### CHAPTER TWO

### 2.0 Literature Review

### With the financial industry's growing reliance on automated decision-making procedures for lending, credit default prediction has emerged as a critical study topic. By providing strong tools for credit default prediction, machine learning (ML) helps financial institutions improve their risk assessment and management plans.

### 2.1 Definition of Credit Default

### The term "credit default" describes a borrower's inability to fulfill their responsibilities under a loan agreement, particularly their incapacity to make scheduled payments. Both borrowers and financial institutions, especially banks, face serious difficulties as a result of this issue. Default can have serious consequences for borrowers, such as collateral loss, legal penalties, harm to creditworthiness, and limited access to future lending choices (Domeher & Abdulai, 2012). Institutional banks experience significant capital and income losses, more expenses for tracking and collecting past-due loans, and decreased sustainability and profitability as a result of an increase in non-performing assets (Ahlin et al., 2011). Credit default has ramifications that go beyond specific lenders and borrowers; it has the potential to undermine the larger financial system, impacting investor confidence, liquidity, and general economic stability.

### 2.2 Importance of Predicting Credit Default for Financial Institutions

### Predicting loan defaults with accuracy is essential for risk mitigation and efficient financial management, which has a big impact on both specific financial institutions and the overall economy. For banks, anticipating borrower defaults is crucial since it guides their lending policies and risk management plans. According to Louzada et al. (2016), accurate default forecasting is essential for capital allocation optimization, reducing financial losses, and preserving a balanced loan portfolio. Predicting defaults necessitates a thorough examination of numerous risk indicators, which emphasizes the necessity for dynamic credit scoring models that can adjust to shifting borrower patterns and market dynamics.

### The variety of borrower profiles and the increasing complexity of financial markets provide significant challenges for traditional credit scoring methods. Financial organisations have historically found it challenging to manually assess a client's creditworthiness, especially during periods when their clientele is rapidly expanding. The delays in loan processing brought on by this human assessment may impair the effectiveness of banks and other financial organisations. Automating credit assessment procedures with machine learning and artificial intelligence has become essential to increase loan processing speed, accuracy, and economy (Albastaki & RIT Libraries, 2022). By employing AI-driven models, financial institutions may more precisely predict the likelihood of loan defaults, allowing for timely interventions and informed lending decisions. Actions should be taken to stop criminals because credit card fraud is increasing quickly on a global scale. Customers would benefit by setting a limit on those activities as their money would be reclaimed and returned to their accounts, and they wouldn't be billed for goods or services they didn't buy (AlEmad & RIT Libraries, 2022).

**2.3 Brief History of ML Applications in Finance**

The application of machine learning (ML) in the finance sector has a rich history, beginning in the early 1980s. One of the first instances was Apex’s launch of PlanPower in 1982, an AI program designed to provide tax and financial advice to clients with high incomes. The subsequent launch of the Personal Financial Planning System by Chase Lincoln First Bank in 1987 marked another milestone. In 1989, the introduction of the FICO Score revolutionized credit scoring, utilizing algorithms that remain foundational in today's banking practices (Sharma, 2023).

Over the decades, financial institutions have increasingly adopted machine learning techniques, particularly for applications such as fraud detection in credit card transactions. The British fund manager Man Group began employing ML for investment strategies in 2014, and by 2016, Bank of America launched its AI chatbot, Erica, enhancing customer interaction. The momentum continued in 2018 with various financial institutions developing recommendation systems to improve customer engagement and service (Sharma, 2023).

A recent survey conducted by the Bank of England (BoE) and the Financial Conduct Authority (FCA) revealed a growing adoption of machine learning within UK financial services, with respondents anticipating significant growth in ML applications over the coming years. The survey found that the median respondent expects their number of ML applications to more than double within three years, with banks and insurance firms predicting an even larger increase (Machine Learning in UK Financial Services, 2024). This trend indicates a strong recognition of the value that machine learning brings to financial operations, enhancing efficiency and decision-making.

## **2.4 How Machine Learning is Used in Finance**

Machine learning applications in finance are diverse, encompassing various functions that improve operational efficiency and decision-making. Some of the prominent use cases include:

### 2.4.1 Algorithmic Trading

Algorithmic trading leverages machine learning algorithms to enhance trading decisions by analyzing market data in real-time. Traders develop mathematical models that monitor business news and trading activities to identify factors that could influence security prices. These models operate under predetermined sets of parameters—such as timing, price, and quantity—allowing for automated trading without human intervention (Team, 2024). Unlike human traders, algorithmic trading can analyze vast datasets simultaneously, making rapid trading decisions that provide a competitive edge in the market. The emotional neutrality of algorithms also reduces the risk of poor trading decisions driven by psychological biases.

### 2.4.2 Fraud Detection and Prevention

Fraud detection remains a critical concern for banking institutions, leading to billions of dollars in losses annually. Traditional fraud detection systems often relied on rule-based approaches, which could easily be circumvented by sophisticated fraudsters. Modern financial services now utilize machine learning to detect unusual activities and anomalies within large datasets, significantly enhancing fraud prevention efforts (Leyden, 2024). By analyzing transaction data against historical account behaviors, machine learning systems can identify potential frauds in real-time, allowing financial institutions to act swiftly to mitigate losses.

### 2.4.3 Portfolio Management (Robo-Advisors)

Robo-advisors represent another innovative application of machine learning, offering automated financial advice to investors. These online platforms use algorithms to create personalized investment portfolios based on individual goals and risk tolerances (Team, 2024). With lower account minimums and fees compared to traditional portfolio managers, robo-advisors democratize access to investment strategies. By inputting their financial goals, investors can leverage these systems to optimize their portfolios across various asset classes, enhancing long-term financial outcomes.

### 2.4.4 Loan Underwriting

In the banking sector, machine learning streamlines the underwriting process by enabling rapid assessments of loan applications. Algorithms analyze extensive consumer data, such as age, income, occupation, and credit history, to make informed decisions regarding loan approvals (Team, 2024). By training machine learning models to recognize patterns and exceptions in consumer data, financial institutions can significantly reduce the time and resources spent on manual evaluations, leading to more efficient loan processing.

**2.5 Related Works**

The application of machine learning to credit risk assessment has gained traction due to its potential to analyze large datasets more effectively than traditional statistical methods. For instance, Rahman et al. (2018) developed a model that utilizes multiple machine learning techniques to predict loan defaults using a modified version of the German credit dataset. The authors selected 23 features through rigorous feature selection methods and employed algorithms including Logistic Regression, Naïve Bayes, Decision Trees, and Random Forests. Their findings indicated that the Decision Tree model was particularly effective for predicting credit defaults, underscoring the significance of feature selection in enhancing predictive accuracy.

Similarly, Uddin and Rahman (2024) conducted a comparative analysis of various machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest. Their study demonstrated that Random Forest and Decision Tree models achieved the highest accuracy rates of 92% and 94%, respectively, when trained on a dataset from a private bank in Dhaka, Bangladesh. This reinforces the idea that certain algorithms, especially ensemble methods like Random Forest, can significantly improve predictive performance in credit default scenarios.

A central theme in the literature is the comparative analysis of machine learning algorithms. Bazzana et al. (2024) explored this by evaluating various classifiers on a dataset comprising small Italian companies. Their study found that machine learning models generally outperformed traditional logistic regression in predicting defaults. However, the performance gains were modest, suggesting that while machine learning offers improvements, classical models remain relevant due to their interpretability and ease of implementation.

Similarly, Liu et al. (2022) applied k-nearest neighbor, SVM, and Random Forest algorithms to predict default probabilities in China's online credit market. Their results indicated that machine learning models significantly outperformed logistic regression based on metrics such as accuracy and area under the ROC curve (AUC), thus demonstrating the enhanced predictive capabilities of these algorithms.

Athreyas et al. (2022) also contributed to this comparative discourse by analyzing various machine learning techniques to determine their efficacy in predicting loan eligibility and default. Their findings indicated that different algorithms showed varying levels of success, emphasizing the need for context-specific model selection.

Feature selection is a critical component in building effective machine learning models for credit default prediction. Wu (2022) employed feature engineering techniques such as variance threshold and Variance Inflation Factor (VIF) to filter out irrelevant features before applying Random Forest and XGBoost algorithms. Both models achieved high accuracy rates, demonstrating that effective feature selection can enhance the performance of ML algorithms significantly.

In the context of the Bank of Taiwan, Arora et al. (2022) utilized historical data to identify patterns related to loan defaults. Their analysis revealed that logistic regression provided valuable insights into the relationships between various independent variables and the likelihood of default, emphasizing the importance of interpretability in machine learning applications.

Beyond predictive accuracy, the economic implications of adopting machine learning for credit default prediction are profound. Alonso et al. (2020) investigated the economic benefits of using machine learning models like XGBoost compared to traditional models such as logistic regression. Their simulations indicated substantial savings in regulatory capital, suggesting that implementing advanced ML models could result in significant financial advantages for banks.

Chen (2022) focused on the practical application of linear regression and neural networks to predict loan default behavior. The study highlighted that improved prediction models could effectively reduce default risks, thereby enhancing the overall risk management framework for financial institutions.

Despite the promising results, the literature also highlights several challenges associated with implementing machine learning models for credit default prediction. Neema et al. (2017) emphasized the issue of imbalanced datasets, where non-defaulting customers far outnumber defaulters, complicating the classification task. Their study explored cost-sensitive learning approaches to improve prediction accuracy while maintaining a balance between false positives and false negatives.

The table below summarizes the related works on machine learning approaches for credit default prediction

| **Author(s) with Year** | **Title** | **ML Techniques Used** | **Key Findings** | **Limitations** |
| --- | --- | --- | --- | --- |
| Rahman et al. (2018) | A Machine Learning Approach to Credit Default Prediction and Individual Credit Scoring | Logistic Regression, Naïve Bayes, Decision Trees, Random Forest | Developed a model to output a credit score; Decision Tree was most effective in predicting loan defaults. | Limited generalizability due to dataset modification. |
| Uddin & Rahman (2024) | A Comparative Study of Machine Learning Algorithms for Enhanced Credit Default Prediction | SVM, K-NN, Logistic Regression, Decision Tree, Random Forest | Random Forest and Decision Tree achieved highest accuracy (92% and 94%). | Dataset may not represent all types of credit markets. |
| Arora et al. (2022) | Prediction of credit card defaults through data analysis and machine learning techniques | Logistic Regression, Decision Trees, Random Forest | Found that ML algorithms can accurately predict defaults, with logistic regression providing valuable exploratory insights. | Relied on historical data, which may not reflect current trends. |
| Wu (2022) | Machine Learning Approaches to Predict Loan Default | Random Forest, XGBoost | Random Forest and XGBoost showed similar high accuracy (~90%) in predicting defaults after effective feature selection. | Limited comparison with other advanced ML techniques. |
| Bazzana et al. (2024) | Machine learning techniques for default prediction: an application to small Italian companies | Various ML classifiers, Logistic Regression | ML models slightly outperformed logistic regression, but the performance gain was modest; classical models remain valuable for their interpretability. | Small sample size may limit broader applicability. |
| Liu et al. (2022) | Applying machine learning algorithms to predict default probability in the online credit market | K-NN, SVM, Random Forest | Machine learning models outperformed logistic regression, providing significant benefits to investors based on predictive accuracy. | Focused on a specific market segment, limiting generalization. |
| Neema et al. (2017) | The comparison of machine learning methods to achieve most cost-effective prediction for credit card default | Various ML methods, cost-sensitive approaches | Explored cost-effective predictions by balancing accuracy with costs associated with misclassifications; found Random Forest performed best in cost vs. accuracy. | Imbalance in dataset could skew results. |
| Alonso et al. (2020) | Understanding the performance of machine learning models to predict credit default | Logistic Regression, Lasso, CART, Random Forest, XGBoost, Deep NN | ML models outperformed traditional models in classification and calibration, suggesting significant economic benefits in regulatory capital requirements. | Complexity of models may hinder implementation in smaller institutions. |
| Chen (2022) | Prediction and Analysis of Financial Default Loan Behavior Based on Machine Learning Model | Linear Regression, Neural Networks | Improved logistic regression and neural networks showed advantages in predicting default risk effectively. | May require extensive data preprocessing for optimal performance. |
| Athreyas et al. (2022) | A Comparative Study of Machine Learning Algorithms for Predicting Loan Default and Eligibility | Various ML algorithms | Aimed to determine effective algorithms for predicting loan repayment capacity, highlighting the need for context-specific model selection. | Limited by the availability of relevant training data. |

The application of machine learning in credit default prediction has significantly advanced the field of financial risk assessment. Studies consistently demonstrate that machine learning techniques, particularly ensemble methods, outperform traditional models in accuracy and predictive power. However, challenges such as feature selection, data imbalance, and interpretability remain pertinent. Continued research is essential to navigate these challenges and to harness the full potential of machine learning in credit risk management.

The evolving nature of credit markets necessitates this research to refine machine learning techniques for credit default prediction. There is limited exploration of how different features interact and impact model performance, particularly among credit card consumers. This research will address these gaps by conducting a thorough comparative analysis of these three models (Logistic Regression, Random Forest, and XGBoost), evaluating their predictive accuracy, interpretability, and robustness in handling various feature interactions. By focusing specifically on credit card consumers, this study aims to provide practical insights for decision-makers in banking, enhancing the understanding of model performance in real-world applications.

**References**

Ahlin, C., Lin, J., & Maio, M. (2011). Where does microfinance flourish? Microfinance institution performance in macroeconomic context. Journal of Development Economics, 95(2), 105–120. https://doi.org/10.1016/j.jdeveco.2010.04.004

AlEmad, M. & RIT Libraries. (2022). Credit Card Fraud Detection Using Machine Learning (By Rochester Institute of Technology & RIT Digital Institutional Repository) [Rochester Institute of Technology]. <https://repository.rit.edu/theses>

Albastaki, A. A. A. & RIT Libraries. (2022). Loan Default Prediction System [Rochester Institute of Technology]. In Rochester Institute of Technology, Rochester Institute of Technology, RIT Digital Institutional Repository, RIT Digital Institutional Repository, & RIT, Theses. https://repository.rit.edu/theses

Alonso, A., Carbó, J. M., Banco de España, José Manuel Marqués, Ana Fernández, & Sergio Gorjón. (2020). Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation [Journal-article].

Arora, S., Bindra, S., Singh, S., & Nassa, V. K. (2022). Prediction of credit card defaults through data analysis and machine learning techniques. Materials Today: Proceedings, 51(1), 110-117. <https://doi.org/10.1016/j.matpr.2021.04.588>

Athreyas, S. S., K, T. B., S, V. K., Bairy, S. S., & K, H. (2022). A Comparative Study of Machine Learning Algorithms for Predicting Loan Default and Eligibility. Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES) – an International Journal, 116–117. https://www.pices-journal.com

Bazzana, F., Bee, M. & Hussin Adam Khatir, A.A. Machine learning techniques for default prediction: an application to small Italian companies. Risk Manag 26, 1 (2024). https://doi.org/10.1057/s41283-023-00132-2

Chen, H. (2022). Prediction and Analysis of Financial Default Loan Behavior Based on Machine Learning Model. Computational Intelligence and Neuroscience, 2022, 1–10. https://doi.org/10.1155/2022/7907210

Domeher, D., & Abdulai, R. (2011). Access to Credit in the Developing World: does land registration matter? Third World Quarterly, 33(1), 161–175. https://doi.org/10.1080/01436597.2012.627254

Leyden, M. H. D. S. a. J. (2024, September 12). The 18 biggest data breaches of the 21st century. CSO Online. https://www.csoonline.com/article/534628/the-biggest-data-breaches-of-the-21st-century.html

Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., & Li, A. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. International Review of Financial Analysis, 79, 101971. https://doi.org/10.1016/j.irfa.2021.101971

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science, 21(2), 117–134. https://doi.org/10.1016/j.sorms.2016.10.001

Machine learning in UK financial services. (2024, October 24). Bank of England. https://www.bankofengland.co.uk/report/2019/machine-learning-in-uk-financial-services

Neema, S., Soibam, B., & Department of Computer Science and Engineering Technology, University of Houston – Downtown. (2017). The comparison of machine learning methods to achieve most cost-effective prediction for credit card default. Journal of Management Science and Business Intelligence, 36–41. https://doi.org/10.5281/zenodo.851527

Rahman, Md Jaber & Ahmed, Hasib & Alam, A. N. M. Sajedul. (2018). A Machine Learning Approach to Credit Default Prediction and Individual Credit Scoring. 10.13140/RG.2.2.27486.92486.

Sharma, K. (2023, August 23). In its analysis of over 1,400 use cases from “Eye on Innovation” in Financial Services Awards, Gartner found that machine learning (ML) is the top technology used to empower innovations at financial services firms, with operational efficiency and cost optimisation as key intended business outcomes.  ML is a branch of artificial intelligen [. . .]. Ubuntu. https://ubuntu.com/blog/machine-learning-in-finance-history-technologies-and-outlook

Team, C. (2024, July 8). Machine Learning (in Finance). Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/machine-learning-in-finance/

Uddin, Mohammad Salah & Rahman, Md. (2024). A Comparative Study of Machine Learning Algorithms for Enhanced Credit Default Prediction. 10.1007/978-981-99-8438-1\_15.

Wu, W. (2022) Machine Learning Approaches to Predict Loan Default. Intelligent Information Management, 14, 157-164. doi: 10.4236/iim.2022.145011.