

Credit Risk Assessment and Explainability of Machine Learning

Qihang Yang

*Business School, Suzhou Institute of Technology, Jiangsu University of Science and Technology,
Zhangjiagang, China
austin051104@outlook.com*

Abstract. With the increase in online development of credit business, the customer information that financial institutions can use is also increasing, and the traditional credit scoring model has limitations in fitting high-dimensional characteristics, non-linear correlation, and multi-source heterogeneous data. This paper focuses on traditional statistical methods, machine learning models, and interpretable artificial intelligence in credit risk assessment, focusing on the applicable characteristics of logistic regression, random forest, XGBoost, and neural networks, and discusses the role of interpretation tools such as SHAP and lime in credit audit. The results show that machine learning algorithms have better classification and recognition ability for complex data samples, but the effect of the model will be affected by factors such as sample quality, default caliber, index selection, and business scenarios. At present, there is no single optimal model for all credit tasks. Traditional methods still have irreplaceable value in transparency, stability, and regulatory communication. This paper believes that the credit risk model should choose between forecasting ability, interpretation requirements, misjudgment cost, and compliance conditions, and form a more stable application path through traditional model benchmarks, machine learning assistance, interpretation tools, and manual review.

Keywords: machine learning, credit risk assessment, interpretable artificial intelligence, credit score

1. Introduction

Inclusive Finance and digital credit business have been rapidly promoted, and the credit risk control mode has gradually shifted from manual experience audit to data-driven decision-making. The source of customer information of financial institutions is no longer limited to financial statements and central bank credit data. Multi-source data such as transaction flow, consumption behavior, and third-party external portraits are included in the evaluation system. With the increase in data richness, the complexity of data structure has increased significantly. Traditional credit scoring cards and logistic regression models have strong explanatory results, but it is difficult to characterize high-dimensional characteristics, variable interaction effects, and risk information contained in unstructured data. Machine learning algorithms are thus widely used in credit scoring, default probability prediction, and pre-loan qualification audit, but the increase in model complexity will

synchronously reduce the traceability of decision logic. How to balance recognition ability and transparency has become an important issue in intelligent credit risk control.

The existing literature mainly focuses on traditional statistical methods, machine learning prediction, and model interpretation. SHAP can help identify key factors in the default of small and micro enterprises and improve the readability of complex model results [1]. XGBoost has good classification ability in the identification of corporate default risk, but the performance of the model still depends on the characteristics and sample conditions [2]. Adjusting the loss function for category imbalance helps to improve the identification effect of a small number of default samples [3]. Introducing text information and regularization methods into traditional models can improve the screening and classification performance of loan risk variables [4]. The optimal classifiers on different credit score data are not completely consistent, and the model evaluation cannot rely only on a single data set [5]. Financial services subject to regulatory constraints need to be able to explain model decisions, and simple black-box forecasting is difficult to land directly [6]. In addition to accuracy, credit scoring research should also pay attention to sample imbalance, misjudgment cost, and model stability [7]. XAI helps to improve the transparency of credit decisions, but the interpretation of results still needs to be combined with business processes and risk governance [8].

The existing literature confirms that machine learning enriches the technical tool set of credit risk assessment, but there are still some research problems that have not yet formed a unified conclusion: the data sets, default criteria, and model evaluation indicators used in different studies are not unified, resulting in the performance of various algorithms not being able to be directly compared horizontally. Some papers focus on accuracy or AUC, and lack discussion on misjudgment cost, cross-period stability, and deployment conditions. SHAP, lime, and other tools can only quantify the contribution of features to the prediction results, cannot identify the real causal correlation between variables, and the stability of the interpretation results is insufficient. This paper brings traditional statistical models, machine learning algorithms, and interpretable artificial intelligence into a unified analysis framework, compares them from the dimensions of data adaptability, predictive performance, interpretability, landing cost, and regulatory adaptability, and focuses on the conditions under which different methods are more suitable.

2. Theoretical basis of credit risk assessment

The core of credit risk assessment is to judge the possibility of future default according to the historical information and current characteristics of borrowers. Traditional credit scoring divides borrowers into default and normal samples, which essentially belongs to the task of supervised classification. The linear correlation between quantifiable characteristics of logistic regression and default probability, and the sign of the regression coefficient, intuitively reflects the direction of income, liabilities, historical credit investigation, and other variables on credit risk. The scorecard then converts the statistical results into scores and grades for approval, credit, and post-loan management. Such models adapt to standardized structured data to facilitate the review and verification of regulators, but the model needs to preset the form of variable function, which makes it difficult to capture the complex interaction and nonlinear risk correlation between features.

Machine learning is also based on historical samples, but emphasizes automatic search for classification rules. Random forests vote through multiple decision trees to reduce instability, XGBoost corrects errors round by round, and neural networks use multi-layer structures to fit complex relationships. The above algorithms are suitable for high-dimensional, multi-source heterogeneous feature data sets, but higher classification accuracy does not mean higher business landing value. Credit modeling needs to balance two types of misjudgment costs: missed judgment

of high-risk customers will lead to credit bad debt losses, and misjudgment of high-quality customers will reduce the profit margin of institutions. Therefore, the evaluation model should not only look at the accuracy rate, but also combine the recall rate AUC, Miscalculation of cost, and stability.

Global interpretation is used to identify the core risk characteristics of the model as a whole, and local interpretation is used to disassemble the risk judgment logic for a single individual. By quantifying the contribution value of characteristics, SHAP can clearly characterize the direction and extent of influence of various variables on the prediction results, and can realize the summary of global laws and the personalized decision-making interpretation of a single business at the same time. LIME constructs a simpler approximate model around a sample to interpret individual prediction results. The implementation logic of characteristic importance calculation is simple, which can quickly locate the core input variables of the model, but it cannot explain the basis of risk judgment for a single borrower. Various interpretable tools merely enhance the readability of model decision outputs and cannot be equated with causal inference analysis, nor can they replace the model validity test and manual business review process.

3. Machine learning methods and credit applications

Random forest, XGBoost and neural networks are mainstream machine learning tools widely adopted in credit risk modeling. Random forest simultaneously processes continuous and categorical variables and captures partial nonlinear correlations. Ensemble voting across multiple decision trees stabilizes model output, fitting mass customer credit scoring with complex variable sources. Yet deeper, more numerous tree layers complicate model structure and bias feature importance toward multi-valued variables. XGBoost iteratively corrects residual fitting and suits corporate default prediction, micro-and-small lending and personal credit scoring. It elevates identification of scarce default samples via customized sample weights, loss functions and classification thresholds. Its rich hyperparameters heavily rely on learning rate, tree depth and sample segmentation rules. Lacking cross-validation and out-of-sample generalization tests easily triggers overfitting and overestimated performance, only qualifying as auxiliary tools for traditional benchmark models. Neural networks fit intricate variable interactions via multilayer architectures and excel at unstructured data including texts, images and transaction sequences, expanding credit data sources while demanding larger samples and stronger computing resources. Complex neural networks gain no edge for limited-sample, low-variable routine scoring tasks, so data types and business goals must top model selection priorities.

As seen in Table 1, credit datasets feature severe class imbalance where normal borrowers vastly outnumber defaulters, making pure accuracy misleading even with high values, as the model may fail to detect real high-risk clients. Researchers must simultaneously report AUC, recall, accuracy and confusion matrix, alongside threshold impacts on business returns. Unified sample splitting and evaluation metrics are required for horizontal model comparison, otherwise experimental settings distort research conclusions. Machine learning and traditional statistical models act as complementary rather than substitutive tools. Logistic regression benchmarks boast superior structural transparency, stable operation and convenient regulatory communication, while machine learning outperforms in mining intricate data correlations. Practical workflows first build logistic regression benchmarks, then test stable performance improvements brought by random forest, XGBoost or neural networks. Performance advantages confined to single datasets or single indicators cannot prove complex models superior. Only algorithms sustaining steady effects across

out-of-sample data, multi-period observations and core business indicators deserve formal deployment in real credit risk systems.

Table 1. Comparison between traditional statistical methods and machine learning methods

Comparison direction	Traditional statistical methods	Machine learning methods
Representative method	Logistic regression, credit score card	Random forest, XGBoost, neural network
Data adaptability	Structured and stable data	High-dimensional, multi-source, multi-source and non-linear data
Key strengths	Transparent and stable, convenient for review and regulatory communication	Strong ability to identify and predict complex relationships
Major limitations	The description of complex relations is limited and relies on manual setting of characteristics.	The degree of black box is higher, and the cost of regulation and participation in governance is higher.
App focus	Routine scoring, compliance interpretation and benchmarking	Complex scene recognition and assistant decision-making

4. Role and limitations of artificial intelligence

The result of credit approval directly determines the borrower's credit qualification and financing cost, and the model only outputs a single risk score, which cannot meet the needs of many parties: risk control personnel need to clarify the core risk driving factors, rejected customers have the right to know the basis for rejection of the audit, and regulators need to verify whether the model has discriminatory feature dependence. SHAP can break down the forecast results into the contributions of different characteristics and help identify the impact of factors such as income, liabilities, historical overdue or operating conditions on the results. Black-box A.I. is not suitable for direct use in heavily regulated financial services [9]. Interpretation tools can enhance the transparency of credit assessment and help policymakers understand model output [10].

The LIME algorithm specializes in local interpretation for individual borrowers: it trains a simple linear model on disturbed samples near target cases to approximate the original model's local decision boundary. Its interpretation logic is intuitive, yet output results are highly susceptible to disturbance intervals and parameter configurations. Identical customer cases may yield inconsistent explanations under different parameter setups, requiring separate stability testing. The algorithm swiftly computes feature importance to help analysts quickly grasp the model's core dependent variables in preliminary overall observation, yet it cannot clarify whether a single variable raises or lowers risk for specific borrowers, nor can it deduce causal relationships between variables. As seen in Table 2, SHAP balances global and local interpretation with comprehensive information but carries heavy computation load, while LIME delivers flexible individual explanations with weak stability and feature importance only offers superficial overall analysis. In real credit audits, staff shall combine global model logic inspection, local personalized interpretation and manual review, and cross-check interpretation outputs against business logic and compliance rules instead of directly adopting technical results as final judgments.

Table 2. Uses and limitations of common interpretable methods

Method	Scope of interpretation	Primary use	Problems needing attention
SHAP	Global and local	Explain the impact of variables on the overall model or single prediction	The calculation is large, and the results are affected by the basic model and samples
LIME	local	Explaining single credit judgment with a simple local model	It is sensitive to disturbance mode and parameters, and the results may be different under different settings, with limited stability
Feature importance	overall situation	Quickly identify the variables that the model mainly depends on	It is difficult to explain the forecast results of individual customers, and it is impossible to judge the causal relationship accordingly

5. Future research

The primary defect of existing relevant research is poor horizontal comparability: the default identification standards used in the literature are not unified. Some studies divide default samples by overdue days, and some take bad debt losses and credit rating downgrades as the basis for judgment. At the same time, the data set sources cover open databases, internal data of commercial banks and Internet credit platforms, and the sample size, customer group structure and observation cycle are significantly different. Even with the same model, the results can be different. The follow-up study should explain the data processing rules and evaluation indicators, and use another classic credit data set to test whether the conclusion depends on the original sample. Model evaluation also needs to shift from higher scores to availability. Accuracy and AUC can reflect classification performance, but they cannot directly represent business losses. Financial institutions are more concerned about the loss of high-risk customers, the impact of misjudging normal customers, and whether the model is stable in the next period. Follow-up studies can include cost-sensitive indicators, cross-period testing and stress testing, and continue to monitor model drift. After changes in customer structure, economic environment, or product rules, the model should be reevaluated promptly, rather than following the results of a training session for a long time.

Interpretability studies also need more specific evaluation methods. At present, many papers believe that the model can be explained after showing the results of a snapshot chart or LIME, but whether the explanation is stable, whether different methods are consistent, and whether business personnel can understand it are often lack of tests. Subsequently, the results of various interpretation tools under the same model can be compared, the changes of interpretation in different samples and periods can be observed, and business personnel can be invited to evaluate readability. Algorithmic fairness and data privacy should also be implemented in variable review, authority management, anonymity, and manual appeal mechanisms. From the perspective of the application process, data review, out-of-sample testing, and fairness inspection should be completed before the model is launched, and variable distribution and prediction results should be monitored after the model is launched. When the model drifts, the threshold can be adjusted, retrained, or returned to a stable benchmark model. The interpretation results should also be retained for review and tracking. Incorporating these steps into model governance can improve the reliability of the credit system more than adding complex algorithms alone.

6. Conclusion

This paper systematically combs through the theory and application status of traditional statistical models, machine learning algorithms, and interpretable artificial intelligence under the credit risk assessment system. Studies have confirmed that random forests, XGBoost, and deep neural networks have stronger complex data association mining capabilities and have good application potential in personal consumption loans, small and micro enterprise credit, and enterprise default warning scenarios. However, the effect of the model will be affected by data quality, sample structure, default definition and evaluation method, and there is no absolute optimal model divorced from specific scenarios. Although there are limitations in the identification of complex relationships, logistic regression and credit scoring cards have the advantages of intuitive structure, stable operation, and easy review, and still play a basic role in actual credit management. Machine learning is more suitable for supplementing risk information beyond traditional models than simply replacing existing systems.

The main value of this paper is to discuss model forecasting, interpretation requirements, and business governance together. For financial institutions, the more feasible path is to take the traditional model as the benchmark, use machine learning to process high-dimensional and non-linear characteristics, and then explain the model judgment through tools such as SHAP and LIME, while retaining manual review and continuous monitoring. This paper is still mainly based on the existing literature, without using unified data to directly test different models, and some conclusions need to be verified by empirical analysis. In the follow-up study, two sets of open standard credit data sets can be selected to carry out control experiments to uniformly calculate the prediction performance, misjudgment cost, and interpretation stability of various models. At the same time, it expands the subdivision research directions of algorithm fairness, data privacy security, and dynamic monitoring of model drift. Only in this way can the applicability of intelligent credit models for long-term deployment be verified, rather than merely attaining superior experimental indicators.

References

- [1] Lei, X., Lin, L., Xiao, B., & Yu, H. (2024). Re-exploration of small and micro enterprises' default characteristics based on machine learning models with SHAP. *Chinese Journal of Management Science*, 32(5), 1–12.
- [2] Chi, G., & Wang, S. (2024). Default risk prediction model for Chinese listed companies based on XGBoost. *Journal of Systems & Management*, 33(3), 735–754.
- [3] Yang, L., & Shi, B. (2022). Credit risk evaluation model and demonstration based on focal loss modified cross-entropy loss function. *Chinese Journal of Management Science*, 30(5), 65–75.
- [4] Wang, X., Zhang, Z., & Ma, S. (2021). A loan credit risk model incorporating text prior information. *Chinese Journal of Management Science*, 29(5), 34–44.
- [5] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- [6] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, 57(1), 203–216.
- [7] Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science*, 8, 180–201.
- [8] Nallakuruppan, M. K., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An explainable AI framework for credit evaluation and analysis. *Applied Soft Computing*, 153, 111307.
- [9] Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management*, 16(12), 496.
- [10] Chang, V., Xu, Q. A., Akinloye, S. H., Benson, V., & Hall, K. (2025). Prediction of bank creditworthiness through credit risk analysis: An explainable machine learning study. *Annals of Operations Research*, 354, 247–271.