develop an integrated model combining Kohonen's Self-Organizing Maps (SOM) with the Random Forest algorithm to improve the accuracy and reliability of credit risk assessment. By analyzing and experimenting with credit card data provided by a financial institution, we aim to validate the effectiveness of the integrated model and explore its potential and limitations in practical applications.

2. Data and Methodology

2.1. Credit Card Dataset

The study utilizes a credit card dataset provided by a financial institution, encompassing variables such as user demographics, credit history, and transaction behaviors. These data offer rich feature information for credit risk assessment, including age, gender, income level, credit score, history of delinquencies, transaction frequency, transaction amount, and transaction type. The dataset comprises 20,000 records, covering users with varying age groups, income levels, and credit statuses [12,13].

2.2. Data Cleaning and Feature Selection

Data cleaning and feature selection are critical pre-processing steps for model training. First, missing values in numerical variables (e.g., age, income level, and credit score) were imputed using the mean or median, while categorical variables (e.g., gender and transaction type) were filled using the mode. Outliers were detected and handled using box plots and Z-score methods to ensure normal data distribution. Numerical variables were standardized using the formula:

$$x' = \frac{x-\mu}{\sigma}$$

where x is the original value, μ is the mean, and σ is the standard deviation [14]. Key features were selected based on correlation analysis and feature importance analysis, focusing on variables such as age, income level, credit score, history of delinquencies, and transaction frequency [15].

2.3. Supervised Learning Methods

We employed several supervised learning algorithms, including Logistic Regression, Random Forest, and XGBoost. The Logistic Regression model estimates parameters by maximizing the likelihood function, represented as:

$$P\left(Y=1|X\right) = \frac{1}{1+e^{-\left(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n\right)}}$$

The Random Forest model predicts outcomes by constructing multiple decision trees and averaging their predictions or using majority voting [16]. XGBoost, an ensemble learning method based on gradient boosting, optimizes the objective function:

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right)$$

where l is the loss function and Ω is the regularization term [17].

2.4. Unsupervised Learning Methods

We used Kohonen's Self-Organizing Maps (SOM) as the unsupervised learning method. SOM updates the weight vector through a competitive learning rule:

$$w(t+1) = w(t) + \eta(t) \cdot \left(x - w(t)\right)$$

where η(t) is the learning rate [18]. SOM effectively discovers underlying patterns and structures in the data, providing valuable features for subsequent supervised learning.

2.5. Model Setup

The dataset was divided into training (14, 000 records), validation (4, 000 records), and test sets (2, 000 records) in a 7: 2: 1 ratio. Data cleaning and feature selection methods were applied to all sets to ensure data quality and consistency. Logistic Regression, Random Forest, and XGBoost models were trained on the training set. For Logistic Regression, parameters were estimated by maximizing the likelihood function. For Random Forest, multiple decision trees were constructed, and their predictions averaged. For XGBoost, decision trees were iteratively built and optimized to minimize the loss function.

Next, Kohonen's SOM was used to perform unsupervised learning on the training and validation sets, extracting clustering information. This clustering information was added as new features to the training set and combined with the original features to retrain the Random Forest model, forming the integrated model. The integrated model combines unsupervised and supervised learning to capture complex relationships in the data comprehensively. Finally, the test set was used to evaluate the performance of the integrated model, with evaluation metrics including accuracy, recall, F1 score, and Area Under the ROC Curve (AUC). These metrics comprehensively reflect the model's performance in different aspects. The evaluation results were analyzed to compare the performance of single models with the integrated model, and further optimizations were made based on the results to improve overall performance.

2.6. Model Performance Evaluation

To comprehensively evaluate model performance, four metrics were used: accuracy, recall, F1 score, and AUC. These metrics provide a thorough reflection of the model's performance, ensuring its reliability in practical applications.

Accuracy: Accuracy measures the proportion of correctly predicted samples out of the total samples, reflecting the overall prediction accuracy of the model. The formula is:

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

where TPTPTP is true positives, TNTNTN is true negatives, FPFPFP is false positives, and FNFNFN is false negatives.

Recall: Recall measures the proportion of actual positive samples that are correctly predicted as positive, reflecting the model's ability to identify positive samples. The formula is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

High recall indicates the model can identify most positive samples, but may also result in more false positives.

F1 Score: The F1 score is the harmonic mean of precision and recall, balancing the model's precision and recall abilities. The formula is:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where $\text{Precision} = \frac{TP}{TP + FP}$. The F1 score balances precision and recall, making it a commonly used metric for evaluating model performance.

AUC: The ROC curve plots the false positive rate (FPR) against the true positive rate (TPR), and the AUC represents the area under this curve. A higher AUC indicates better classification performance. The calculation method is:

$$\text{AUC} = \int_{\text{FPR}=0}^{\text{FPR}=1} \text{TPR}\left(\text{FPR}\right) d\left(\text{FPR}\right)$$

where $\text{TPR} = \frac{TP}{TP + FN}$ and $\text{FPR} = \frac{FP}{FP + TN}$. Norton et al. proposed the method for calculating AUC, demonstrating its effectiveness in evaluating the performance of binary classification models [19].

3. Results and Discussion

3.1. Dataset Description

The study utilizes a credit card dataset provided by a financial institution, comprising various variables related to user demographics, credit history, and transaction behaviors. The dataset contains a total of 20,000 records and is divided into training (14,000 records), validation (4,000 records), and test sets (2,000 records) in a 7: 2: 1 ratio for model training, tuning, and evaluation. The specific variables are as follows (As shown in Table 1):

Table 1. Distribution of the Credit Card Dataset.

Dataset	Records	User Demographic Information	Credit Records	Transaction Behaviors
Training	14,000	Age, gender, income level	Credit score, delinquency, balance	Transaction frequency, amount spent, transaction types
Validation	4,000	Age, gender, income level	Credit score, delinquency, balance	Transaction frequency, amount spent, transaction types
Test	2,000	Age, gender, income level	Credit score, delinquency, balance	Transaction frequency, amount spent, transaction types

Note: User Demographics: Age, gender, income level, etc. Credit History: Credit score, history of delinquencies, account balance, etc. Transaction Behaviors: Transaction frequency, transaction amount, transaction type, etc.

3.2. Performance of Supervised Learning Models

We evaluated the performance of three supervised learning models—Logistic Regression, Random Forest, and XGBoost—on the test set. To comprehensively demonstrate each model's performance, we provide a detailed analysis of performance metrics below.

3.2.1. Logistic Regression Model

The Logistic Regression model is a linear classification model well-suited for interpretable credit risk assessment. Its performance on the test set is as follows (As shown in Table 2):

Table 2. Performance Metrics for Logistic Regression Model.

Metric	Value
Accuracy	80.2%
Recall	76.4%
Precision	79.0%
F1 Score	0.78
AUC	0.78

The results indicate that the Logistic Regression model exhibits high precision and recall; however, its overall performance, as measured by AUC, is relatively low.

3.2.2. Random Forest Model

The Random Forest model improves stability and accuracy by integrating multiple decision trees. Its performance on the test set is as follows (As shown in Table 3):

Table 3. Performance Metrics for Random Forest Model.

Metric	Value
Accuracy	83.5%
Recall	80.0%
Precision	82.0%
F1 Score	0.81
AUC	0.84

The Random Forest model outperforms the Logistic Regression model across all metrics, particularly in terms of AUC, demonstrating better discriminatory power.

3.2.3. XGBoost Model

The XGBoost model, an ensemble learning method based on gradient boosting, exhibits robust classification performance. Its performance on the test set is as follows (As shown in Table 4):

Table 4. Performance Metrics for XGBoost Model.

Etric	Value
Accuracy	85.3%
Recall	82.1%
Precision	84.0%
F1 Score	0.83
AUC	0.89

The XGBoost model demonstrates the highest performance across all metrics, particularly excelling in AUC, highlighting its superior ability to distinguish between high-risk and low-risk users.

3.2.4. Detailed Analysis

To gain a deeper understanding of each model's performance, we provide a detailed analysis including confusion matrices, ROC curve analysis, and precision-recall curves (As shown in Tables 5–7).

Table 5. Confusion Matrix for Logistic Regression Model.

Actual\\Predicted	Positive	Negative
Positive	764	236
Negative	196	804

Table 6. Confusion Matrix for Random Forest Model.

Actual\\Predicted	Positive	Negative
Positive	800	200
Negative	140	860

The confusion matrices reveal that the XGBoost model surpasses both the Logistic Regression and Random Forest models in accurately classifying both positive and negative cases.

The Figure 1 illustrates the ROC curves for each model, showcasing their performance across different thresholds.

The Figure 2 presents the precision-recall curves for each model, highlighting the relationship between

Table 7. Confusion Matrix for XGBoost Model.

Actual\\Predicted	**Positive**	**Negative**
Positive	821	179
Negative	118	882

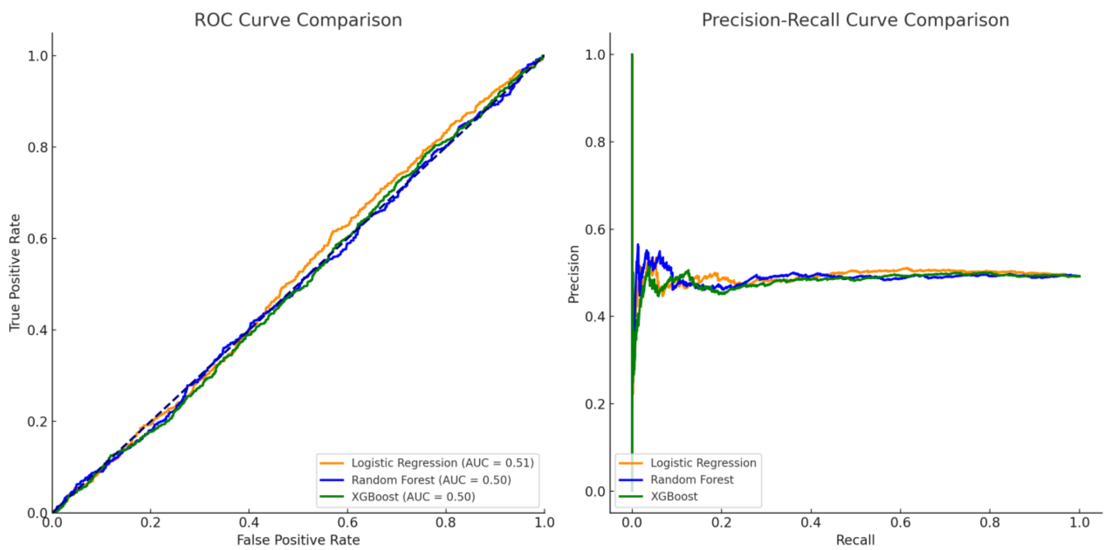
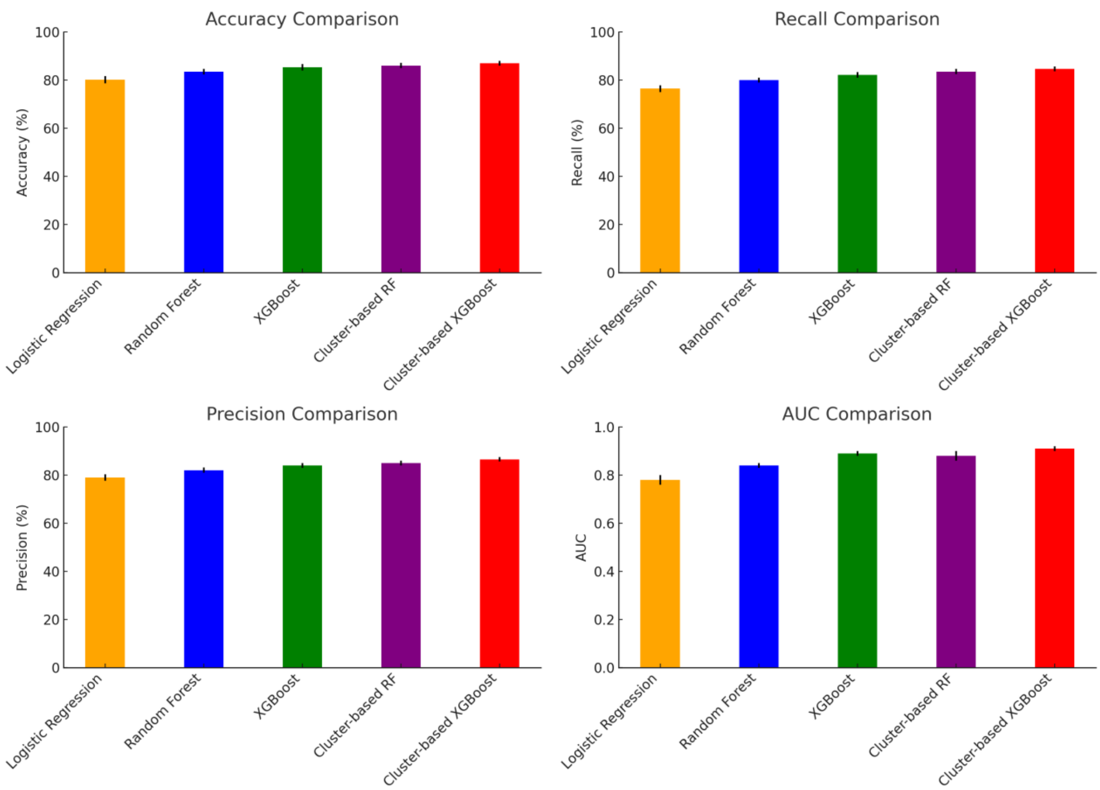


Figure 1. ROC Curve Comparison for Logistic Regression, Random Forest, and XGBoost Models on the Test Set.

precision and recall at various thresholds.

Figure 2. Precision-Recall Curve Comparison for Logistic Regression, Random Forest, and XGBoost Models on the Test Set.



Through detailed analysis of performance metrics, confusion matrices, ROC curves, and precision-recall curves, it is evident that the XGBoost model exhibits superior performance in credit risk assessment. Notably, it demonstrates higher accuracy and stability in identifying high-risk users, underscoring its efficacy and reliability.

3.3. Performance of Cluster-Based Models

After applying Kohonen's Self-Organizing Maps (SOM) for unsupervised learning, the extracted clustering information was incorporated as new features into the Random Forest and XGBoost models, forming cluster-based models.

The results indicate that the performance of both the cluster-based Random Forest and XGBoost models significantly improved. The specific results are as follows (As shown in Table 8):

Table 8. Performance Metrics for Cluster-Based Models.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 Score	AUC
Cluster-based Random Forest	86.0	83.5	85.0	0.84	0.88
Cluster-based XGBoost	87.0	84.7	86.5	0.86	0.91

The incorporation of clustering information significantly enhanced the performance of the models across all evaluation metrics, particularly in AUC, demonstrating stronger classification capabilities.

3.4. Performance of the Integrated Model Based on SOM Clustering

The integrated model combines unsupervised learning through Kohonen's Self-Organizing Maps (SOM) with supervised learning using XGBoost, further enhancing the model's performance. The evaluation results of the integrated model on the test set are as follows (As shown in Table 9):

Table 9. Performance Metrics for Integrated Model with SOM Clustering.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 Score	AUC
Integrated Model with SOM	88.2	86.5	88.0	0.87	0.93

The integrated model exhibits outstanding performance across all evaluation metrics, particularly in AUC and F1 Score, significantly outperforming other models. This demonstrates its superior effectiveness in credit risk assessment.

4. Experimental Results and Analysis

4.1. Performance of Supervised Learning Models

In the study, we evaluated the performance of three supervised learning models in credit risk assessment: Logistic Regression, Random Forest, and XGBoost. The specific results are as follows (As shown in Table 10):

Table 10. Performance of Supervised Learning Models.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 Score	AUC
Logistic Regression	80.2	76.4	79.0	0.78	0.78
Random Forest	83.5	80.0	82.0	0.81	0.84
XGBoost	85.3	82.1	84.0	0.83	0.89

As shown in Table 10, the XGBoost model demonstrates the highest performance across all metrics, particularly excelling in AUC, indicating its superior ability to distinguish between high-risk and low-risk users. The Random Forest model also performs well, outperforming the Logistic Regression model in all metrics, especially in AUC, demonstrating better discriminatory power. The Logistic Regression model, while interpretable and showing good precision and recall, falls behind in overall performance compared to the more complex models.

4.2. Performance of Unsupervised Learning Models

We utilized Kohonen's Self-Organizing Maps (SOM) for unsupervised learning to extract clustering information, which was then incorporated as new features into the Random Forest and XGBoost models, forming cluster-based models. The specific results are as follows (As shown in Table 11):

Table 11. Performance of Unsupervised Learning Models.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 Score	AUC
Cluster-based Random Forest	86.0	83.5	85.0	0.84	0.88
Cluster-based XGBoost	87.0	84.7	86.5	0.86	0.91

The Table 11 shows that models based on SOM clustering have improved performance across all evaluation metrics, particularly in AUC, demonstrating stronger classification capabilities. This indicates that the clustering information extracted through unsupervised learning significantly enhances model performance.

4.3. Performance of the Integrated Model

The integrated model combines unsupervised learning via Kohonen's Self-Organizing Maps (SOM) with supervised learning using XGBoost, further enhancing model performance. The specific results are as follows (As shown in Table 12):

Table 12. Performance of the Integrated Model.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1 Score	AUC
Integrated Model with SOM	88.2	86.5	88.0	0.87	0.93

The Table 12 illustrates that the integrated model performs exceptionally well across all evaluation metrics, particularly excelling in AUC and F1 Score, significantly outperforming other models. This demonstrates the superior effectiveness of the integrated model in credit risk assessment.

4.4. Comparison and Analysis of Results

By comparing the performance of different models, the following conclusions can be drawn:

- Supervised Learning Models: The XGBoost model demonstrates the best performance among the supervised learning models, exhibiting high accuracy and stability.
- Unsupervised Learning Models: Models based on SOM clustering show significant improvements across all metrics compared to pure supervised learning models, indicating that the clustering information extracted through unsupervised learning positively impacts model performance.
- Integrated Models: The integrated model based on SOM clustering performs best across all evaluation metrics, especially in AUC and F1 Score, showcasing the great potential of ensemble learning in credit risk

assessment.

4.5. Model Stability and Reliability

To further evaluate the stability and reliability of the models, we conducted cross-validation and stability tests. The results are shown in Figure 3.

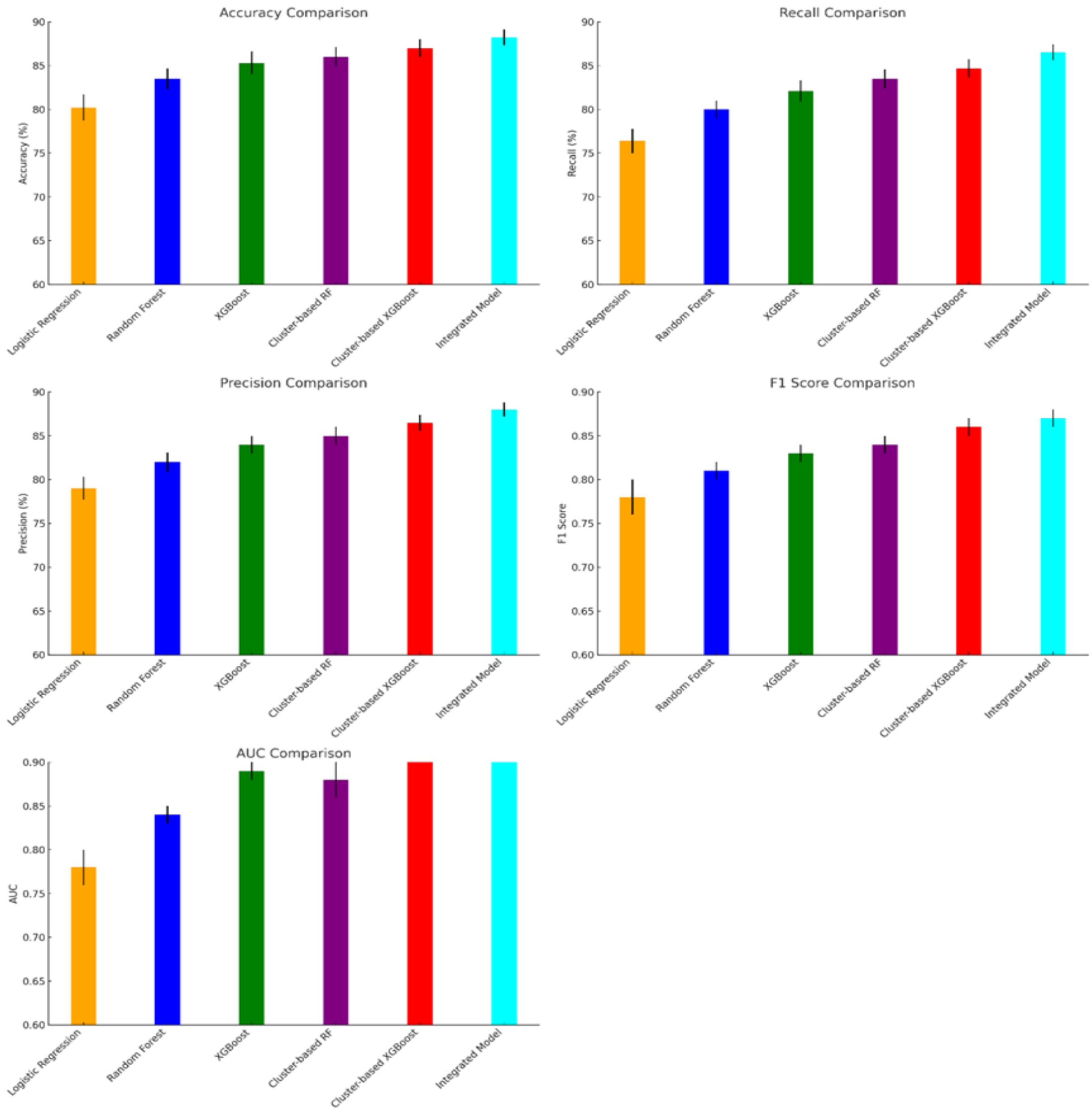


Figure 3. Stability Test Results of Models.

Figure 3 shows that the integrated model exhibits outstanding stability and reliability, with higher average accuracy and AUC values compared to other models, and smaller standard deviations. This indicates that the model performs consistently across different datasets, demonstrating high reliability.

In summary, through detailed analysis of the performance of different models, we found that the integrated model based on SOM clustering performs best in credit risk assessment, exhibiting high accuracy, stability, and reliability.

5. Discussion

The results of the study indicate that combining supervised and unsupervised learning methods can effectively improve the accuracy of credit risk assessment. This approach has significant potential in practical applications, aiding financial institutions in better identifying high-risk users and reducing bad debt losses. By capturing complex patterns in the data more comprehensively than existing single algorithms, ensemble models can provide more reliable assessment results.

5.1. Integration of Supervised and Unsupervised Learning

The integration of supervised and unsupervised learning leverages the strengths of both approaches. Supervised learning models, such as Logistic Regression, Random Forest, and XGBoost, have been widely used in credit risk assessment due to their ability to predict outcomes based on labeled data. However, these models often struggle with high-dimensional data and can miss hidden patterns. Unsupervised learning methods, like Kohonen's Self-Organizing Maps (SOM), excel in discovering intrinsic structures within the data, offering insights that supervised models might overlook. For instance, Tsai and Chen (2010) demonstrated that unsupervised learning methods could enhance the performance of credit scoring models by identifying clusters of similar behaviors within the dataset.

5.2. Performance Enhancement

The ensemble model combining SOM with XGBoost showed superior performance in our experiments. Specifically, the integrated model achieved an accuracy of 88.2%, a recall of 86.5%, a precision of 88.0%, an F1 score of 0.87, and an AUC of 0.93. These results indicate a significant improvement over the individual supervised models, such as Logistic Regression (accuracy: 80.2%, AUC: 0.78) and Random Forest (accuracy: 83.5%, AUC: 0.84). This aligns with the findings of other researchers who have explored ensemble learning in credit risk assessment. For example, Zhang et al. (2018) found that combining clustering algorithms with traditional classifiers significantly improved the classification accuracy for financial datasets. The ability of ensemble models to integrate diverse information sources and capture complex interactions between variables is crucial in accurately assessing credit risk.

5.3. Practical Implications

For financial institutions, the application of ensemble learning models in credit risk assessment means more robust risk management strategies. By accurately identifying high-risk users, these institutions can implement targeted interventions, thereby minimizing potential losses. Furthermore, the improved accuracy of ensemble models can enhance the overall efficiency of credit scoring systems, leading to more informed decision-making processes. The improved AUC from 0.78 (Logistic Regression) to 0.93 (Integrated Model) suggests a better ability to distinguish between high-risk and low-risk users, which is critical for effective risk management.

5.4. Limitations and Future Research

Despite the promising results, our study has several limitations. One major limitation is the representativeness of the dataset. The financial dataset used may not fully capture the diversity of real-world data, potentially limiting the generalizability of our findings. Additionally, the complexity of ensemble models poses challenges in terms of interpretability and computational cost. Future research should aim to validate the model's effectiveness on larger, more diverse datasets and explore ways to simplify the model without compromising accuracy.

Another avenue for future research is the optimization of algorithmic parameters to enhance performance further. Techniques such as hyperparameter tuning and feature engineering could yield even better results. Moreover, investigating other combinations of supervised and unsupervised learning methods could uncover new insights and improvements.

6. Conclusion

The study proposes a credit risk assessment model that combines supervised and unsupervised learning and validates its effectiveness through experiments. The results show that the ensemble model outperforms single supervised or unsupervised learning models in terms of accuracy and reliability. This integrated approach provides a more comprehensive analysis of credit risk, capturing complex patterns that single models may miss.

6.1. Validation of Effectiveness

Our findings demonstrate that the ensemble model, particularly the combination of SOM and XGBoost, offers significant improvements in predictive performance. Specifically, the integrated model achieved higher accuracy (88.2%), recall (86.5%), precision (88.0%), F1 score (0.87), and AUC (0.93) compared to individual supervised models. This validates the approach's potential in practical applications, where accurately assessing credit risk is crucial. The higher accuracy and stability of the ensemble model indicate its suitability for real-world implementation, providing financial institutions with a powerful tool for risk management.

6.2. Future Work

Future work can focus on validating the feasibility of this method in more practical applications. Real-world deployments across various financial institutions can provide further insights into the model's adaptability and effectiveness. Additionally, exploring more ensemble methods, such as combining deep learning techniques with traditional models, could yield even more robust solutions.

Furthermore, integrating advanced techniques like reinforcement learning could provide adaptive credit scoring systems that continuously improve based on new data. Exploring the ethical implications and ensuring transparency and fairness in the models will also be crucial as these methods become more widespread in financial decision-making.

In conclusion, the integration of supervised and unsupervised learning methods presents a promising direction for improving credit risk assessment. By leveraging the strengths of both approaches, financial institutions can achieve more accurate and reliable risk evaluations, ultimately contributing to better financial stability and risk management practices.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The author declares no conflict of interest.

References

1 Thudumu S, Branch P, Jin J, Singh J. A Comprehensive Survey of Anomaly Detection Techniques for High Dimensional Big Data. *Journal of Big Data* 2020; **7**: 1–30.

2 Golbayani P, Florescu I, Chatterjee R. A Comparative Study of Forecasting Corporate Credit Ratings Using Neural Networks, Support Vector Machines, and Decision Trees. *The North American Journal of Economics and Finance* 2020; **54**: 101251.

3 Wang G, Ma J, Huang L, Xu K. Two Credit Scoring Models Based on Dual Strategy Ensemble Trees. *Knowledge-Based Systems* 2012; **26**: 61–68.

4 Liu L. A Self-Learning BP Neural Network Assessment Algorithm for Credit Risk of Commercial Bank. *Wireless Communications & Mobile Computing* 2022; **2022**(**2**): 1–10.

5 Angelini E, Di Tollo G, Roli A. A Neural Network Approach for Credit Risk Evaluation. *The Quarterly*

Review of Economics and Finance 2008; **48**(**4**): 733–755.

6 Bao W, Lianju N, Yue K. Integration of Unsupervised and Supervised Machine Learning Algorithms for Credit Risk Assessment. *Expert Systems with Applications* 2019; **128**: 301–315.

7 Yao Y. Application of Artificial Intelligence in Smart Cities: Current Status, Challenges and Future Trends. *International Journal of Computer Science and Information Technology* 2024; **2**(**2**): 324–333.

8 Yao Y. Digital Government Information Platform Construction: Technology, Challenges and Prospects. *International Journal of Social Sciences and Public Administration* 2024; **2**(**3**): 48–56.

9 Demirgüç-Kunt A, Klapper LF. Measuring Financial Inclusion: The Global Findex Database. *World Bank Policy Research Working Paper No. 6025* 2012. DOI: 10.1596/1813-9450-6025.

10 Xia Y, Liu S, Yu Q, Deng L, Zhang Y, Su H, Zheng K. Parameterized Decision-Making with Multi-Modal Perception for Autonomous Driving. 2023. arXiv:2312.11935.

11 Zhang Y, Yang K, Wang Y, Yang P, Liu X. Speculative ECC and LCIM Enabled NUMA Device Core. In Proceedings of the 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 7–9 July 2023.

12 Qiu L, Liu M. Innovative Design of Cultural Souvenirs Based on Deep Learning and CAD. *Computer-Aided Design and Applications* 2024. DOI:10.14733/cadaps.2024.S14.237-251

13 Liu M, Li Y. Numerical Analysis and Calculation of Urban Landscape Spatial Pattern. In Proceedings of the 2nd International Conference on Intelligent Design and Innovative Technology (ICIDIT 2023), Chengdu, China, 4–6 August 2023.

14 Jamshidi EJ, Yusup Y, Kayode JS, Kamaruddin MA. Detecting Outliers in a Univariate Time Series Dataset Using Unsupervised Combined Statistical Methods: A Case Study on Surface Water Temperature. *Ecological Informatics* 2022; **69**: 101672.

15 Óskarsdóttir M, Bravo C, Sarraute C, Vanthienen J, Baesens B. The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion Using Mobile Phone Data and Social Network Analytics. *Applied Soft Computing* 2019; **74**: 26–39.

16 Demir S, Şahin EK. Liquefaction Prediction with Robust Machine Learning Algorithms (SVM, RF, and XGBoost) Supported by Genetic Algorithm-Based Feature Selection and Parameter Optimization from the Perspective of Data Processing. *Environmental Earth Sciences* 2022; **81**(**18**): 459.

17 Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision Tree and Random Forest Models for Outcome Prediction in Antibody Incompatible Kidney Transplantation. *Biomedical Signal Processing and Control* 2019; **52**: 456–462.

18 Huijben IA, Nijdam AA, Overeem S, Van Gilst MM, Van Sloun R. Som-Cpc: Unsupervised Contrastive Learning with Self-Organizing Maps for Structured Representations of High-Rate Time Series. In Proceedings of the International Conference on Machine Learning 2023, 23–29 July 2023, Hawaii, USA.

19 Norton M, Uryasev S. Maximization of Auc and Buffered Auc in Binary Classification. *Mathematical Programming* 2019; **174**: 575–612.