Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.22400

Application of Data Mining Algorithms in Bank Credit Risk Assessment: Review and Prospect

Yidan Yang

Future Technology College, Nanjing University of Information Science and Technology, Nanjing, China
15371951058@163.com

Abstract: Against the backdrop of intensified competition in the financial market and increased demand for risk management and control, the wave of digital transformation is profoundly reshaping the landscape of the financial industry. The accumulation of massive amounts of data has brought new opportunities and challenges for financial institutions to innovate business models and enhance risk management capabilities. This paper focuses on the application of data mining algorithms in bank credit risk assessment. Through literature review and case analysis, it explores the application status, effects, challenges faced, and looks ahead to future development trends. The findings domenstrate that data mining algorithms can significantly improve the accuracy of credit risk assessment and decision-making efficiency, but there are also issues such as data quality and algorithm interpretability. Future studies should focus onalgorithm integration, enhanced interpretability, and combination with emerging technologies will be the development directions. Banks should actively respond by improving data management and technology application strategies.

Keywords: Data mining algorithms, Bank credit risk assessment, Application status, Challenges, Development trends

1. Introduction

With the rapid development of the financial industry, banks are facingincreasingly complex and diverse credit risks. As one of the core businesses of banks, the accuracy of credit risk assessment in credit operations is directly related to the stable operation and economic benefits of banks. Traditional credit risk assessment methods, such as expert experience methods and credit scoring cards, have limitations when dealing with large-scale and high-dimensional data, and they have difficulty in accurately identifying potential risks. The rise of data mining algorithms has brought new opportunities for bank credit risk assessment. It can extract valuable information from massive amounts of data, reveal the potential patterns and laws behind the data, and provide a more scientific and accurate basis for credit risk assessment. Currently, many domestic and foreign banks have attempted to apply data mining algorithms to the field of credit risk assessment and achieved certain results [1,2].

This paper adapts the literature review to sort out relevant theories and practical achievements, and combines the case analysis to deeply analyze the application examples of domestic and foreign banks. It aims to explore the application status, application effects, challenges faced, and future

DOI: 10.54254/2755-2721/2025.22400

development trends of data mining algorithms in bank credit risk assessment, providing a reference for banks to better utilize data mining algorithms to improve the level of credit risk assessment. This research findings helps to promote the digital transformation of bank credit risk management, improve the efficiency of bank risk management, and ensure the stability of the financial market.

2. Basic Theories of Data Mining Algorithms and Bank Credit Risk Assessment

2.1. Overview of Data Mining Algorithms

2.1.1. Common Types of Data Mining Algorithms

Common types of data mining algorithms are rich and diverse, each representing unique functions but closely related to others, working together to extract valuable information from massive amounts of data. Main types of common data mining algorithms include: Classification algorithms (such as decision trees, random forests, support vector machines) are used to predict discrete labels; clustering algorithms (such as K-means, DBSCAN) are used to discover natural groupings in data; association rule algorithms (such as Apriori, FP-Growth) are used to mine potential associations between item sets; regression algorithms (such as linear regression, logistic regression) are used to predict continuous values; dimensionality reduction algorithms (such as PCA, t-SNE) are used to reduce the dimensionality of data while retaining key information; anomaly detection algorithms (such as isolation forests, One - Class SVM) are used to identify outliers in data; time - series analysis algorithms (such as ARIMA, LSTM) are used to analyze the trends and periodicities of time - related data [3].

2.1.2. Algorithm Selection and Application Scenarios

In the complex system of bank credit business, the rational selection of algorithms plays a decisive role in accurately assessing credit risks. This process needs to be closely combined with multiple business objectives, complex data characteristics, and limited computing resources. Algorithm selection should be combined with business objectives, data characteristics, and computing resources: Classification algorithms are used for default prediction, clustering algorithms for customer segmentation, anomaly detection for identifying fraud, time - series analysis for monitoring dynamic risks, dimensionality reduction techniques for optimizing high-dimensional data, and association rules for mining cross-selling opportunities [4]. In addition, bank credit data often has high-dimensional characteristics, including a large amount of customer information, transaction records, etc. This not only increases the complexity of data processing but may also lead to the "curse of dimensionality" problem, affecting algorithm performance. Dimensionality reduction techniques such as principal component analysis (PCA) can pre-process high-dimensional data, extract main features, reduce the dimensionality of data while retaining key information, reduce the amount of computation, and improve the operating efficiency of algorithms, enabling subsequent risk assessment models to operate more efficiently and accurately. To ensure the continuous effectiveness of algorithms in bank credit risk assessment, model iteration and optimization are indispensable. Some banks use ensemble learning methods, such as the random forest algorithm, which combines the prediction results of multiple decision tree models. The final conclusion is obtained through methods such as voting or averaging, thereby enhancing the generalization ability of the model and reducing the errors and overfitting risks of a single model. Meanwhile, the market environment and customer behavior are in dynamic change. Banks need to regularly update the training data, incorporate new credit business data, customer behavior data, etc. into model training, so that the algorithms can adapt to changes in a timely manner and maintain the ability to accurately assess credit risks.

DOI: 10.54254/2755-2721/2025.22400

2.2. Bank Credit Risk Assessment System

2.2.1. Traditional Assessment Methods and Their Limitations

Traditional bank credit risk assessment mainly relies on expert experience judgment, financial ratio analysis, and static credit scoring card models. Its core limitations are as follows: Strong subjectivity, vulnerable to human factors; single - dimensional data, overly relying on financial statements and historical repayment records, making it difficult to capture dynamic risks such as customer behavior and market fluctuations; model updates are lagging, unable to respond to changes in the economic environment in real-time; in addition, traditional methods have insufficient ability to mine non-linear relationships and potential risk associations, resulting in low-accuracy risk assessments for complex customer groups such as small and micro enterprises and emerging industries, and a lack of forward-looking early-warning capabilities, making it difficult to meet the needs of financial innovation and digital transformation.

2.2.2. Elements of the Modern Assessment System Construction

The modern bank credit risk assessment system takes data-driven as the core. It constructs a panoramic customer profile by integrating multi-dimensional data (customer behavior, third-party credit information, macro-economic indicators, etc.), uses machine learning algorithms (such as random forests, LSTM neural networks) to mine deep-level data associations and build dynamic risk prediction models, combines streaming computing technology to achieve real-time monitoring of the entire process before, during, and after the loan. Additionally, through risk quantification models (such as PD, LGD), it realizes accurate pricing, designs scenario-based assessment plans for different industries, and finally embeds technologies such as federated learning and differential privacy under the compliance framework to ensure data security, achieving intelligent, accurate, and real - time risk assessment [5].

3. Application Status of Data Mining Algorithms in Bank Credit Risk Assessment

3.1. Domestic Application Examples and Achievements

In China, many commercial banks have applied data mining technology to credit risk assessment and achieved remarkable results. For example, Lianshui Rural Commercial Bank used the Apriori association rule algorithm to mine the potential associations among credit applications, review and approval, and credit ratings. It discovered rules such as "customers with high housing provident fund amounts are more likely to obtain high credit lines". Moreover, it combined the C4.5 decision tree algorithm to build a customer credit rating prediction model, achieving a credit rating prediction accuracy rate of 75%. Through rule optimization, the matching success rate of credit products was increased by 20%, and the non-performing loan ratio was reduced by 15% [6].

3.2. Key Technical Links in the Application

Data pre-processing is a critical step that includes data cleaning, data integration, data transformation, and data reduction. Noise data and duplicate data can be removed through cleaning, data from different data sources can be integrated, the data format can be transformed to make it more suitable for algorithm processing, and the dimensionality of data can be reduced through reduction. Model training and optimization are equally important, involving the selection of appropriate training data and optimization algorithms. Adjusting model parameters, such as the learning rate and the number of hidden layer nodes in neural networks, can significantly enhance

DOI: 10.54254/2755-2721/2025.22400

model performance. Model evaluation uses indicators such as accuracy, recall rate, and F1 - value to ensure the reliability of the model [7].

4. Application Effects and Challenges of Data Mining Algorithms

4.1. Analysis of Application Effects

4.1.1. Improvement in Risk Identification Accuracy

Data mining algorithms can process multi-source and high-dimensional data, mine potential risk features, and improve the accuracy of risk identification. For example, Cai Shousong et al. could accurately predict the credit status of borrowers based on the decision tree algorithm, and the binomial logistic regression algorithm could perform well. The combination of the two could truly estimate the credit status of lenders and improve transaction efficiency [8].

4.1.2. Improvement in Decision-Making Efficiency and Quality

In the digital age today, algorithms play a transformative role in the bank credit field, greatly optimizing the credit approval process. With their powerful data-processing capabilities, algorithms can analyze and calculate massive amounts of credit data at an extremely fast speed. This data covers multiple dimensions, including customers' basic information such as age, occupation, and income level; credit records, including past loan repayment situations and credit card usage records; and financial status, such as balance sheets and cash flow data. Traditional credit approval methods rely on manual review of these complex data one by one, which is inefficient and prone to omissions. Algorithms can quickly integrate and deeply mine this massive amount of data, quickly sort out key information, and generate accurate credit risk assessment results based on pre-set models and rules.

This algorithm-based assessment model significantly shortens the credit approval time. In the past, the bank credit approval process was cumbersome. Starting from when customers submitted application materials, credit officers had to manually check and verify various information, and then conduct risk assessments based on their own experience and limited reference standards. The entire process took a long time. For example, in the traditional approval model, the approval time for a credit business was usually 3-5 days. During this period, customers might choose other financial institutions due to the long waiting time, resulting in the loss of potential customers for banks. However, with the introduction of automated assessment systems, algorithms fully take over data processing and risk assessment. After receiving a customer's application, the system instantly starts the data reading and analysis program, can complete the processing of a large amount of data in just a few minutes, and then quickly generates an assessment report. Practice has proven that banks using automated assessment systems have significantly shortened the credit approval time to 1-2 days, greatly enhancing the customer experience and the bank's competitiveness in the market.

4.2. Challenges and Problems Faced

4.2.1. Data Quality and Security Issues

In terms of data quality, problems such as missing values, outliers, redundant variables, and multicollinearity among variables commonly existing in high-dimensional data directly affect the accuracy and stability of model training. The diversity and heterogeneity of data sources lead to inconsistent data formats and ambiguous semantics, exacerbating the complexity of data cleaning and integration. In addition, error values and low-variance features generated during the data

DOI: 10.54254/2755-2721/2025.22400

collection process due to business system design flaws or human operation errors further reduce the usability of data.

Concerning data security, customer privacy information faces the risk of leakage during storage, transmission, and analysis. Especially in cross-platform data cooperation scenarios, blurred data permission boundaries and loopholes in compliance management may lead to legal disputes. Moreover, the lag in the data security technical protection system, such as insufficient encryption algorithm strength and imperfect access control mechanisms, makes it difficult to effectively resist external attacks and internal abuse risks, posing a potential threat to the sustainability of data mining applications.

4.2.2. Algorithm Complexity and Interpretability

Modern models such as deep neural networks and ensemble learning frameworks have significantly improved prediction accuracy, but their highly non-linear structures and complex parameter spaces make model tuning difficult, requiring a large amount of computing resources and time costs. The "black-box" nature of the algorithm decision-making process cause difficulty in business personnels; understanding of the model logic. Especially under regulatory compliance requirements, the inability to clearly explain the basis for risk identification may lead to regulatory questions and customer trust crises. In addition, the sensitivity of complex models to changes in data distribution exacerbates the adaptability problem in dynamic scenarios. The concept drift phenomenon may cause the model's effectiveness to decay rapidly. To address the above problems, it is necessary to develop interpretability-enhancing technologies, construct rule engines and model result verification mechanisms combined with domain knowledge, explore lightweight model architectures to balance prediction ability and interpretability, and improve the dynamic robustness of models through online learning and meta-learning technologies.

5. Future Development Trends and Suggestions

5.1. Technical Development Trends

In the field of bank credit risk assessment, the future development of data mining technology lies in multi-dimensional integration: Relying on big data and cloud computing platforms to achieve efficient processing of massive financial data, combining deep learning and neural network models to break through the bottleneck of complex non-linear relationship modeling, and promoting the in-depth application of unstructured data such as customer behavior patterns and social networks; automated machine learning will simplify the development process of credit risk control models and improve model iteration efficiency; the maturity of federated learning and privacy-computing technologies will promote cross-institutional data collaboration and build more accurate joint risk control models while protecting user privacy; the development of interpretable AI technologies will enhance model transparency and meet financial regulatory requirements through visualizing decision-making paths; interdisciplinary integration (such as the combination of financial engineering and artificial intelligence) will give birth to new risk assessment index systems, and breakthroughs in real-time streaming data mining technology will promote the evolution of credit approval to second-level response, comprehensively enabling the dynamic and accurate upgrade of intelligent risk control systems [9].

5.2. Bank Response Strategies and Suggestions

Banks should strengthen the full-process risk control mechanism, incorporate key indicators such as customers' overdue records in the past year and job characteristics into the pre-loan access

DOI: 10.54254/2755-2721/2025.22400

assessment, and achieve accurate risk identification combined with post-loan dynamic monitoring; promote data governance and technology platform construction, build an intelligent risk control analysis platform by refining data standards, optimizing system functions, and strengthening the cultivation of compound talents; deepen cross-platform data cooperation, access government public data and Internet behavior data to solve the problem of information asymmetry; optimize the functions of risk control systems, embed data mining algorithms into business processes, and improve automated approval and real-time early-warning capabilities; improve the employee compliance management mechanism, prevent moral risks through "compliance portraits", violation points, and the linkage between incentives and constraints, and form a comprehensive risk management system that combines systems, technologies, and personnel [10].

6. Conclusion

This paper explores the application of data mining algorithms in bank credit risk assessment and finds that this technology has significant effects in improving risk identification accuracy and decision-making efficiency. However, challenges such as data quality and algorithm interpretability are remains in the application process. In the future, algorithm integration, enhanced interpretability, and combination with emerging technologies are the development trends. This research has certain limitations. The analysis of some emerging algorithm application cases is not in-depth enough. Future research should focus on the practical application effects and optimization strategies of emerging algorithms in bank credit risk assessment to provide more targeted guidance for bank credit risk management.

References

[1] *Jin, J. Y. (2022). Research on Credit Risk Assessment Method for Bank Users Based on Data Mining(Master's thesis, Shenyang University of Technology). https://link.cnki.net/doi/10.27322/d.cnki.gsgyu.2022.001117*

[2] *Xiong, Y. (2023, July 13). Application and risk control of big data technology in the financial industry. Finance and Accounting News, 006. https://doi.org/10.28104/n.cnki.nckxb.2023.000319*

[3] *Lu, Y. Y. (2017). Design and Implementation of Data Mining Algorithms under Big Data Platform(Master's thesis, China University of Petroleum (Beijing)). https://cnki2.699wx.cn/kcms2/article/abstract?v=Kk8bzUe9ukp_ONBgdztkpDoNAmQFCSGt4Tpqd_7OkbFuF5qVF6TP3AUbPGUUvpOtghBE-DTDDkZRFY4FBt8ahi-nMJhOOf1ZoJNpotTeAhFcydJqGterNjTHdAWir-hx8EoYFdzDaYBRgBhSYECd7jVE1uQyEPoo0dQfFZYBlPFRoHt4rve17r8eXuBb2xgQ9ImDFIT3dNU=&uniplatform=NZKPT&language=CHS*

[4] *Pan, H. L. (2022, March 14). Prudent grasp of algorithm application scenarios. International Finance News, 014. https://doi.org/10.28403/n.cnki.nifnb.2022.000227*

[5] *Zhang, Y. (2022). Research on the Construction and Implementation of Big Data Governance System for Credit Enterprises (Master's thesis, Shanghai University of Finance and Economics). https://link.cnki.net/doi/10.27296/d.cnki.gshcu.2022.000667*

[6] *Huangfu, L. Y. (2019). Implementation of Credit Risk Management System Based on Data Mining (Doctoral dissertation, Jiangsu University of Science and Technology). https://doi.org/10.27171/d.cnki.ghdcc.2019.000021*

[7] *Ye, Y. J., & Li, C. (n.d.). Application risks and governance paths of synthetic data in artificial intelligence model training. Information Studies: Theory & Application, 1-11.*

[8] *Cai, S., & Zhang, J. (2020). Exploration of credit risk of P2P platform based on data mining technology. Journal of Computational and Applied Mathematics, 372. https://doi.org/10.1016/j.cam.2020.112718*

[9] *Xu, Y. G. (2024). A review of data mining algorithm research. Computer Knowledge and Technology, 20(24), 64-66+69. https://doi.org/10.14004/j.cnki.ckt.2024.1239*

[10] *Sun, C. (2021). Research on Credit Risk Management of A Rural Commercial Bank Based on Data Mining (Master's thesis, Nanjing University of Posts and Telecommunications). https://doi.org/10.27251/d.cnki.gnjdc.2021.001415*