Research on Bank Credit Risk and Default Customer Identification Integrating Machine Learning Techniques

Haoran Deng
Shanghai Pudong Development Bank Shenzhen Branch, Shenzhen, Guangdong, China

Abstract: This paper aims to explore the application of machine learning technology in bank credit risk management and default customer identification. By outlining the fundamental theoretical framework of credit risk, the concepts, classifications, and key influencing factors are clarified. The analysis focuses on the application of machine learning in credit risk identification, covering its suitability, core algorithms, and practical advantages and limitations. Addressing the specific needs of default customer identification, the paper further examines how machine learning enhances the accuracy of identifying default customers and proposes various model optimization strategies. Through in-depth analysis of relevant theories and technologies, this paper seeks to provide financial institutions with effective technical support and theoretical foundations for managing credit risks.

Keywords: Credit Risk; Machine Learning; Risk Assessment

1. Introduction

In today's continuously evolving global economic environment, banks, as key players in the financial market, face increasingly complex credit risk management challenges. Traditional credit risk assessment methods often rely on qualitative analysis and historical data. While these methods can reveal a client's repayment ability and credit risk to some extent, their low data processing efficiency and limitations in predictive capability have become insufficient to meet the demands of the modern financial system. With the widespread adoption of big data technology and the rapid development of artificial intelligence, machine learning presents new solutions for credit risk management. Its robust data processing capacity and adaptive learning models not only enable more accurate prediction of credit default risk but also enhance the risk control efficiency of financial institutions on a broader scale. The objective of this paper is to explore the application of this emerging technology, offering innovative approaches and technical support for bank credit risk management.

2. The Fundamental Theoretical Framework of Bank Credit Risk

2.1 The Concept and Classification of Credit Risk

Credit risk represents one of the core risks faced by banks and other financial institutions when providing loans to individuals or businesses, involving the potential for borrowers to default or fail to repay their debts as scheduled. Credit risk is not only reflected in individual borrower defaults but also encompasses systemic risks arising from changes in the economic environment and industry fluctuations. The concept of credit risk originates from the information asymmetry inherent in financial markets, as banks, when extending loans, cannot fully ascertain the financial condition of the borrower, future income levels, or the dynamic shifts in the economic environment. This asymmetry makes it difficult to completely eliminate credit risk.

Credit risk is typically categorized into default risk, credit migration risk, and concentration risk[1]. Default risk, the most

common type, refers to the failure of the borrower to fulfill their repayment obligations. This risk is closely tied to the borrower's credit score, cash flow, and operational status. Credit migration risk arises when the borrower's credit rating changes during the loan period, often manifesting as a downgrade, leading banks to reassess the risk of their asset portfolios. Concentration risk pertains to the excessive focus of loans in a specific industry or region, which exposes the entire loan portfolio to systemic shocks when external economic conditions change. These various types of credit risk are interconnected, creating a complex risk management structure that requires multi-layered risk mitigation measures in credit decision-making.

2.2 The Main Theoretical Foundations of Credit Risk Management

The theoretical foundations of credit risk management derive from modern financial theory, primarily encompassing credit risk evaluation, capital allocation theory, and portfolio management theory. Credit risk evaluation theory focuses on predicting the probability of borrower default using quantitative metrics. Typical methods include credit scoring models, probability of default models, and market-based default analysis tools. Credit scoring models, such as the Z-score model and FICO scores, have been widely applied historically. These tools analyze historical data to estimate default probabilities, providing support for credit decision-making[2].

Capital allocation theory plays a significant role in credit risk management. The Basel Accords introduced capital adequacy requirements, mandating banks to allocate sufficient capital based on their credit risk levels to prevent systemic collapse caused by borrower defaults. This theory underscores the importance of ensuring that banks hold adequate capital reserves to cover potential credit losses when issuing loans. Dynamic capital regulation strategies, which adjust capital buffers according to macroeconomic fluctuations, have gained increasing attention, as they enhance banks' ability to withstand economic cycle variations.

Portfolio management theory emphasizes risk diversification to optimize credit asset portfolios. According to modern portfolio theory, financial institutions can reduce the impact of single risk factors on the entire portfolio by distributing credit assets across different industries, regions, and borrower groups. Credit risk management involves not only the assessment of individual borrowers but also the systemic optimization of the entire loan portfolio. By incorporating diversification and hedging tools, overall risk exposure is minimized.

2.3 Key Factors in Credit Risk Assessment

The process of credit risk assessment involves the comprehensive consideration of multiple dimensions and factors. The borrower's financial condition and credit history serve as the core basis for evaluating credit risk. Financial condition indicators, such as liquidity ratios and debt-to-asset ratios, reflect the borrower's ability to repay and manage cash flow. Credit history, including past repayment behavior and records of credit defaults, provides insight into the borrower's historical credit performance.

Additionally, the borrower's industry background and the macroeconomic environment play a crucial role in credit risk assessment. Certain industries are more susceptible to cyclical risks, being particularly affected by economic fluctuations, such as the real estate or energy sectors. During economic downturns, businesses in these industries face greater operational pressures, which in turn increase default risk. Financial institutions must adjust their lending dynamically, considering industry-specific cyclical characteristics to prevent excessive risk concentration.

Macroeconomic indicators also constitute an essential component of credit risk evaluation. Factors such as interest rate changes, inflation rates, and monetary policy directly influence the borrower's repayment ability and the broader credit environment. Rising interest rates, for instance, increase the repayment burden on borrowers, particularly those relying on short-term loans. Tightening monetary policy could also constrain banks' lending capacity, potentially impacting overall credit market stability. Therefore, in credit risk assessment, banks must utilize a robust indicator system, integrating micro-level borrower behavior with macro-level economic conditions to conduct a comprehensive and in-depth analysis, ensuring the safety and stability of loan decisions.

3. Application of Machine Learning Technology in Credit Risk Identification

3.1 Overview of Machine Learning Technology and Its Applicability in the Financial Sector

Machine learning technology, essentially a data-driven algorithmic system, can automatically extract underlying patterns and features from vast amounts of data, facilitating prediction and decision-making[3]. Unlike traditional rule-based systems, machine learning relies on statistical methods and probability theory, making it capable of handling nonlinear relationships and high-dimensional data. The financial sector, particularly in the context of bank credit risk management, is characterized by high complexity and dynamism, where traditional statistical approaches often struggle to meet the demands posed by massive datasets and diverse feature variables. Machine learning, with its advantages in big data processing, real-time learning, and model updating, has gradually become a crucial tool in financial risk management. Financial data come from a wide range of sources, including customer financial status, behavioral patterns, and market trends. These datasets are often highly complex and noisy, but machine learning techniques demonstrate strong adaptability and robustness in processing such high-dimensional data. In credit risk management, financial institutions must identify potential high-risk clients from large datasets and predict future default behaviors. Traditional statistical models face limitations when dealing with unstructured data, while the broad application of machine learning effectively addresses these challenges. Its applicability in the financial domain is evident not only in its capacity for feature extraction but also in its ability to handle large volumes of complex data within short time frames and make effective decisions.

3.2 Core Algorithms of Machine Learning in Credit Risk Prediction

The successful application of machine learning technology in credit risk prediction depends on a series of core algorithms. These algorithms are capable of extracting valuable features from the data while accurately predicting future default risks. Logistic regression, a classic machine learning algorithm, is widely used in the field of credit risk prediction. It analyzes the linear relationship between input variables and the output risk of default, resulting in a probability of default. Although simple and effective, it shows certain limitations when dealing with nonlinear data. Random forest, which constructs multiple decision trees, uses the ensemble learning approach to enhance prediction accuracy and stability. When dealing with highly noisy or complex interactive datasets, random forests exhibit strong robustness. Gradient boosting decision trees (GBDT) further enhance the effect of ensemble learning by gradually optimizing the weights of each tree, eventually forming a powerful prediction model. Support vector machines (SVM) excel in handling high-dimensional data, especially in cases where data samples are limited. SVM builds a maximum-margin

hyperplane to classify data points accurately as either "default" or "non-default." Neural networks, particularly deep learning models, through their complex network structure of multi-layer neurons, can extract deep feature patterns from vast amounts of data, making them suitable for processing highly nonlinear and complex credit risk data. Each of these core algorithms has distinct characteristics, and financial institutions can select the most appropriate algorithm for credit risk prediction based on their data structures and risk management needs.

3.3 Advantages and Challenges of Machine Learning in Credit Risk Assessment

The application of machine learning in credit risk assessment offers numerous advantages. Machine learning can extract valuable features from complex, multi-source, and unstructured data, significantly improving data processing efficiency and accuracy. The data sources for bank clients are extensive, including traditional financial data as well as behavioral data and social network data. Machine learning can unify the processing of these different types of data, uncovering hidden risk features. Machine learning has adaptive learning capabilities, allowing models to update continuously based on new data, thus improving credit risk prediction. Models can quickly respond to changes in the external economic environment and dynamically adjust risk assessment strategies, making prediction results more accurate. Machine learning also demonstrates excellent predictive capabilities when dealing with nonlinear data relationships and high-dimensional data. Traditional statistical models struggle to capture complex data characteristics effectively, whereas machine learning technologies exhibit significant advantages in this area[4].

Nevertheless, the application of machine learning in credit risk assessment also faces certain challenges. The "black box" nature of these models makes them difficult to interpret, which may hinder financial institutions from explaining the rationale behind decisions to regulators or clients. Machine learning is highly dependent on data quality; noise, missing values, or biases in the data can negatively impact the model's predictive outcomes. In credit risk management, financial institutions not only need to incorporate machine learning technologies but also must establish comprehensive data governance frameworks to ensure model stability and reliability.

4. Strategies for Identifying Default Customers Using Machine Learning Technology

4.1 Theoretical Foundations and Challenges in Identifying Default Customers

Identifying default customers is a crucial aspect of bank credit risk management, focusing on analyzing customers' historical data and behavioral patterns to assess the likelihood of future defaults. Traditional default identification relies on static financial data and credit scoring models. While these methods provide a framework for assessing default risk to some extent, they often struggle to handle the complexity of data and the dynamic nature of market environments. The theoretical foundation for default customer identification is rooted in the measurement of credit risk, primarily including default probability models, liquidity risk theory, and studies in behavioral economics regarding irrational decision-making. Default probability models typically apply statistical methods to establish the probability distribution of borrower default risks based on historical data and feature variables. Liquidity risk theory emphasizes that borrowers facing liquidity constraints may default, limiting their debt repayment capacity. Research in behavioral economics shows that a customer's default behavior is influenced not only by financial conditions but also by

psychological expectations and market signals. Default customer identification faces multiple levels of challenges, including how to effectively integrate multi-source data, capture nonlinear relationships, and accurately identify potential high-risk customers in complex economic environments.

4.2 Application Characteristics of Machine Learning in Identifying Default Customers

The application of machine learning technology brings new methodologies and technical advantages to identifying default customers. Unlike traditional statistical models, machine learning algorithms can extract valuable patterns from large, complex, nonlinear, and high-dimensional datasets, enabling more efficient risk identification[5]. Machine learning models based on big data have several key application characteristics in identifying default customers. First, machine learning can handle unstructured and semi-structured data sources, such as customer social media activities, payment behavior records, etc., thereby enhancing the comprehensiveness of models. Second, machine learning algorithms exhibit strong adaptability, dynamically updating default risk assessment models and continuously improving prediction accuracy over time. Third, unlike traditional linear models, machine learning, especially deep learning algorithms, displays high predictive accuracy when dealing with highly complex nonlinear relationships. Commonly used machine learning algorithms include support vector machines, random forests, gradient boosting decision trees, and neural networks. These models can automatically learn risk features from large datasets and optimize default customer identification results. Another characteristic of machine learning in practice is its scalability and flexibility, allowing model adjustments to meet different risk preferences and strategic needs, providing customized risk prediction solutions.

4.3 Model Optimization to Improve the Accuracy of Default Customer Identification

Model optimization is a critical step in enhancing the accuracy of default customer identification. By improving algorithm design, feature selection, and parameter tuning, significant improvements in the predictive performance of machine learning models can be achieved. In the feature selection phase, identifying variables strongly correlated with default behavior can greatly reduce model complexity while enhancing prediction efficiency. Common feature selection methods include recursive feature elimination and information gain-based ranking methods. By retaining the most predictive features, models can maintain high accuracy while lowering computational costs. Hyperparameter adjustment is crucial during model training, and careful setting of learning rates, decision tree depths, and regularization parameters can significantly enhance the model's generalization ability and stability. Ensemble learning methods, such as random forests and gradient boosting decision trees, are widely used in the field of default customer identification. These methods combine multiple weak classifiers to enhance model robustness and noise resistance. Model optimization also involves the application of model integration techniques such as Bagging and Boosting, which construct multiple different sub-models and average their results to improve overall model performance. Selecting appropriate evaluation metrics is equally important. Commonly used metrics include the area under the ROC curve (AUC), precision, and recall rates. These metrics help analyze model performance in real-world application scenarios, further refining the model structure and parameter settings to ensure higher predictive capability and adaptability when facing complex data and economic environments.

5. Conclusion

This paper provides an in-depth discussion of the theoretical framework of credit risk management and the application of machine learning technology in this field, highlighting the importance of integrating artificial intelligence in risk management. The analysis of machine learning technology's application in credit risk assessment and default customer identification shows that this technology, with its powerful data processing capabilities and flexible model design, offers financial institutions more precise and efficient risk control methods. Although the application of machine learning still faces certain technical and practical challenges, its potential remains immense and far-reaching.

References

[1] Lei Xinnan, Lin Lefan, Xiao Binqing, et al. Re-exploring the Default Characteristics of Small and Micro Enterprises: A Machine Learning Model Based on SHAP Interpretation [J]. Chinese Management Science, 2024(5):61-65.

[2] Lei Cheng, Zhang Lin. A Federated Learning Model Based on Update Quality Detection and Malicious Client Identification [J]. Computer Science, 2024(4):21-27.

[3]Li Chaojie. Research on Intelligent Decision-Making System for Credit Supply to Small and Micro Enterprises Based on Machine Learning: A Case Study of Shared Tax Data from the "Bank-Tax Interaction" Policy [J]. Journal of Hohai University: Philosophy and Social Sciences Edition, 2022, 24(4):66-74.

[4] Lu Rongwei, Huang Changer, Xie Jiuhui. Research on Credit Card Overdue Prediction Based on Machine Learning [J]. Science and Technology Innovation, 2024(006):000.

[5] Gu Zhouyi, Hu Lijuan. Research on Credit Risk Assessment of Commercial Bank Customers from the Perspective of Machine Learning [J]. Financial Development Research, 2022(1):6.

