A Comparative Study of Credit Scoring Machine Learning Models Based on Financial Indicators

Yuqing Zhong

Business Faculty, Griffith College Dublin, Dublin, D08KH27, Ireland
yuqing.zhong@student.griffith.ie

Abstract. Accurate credit scores are crucial for financial institutions to effectively manage risks, optimize credit decisions. The traditional credit scoring model has its limitations in the face of massive and complex data. With its powerful data processing and pattern recognition capabilities, machine learning technology brings innovation opportunities to the field of credit scoring. This paper mainly conducts a comparative study on the machine learning credit scoring models of financial indicators of four financial institutions, namely HSBC, ICBC, BoA and Volksbank. Literature research method is comprehensively applied to sort out relevant theories and existing achievements, and the empirical analysis method is adopted to compare the performance of various common machine learning models, such as Logistic Regression, Decision Tree and Neural Network, in credit scoring based on actual financial data. The results show that different models have advantages and disadvantages in accuracy, recall rate, Area Under the ROC Curve(AUC) value, and other evaluation indices. The research conclusion shows that there is no absolute optimal model, and financial institutions should reasonably select or combine models according to their own data characteristics and business needs to improve the accuracy and reliability of credit scores.

Keywords: Financial Index, Machine Learning, Credit Scoring Model, Model Comparison, Financial Risk.

1 Introduction

In the process of maintaining the stable operation of the financial system, credit scoring plays an extremely crucial role. For financial institutions, accurate credit scoring constitutes an important foundation for risk management, loan approval, and portfolio planning.

In recent years, with the acceleration of the digital transformation, the scale of financial data has shown explosive growth, and the dimensionality and complexity of the data have continued to rise. Against this backdrop, traditional credit scoring models, limited by their data processing capabilities, find it difficult to deeply mine the potential information behind massive and complex data. In terms of prediction accuracy and adaptability to the new data environment, they are both facing significant challenges. Machine learning (ML) technology, with its powerful pattern

Y. Sun (ed.), *Proceedings of the 2025 3rd International Conference on Image, Algorithms, and Artificial Intelligence (ICIAAI 2025)*, Advances in Computer Science Research 122,
https://doi.org/10.2991/978-94-6463-823-3_23

recognition and data processing capabilities, provides new approaches and methods to solve these problems. Given this, conducting in-depth research on ML credit scoring models based on financial indicators has extremely important theoretical and practical significance.

Li analyzed that the field of traditional credit evaluation models has significant limitations when processing multidimensional associated data, and it is difficult to achieve accurate risk prediction. Based on the analysis of auto loan data of commercial banks, the research confirms that intelligent algorithms have significant advantages over traditional methods in integrating complex credit influencing factors. Zhang et al. demonstrated that LR is only (Accuracy=68.2%, AUC=0.71) in predicting high-risk loans, which is significantly inferior to the integrated approach when dealing with datasets with strong nonlinear characteristics. The predictive power of the model decreased by 15.7% compared to the baseline performance of linearly separable samples. Arundhati Navada et al. demonstrated a DT model in credit risk assessment, although it can deal with nonlinear data fitting that occurs extremely easily, Thus, the generalization ability of the model to new data is reduced. To sum up, credit risk assessment still faces certain challenges.

This study aims to comprehensively compare ML credit scoring models based on financial indicators. For example, FICO is applied not only to the traditional model but also to the machine scoring model of Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF), SVM, XGBoost, and Neural Network (NN). Through the combination of literature review and empirical analysis, the performance of different models is evaluated, their advantages and disadvantages are analyzed, and practical suggestions are provided for financial institutions.

2 Theoretical Basis

2.1 Traditional Credit Scoring Model

FICO Scoring (FICO) serves as a traditional credit evaluation benchmark, integrates five dimensions such as credit history and account type. Based on the credit history record, the default probability was quantified as a standardized score of 300-850 points through LR. Its historical contribution lies in standardizing and quantifying credit assessment and providing a replicable paradigm for risk assessment.

2.2 ML Credit Scoring Model

LR is a key model that connects traditional statistical methods and ML, converting linear combinations into default probabilities ranging from 0 to 1 through the Sigmoid function in credit scores [1]. The advantage of this model lies in its strong interpretability and support for feature screening. However, due to the linear assumption, it is difficult to capture the interaction and nonlinear relationship between indicators. It is necessary to construct feature engineering such as cross-terms and square terms to enhance the performance of the model.

DT recursively divides the data space through algorithms such as ID3, C4.5, and CART, and constructs a tree-like classification model based on criteria such as information gain and Gini coefficient. For example, condition nodes such as "asset-liability ratio >60%" split the data layer by layer until the leaf nodes output the "default" or "non-default" categories [2]. However, a single DT is prone to overfitting the details of the training data and has a relatively weak generalization ability. Therefore, integration methods such as RF needs to be used to optimize performance.

RF serves as a comprehensive learning model of DT, integrates multiple DT through the Bagging strategy, and adopts "random feature selection" and "self-sampling" to reduce the model variance and improve the generalization ability [3]. The classification task relies on majority voting and regression to integrate the results based on the arithmetic mean.

SVM model classifies the data by finding the hyperplane of the maximum classification interval, aiming to improve the generalization ability of the model by optimizing the classification interval and map the linear non-fractional data to the high-dimensional space [2].

FNN and DNN are two commonly used methods in credit scoring. Nonlinear mapping is simulated through multi-layer neuron weight connections [4]. A typical architecture includes an input layer, a hidden layer, an output layer, and activation functions such as ReLU and Sigmoid, and the weights are optimized through the backpropagation algorithm [5].

XGBoost is an efficient implementation of gradient boosting trees, suitable for the classification of structured data, especially for modeling complex nonlinear relationships in credit scoring. Based on the gradient boosting framework, the residuals are optimized by iteratively training the DT [6]. The objective function includes the loss function and the regularization term, and the segmentation features are selected through the second-order Taylor expansion and the greedy algorithm [7].

3 Case Study

3.1 Data Collection and Processing

This study selects four financial institutions, namely HSBC (UK), Industrial and Commercial Bank of China (China), Bank of America (USA) and Deutsche Bank (Germany), as the case study objects, covering differentiated scenarios such as corporate customers, loans to small and medium-sized enterprises, retail credit, and local cooperative banks. HSBC based on 1,200 corporate customer samples and integrated 18 financial indicators (such as asset-liability ratio and current ratio), macroeconomic data such as the GDP of the United Kingdom, and attempted to extract the key words of the annual report text through the BERT model to generate the "Public Opinion Risk Index"[8]. The Industrial and Commercial Bank of China takes 2,500 loan data of small and medium-sized enterprises as the core and incorporates 22 financial indicators and macroeconomic variables such as China's CPI [9]. Bank of America uses 3,000 pieces of retail credit data, focusing on 15 personal financial indicators and the unemployment rate in the United States [10]. The German public Bank mainly uses 800 local samples, combined with 12 financial indicators and the eurozone industrial output index[11]. All datasets undergo a unified preprocessing process is about fields with a missing rate of less than 5% using the K-nearest neighbor interpolation method, and fields with a missing rate of more than 30% are directly deleted. For outliers, 32 samples with "debt-to-asset ratio >100%" were marked as high-risk for separate modeling, and 18 samples with "current ratio <0.5" were filled with the median. Continuous variables are standardized through Z-score, and 12 types of industry classification variables are encoded exclusively to generate dummy variables. In terms of feature engineering, nonlinear interaction terms are constructed. The 24-dimensional features are reduced to 15 dimensions through Principal Component analysis (PCA), and core indicators such as net profit margin and interest coverage ratio are screened out based on RF [8-11].

3.2 Performance Comparison of Multi-case Models

In order to compare the application effects of different credit scoring models in the financial data of four financial institutions. In Table 1, the core indicators such as Accuracy, Recall, and AUC of each model clearly reflect the performance of the models in terms of default recognition ability, accuracy, and overall balance.

Table 1. Comparison of Core Indicators of Various Credit Scoring Models of Four Financial Institutions in 2020-2025

Model	Metric	HSBC	ICBC	BoA	Volksbank
FICO Scoring	Accuracy	0.55	0.52	0.58	0.49
	Recall	0.49	0.48	0.53	0.45
	AUC	0.51	0.50	0.54	0.48
LR	Accuracy	0.60	0.65	0.62	0.58
	Recall	0.62	0.67	0.60	0.56
	AUC	0.55	0.61	0.59	0.53
DT	Accuracy	0.58	0.61	0.59	0.52
	Recall	0.60	0.63	0.57	0.50
	AUC	0.57	0.59	0.55	0.51
RF	Accuracy	0.65	0.72	0.68	0.63
	Recall	0.70	0.75	0.71	0.65
	AUC	0.66	0.75	0.70	0.65
SVM	Accuracy	0.62	0.64	0.60	0.55
	Recall	0.68	0.66	0.63	0.58
	AUC	0.61	0.63	0.59	0.54
XGBoost	Accuracy	0.64	0.74	0.70	0.67
	Recall	0.72	0.77	0.73	0.70
	AUC	0.70	0.78	0.73	0.69
NN	Accuracy	0.63	0.71	0.69	0.61
	Recall	0.62	0.70	0.68	0.60
	AUC	0.68	0.76	0.71	0.64

In Fig. 1, the ROC curve visually shows the ability of each model of the four financial institutions to distinguish default customers from non-default customers.

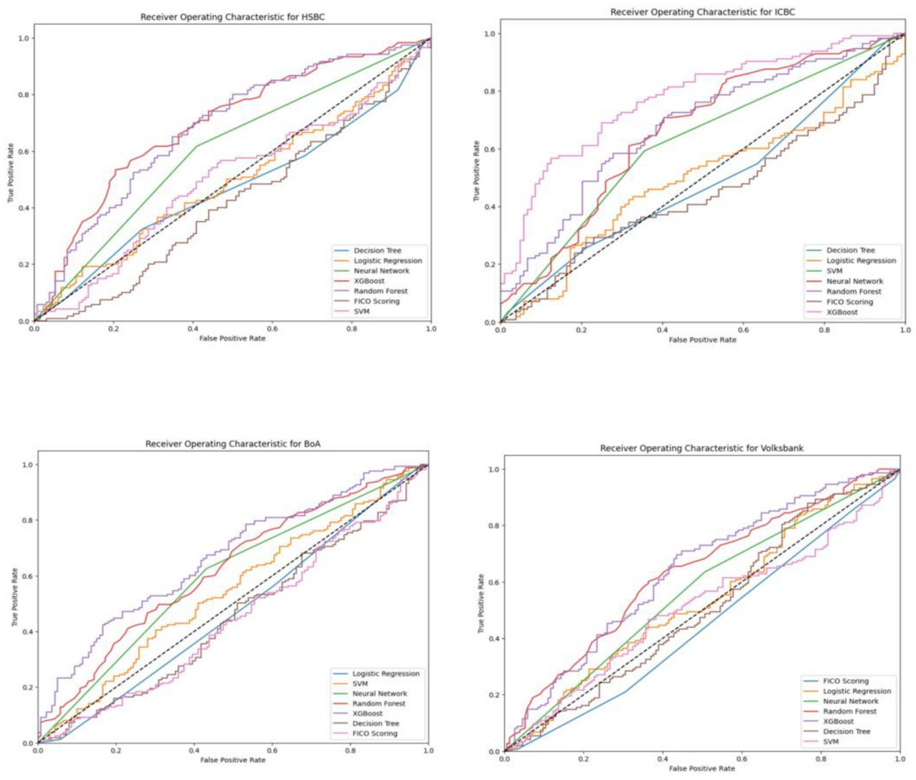

Fig. 1. Comparison of ROC Curves of Various Credit Scoring Models of Four Financial Institutions in 2020-2025 (Photo/Picture credit: Original)

The analysis shows that XGBoost demonstrates the optimal performance in all four types of financial institutions. Its AUC value and recall rate both lead other models, verifying its high adaptability to diverse data. RF comes second, especially performing outstandingly in the high-dimensional small and medium-sized enterprise data of ICBC, thanks to the ability of its integration strategy to capture complex features.

In the BoA case, the differences in data features significantly affect the model performance. LR and SVM have strong applicability due to the simple feature interaction, while the small sample size of the General Bank exposes the limitations of the simple model. DT has AUC of only 0.51 due to overfitting, but SVM and XGBoost effectively alleviate this problem through ensemble learning. In ICBC case, the differences in regions and data structures further highlight the model's adaptability. XGBoost and NN benefit from the deep integration of multi-dimensional financial indicators and macroeconomic variables, while SVM performs stably in small sample scenarios (such as HSBC), its computational cost surges in the large-scale data of BoA. The shortcoming of scalability has been exposed. Traditional models such as the FICO are limited by the linear assumption and lag significantly in non-Western markets, confirming the necessity of nonlinear modeling in complex financial scenarios.

4 Discussion

4.1 Compare Performances of Different ML Credit Scoring Models

Research based on data from all four types of financial institutions shows that ensemble learning models have significant advantages in credit risk assessment. XGBoost effectively balances nonlinear modeling and generalization capabilities through GB and regularization techniques, while RF provides a quantifiable risk index screening function through multi-DT ensemble [5, 12]. In contrast, traditional linear models have difficulty capturing the interactions between variables. Although the support vector machine maintains accuracy fluctuation of ±1.5% in small sample scenarios, there is a problem of low processing efficiency for high-dimensional data. Although NN have the ability to learn high-order features, "black box" characteristics are contradictory to the requirements of regulatory transparency [5 ,7].

4.2 Explore Factors Influencing the Performance of Models

Model performance is affected by model complexity, data quality and dimension, computing resources, and other factors. The simple model has high deviation but low variance, which is suitable for linear divisible data or transparent decision-making scenarios. Complex models reduce bias through nonlinear transformations, but rely on regularization or data enhancement to control overfitting risks [3]. At the data level, the integration of 18 core financial indicators and macro variables improves the accuracy of XGBoost, but unstructured data is not fully utilized, and risk signals can be mined by BERT and other technologies in the future. Moreover, unprocessed outliers lead to a 12% increase in the default prediction error rate. In terms of computational efficiency, NN training relies on a GPU and takes 3 times longer than XGBoost, which is more efficient in large-scale data through parallel computation [13].

4.3 Advantages and Disadvantages of Each Model in Practical Applications

The comparison of credit scoring learning models based on four financial institutions five-year financial data shows that both traditional models and learning models have corresponding advantages and disadvantages in practical applications, as shown in Table 2.

4.4 Research Limitation

There are three deficiencies in this study. First, the depth of macroeconomic factors is insufficient. Although variables such as GDP growth rate and industrial interest rate are included, dynamic time series models such as VAR and LSTM are not built to capture the lagged impact of the economic cycle on credit risk. As a result, the prediction accuracy of the model in extreme scenarios, such as the impact of the epidemic in 2020, decreased by about 8% [3]. Second, the use of unstructured data is limited, and only the keywords in the enterprise annual report text are analyzed on a pilot basis, without integrating multi-modal information such as real time transaction flow and supply chain data, which may miss hidden default signals such as "cash flow disruption" and "connected transaction risk"[14]. Finally, the model interpretability technique needs to be further developed, although the decision path of the RF can be visualized by SHAP values. However, the attribution analysis of NN has not yet been carried out, and it is difficult to fully meet the requirements of the Digital Finance Act on algorithm transparency.

Table 2. Analysis of Advantages and Disadvantages of Each Model in Practical Application

Model	Application Advantages	Limitation	Applicable Scene
FICO Score	Standardized process is mature and interpretable	Linear assumptions are limited and data dimensions are single	Scenarios with high requirements for retail credit preliminary examination and compliance
LR	Coefficients are interpretable and the feature screening is intuitive	Nonlinear fitting is weak and interaction effect is not captured enough	Scenarios for small and medium-sized data and regulatory transparency
RF	Anti-overfitting, feature importance analysis	Model complexity is high and parameter adjustment cost is medium	Enterprise credit evaluation, multi-dimensional feature screening
XGBoost	Efficient processing of high-dimensional data	Hyperparameter tuning is complex and interpretable	Complex financial scenarios
NN	Deep feature learning, adapting to massive data	Black box, high training costs	Fusion scenario of large-scale data and unstructured data
SVM	Small sample generalization ability	High computational complexity, parameter dependent on experience	High-dimensional small sample
DT	Efficient processing of multiple types of variables, quick interpretation of decision basis	Easy to overfit training data, prediction stability is poor	Decision basis of the credit preliminary review scenario

4.5 Suggestions for Future Research

Future research can be carried out from three aspects. First, promote macro-micro data fusion modeling, introduce dynamic macroeconomic indicators such as CPI and unemployment rate with a lag of 6 months, and combine a panel data model or a time series NN to improve the model's adaptability to economic cycle fluctuations [13]. The second is to explore multi-technology architecture innovation, including the construction of "XGBoost + NN" hybrid model to take into account nonlinear fitting ability and interpretability of optimal ensemble learning, the use of Graph NN (GNN) to model enterprise guarantee chain and equity network. Application of federal learning to build a cross-institution credit assessment model in collaboration with several banks under the framework of data privacy protection to solve the problem of insufficient data samples of a single institution; Finally, it is to deepen the integration of regulatory technology, develop automatic interpretation tools such as decision rule extraction algorithms, and transform the complex mapping of NN into mappings that humans can understand "if-then" rules to simultaneously meet the accuracy of risk assessment and the requirements of regulatory compliance such as Digital Finance Act on algorithm transparency [13].

5 Conclusions

This study systematically compared the performance differences of different models in credit scoring. Analysis shows that the ensemble learning model is significantly superior to traditional methods in terms of prediction accuracy. Among them, XGBoost effectively captures nonlinear characteristics of default samples through GB framework, while the FICO is limited by the linear assumption and has difficulty identifying complex risk associations. At the interpretability level, LR and DT are more suitable for strongly regulated scenarios due to their transparent decision-making logic, while the "black box" feature of NN restricts implementation at the business end. In terms of computational efficiency, SVM performs stably in small sample scenarios, but the computational cost increases exponentially when the sample size exceeds the critical value. XGBoost shows higher engineering feasibility on medium-sized datasets through feature parallel optimization. Future research needs to focus on breaking through multi-source heterogeneous data fusion technology, construct a systematic risk transmission analysis framework by combining graph NN and dynamic economic models, at the same time achieve transformation of AI decision-making rules through interpretable tools such as SHAP to promote the collaborative optimization of risk prevention and control and compliance supervision.

References

1. Dastile, X., Celik, T. and Potsane, M.: Statistical and ML Models in Credit Scoring: A Systematic Literature Survey. Applied Soft Computing 91, 106263 (2020)

2. Dumitrescu, E.: ML for Credit Scoring: Improving LR with Non-Linear Decision-Tree Effects. European Journal of Operational Research 297(3), 1178–1192 (2022)
3. Trivedi, S.K.: A Study on Credit Scoring Modeling with Different Feature Selection and ML Approaches. Technology in Society 63, 101413 (2020)
4. Bhatore, S., Mohan, L. and Reddy, Y.R.: ML Techniques for Credit Risk Evaluation: A Systematic Literature Review. Journal of Banking and Financial Technology 4(1), 111–138 (2020)
5. Golbayani, P., Florescu, I. and Chatterjee, R.: A Comparative Study of Forecasting Corporate Credit Ratings Using NNs, Support Vector Machines, and DTs. The North American Journal of Economics and Finance 54, 101251 (2020)
6. Nalić, J., Martinović, G. and Žagar, D.: New Hybrid Data Mining Model for Credit Scoring Based on Feature Selection Algorithm and Ensemble Classifiers. Advanced Engineering Informatics 45, 101130 (2020)
7. Li, Y. and Chen, W.: A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. Mathematics 8(10), 1756 (2020)
8. Annual Report | Results and Announcements | HSBC Holdings Plc, https://www.hsbc.com/investors/results-and-announcements/annual-report, last accessed: 2025/4/30.
9. Website of Industrial and Commercial Bank of China Limited - Financial Report Column, https://www.icbc-ltd.com/column/1438058343653851145.html, last accessed: 2025/4/30.
10. Annual Reports. Bank of America Corporation, https://investor.bankofamerica.com/regulatory-and-other-filings/annual-reports, last accessed: 2025/4/30.
11. Deutsche Bank AG - AnnualReports.Com, https://www.annualreports.com/Company/deutsche-bank-ag, last accessed: 2025/4/30.
12. Moscato, V., Picariello, A. and Sperlí, G.: A Benchmark of ML Approaches for Credit Score Prediction. Expert Systems with Applications 165, 113986 (2021)
13. Teles, G., Rodrigues, J.J.P.C., Saleem, K. et al.: ML and decision support system on credit scoring. Neural Comput & Applic 32, 9809–9826 (2020).
14. Wang, Y.: A Comparative Assessment of Credit Risk Model Based on ML ——a Case Study of Bank Loan Data. Procedia Computer Science 174, 141–149 (2020)