Asia Pacific Science Press

Asia Pacific Economic and Management Review
ISSN(O):3005-9275
ISSN(P):3005-9267
2026 International Conference on Digital Economy and Artificial Intelligence

Research on the Application of Big Data Technology in Bank Credit Risk Assessment

Jiachen Yan*

School of Finance and Economics, Xizang University, Lhasa, 850000, Xizang, China

Corresponding author: Jiachen Yan, vesper_clay@outlook.com

Copyright: 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY-NC 4.0), permitting distribution and reproduction in any medium, provided the original author and source are credited, and explicitly prohibiting its use for commercial purposes.

Abstract: With the rapid development of FinTech, commercial banks are facing an increasingly complex credit environment. Traditional credit risk assessment models struggle to meet the processing demands of massive and multi-dimensional data. Big data technology provides new perspectives and tools for bank risk control. Based on the theoretical mechanism of big data risk control, this paper constructs a multi-dimensional indicator system including demographic characteristics, asset status, and behavioral preferences. Using the personal credit dataset of a domestic commercial bank, Logistic Regression (LR) model and XGBoost machine learning model were established for empirical comparative analysis. The results show that introducing big data variables and adopting the XGBoost model can signifi cantly improve the accuracy and AUC value of default prediction, eff ectively reducing the credit default risk of commercial banks. Finally, countermeasures are proposed to address problems such as data silos, weak model interpretability, and privacy protection.

Keywords: Big Data; Commercial Bank; Credit Risk Assessment; XGBoost; Empirical Analysis

Published: Aug 28, 2026

DOI: https://doi.org/10.62177/apemr.v3i7.1670

1. Introduction

Commercial banks are the core of the fi nancial model, and managing risks is their primary business content. Credit risk has long occupied the top position in the bank's risk exposure. Traditional credit risk assessment mostly relies on expert scoring methods under the framework of the Basel Accord or basic statistical models. The 5C principle belongs to such models. The assessment process places more emphasis on hard information such as the borrower's financial statements and the value of collateral. With the rise of Internet fi nance and the continuous implementation of inclusive fi nance policies, the bank's service customer base has gradually extended to long-tail customers. Most of these customers have no complete credit records and cannot provide sufficient collateral of value. Information asymmetry has led to adverse selection and moral hazard, and related issues have gradually become prominent. The scale of big data technology has continued to expand, creating conditions for solving this problem. Big data has the characteristics of Volume, Velocity, Variety, and Value. Banks can analyze customer funds flow and incorporate e-commerce transaction data, social behavior, and operator data, etc., to draw multi-dimensional customer portraits. Existing literature mostly conducts theoretical discussions on big data, but there are few empirical comparisons based on specifi c machine learning algorithms[1]. This paper sorts out empirical data and analyzes the practical application eff ect of big data technology in bank credit rating within the framework of integrated learning algorithms, to provide theoretical support for the digital transformation of commercial banks and to clarify practical directions.

2. Theoretical Foundation of Big Data in Credit Risk Assessment

2.1 The Connotation of Big Data Credit Risk Control

Big data credit risk assessment is a process where commercial banks utilize technologies such as cloud computing and machine learning to collect, clean, and mine large-scale structured and unstructured data. Structured data includes account transactions and credit reports, while unstructured data includes personal consumption habits and APP usage trajectories. This process ultimately converts the borrower's default probability into specific numerical values. Under the traditional risk control model, the available data range is limited, and the model algorithm's ability to handle complex correlations is limited. The credit risk assessment involving big data, in terms of data dimension coverage and model algorithm processing capabilities, has advantages different from those in the past.

2.2 Mechanism of Action

Information Screening Mechanism: Big data can utilize association rule mining methods to identify fraudulent groups hidden in complex network relationships. This can compensate for the relatively low coverage of traditional credit reporting[2]. Dynamic Monitoring Mechanism: Traditional rating methods are mostly static, while big data technology can achieve real-time risk control in milliseconds, enabling timely detection of abnormal behaviors of customers, such as suddenly starting frequent small loans or frequently changing contact information. Long-tail Coverage Mechanism: Analyzing fragmented behavioral data can establish credit scores for those "white clients" without credit records, thereby expanding the service scope of the bank.

3. Empirical Research Design

To verify the effectiveness of big data technology in risk assessment, this paper compares the XGBoost algorithm, which performs well in machine learning, with traditional logistic regression.

3.1 Data Sources and Preprocessing

The data used in this paper comes from a de-identified dataset of the personal consumer credit department of a domestic commercial bank. The original sample consists of 12,000 entries, with a time span from January 2022 to June 2023. The target variable Label defines "default" as overdue for more than 90 days M3+, marked as 1, and other situations as normal performance, marked as 0. The default samples account for approximately 8.5% of the total[3]. There is a problem of class imbalance. Data cleaning operations are as follows: fields with a missing rate exceeding 50% are removed, numerical missing values are filled with the median, and the SMOTE algorithm is used to handle the training set to balance the positive and negative sample ratios.

3.2 Construction of the Index System

Based on the logic of big data risk control, this paper has constructed an evaluation system consisting of 3 first-level indicators and 12 second-level indicators (see Table 1).

Table 1: Credit Risk Assessment Indicator System

Primary Indicator	Secondary Indicator	Indicator Description	Variable Type
Basic Attributes	Age	Borrower's Age	Continuous
	Gender	0-Female, 1-Male	Discrete
	Education Level	1-High School or Below, 2-Bachelor's Degree, 3-Graduate Degree (Master/PhD)	Discrete
	Marital Status	0-Unmarried, 1-Married, 2-Divorced	Discrete
Assets & Liabilities	Monthly Income	Post-Tax Average Monthly Income	Continuous
	Debt-to-Income Ratio (DTI)	Monthly Repayment Amount / Monthly Income	Continuous
	Housing Property Type	0-Renting, 1-Own Home with Mortgage, 2-Own Home without Mortgage	Discrete
	Credit Card Utilization Rate	Used Credit Limit / Total Credit Limit	Continuous

Primary Indicator	Secondary Indicator	Indicator Description	Variable Type
Behavior & Credit History	Credit Inquiry Count (Last 6 Months)	Number of Hard Inquiries (Institutional Inquiries)	Continuous
	Multi-Lending Count	Number of Loan Apps Installed Across Different Platforms	Discrete
	Mobile Network Tenure	Mobile Phone Number Tenure (Months)	Continuous
	Social Graph Stability	Overdue Contact Ratio (Social Graph)	Continuous

3.3 Model Construction and Parameter Setting

To confirm the superiority of the big data algorithm, this study first established a traditional logistic regression model, and then an XGBoost model[4]. 1. The logistic regression model, or LR model, has a simple structure, clear variable coefficients, and strong interpretability. It has always been the mainstream algorithm for the development of bank scoring cards. Its basic formula is:

$$P(Y=1|X)=\frac{1}{1+e^{-(w^Tx+b)}}$$

Among them, x represents the input feature vector and w represents the weight coefficient. Although Logistic Regression is stable, its essence is a linear classifier and is difficult to capture the non-linear relationships and feature interactions that are ubiquitous in large datasets.

XGBoost was proposed by Chen et al. and belongs to the ensemble learning algorithm derived from gradient boosting trees. Compared with traditional gradient boosting trees, XGBoost incorporates regularization terms to control overfitting and also supports parallel computing. The core logic of the algorithm is to add a new decision tree each time, fitting the residuals generated by the previous round of predictions. The objective function is:

$$Obj=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k)$$

Among them, l represents the loss function and Ω represents the regularization term. This paper uses the xgboost library in Python for modeling, and the main hyperparameters are set as follows: learning_rate = 0.05, max_depth = 6, n_estimators = 500, subsample = 0.8.

3.4 Analysis of Empirical Results

We divided the preprocessed data into a training set and a test set in a 7:3 ratio. To objectively evaluate the model performance, accuracy rate, AUC value (area under the ROC curve), and KS value were selected as evaluation indicators[5].

3.4.1 Variable Selection and Importance Analysis

Before processing the input data of the model, feature selection was conducted using information values. After the selection was completed, it could be confirmed that the IV value of the number of queries in the past 6 months was 0.42, the IV value of the number of multiple borrowing times was 0.38, and the IV value of the debt-to-income ratio was 0.31. The predictive capabilities of these three were ranked at the top. This result verifies the viewpoint of this paper, that the behavioral trajectories within the scope of big data, including multiple borrowing behavior and query records, reflect the current credit risk of customers, and are superior to static demographic characteristics.

3.4.2 Model Performance Comparison

The trained LR model and XGBoost model were verified on the test set, and the results are shown in Table 2.

Table 2: Model Performance Comparison Results

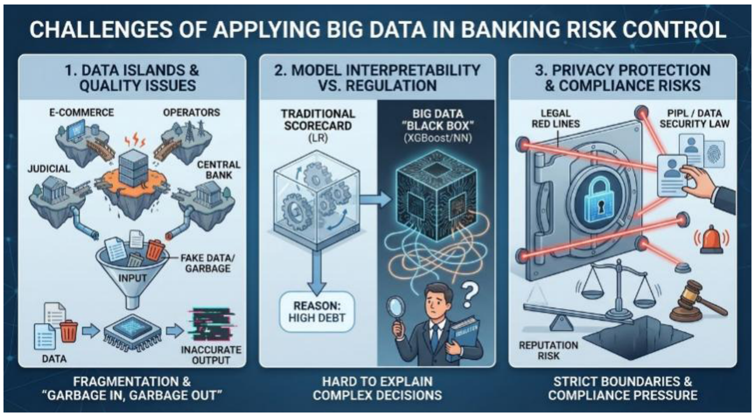
Model Name	Accuracy	Recall	AUC	KS Statistic
Logistic Regression (LR)	0.784	0.652	0.741	0.38
XGBoost	0.892	0.815	0.886	0.54

In terms of overall predictive ability, the XGBoost model outperforms the LR model in all indicators. The AUC value of the XGBoost model is 0.886, while that of the LR model is 0.741. The XGBoost model has a higher value. After introducing multi-dimensional big data features, the ensemble learning algorithm can discover the nonlinear patterns in the data and make more accurate judgments. The risk discrimination ability is measured by the KS value. In the banking industry, the model's KS value must exceed 0.3, and an excellent model's KS value must exceed 0.45. The KS value of the XGBoost model is 0.54[6]. This model can widen the score gap between "good customers" and "bad customers". The bank can rely on the model output to formulate credit approval Cut-off strategies, which are more accurate. The capture ability of the model for default customers is measured by the Recall value. In the risk control scenario, the cost of misclassifying bad customers is higher than that of misclassifying good customers. The recall rate of the XGBoost model is 0.815, which can identify more potential default subjects and directly reduce the bank's bad debt losses.

4. Challenges Faced by Big Data Technology in Bank Risk Control Applications

Although empirical analysis has proven the effectiveness of big data technology, in the actual implementation process, commercial banks still encounter many challenges.

Figure 1 Challenges of applying big data in banking risk cotrol



4.1 Data Isolation and Data Quality Issues at Present

The data of various institutions within the credit reporting model in China has not been fully connected. The coverage of the People's Bank of China's credit reporting center mainly consists of banking business data, while external data from fields such as e-commerce operators and judicial affairs are held by different institutions separately, forming "data islands". During the process of banks accessing external data, they will encounter issues such as inconsistent interface standards, untimely data updates, and even data fraud. Fraudulent activities such as "blackstreaming" are common types of data fraud. The "garbage data" input into the model will lower the accuracy of the output results.

4.2 Model Explainability and Regulatory Conflict

The regulatory rules in the financial industry are strict, and there are clear requirements for model explainability. Traditional scoring cards rely on the LR algorithm and can directly inform loan applicants of the reasons for denial, which are due to their excessive debt. Big data models represented by XGBoost and neural networks are often regarded as "black boxes". The complex tree structure inside the model makes it impossible to explain the specific denial judgment logic to regulatory authorities and loan applicants[7]. The application scope of advanced machine learning models cannot cover the core links of credit approval.

4.3 Privacy Protection and Compliance Risks

After the promulgation of the Personal Information Protection Law and the Data Security Law，data collection activities are subject to clear boundary constraints. During the operation of big data risk control, banks need to obtain multi-dimensional user data. Data mining requires meeting two conditions: not infringing on user privacy and obtaining full authorization from users. This condition is a legal compliance requirement that banks must abide by. Data mining beyond a reasonable range may trigger reputational risks.

5. Countermeasures and Suggestions for Enhancing the Bank's Big Data Risk Control Capabilities

5.1 Build a Comprehensive Data Ecosystem and Break Down Information Barriers

For banks to make breakthroughs in big data risk control, the primary task is to establish an "internal-external integration" model of a comprehensive data ecosystem, breaking through the "data silo" effect that has long restricted the operation of risk assessment models. Starting from internal integration, banks can promote cross-departmental data governance. Under the traditional banking model, data such as credit cards, personal loans, savings and investment, and settlement accounts are mostly scattered and stored in different business lines and IT systems. This fragmented state prevents banks from having a complete picture of the customer's real assets and liabilities. Banks can build an enterprise-level data warehouse or data lake, integrating structured transaction data with unstructured customer service logs and online banking operation trajectories. This breaks down the data barriers between departments and generates a unified customer view. After cleaning and standardization processing, it can uncover the patterns of funds accumulation and consumption preferences during the customer's life cycle, providing data support for accurately identifying hidden debts and multiple loans. In terms of external cooperation, it is necessary to actively expand multi-dimensional data sources and apply cutting-edge privacy computing technologies. Banks, relying solely on internal data, cannot establish profiles of "long-tail customers". Banks can actively connect with licensed institutions such as the China National Credit Information Service, introducing alternative data from tax, social security, housing provident fund, and e-commerce operators in accordance with compliance requirements. During external data cooperation, there are risks of data leakage and challenges in protecting commercial secrets.

5.2 Strengthen Model Integration and Balance Accuracy and Interpretability

During the implementation of fintech applications, there is a natural contradiction in machine learning models: the higher the prediction accuracy, the lower the interpretability. This contradiction is also known as the black box paradox, which has become the main limitation for the wide application of advanced algorithms such as XGBoost and neural networks in the core credit granting process. To address this contradiction, banks can simultaneously promote model integration and enhance interpretability. Implement dual-mode operation and adopt a phased mixed deployment plan. In the entire process of credit business, the requirements for interpretability vary in different stages. In the pre-loan screening stage, the risk tolerance is low, and the work direction is to quickly screen out high-risk individuals. This stage can be fully entrusted to high-dimensional nonlinear models such as XGBoost and Random Forest for hard filtering. By leveraging the ability of these models to capture abnormal behavior patterns, the possibility of default can be controlled. In the credit approval and pricing stage, while meeting regulatory requirements, to explain the reasons for loan rejection or pricing, the output results of the machine learning model can be used as the core variable and input into interpretable models such as logistic regression LR or decision trees. Alternatively, machine learning can be used to complete feature engineering screening. Finally, the scorecard model outputs the final score. This hybrid architecture is based on black box models and outputs results from white box models. While maintaining the accuracy advantage of big data algorithms, it also ensures compliance and logicality in business aspects[8]. By introducing post-explanation tools such as SHAP values and LIME, these are applied to the business of financial institutions. When dealing with complex ensemble learning models, the technical team of banks will not merely focus on increasing the AUC value. Instead, they will actively attempt to dissect the internal structure of the black box and invoke the SHAP algorithm under the framework of game theory.

5.3 Improve Data Governance and Privacy Protection Mechanisms

The "Data Security Law" and the "Personal Information Protection Law" were successively promulgated and implemented. Data compliance has become the core support for big data risk control in commercial banks. Banks need to establish a full life cycle data security management system covering all aspects of data collection, transmission, storage, and use to ensure that the application of technology complies with legal and ethical norms. In the data collection stage, all parties need to strictly follow the two principles of "minimum necessity" and "authorization and consent". When collecting customer device information, location information, and behavioral data through APPs or web pages, the purpose, method, and scope of collection should be clearly informed. The "one-size-fits-all authorization" and "tyrannical terms" should be eliminated.

Personal privacy data that does not fall within the necessary scope of business shall not be collected. During the process of introducing third-party external data, banks need to establish a strict supplier admission review mechanism to verify the legality of the data source and avoid the potential legal liabilities caused by the illegal crawling of third parties or the inflow of black market data. In the data processing and storage stage, banks need to build a technical protection framework, implement a data classification and grading management system, and de-identify or encrypt sensitive personal information. In the data invocation stage, strict permission control should be enabled, combined with watermark traceability technology, to prevent internal personnel from illegally querying or disclosing data. Regular data security audits and penetration tests should be conducted to build a multi-layer security defense line covering pre-event prevention, in-event monitoring, and post-event traceability. In the aspect of data application and ethics, a technology ethics committee should be established to prevent algorithm discrimination.

6. Conclusion

This paper selects the credit data of a domestic commercial bank to analyze the practical value of big data technology applied to credit risk assessment. By comparing the operation results of logistic regression and XGBoost models, it can be seen that incorporating multi-dimensional data features such as social behavior and cooperating with XGBoost ensemble learning algorithms can improve the bank's ability to identify default risks. The AUC value rises to 0.886. The technology implementation cannot be completed in a short period of time. Currently, commercial banks face data silos, insufficient model interpretability, and multiple issues such as privacy compliance. Adjustments can be made from the two directions of technical iteration and system establishment. Technical iteration involves privacy computing, explainable AI, and system establishment covers data governance, compliance review. In the future, the continuous development of fintech will lead to the separation of big data risk control from the simple risk prevention tool positioning and its integration with business expansion tools for banks to tap high-quality customers and implement differentiated pricing, promoting the completion of high-quality digital transformation of the banking industry.

Funding

No

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Deep, G., Deep, A., Rachev, T. S., et al. (2026). Google Trends—Augmented XGBoost for market volatility prediction: A machine learning early warning system. Journal of Behavioral and Experimental Finance, 49, 101159.

[2] Thommandru, A., & Mone, V. (2026). Datafication of mourning: Emotional labour, memory politics, and ethical innovation in digital grief technologies. Journal of Responsible Technology, 25, 100159.

[3] Kumar, S., Ranjan, A., Saxena, S., et al. (in press). ANFISBased MultiSensor Data Fusion Model for optimized autonomous vehicle navigation using big data and filtering techniques. Mobile Networks and Applications, 1-15.

[4] Mekonen, M. T., & Japee, P. G. (2026). Green credit management and perceived lending outcomes in Ethiopia: The mediating role of risk management in banks and microfinance institutions. Journal of Banking Regulation, 27(2), 13.

[5] Sundaram, S., & Mary, P. P. J. (2026). Heuristicderived improved triple data encryption standard (ITDES) algorithm for multihop concurrent big data sharing using blockchain technology. Knowledge and Information Systems, 68(1), 79.

[6] Saeed, M. M., & Donkoh, E. (2026). Credit risk management, bankspecific factors, and financial performance of banks: Insights from an emerging economy. Social Sciences & Humanities Open, 13, 102523.

[7] Ariefianto, D. M., Nur, T., & Meivitawanli, B. (2026). Credit risk management dynamics: Evidence from Indonesian rural banks. Risks, 14(1), 9.

[8] Ally, J. O., Kulindwa, J. Y., & Mataba, L. (2025). Financial technology and credit risk management: The case of nonperforming loans in Tanzanian banks. Cogent Economics & Finance, 13(1).