**CHAPTER ONE**

**1.0 Introduction**

The banking industry is one of the most dynamic yet inherently risky sectors of the global economy. A central feature of banking operations is lending, including services like personal loans, mortgages, and credit card loans, which have been instrumental in fueling economic growth over the past few decades. Loans, however, come with inherent risks, most notably the risk of default, where borrowers fail to meet their repayment obligations. The ability to accurately predict and manage this risk is crucial for the sustainability of financial institutions and the broader economy (Steenackers & Goovaerts, 1989; Dornadula & Geetha, 2019).

In recent years, the exponential increase in credit card usage has posed unique challenges for financial institutions. As credit becomes more accessible, the rate of credit card defaults has also risen, placing strain on banking resources and underscoring the need for efficient risk management strategies. Banks have increasingly turned to advanced data analytics and machine learning (ML) techniques to predict credit card defaults and other loan repayment issues. ML models can identify patterns in borrower behavior and predict the likelihood of default, thus allowing financial institutions to take preventive actions to mitigate risks. This project seeks to provide a comparative analysis of various ML approaches—namely, Logistic Regression, Random Forest, and XGBoost—in predicting credit card defaults, evaluating the performance and robustness of each model in real-world applications (Butaru et al., 2016; Madaan et al., 2021).

**1.1 Background**

Credit risk refers to the potential that a borrower may fail to meet their financial obligations as stipulated in a loan agreement, leading to a "default." Effective credit risk management is essential, as high default rates can threaten the solvency of financial institutions, increase borrowing costs, and destabilize the broader economy. Given the importance of accurately predicting credit defaults, financial institutions have traditionally relied on statistical models such as discriminant analysis and logistic regression. However, recent advancements in ML have led to more accurate and adaptable solutions for credit risk assessment (Lai, 2020; Malekipirbazari & Aksakalli, 2015).

In traditional risk assessment, the applicant’s financial status, loan amount, age, and profession are some of the factors considered to determine loan eligibility. With the advent of ML, predictive models have become highly sophisticated, allowing financial institutions to harness large volumes of data to make more informed lending decisions. The evolution of credit scoring systems from purely statistical approaches to complex ML-based models reflects the industry's need for enhanced accuracy and efficiency in credit risk evaluation (van Liebergen, 2017; Bank of England & Financial Conduct Authority, 2019).

Machine learning has brought transformative changes to credit risk management by enabling models that learn from historical data to predict the likelihood of future defaults. ML models allow for greater accuracy by dynamically incorporating a vast array of variables and interactions that were previously challenging to analyze using traditional methods. For instance, as digital banking grows, financial institutions can now collect extensive transaction data, enabling the application of ML models to recognize patterns in spending behavior, employment history, and debt levels to predict credit risk. These models can adapt to evolving borrower behaviors and economic conditions, making them an invaluable tool for credit risk assessment (Li, 2019; Chou et al., 2018).

In addition to enhancing predictive accuracy, ML models support scalability, enabling financial institutions to process vast amounts of data and evaluate thousands of loan applications in real-time. This capability is especially valuable given the rapid growth of online shopping and mobile payments, which have generated massive amounts of transactional data. ML models such as Random Forests and XGBoost have demonstrated superior performance in handling large datasets and dealing with complex patterns in the data. Random Forest, for example, has outperformed traditional models in some cases due to its ensemble-based approach, which reduces variance and improves accuracy (Malekipirbazari & Aksakalli, 2015). XGBoost, a powerful gradient-boosting algorithm, has also gained traction due to its efficiency and predictive strength, particularly in handling high-dimensional data (Ma et al., 2018).

When choosing an ML model for credit default prediction, it is essential to balance accuracy with interpretability. Logistic regression, a statistical model, is widely used for binary classification problems and offers a relatively interpretable approach to credit risk prediction. However, it may lack the flexibility needed to capture complex patterns in data. On the other hand, Random Forests and XGBoost provide high accuracy but are generally less interpretable, as they involve complex ensembles of decision trees (Ma et al., 2018; Leo et al., 2019).

Random Forest, an ensemble learning method, is particularly effective for high-dimensional data as it aggregates multiple decision trees to improve prediction accuracy and reduce overfitting. Its robustness and ability to handle large datasets make it a preferred choice for credit default prediction, as seen in applications across various financial institutions (Chou, 2015; Madaan et al., 2021). XGBoost, meanwhile, has been designed for efficiency and scalability. This algorithm iteratively improves predictions by minimizing the error of previous models, making it especially suitable for large-scale and complex datasets where accuracy is paramount.

This study’s findings will have significant implications for financial institutions seeking to enhance credit risk assessment using ML techniques. By providing a clear understanding of the strengths and weaknesses of each model, this research aims to help banks and lending institutions make informed choices about their credit scoring systems.

Although much research has been conducted on credit risk prediction using machine learning, there is still a need for comparative studies evaluating model performance in specific contexts, such as credit card defaults. Additionally, while Random Forest and XGBoost offer high predictive accuracy, their complex structures make them less interpretable than simpler models, posing challenges for regulatory compliance and decision-maker trust. Future research could explore methods for improving the interpretability of these advanced models or integrating explainable AI techniques into the credit scoring process (Ribeiro et al., 2016; Vidovic & Yue, 2020).

**1.2 Problem Statement**

The rise in consumer credit has elevated the importance of accurately predicting credit default to maintain financial stability. With the expansion of credit card use, financial institutions face challenges in identifying potential defaulters, resulting in increased credit losses and economic instability (Lai, 2020). Traditional credit scoring systems, while effective to a degree, struggle to accommodate the complexity of modern credit data, prompting institutions to adopt machine learning (ML) approaches for predictive accuracy and efficiency (Madaan et al., 2021). However, the diversity of ML models, such as Logistic Regression, Random Forest, and XGBoost, presents a challenge in determining the most effective method for accurately predicting credit default. Each model offers unique advantages in terms of interpretability, accuracy, and feature handling, but their comparative effectiveness for this purpose remains underexplored (Malekipirbazari & Aksakalli, 2015). Additionally, despite advancements, a significant need remains to identify key factors influencing credit default, which would enable financial institutions to tailor risk management strategies effectively (Li, 2019). This study aims to evaluate and compare the predictive capabilities of various ML models for credit default prediction, seeking to identify the model that offers the highest predictive accuracy while highlighting the most impactful features influencing creditworthiness.

**1.3 Motivation**

The motivation behind this study arises from the growing prevalence of credit defaults, a critical risk factor for financial institutions as credit usage surges globally (Lai, 2020). Traditional methods of assessing credit risk, which often involve manual or statistical approaches, struggle to accurately capture complex, modern data patterns, leading to potential inaccuracies in default predictions (Madaan et al., 2021). Machine learning (ML) methods, such as Logistic Regression, Random Forest, and XGBoost, offer a promising alternative, providing enhanced accuracy by uncovering subtle insights in vast data. However, determining the most effective ML model for credit default prediction remains challenging (Malekipirbazari & Aksakalli, 2015). This study aims to compare these models, offering valuable guidance to improve predictive capabilities and support more robust risk management in lending institutions.

**1.4 Aim and Objectives**

This research aims to provide a comparative analysis of Logistic Regression, Random Forest, and XGBoost for credit default prediction, focusing on credit card consumers. The study will evaluate each model’s accuracy, interpretability, and robustness, examining how well they handle various feature interactions and respond to changes in data patterns. Key objectives include:

1. **Comparing Predictive Accuracy**: To measure and compare the accuracy of Logistic Regression, Random Forest, and XGBoost models in predicting credit defaults.
2. **Feature Impact Analysis**: To analyze how key features—such as financial history, income level, credit score, and spending behavior—impact model performance and prediction accuracy.
3. **Model Interpretability and Practicality**: To assess the interpretability and ease of use of each model in real-world banking scenarios, especially for decision-makers who may need to understand the reasons behind specific predictions.
4. **Recommendation of Best-Fit Model**: To identify and recommend the most effective model for credit default prediction based on the analysis.

**1.5 Research Questions**

1. How do machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost compare in terms of predictive accuracy for credit default prediction?
2. What is the effect of key feature selections on the performance of machine learning models?

**1.6 Outlines for the dissertation**

This dissertation will analyze and compare machine learning approaches for credit default prediction. Chapter 2 will review existing literature on credit risk assessment, highlighting the evolution from traditional statistical methods to contemporary machine learning techniques. Chapter 3 will outline the methodology, detailing data collection, preprocessing, and the chosen algorithms: Logistic Regression, Random Forest, and XGBoost. Chapter 4 will present the results, showcasing model performance metrics. Finally, Chapter 5 will discuss the findings, linking them to prior research, and will conclude with recommendations for financial institutions on utilizing machine learning for credit risk management, along with suggestions for future research directions.