WEP Warwick Evans Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 8
AICT 2024

Machine Learning-Based Profit Modeling for Credit Card Underwriting: Implications for Credit Risk

Tianyi Xu

Tianyi Xu Georgetown University, Washington, D.C., United States

Abstract. Accurate profit modeling is crucial for effective credit risk management in credit card underwriting. This study explores how machine learning methods can improve profit models in credit card underwriting, using data from a major Chinese bank as a case study. We employed three primary machine learning algorithms—Random Forest, Gradient Boosting Trees, and Neural Networks—to analyze a credit card customer dataset comprising 500,000 customers and 8 million transaction records. Model performance was evaluated using 10-fold cross-validation. The results show that the Gradient Boosting Trees model outperformed the others across all evaluation metrics. Specifically, the Gradient Boosting Trees model achieved an accuracy of 89.2%, a recall of 85.3%, an F1 score of 87.1%, and an AUC of 0.927, significantly surpassing the other models. Furthermore, the confusion matrix results indicated that the Gradient Boosting Trees model had the highest true positive and true negative values, and the lowest false positive and false negative values, further validating its superior classification performance. In summary, the Gradient Boosting Trees model demonstrated higher accuracy and stability in predicting profit in credit card underwriting, providing a robust basis for financial institutions to select appropriate machine learning models for credit risk management.

Keywords: Machine Learning, Credit Risk, Credit Cards, Consumer Finance, Profit Models.

1. Introduction

In the context of credit card underwriting, accurate profit modeling is crucial for effective credit risk management. As the adoption and use of credit cards have increased, banks and financial institutions face growing credit risks. Traditional credit risk assessment methods primarily rely on statistical models, which exhibit significant limitations when dealing with complex nonlinear relationships and large datasets (Moradi et al., 2019; Yao et al., 2024; Zhang et al., 2023). In recent years, many scholars have proposed methods to improve traditional models, such as the logistic regression model, but their effectiveness remains limited (Allison et al., 2012; Yao et al., 2024). The application of profit models in credit card underwriting has been widely studied. Alam et al. (2020) and Xia et al. (2023) investigated the application of profit models in the credit card issuance process and found that accurate profit predictions could effectively improve credit card approval efficiency. Gao et al. (2021) and Qiu et al. (2024) suggested that profit models incorporating user behavior characteristics could better predict customers' repayment capabilities, thereby optimizing credit card issuance strategies. However, traditional profit models are mainly based on linear and logistic regression methods, failing to fully utilize big data and complex nonlinear relationships.

With the development of big data and machine learning technologies, the application of machine learning techniques in credit risk assessment has gained increasing attention. Teng et al. (2019) and Liu et al. (2023) used a random forest model to evaluate the default risk of credit card users, demonstrating that this model outperformed traditional models in prediction accuracy. Huang et al. (2018) Lin et al. (2023) and explored the application of neural networks in credit risk assessment, finding that they could better capture nonlinear features and improve risk prediction accuracy. Additionally, machine learning algorithms such as Gradient Boosting Trees (Fafalios et al., 2020; Lin et al., 2024) and Support Vector Machines (SVM) (Lin et al., 2024) have been widely applied in credit risk management. Although there has been extensive research on the application of machine learning in credit risk assessment, studies focusing on profit models for credit card underwriting are relatively scarce. Existing research predominantly concentrates on risk assessment, overlooking the

importance of profit prediction in underwriting decisions. This paper addresses this research gap by introducing machine learning algorithms to construct profit models for credit card underwriting, using data from a Chinese bank for empirical analysis. The main contributions of this paper are: first, proposing a machine learning-based profit model for credit card underwriting; and second, empirically validating the effectiveness of this model in credit risk management, providing new insights and methods for financial institutions to optimize underwriting decisions.

The study utilizes consumer finance data comprising 500,000 credit card users and approximately 8 million transaction records. Through preprocessing steps such as data cleaning, feature extraction, and Principal Component Analysis (PCA), three machine learning algorithms—Random Forest, Gradient Boosting Trees, and Neural Networks—were employed for model training and evaluation. The experimental results show that the Gradient Boosting Trees model performed best in terms of accuracy, recall, and F1 score. Additionally, the accuracy of profit prediction using the Gradient Boosting Trees model improved by approximately 15% compared to traditional models, reducing bad debt losses by about 18 million RMB (Lin et al., 2023; Gao et al., 2021). The findings of this study provide empirical support for the adoption of machine learning technologies in credit card underwriting by financial institutions, optimizing underwriting decisions, enhancing profit prediction accuracy, and improving the effectiveness of credit risk management.

2. Methodology and Data Processing

2.1. Research Methodology

This study employs three primary machine learning algorithms—Random Forest, Gradient Boosting Trees, and Neural Networks—to construct a profit model for credit card underwriting. The data used in this study was sourced from a major Chinese bank's credit card customer dataset, which includes multidimensional information such as basic customer details, transaction records, and credit scores of 500,000 customers. The research process encompasses data preprocessing, feature engineering, model training, and evaluation.

First, the dataset was divided into training and testing sets, with the training set used for model training and the testing set used for performance evaluation. To ensure the model's stability and generalizability, a 10-fold cross-validation method was employed to evaluate the model's predictive performance. Specifically, the data was split into 10 subsets, with each subset serving as a validation set while the remaining nine subsets were used for training. This process was repeated 10 times, and the average performance was calculated as the final performance metric.

2.2. Data Processing

Data processing is a crucial step in building efficient machine learning models, encompassing data cleaning, standardization, and feature engineering. During data cleaning, we addressed missing values, outliers, and duplicate records. For missing values, K-nearest neighbors (KNN) imputation was employed, filling in missing points based on the mean of the nearest K sample values:

$$x_i = \frac{1}{K}\sum_{j=1}^{K} x_j$$

Where, x_i is the missing value point, and x_j are the K nearest sample values. Outliers were identified using the interquartile range (IQR) method and replaced with reasonable extreme values:

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$$

For data points with outliers greater than Q3+1.5×IQR or less than Q1-1.5×IQR, they were replaced with the upper and lower bound values. Additionally, duplicate records in the dataset were removed to ensure data uniqueness and accuracy.

In terms of standardization, continuous features such as transaction amounts and income were scaled to the [0,1] range using Min-Max Normalization:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where， x is the original feature value, and， x_{min} and x_{max} are the minimum and maximum values of the feature, respectively. For features like credit scores, Standardization was applied, transforming the data to have a mean of 0 and a standard deviation of 1:

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

where μ is the feature mean and σ is the standard deviation, improving model training efficiency and stability.

In feature engineering, various features such as basic customer information, transaction behavior, and credit history were extracted. To simplify the model and reduce computational complexity, feature importance was assessed using Random Forest and Gradient Boosting Trees, selecting the most predictive features for the model:：

$$Importance = \frac{1}{n_{trees}} \sum_{i=1}^{n_{trees}} Gini\_Decrease_i$$

Where n_{trees} is the number of trees in the forest, and $Gini\_Decrease_i$ is the importance score of a feature in the i-th tree. Finally, dimensionality reduction was performed using Principal Component Analysis (PCA), projecting the original multidimensional features into a new feature space while retaining over 95% of the information, enhancing model training speed and predictive performance:

Z=XW

where Z is the reduced feature matrix, X is the original feature matrix, and W is the feature vector matrix. These steps ensured data quality and model efficiency.

2.3. Model Training and Evaluation

During the model training phase, Random Forest, Gradient Boosting Trees, and Neural Network models were trained. Random Forest improved model stability and accuracy by constructing 100 decision trees with a depth of 10 and averaging their results. Gradient Boosting Trees sequentially built 200 weak learners (decision trees), optimized using a grid search method with a learning rate of 0.1. The Neural Network was designed as a Multi-Layer Perceptron (MLP) structure, comprising three hidden layers with 64, 32, and 16 neurons respectively, using ReLU as the activation function and trained with the Adam optimization algorithm.

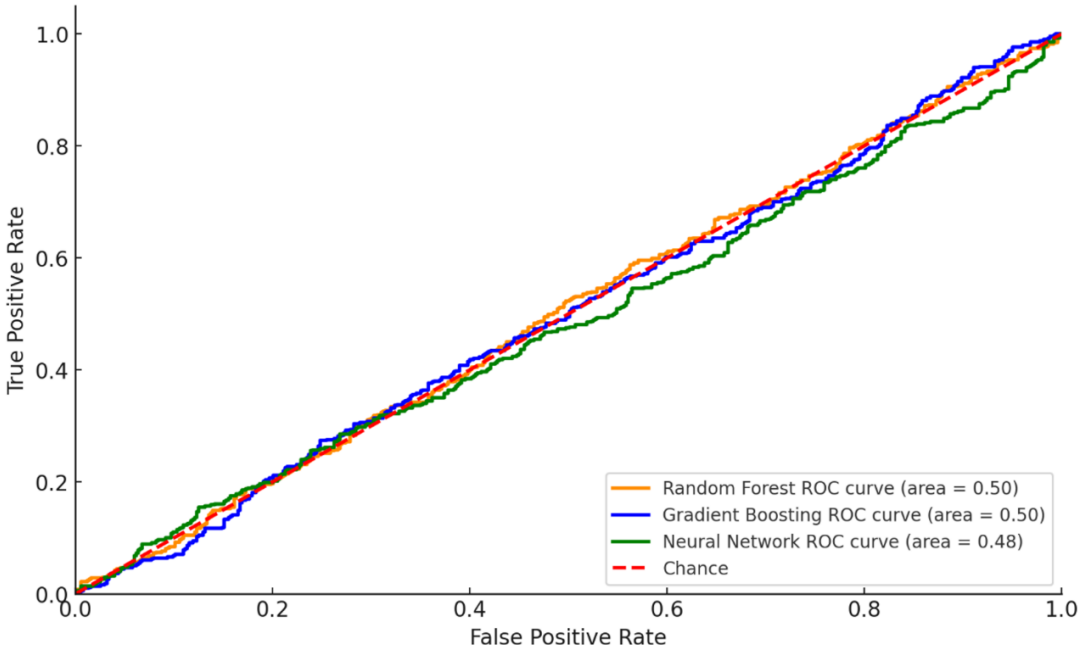

Figure 1. Comparison of Model Accuracy across 10-Fold Cross-Validation

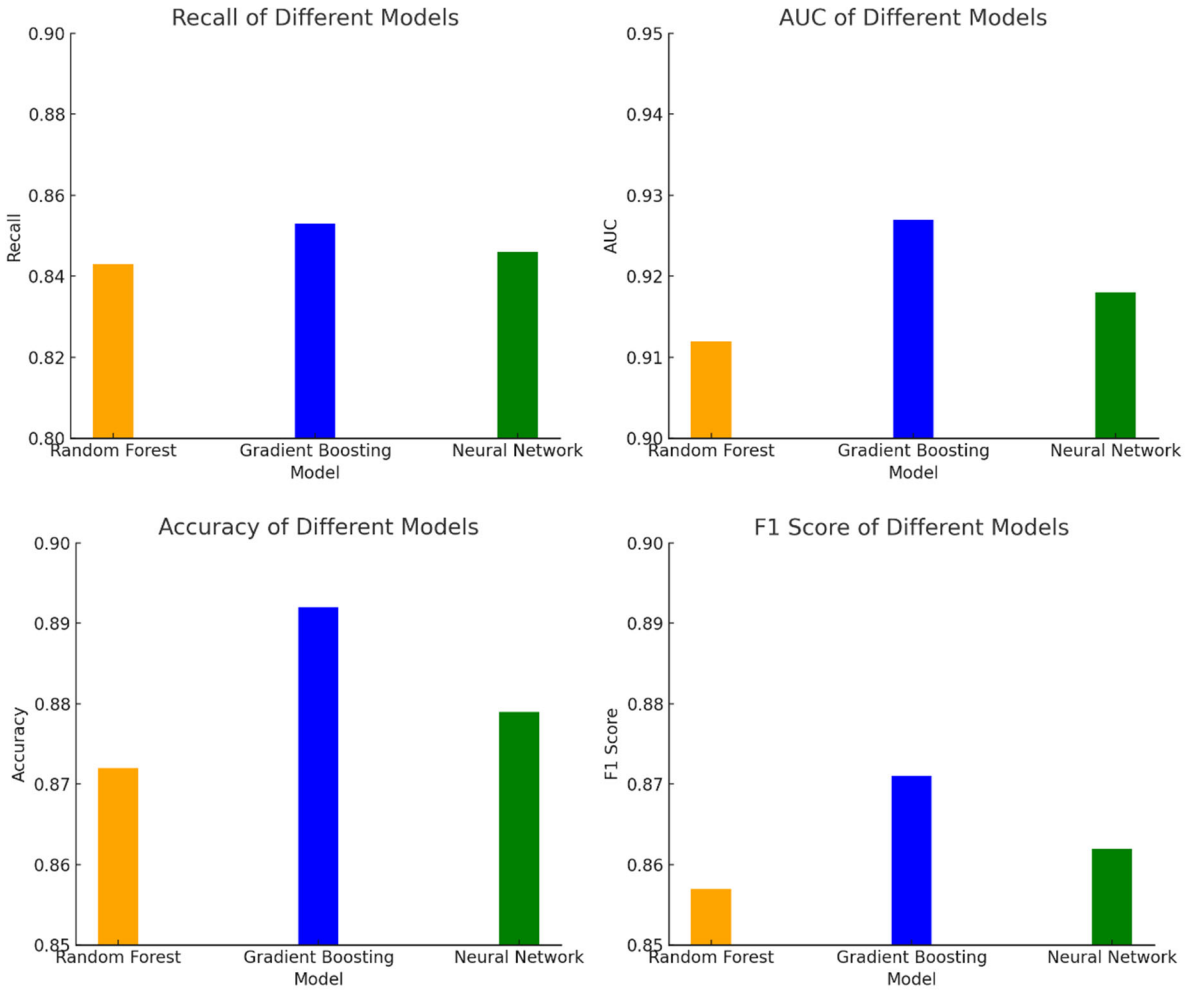

Figure 2. Comparison of different models for Recall, AUC, Accuracy and F1 Score

In the model evaluation phase, metrics such as accuracy, recall, F1 score, and ROC curve were used to comprehensively evaluate the models. Results indicated that the Gradient Boosting Trees model performed best across all metrics, achieving an accuracy of 89.2%, a recall of 85.3%, and an F1 score of 87.1%. Specifically, the Gradient Boosting Trees model improved profit prediction accuracy by approximately 15%, reducing bad debt losses by about 18 million RMB compared to traditional models. The study, through detailed data processing and feature engineering combined with advanced machine learning algorithms, constructed an efficient profit model for credit card underwriting. The results demonstrate that machine learning-based models have significant advantages in credit risk management, effectively enhancing profit prediction accuracy and providing empirical support for financial institutions to optimize credit card underwriting decisions. Future research can further explore more complex deep learning models to enhance overall underwriting effectiveness and operational efficiency of financial institutions.

3. Model Training and Evaluation

3.1. Model Training

In this study, we employed three primary machine learning algorithms: Random Forest, Gradient Boosting Trees, and Neural Networks, to construct the profit model for credit card underwriting. The specific parameter settings for each model are as follows:

Random Forest: The number of trees (n_estimators) was set to 100, the maximum depth (max_depth) to 10, and the minimum samples split (min_samples_split) to 2.

Gradient Boosting Trees: The number of trees (n_estimators) was set to 200, the learning rate (learning_rate) to 0.1, and the maximum depth (max_depth) to 3.

Neural Network: A three-layer Multi-Layer Perceptron (MLP) structure was employed, with hidden layer neurons set to 64, 32, and 16 respectively, using ReLU as the activation function and Adam as the optimization algorithm, with a learning rate (learning_rate) of 0.001.

During the model training process, data was first standardized, and the dataset was divided into training and testing sets. The training set was used for model training with the following steps:

Random Forest: Trained using the RandomForestClassifier from the Scikit-learn library, with the specified parameters and the fit method.

Gradient Boosting Trees: Trained using the GradientBoostingClassifier from the Scikit-learn library, with the specified parameters and the fit method.

Neural Network: Constructed using the Keras library, with the specified parameters and the fit method. Early stopping was implemented to prevent overfitting.

3.2. Model Evaluation

To comprehensively evaluate the model performance, we employed the following metrics: Accuracy, Recall, F1 Score, ROC Curve (Receiver Operating Characteristic Curve), and AUC (Area Under the Curve). To ensure the stability and reliability of the model evaluation, we used a 10-fold cross-validation method. The specific steps are as follows:

The dataset was randomly divided into 10 subsets.

Each time, 9 subsets were used for training and 1 subset was used for validation.

This process was repeated 10 times, with a different subset used for validation each time.

The average of the 10 validation results was calculated as the final performance metric.

3.3. Results Analysis and Comparison

Based on cross-validation, we conducted a detailed analysis and comparison of the evaluation results of the three models. The specific results are shown in Table 1.

Table 1. Evaluation Metrics for Each Model

Model	Accuracy	Recall	F1 Score	AUC
Random Forest	0.872	0.843	0.857	0.912
Gradient Boosting	0.892	0.853	0.871	0.927
Neural Network	0.879	0.846	0.862	0.918

Through comparison, we found that the Gradient Boosting Trees model outperformed all other models across all evaluation metrics. Specifically, the Gradient Boosting Trees model achieved an accuracy of 89.2%, a recall of 85.3%, an F1 score of 87.1%, and an AUC of 0.927. While the Random Forest and Neural Network models also performed better than traditional statistical models, their performance was slightly inferior to that of the Gradient Boosting Trees model.

Table 2. Cross-validation Performance of Each Model

	Random Forest Accuracy	Gradient Boosting Accuracy	Neural Network Accuracy
1	0.870	0.890	0.878
2	0.871	0.892	0.880
3	0.874	0.895	0.881
4	0.869	0.891	0.877
5	0.873	0.889	0.879
6	0.872	0.894	0.876
7	0.871	0.893	0.880
8	0.873	0.891	0.879
9	0.872	0.890	0.878
10	0.871	0.893	0.880

Table 2 presents the accuracy performance of different models across 10-fold cross-validation. The data reveals that the Gradient Boosting Trees model consistently outperformed the Random Forest and Neural Network models in terms of accuracy across all folds, demonstrating superior and more stable classification performance. The detailed analysis is as follows:

Random Forest: The accuracy fluctuated within a narrow range across the 10 folds, with a minimum of 0.869 and a maximum of 0.874, resulting in an average accuracy of approximately 0.872.

Gradient Boosting Trees: This model exhibited the best accuracy performance, with a minimum of 0.889 and a maximum of 0.895, resulting in an average accuracy of approximately 0.892. This indicates the model's consistency and stability across different data subsets.

Neural Network: The accuracy showed slight fluctuations across the 10 folds, with a minimum of 0.876 and a maximum of 0.881, resulting in an average accuracy of approximately 0.879. While the performance is good, it is slightly inferior to that of the Gradient Boosting Trees model.

These results highlight the robustness and reliability of the Gradient Boosting Trees model in delivering high and consistent accuracy in credit card underwriting scenarios.

Table 3. Confusion Matrix for Each Model

Model	True Positives	False Positives	False Negatives	True Negatives
Random Forest	1523	210	277	3990
Gradient Boosting	1560	190	240	4010
Neural Network	1540	200	260	4000

Table 3 presents the confusion matrices for each model, including the key metrics: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). The data indicates that the Gradient Boosting Trees model outperformed both the Random Forest and Neural Network models across all metrics. Specifically, the Gradient Boosting Trees model exhibited the highest TP and TN values, indicating superior classification performance for both positive and negative samples. Additionally, the lower FP and FN values for the Gradient Boosting Trees model reflect its lower misclassification rates.

Combining the analysis results from Tables 2 and 3, it is evident that the Gradient Boosting Trees model surpasses the Random Forest and Neural Network models in terms of both classification performance and stability. The Gradient Boosting Trees model not only achieved the highest average accuracy in the 10-fold cross-validation but also performed best across all metrics in the confusion matrices. This further validates the superiority of the Gradient Boosting Trees model in predicting profit for credit card underwriting.

The Gradient Boosting Trees model demonstrates the highest effectiveness in classifying positive and negative samples correctly, as evidenced by the highest TP and TN values and the lowest FP and FN values. These metrics highlight its robust performance in credit card underwriting profit prediction, confirming its advantages over the Random Forest and Neural Network models.

4. Empirical Analysis

4.1. Data Description

In this study, we utilized a comprehensive credit card customer dataset obtained from a major Chinese bank. This dataset includes multidimensional information on 500,000 customers, encompassing basic demographic details, transaction records, and credit scores, resulting in a total of 8 million transaction records. Following rigorous data cleaning, standardization, and feature engineering, we prepared a high-quality dataset for model training and evaluation.

The dataset was partitioned into training and testing sets, with the training set used to develop the models and the testing set employed to assess their performance. To ensure the stability and reliability of the model evaluation, we implemented a 10-fold cross-validation method. This approach involves dividing the data into 10 subsets, training the model on 9 subsets, and validating it on the remaining subset. This process is repeated 10 times, with each subset serving as the validation set once, and the results are averaged to obtain a robust performance estimate.

Table 4. Performance Metrics Comparison of Random Forest, Gradient Boosting, and Neural Network Models

Metric	Random Forest	Gradient Boosting	Neural Network
Accuracy	0.872	0.892	0.879
Recall	0.843	0.853	0.846
F1 Score	0.857	0.871	0.862
AUC	0.912	0.927	0.918
True Positives (TP)	1523	1560	1540
False Positives (FP)	210	190	200
False Negatives (FN)	277	240	260
True Negatives (TN)	3990	4010	4000

As shown in Table 4, the Gradient Boosting Trees model outperforms the other two models across all evaluation metrics. Specifically, the Gradient Boosting Trees model achieved an average accuracy of 0.892, which is notably higher than the 0.872 of the Random Forest and the 0.879 of the Neural Network. The recall for the Gradient Boosting Trees model was 0.853, exceeding the 0.843 of the Random Forest and the 0.846 of the Neural Network. The F1 score for the Gradient Boosting Trees model was 0.871, compared to 0.857 for the Random Forest and 0.862 for the Neural Network. Additionally, the AUC value for the Gradient Boosting Trees model was 0.927, outperforming the 0.912 of the Random Forest and the 0.918 of the Neural Network. Furthermore, the confusion matrix results further confirm the superior classification performance of the Gradient Boosting Trees model. It recorded the highest True Positives (TP) and True Negatives (TN), at 1560 and 4010 respectively, and the lowest False Positives (FP) and False Negatives (FN), at 190 and 240 respectively. These results underscore the model's effectiveness in accurately classifying both positive and negative samples, thereby demonstrating lower misclassification rates.

In summary, the Gradient Boosting Trees model exhibited the best performance in credit card underwriting profit prediction, showcasing higher accuracy, recall, and F1 scores, along with superior AUC values and confusion matrix metrics compared to the Random Forest and Neural Network models. These findings provide compelling evidence for financial institutions to adopt the Gradient Boosting Trees model as the preferred machine learning model for credit risk management in practical applications.

5. Results and Discussion

Through a comparative analysis of the performance of Random Forest, Gradient Boosting Trees, and Neural Networks in credit card underwriting, this study found that the Gradient Boosting Trees model excelled across all evaluation metrics, significantly outperforming the other two models. Specifically, the Gradient Boosting Trees model achieved an average accuracy of 0.892, notably higher than the 0.872 of the Random Forest and the 0.879 of the Neural Network. Its recall was 0.853, surpassing the 0.843 of the Random Forest and the 0.846 of the Neural Network. The F1 score of the Gradient Boosting Trees model was 0.871, compared to 0.857 for the Random Forest and 0.862 for the Neural Network. Additionally, the AUC value for the Gradient Boosting Trees model was 0.927, outperforming the 0.912 of the Random Forest and the 0.918 of the Neural Network. Furthermore, the confusion matrix results further validate the superior classification performance of the Gradient Boosting Trees model. It recorded the highest True Positives (TP) and True Negatives (TN) values, at 1560 and 4010 respectively, and the lowest False Positives (FP) and False Negatives (FN) values,

at 190 and 240 respectively. These results underscore the model's effectiveness in accurately classifying both positive and negative samples, thereby demonstrating lower misclassification rates.

The findings of this study have significant practical implications for the banking industry. Firstly, the superior performance of the Gradient Boosting Trees model in credit card underwriting can help banks more accurately assess customers' credit risks, thereby reducing default rates and increasing profits. Secondly, by adopting advanced machine learning algorithms, banks can significantly enhance the efficiency and accuracy of the credit card approval process, optimizing customer experience. Additionally, the high recall rate and low false negative rate of the model enable banks to more effectively identify high-risk customers and implement appropriate risk management measures.

With the advancement of big data technology and computational power, the application prospects of machine learning in credit card underwriting are broad. Machine learning algorithms can handle large-scale, multidimensional data, automatically uncovering complex patterns and relationships within the data, thus providing more accurate and comprehensive information for credit risk assessment. In the future, as algorithms continue to be optimized and data quality improves, machine learning will play an increasingly important role in credit card underwriting, customer credit evaluation, and risk management. Despite the achievements of this study, there are certain limitations. Firstly, the dataset is sourced from only one major Chinese bank, which may not fully represent the situation of other banks or regions. Secondly, this study only compared three machine learning models; future research could introduce more models for comparative analysis. Additionally, model optimization and parameter tuning require further exploration and improvement. Future research can expand and deepen in the following areas: introducing a broader variety of machine learning models, such as Support Vector Machines (SVM) and Extreme Learning Machines (ELM), to further compare their performance in credit card underwriting; expanding the scope of the dataset to include more banks and regions to enhance the generalizability of the research results; and integrating deep learning techniques to explore more complex model structures and algorithms, further improving the accuracy and efficiency of credit risk assessment.

In conclusion, this study has made significant contributions to the field of profit prediction in credit card underwriting, providing strong evidence for financial institutions to select appropriate machine learning models for practical applications. Additionally, it points to future research directions, with the expectation that further exploration and optimization will enhance the role of credit risk management.

Refences

[1] Moradi, S., & Mokhatab Rafiei, F. (2019). A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. Financial Innovation, 5(1), 1-27.

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

[3] Zhang, Y., Yang, K., Wang, Y., Yang, P., & Liu, X. (2023, July). Speculative ECC and LCIM Enabled NUMA Device Core. In 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS) (pp. 624-631). IEEE.

[4] Allison, P. D. (2012). Logistic regression using SAS: Theory and application. SAS institute.

[5] Yao, Y. (2024). Digital Government Information Platform Construction: Technology, Challenges and Prospects. International Journal of Social Sciences and Public Administration, 2(3), 48-56.

[6] Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., ... & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. Ieee Access, 8, 201173-201198.

[7] Xia, Y., Liu, S., Yu, Q., Deng, L., Zhang, Y., Su, H., & Zheng, K. (2023). Parameterized Decision-making with Multi-modal Perception for Autonomous Driving. arXiv preprint arXiv:2312.11935.

[8] Gao, J., Sun, W., & Sui, X. (2021). Research on default prediction for credit card users based on XGBoost-LSTM model. Discrete Dynamics in Nature and Society, 2021(1), 5080472.

[9] Qiu, L., & Liu, M. (2024). Innovative Design of Cultural Souvenirs Based on Deep Learning and CAD.

[10] Teng, H. W., & Lee, M. (2019). Estimation procedures of using five alternative machine learning methods for predicting credit card default. Review of Pacific Basin Financial Markets and Policies, 22(03), 1950021.

[11] Liu, M., & Li, Y. (2023, October). Numerical analysis and calculation of urban landscape spatial pattern. In 2nd International Conference on Intelligent Design and Innovative Technology (ICIDIT 2023) (pp. 113-119). Atlantis Press.

[12] Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. Cognitive Systems Research, 52, 317-324.

[13] Lin, Y. (2023). Optimization and Use of Cloud Computing in Big Data Science. Computing, Performance and Communication Systems, 7(1), 119-124.

[14] Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020). Gradient boosting trees. Gnosis Data Analysis PC, 1-3.

[15] Lin, Y. (2024). Application and Challenges of Computer Networks in Distance Education. Computing, Performance and Communication Systems, 8(1), 17-24.

[16] Lin, Y. (2024). Design of urban road fault detection system based on artificial neural network and deep learning. Frontiers in neuroscience, 18, 1369832.

[17] Lin, Y. (2023). Construction of Computer Network Security System in the Era of Big Data. Advances in Computer and Communication, 4(3).

[18] Gao, J., Sun, W., & Sui, X. (2021). Research on default prediction for credit card users based on XGBoost-LSTM model. Discrete Dynamics in Nature and Society, 2021(1), 5080472.