Journal of Economic Theory and Business Management
ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online) | Vol. 1, No. 3, 2024

Deep Learning Applications in Personal Credit Risk Assessment: Insights from Big Data in Banking

GUPTA, Neha [1*] SHARMA, Kritika [2] VERMA, Siddharth [3]

[1] Indian School of Business, India
[2] Indian Institute of Management, India
[3] Indian Institute of Technology, India

**** GUPTA, Neha is the corresponding author, E-mail: Gupta889@gmail.com***

Abstract: This study explores integrating big data and advanced deep learning techniques for enhancing personal credit risk assessment in commercial banks. Traditional methods must be improved in high-dimensional, sparse, and noisy big data environments. Key challenges include data source diversity, variable selection complexity, and methodological differences in modeling. By leveraging deep learning approaches like Stack Denoising Autoencoder Neural Networks (SDAE-NN) and addressing imbalanced data using Generative Adversarial Networks (GANs), this research aims to develop robust frameworks that improve the accuracy and efficiency of credit risk evaluation.

Keywords: Big Data, Credit Risk Assessment, Deep Learning, Financial Institutions.

DOI: https://doi.org/10.5281/zenodo.11637270

1 Introduction

In recent years, the increasing demand for personal loans has corresponded with a rise in default risks, posing a significant challenge not only to commercial banks but also as a crucial destabilizing factor in the entire financial system. The comprehensive and accurate assessment of personal credit risk remains pivotal for financial institutions, including commercial banks, enhancing their risk management capabilities amidst rapid growth in personal lending. [1] Traditional credit risk assessment overly relies on individual credit reports, which often need to catch up on data timeliness, comprehensiveness, and diversity, failing to meet the evolving needs of today's banking sector.

The advent of the significant data era has enriched personal credit profiles with diverse data sources, mainly as commercial banks accumulate ample data resources. [2,3,4] However, leveraging big data from banks for more comprehensive personal credit risk assessment presents challenges. Big data's high-dimensional and sparse nature complicates feature selection, rendering traditional credit risk assessment methods less suitable. Moreover, the prevalence of high-noise data in big data environments demands effective solutions for accurate personal credit risk evaluation. [5] Addressing imbalanced data samples remains another critical issue that directly impacts the efficacy of risk assessment models.

To harness bank big data effectively for personal credit risk assessment and overcome challenges posed by high-dimensional, sparse, and noisy data, as well as imbalanced datasets, advanced deep learning techniques represent a promising avenue. This study integrates cutting-edge deep learning technologies from artificial intelligence with statistical analysis of bank big data to construct robust frameworks for assessing personal credit risk.

2 Related Work

2.1 Financial credit evaluation

With the advent of the significant data era, data volume and dimensionality have witnessed explosive growth (Ma Shilong, 2016). With internet finance's rise, internet companies and fintech firms continuously compete to bank customer resources, especially long-tail users, leveraging big data technologies to integrate data from various sources. [4,6,7,8] This enables them to uncover patterns that reflect individual behaviors, meeting diverse personalized demands while utilizing big data for risk assessment and control (Ba Shusong, 2016). For commercial banks, relying solely on traditional credit metrics for personal credit risk assessment no longer meets the dynamic financial demands of the big data environment. [3] The challenge lies in effectively harnessing ample data resources to establish risk assessment models that unearth rich, multi-source information embedded within big data, thereby enhancing the evaluation capabilities of personal credit risk from heavy asset-based credit data to lightweight data resources expressed through vast amounts of big data. This transition has gradually become a key factor for commercial banks to enhance their core competitiveness in the fintech era (Jiang Zengming et al., 2019).

Regarding assessment methods, existing research literature and industry practices have proposed and applied a series of theories and methods to address credit risk assessment issues. [9] Broadly categorized, these methods include traditional statistical learning approaches and machine learning-based methods (Chen et al., 2015; Lin, 2012). However, challenges such as high feature dimensionality, data sparsity, and noise are prevalent in the big data environment. Traditional methods must be revised to meet the current demands of credit assessment in big data environments, particularly inefficiently and reasonably describing user characteristics based on domain expert experience, thereby affecting final classification outcomes. [10,11,12,13] With the advancement of artificial intelligence, increasing attention has been directed toward applying cutting-edge intelligent algorithms to financial domains, including credit risk assessment. In recent years, deep learning, which has made significant strides in fields like image, video, speech, and natural language processing, stands out. Exploring how advanced machine learning methods can be integrated with the big data environment of the financial sector for personal credit risk assessment represents a promising research direction.

2.2 Learning Credit Imbalanced Data Samples Using Generative Adversarial Networks

Imbalanced data is a pervasive issue in credit risk assessment, where the number of high credit samples significantly outweighs low credit or default samples. [14] Typically, there are far fewer instances of low credit or default cases compared to high credit cases. Machine learning methods for model learning and prediction typically require a balanced distribution of samples across different labels, especially in supervised learning tasks. Severe imbalance or skewness in data often leads models to favor the majority class, making it challenging to extract and learn crucial information from minority samples. [15] Sometimes, these minority samples might be treated as outliers or excluded altogether. Consequently, models fail to discern patterns across labels, reducing prediction accuracy or failure. Therefore, before conducting data analysis and modeling on personal credit datasets constructed from bank big data, it is essential to appropriately address imbalance issues to extract vital insights from relatively fewer samples, facilitating subsequent model development and evaluation.

This chapter first reviews and summarizes existing methods and their effectiveness in handling imbalanced samples. It identifies and analyzes issues with current approaches and proposes an improved strategy based on Generative Adversarial Networks (GANs) [16]. Building on relevant theories, this chapter designs and introduces enhanced learning methods using GANs. Experimental validation and comparison with existing methods demonstrate the effectiveness and broader applicability of the proposed approach.

2.3 Deep Learning-Based Personal Credit Risk Assessment

In the realm of personal credit risk assessment using bank big data, models' effective learning and prediction capabilities are crucial considerations. Unlike traditional small datasets or sample sets, extensive data collections present challenges such as high dimensionality, sparse features, and sheer volume. Addressing these challenges requires leveraging advanced deep-learning techniques alongside robust feature selection methods.

This chapter builds upon the theoretical foundations of deep learning and feature selection, focusing on their application in credit risk assessment. Given the characteristics of bank big data utilized in this study, the chapter explores state-of-the-art deep learning methodologies from artificial intelligence. Specifically, it introduces and designs a Stack Denoising Autoencoder Neural Network (SDAE-NN) to enhance the processing of high-dimensional and sparse features in bank big data. [17,18,19,20,21] The SDAE-NN framework not only extracts meaningful representations from complex data but also mitigates noise inherent in large datasets, thereby improving the accuracy and robustness of credit risk assessment models. Experimental examples and expanded content illustrate the practical implementation and efficacy of the proposed approach in handling the unique challenges posed by big data environments for personal credit risk evaluation.

1. methodology

3.1 Personal Credit Risk Theory

Credit risk can be broadly categorized into general and specific definitions. In a broader sense, credit risk refers to the potential loss incurred by a counterparty defaulting on a credit transaction, encompassing any credit-related activity. Conversely, in a narrower context, credit risk specifically denotes the risk to creditors resulting from debtors defaulting on their obligations. From the perspective of commercial banks, credit risk associated with clients is often referred to as default risk or loan risk. [22] The occurrence of credit risk is influenced not only by the borrower's repayment capacity but also by their willingness to repay and ethical considerations, compounded by market risks such as interest rate fluctuations and exchange rate volatility.

Credit risk assessment involves financial institutions and rating agencies objectively and fairly evaluating an entity's ability to fulfill financial obligations and their trustworthiness using rigorous analytical methods and assessment systems. [23] This evaluation results in the assignment of a credit rating or credit risk determination, which guides economic activities. Within the framework of commercial banks, assessment of individual credit risk, also known as personal loan risk or individual default risk, evaluates the foreseeable risk of default based on an

individual's existing and historical credit records, repayment capability, willingness to repay, and potential default risks associated with various credit transactions.

In a market economy, all economic entities are interconnected through credit relationships. [24,25,26] Commercial banks, as significant participants in the market economy, fundamentally operate based on credit. Market participants benefit economically and transactionally from favorable credit standings. Thus, an individual's creditworthiness represents accumulated credit capital, reflecting not only their behavior and performance in credit borrowing but also their comprehensive societal, ethical, and legal aspects.

3.2 Causes and Factors Influencing Personal Credit Risk

Several factors contribute to personal credit risk, which can be broadly categorized into repayment capacity, repayment willingness and moral hazard, credit asymmetry, imperfect credit systems, and counterparty credit risk.

1. Repayment Capacity: The primary aspect of assessing an individual's creditworthiness is whether they have sufficient income sources or collateral (such as mortgages or pledges) to repay credit assets. Even if other conditions are favorable, lacking quantifiable funds or reserve guarantees for future loan repayments can hinder a positive credit evaluation.

2. Repayment Willingness and Moral Hazard [27]: Repayment willingness refers to an individual's tendency to default. Generally, moral hazard influences repayment willingness; lower moral hazard indicates higher repayment willingness. Moral hazard encompasses behaviors like concealing or transferring assets to avoid fulfilling credit contracts or deviating from loan purposes. Changing loan purposes or investing in higher-risk projects can drive moral hazard.

3. Credit Asymmetry [28]: Information asymmetry is prevalent in financial market activities, particularly in personal credit-related transactions where individuals have an information advantage over commercial banks or credit institutions. Limited access to comprehensive and accurate personal credit information restricts institutions from making objective assessments, potentially leading to opportunistic defaults due to the individual's informational advantage.

4. Credit System Issues: Issues within credit systems, such as incomplete credit support systems (e.g., inadequate personal credit records or asset declaration systems), hinder the establishment of robust credit regulations—inadequate legal frameworks related to personal credit delay the utilization and development of personal credit data. Moreover, the absence of credit information sharing mechanisms and effective mechanisms for cultivating and using credit products further constrains credit risk assessment, fostering the propagation of credit risk [29].

In summary, the primary contributors to personal credit risk are an individual's repayment capacity and willingness. Assessment indicators for personal credit risk typically qualitatively analyze these factors.

3.3 Construction of individual credit risk assessment characteristics of bank big data

Utilizing Big Data for Credit Risk Assessment presents several notable differences compared to traditional credit scoring:

1. Differences in Data Sources and Features: Traditional credit risk assessment models primarily rely on personal application information submitted by customers when applying for credit products, internal credit transaction data within financial institutions, and data from external credit bureaus like the People's Bank of China Credit Reference Center. [30] These data sources typically exhibit high-value density and quality, albeit with relatively fewer dimensions—usually fewer than 30 variables for modeling. In contrast, the era of big data introduces a more diverse array of personal data sources. Internally, within assessment institutions, this includes not only credit-related data but also other relevant data accumulated about subjects, as well as non-credit business data and external transactional data interfacing with the institution. Externally, data is acquired through the Internet and partnerships with other organizations, such as consumer e-commerce and social network data. Characteristics of big data include high data dimensions due to its wide-ranging sources, significant data sparsity due to varied collection channels and inconsistent standards, and low single-factor value density. This necessitates aggregation, summarization, integration, and algorithmic processing to enhance data differentiation and value in credit assessment.

2. Differences in Variable Selection Methods [27]: The selection of variables directly influences model learning and evaluation outcomes. Traditional credit risk assessment involves fewer variables, typically under several dozen, selected through statistical analysis and variable selection methods to identify crucial credit-related modeling variables with clear interpretability. In contrast, big data introduces hundreds to thousands of dimensions, effectively making it challenging to screen variables through rule-based or manual feature selection methods. Despite the low contribution of individual variables to credit assessment due to their low-value density, combining and altering features can enhance evaluation capabilities. Traditional feature selection methods need help to achieve effective or satisfactory results in this context.

3. Differences in Modeling Methods: Traditional credit assessment models predominantly utilize statistical methods and simpler data mining algorithms such as logistic regression and decision tree models. In contrast, credit risk models in the big data environment primarily rely on

machine learning methods that prioritize data-driven insights over investigating causal relationships of credit risk. [3,4,5,31] While individual variable correlations in big data may be weak, expanding data dimensions and feature combinations strengthen variable description and differentiation capabilities.

4. Other Aspects: Beyond these primary differences, application architecture, and model development operational efficiency are influenced by data usage and modeling methods disparities, impacting engineering practices and operational outcomes.

The divergence between big data and traditional credit scoring is evident across data sources and features, variable selection methods, modeling techniques, and practical application frameworks. [6,32] These distinctions highlight the evolution from a data-scarce, high-value density environment to a data-rich, high-dimensionality context requiring advanced processing techniques for practical credit risk assessment.

4 Conclusion

The issue of data imbalance is prevalent in credit risk assessment, where the number of high-credit samples (termed "good samples") far outweighs low-credit or default samples (termed "bad samples"). Typically, the quantity of default cases is significantly smaller, reflecting real-world financial scenarios where default rates generally hover around a few percent. However, default incidents can profoundly impact financial stability even at these levels, necessitating effective control measures. Machine learning methods for model learning and prediction require balanced sample quantities across different labels, especially in supervised learning tasks. Severe imbalance or skewness in data often biases models towards fitting the majority class, making it challenging to extract and learn critical insights from minority samples. Inadequate learning from these minority samples may result in misclassification or exclusion, reducing model accuracy and effectiveness. For instance, in personal credit risk assessment, the imbalance between good and bad samples can prevent models from discerning distinguishing patterns, potentially mislabeling credit default clients as good credit clients. Such errors directly elevate the risk of credit asset losses, causing direct losses to financial institutions or borrowers. Therefore, addressing data imbalance is crucial before utilizing features derived from bank big data for risk modeling, ensuring effective learning from relatively fewer data samples to enhance subsequent model development and evaluation efforts.

Acknowledgments

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author Contributions

Not applicable.

About the Authors

GUPTA, Neha

Financial Risk, Indian School of Business (ISB), Hyderabad, India.

SHARMA, Kritika

Business Administration, Indian Institute of Management (IIM) Bangalore, India.

VERMA, Siddharth

Electronic information engineering, Indian Institute of Technology (IIT) Kanpur, India.

References

[1] Ni, C., Zhang, C., Lu, W., Wang, H., & Wu, J. (2024).

Enabling Intelligent Decision Making and Optimization in Enterprises through Data Pipelines.

[2] Bao, Q., Wei, K., Xu, J., & Jiang, W. (2024). Application of Deep Learning in Financial Credit Card Fraud Detection. Journal of Economic Theory and Business Management, 1(2), 51-57.

[3] Bai, X., Jiang, W., & Xu, J. (2024). Development Trends in AI-Based Financial Risk Monitoring Technologies. Journal of Economic Theory and Business Management, 1(2), 58-63.

[4] Wang, Y., Zhan, X., Zhan, T., Xu, J., & Bai, X. (2024). Machine Learning-Based Facial Recognition for Financial Fraud Prevention. Journal of Computer Technology and Applied Mathematics, 1(1), 77-84.

[5] Yang, P., Shui, Z., Chen, Z., Baoming, W., & Lei, H. (2024). Integrated Management of Potential Financial Risks Based on Data Warehouse. Journal of Economic Theory and Business Management, 1(2), 64-70.

[6] Wang, J. (2024). Fraud Detection in Digital Payment Technologies Using Machine Learning. Journal of Economic Theory and Business Management, 1(2), 1-6.

[7] Xu, Jiahao, et al. "AI-BASED RISK PREDICTION AND MONITORING IN FINANCIAL FUTURES AND SECURITIES MARKETS." The 13th International scientific and practical conference “Information and innovative technologies in the development of society”(April 02–05, 2024) Athens, Greece. International Science Group. 2024. 321 p.. 2024.

[8] Song, Jintong, et al. "LSTM-Based Deep Learning Model for Financial Market Stock Price Prediction." Journal of Economic Theory and Business Management 1.2 (2024): 43-50.

[9] Jiang, W., Yang, T., Li, A., Lin, Y., & Bai, X. (2024). The Application of Generative Artificial Intelligence in Virtual Financial Advisor and Capital Market Analysis. Academic Journal of Sociology and Management, 2(3), 40-46.

[10] Li, H., Wang, X., Feng, Y., Qi, Y., & Tian, J. (2024). Driving Intelligent IoT Monitoring and Control through Cloud Computing and Machine Learning.arXiv preprint arXiv:2403.18100.

[11] Qi, Y., Wang, X., Li, H., & Tian, J. (2024). Leveraging Federated Learning and Edge Computing for Recommendation Systems within Cloud Computing Networks. arXiv preprint arXiv:2403.03165.

[12] Qi, Y., Feng, Y., Tian, J., Wang, X., & Li, H. (2024). Application of AI-based Data Analysis and Processing Technology in Process Industry.Journal of Computer Technology and Applied Mathematics,1(1), 54-62.

[13] Tian, J., Qi, Y., Li, H., Feng, Y., & Wang, X. (2024). Deep Learning Algorithms Based on Computer Vision Technology and Large-Scale Image Data.Journal of Computer Technology and Applied Mathematics,1(1), 109-115.

[14] Wang, X., Tian, J., Qi, Y., Li, H., & Feng, Y. (2024). Short-Term Passenger Flow Prediction for Urban Rail Transit Based on Machine Learning.Journal of Computer Technology and Applied Mathematics,1(1), 63-69.

[15] Feng, Y., Li, H., Wang, X., Tian, J., & Qi, Y. (2024). Application of Machine Learning Decision Tree Algorithm Based on Big Data in Intelligent Procurement.

[16] Tian, J., Li, H., Qi, Y., Wang, X., & Feng, Y. Intelligent Medical Detection and Diagnosis Assisted by Deep Learning.

[17] Ding, W., Zhou, H., Tan, H., Li, Z., & Fan, C. (2024). Automated Compatibility Testing Method for Distributed Software Systems in Cloud Computing.

[18] Qian, K., Fan, C., Li, Z., Zhou, H., & Ding, W. (2024). Implementation of Artificial Intelligence in Investment Decision-making in the Chinese A-share Market. Journal of Economic Theory and Business Management, 1(2), 36-42.

[19] Li, Zihan, et al. "Robot Navigation and Map Construction Based on SLAM Technology." (2024).

[20] Fan, C., Ding, W., Qian, K., Tan, H., & Li, Z. (2024). Cueing Flight Object Trajectory and Safety Prediction Based on SLAM Technology. Journal of Theory and Practice of Engineering Science, 4(05), 1-8.

[21] Fan, C., Li, Z., Ding, W., Zhou, H., & Qian, K. Integrating Artificial Intelligence with SLAM Technology for Robotic Navigation and Localization in Unknown Environments.

[22] Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. Immediate Traffic Flow Monitoring and Management Based on Multimodal Data in Cloud Computing.

[23] Wang, Y.,Cao, J., Ku, D., Du, J., Ng, V., & Dong, W. "A Structurally Enhanced, Ergonomically and Human–Computer Interaction Improved Intelligent Seat’s System," Designs, vol. 1, no. 2, 2017, p. 11, doi: 10.3390/designs1020011.

[24] Xu, J., Wu, B., Huang, J., Gong, Y., Zhang, Y., & Liu, B. (2024). Practical Applications of Advanced Cloud Services and Generative AI Systems in Medical Image Analysis. arXiv preprint arXiv:2403.17549.

[25] Zhang, Y., Liu, B., Gong, Y., Huang, J., Xu, J., & Wan, W. (2024). Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management. arXiv preprint arXiv:2402.17216.

[26] Zhan, X., Shi, C., Xu, K., Li, L., & Zheng, H. (2024). Aspect category sentiment analysis based on multiple attention mechanisms and pre-trained models. Applied and Computational Engineering, 71, 21-26.

[27] He, Zheng, et al. "Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering."

[28] Zhou, Y., Zhan, T., Wu, Y., Song, B., & Shi, C. RNA Secondary Structure Prediction Using Transformer-Based Deep Learning Models.

[29] Lin, T., & Cao, J. "Touch Interactive System Design with Intelligent Vase of Psychotherapy for Alzheimer's Disease," Designs, 2020, 4(3), 28. Journals Designs Volume 4 Issue 3 10.3390/designs4030028

[30] Liu, B., Cai, G., Ling, Z., Qian, J., & Zhang, Q. Precise Positioning and Prediction System for Autonomous Driving Based on Generative Artificial Intelligence.

[31] Cui, Z., Lin, L., Zong, Y., Chen, Y., & Wang, S. Precision Gene Editing Using Deep Learning: A Case Study of the CRISPR-Cas9 Editor.

[32] Bai, X., & Braganza, M. (2024). Application of Artificial Intelligence Technology in Financial Risk Control. Academic Journal of Sociology and Management, 2(3), 47-53.