**UNIVERSITY OF DAR ES SALAAM**

**RESEARCH PROPOSAL FOR PARTIAL FULFILLMENT OF MASTER OF SCIENCE**

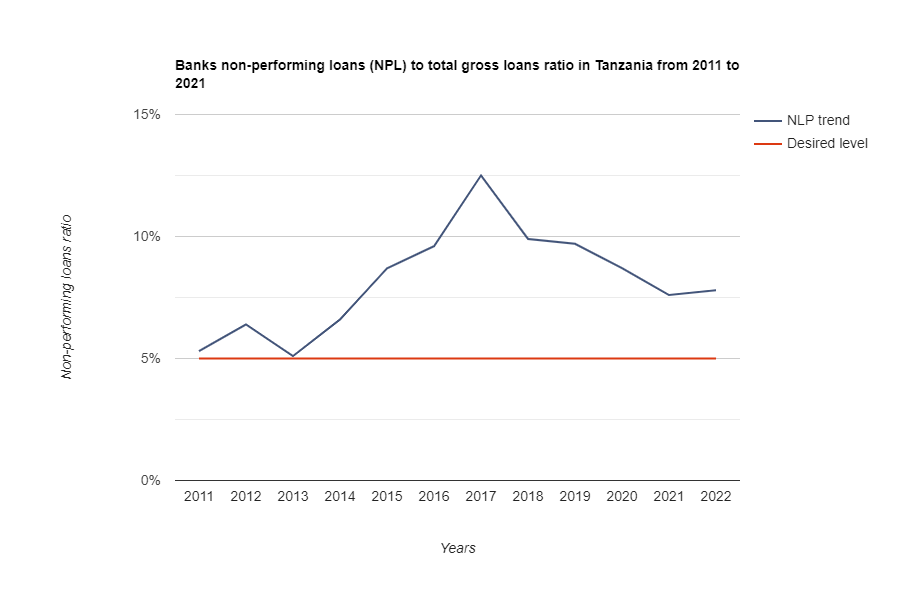
**DEGREE IN DATA SCIENCE BY COURSEWORK AND DISSERTATION.**

|  |  |
| --- | --- |
| **Name of Candidate:** | Mntambo, Stephen W  2022-06-01734 |
| **Name of Supervisor:** | Dr. Wilfred Senyoni |
| **Department /College:** | College Of Information and Communication Technology |
| **Proposed Degree:** | Master Of Science in Data Science |
| **Proposed Title:** | Enhancing Credit Risk Management in Banking Using Machine Learning: Predictive Modeling for Improved Credit Risk Management. |

# CHAPTER ONE INTRODUCTION

## General Introduction

Credit risk management plays a vital role in the stability and profitability of the banking sector. Effective credit risk management ensures that financial institutions are able to assess and mitigate the potential risks associated with lending activities. In recent years, the banking industry in Tanzania has experienced significant growth and transformation, driven by advancements in technology and evolving customer demands. As the sector continues to expand, it becomes increasingly important for banks in Tanzania to adopt innovative approaches to credit risk management to maintain a sustainable and secure lending environment. However, the overall bank's non-performing loans to gross loans ratio have been above the desired ratio that is 5 percent as by the Bank of Tanzania.

Figure 1. The overall NPL ratio trends for Tanzania Banks

Hence there comes a need to leverage machine learning techniques to reveal patterns and insights from past data consisting of defaulters and non-defaulters to be able to predict a new lending customer either to be creditable or not, thus maintaining a low NPL rate.

Machine learning, a subset of artificial intelligence, has emerged as a powerful tool for data analysis and decision-making across various industries. Its ability to process vast amounts of data and identify patterns and trends makes it particularly promising for enhancing credit risk management in the banking sector. By leveraging predictive modeling techniques, machine learning can assist financial institutions in making more accurate credit risk assessments, thereby improving their lending practices and minimizing potential losses.

This research aims to explore the application of machine learning and predictive modeling in the context of credit risk management within the Tanzanian banking sector. The study will investigate how machine learning algorithms can be utilized to analyze historical credit data, identify key risk factors, and develop robust models for predicting credit default probabilities. Furthermore, the research will examine the potential benefits and challenges associated with implementing machine learning-based credit risk management systems in Tanzanian banks (*Monetary Policy Statement - 2023.Pdf*, n.d.).

The findings of this study are expected to contribute to the existing body of knowledge on credit risk management and provide insights into the feasibility and effectiveness of machine learning approaches in the Tanzanian banking industry. By enhancing credit risk management practices, Tanzanian banks can mitigate potential credit losses, improve their overall financial stability, and foster sustainable economic growth.

## Statement of the problem

Credit risk management is a critical concern for banking institutions, as the accurate assessment and prediction of credit risk directly impact their financial stability and profitability. Traditional credit risk management techniques often rely on manual processes and limited statistical models, which may not fully capture the complexity and dynamics of credit portfolios.

This research aims to address the limitations of credit scoring modeling by leveraging machine learning techniques to enhance credit risk management in the banking sector. Specifically, the study focuses on the utilization of alternative data sources, integration of economic effects, and improvement of model explainability to develop more accurate, robust, and transparent credit risk assessment models. By bridging the research gap, this research contributes to the advancement of credit risk management practices, enabling financial institutions to make informed lending decisions and mitigate potential risks effectively

This research aims to address the following key aspects:

1. Development of predictive models: The study aims to develop robust machine learning-based predictive models that can effectively forecast credit risk. These models will leverage a wide range of customer and financial data to capture the underlying patterns and factors contributing to credit risk.
2. Identification of significant risk determinants: By analyzing vast amounts of historical data, the research seeks to identify the key determinants of credit risk, including both traditional factors (e.g., credit history, income, and collateral) and non-traditional factors (e.g., social media data, transactional patterns, and macroeconomic indicators). This will enable a more comprehensive and accurate assessment of creditworthiness.
3. Integration of machine learning into existing credit risk frameworks: The study aims to propose a framework for integrating machine learning techniques seamlessly into the existing credit risk management processes of banking institutions. This integration should be designed to enhance the effectiveness and efficiency of credit risk assessments while ensuring compliance with regulatory requirements.
4. **Objectives**
5. **Main Objective**

The main objective of this research is to enhance credit risk management in the banking sector in Tanzania by leveraging machine learning techniques and predictive modeling by developing robust models that can accurately assess credit risk, predict credit default probabilities, and improve the overall credit risk management practices within Tanzanian banks.

* + 1. **Specific Objectives**

1. To analyze the current credit risk management practices in Tanzanian banks, including the identification and assessment of risk factors, credit scoring methods, and risk mitigation strategies.

2. To explore the potential applications of machine learning algorithms in credit risk management and identify specific areas where predictive modeling can enhance existing practices.

3. To collect and analyze historical credit data from Tanzanian banks, including loan performance, borrower information, and relevant financial indicators, to build a comprehensive dataset for model development.

4. To evaluate different machine learning algorithms and techniques for credit risk assessment and identify the most effective models in predicting credit default probabilities.

5. To develop and validate predictive models for credit risk management using machine learning techniques, taking into consideration the unique characteristics and challenges of the Tanzanian banking sector.

6. To assess the feasibility and implementation challenges of integrating machine learning-based credit risk management systems within Tanzanian banks, including considerations of data availability, infrastructure requirements, and regulatory compliance.

7. To measure the impact of enhanced credit risk management through machine learning on the financial performance and stability of Tanzanian banks, including a reduction in

non-performing loans, improved loan portfolio quality, and overall risk reduction.

* 1. **Significance of the Study**

The study's contributions to the research problem of enhancing credit risk management in the Tanzanian banking sector using machine learning can be summarized as follows:

1. Advancement in Credit Risk Management Practices: The research will contribute to the advancement of credit risk management practices in Tanzanian banks by introducing and evaluating the application of machine learning and predictive modeling techniques. The study will provide insights into how these advanced technologies can be effectively utilized to assess and manage credit risks, thereby enhancing the overall risk management framework within the Tanzanian banking industry.

2. Development of Tailored Credit Risk Models: The study aims to develop robust and tailored credit risk models specific to the Tanzanian banking context. By analyzing historical credit data and employing machine learning algorithms, the research will contribute to the development of accurate and reliable predictive models for credit risk assessment. These models can assist banks in making informed lending decisions, identifying potential defaults, and implementing proactive risk mitigation strategies.

3. Improved Decision-Making and Risk Mitigation: Through the implementation of machine learning-based credit risk management systems, the research will enable Tanzanian banks to make more accurate and data-driven decisions regarding credit approvals, loan pricing, and risk mitigation. By harnessing the power of machine learning algorithms, banks can optimize their credit risk management practices, minimize potential losses, and improve the overall profitability and financial stability of the banking sector.

4. Enhanced Financial Stability: Effective credit risk management is essential for maintaining financial stability within the banking sector. The study's findings and recommendations will contribute to improving the overall financial stability of Tanzanian banks by reducing non-performing loans, enhancing the quality of loan portfolios, and mitigating credit risks. This, in turn, can lead to a more resilient banking sector and contribute to the overall economic stability of Tanzania.

5. Industry Guidance and Best Practices: The research will provide valuable industry guidance and best practices for the adoption and implementation of machine learning-based credit risk management systems in Tanzanian banks. The study will identify the feasibility, challenges, and considerations involved in implementing these systems, including data availability, infrastructure requirements, and regulatory compliance. By offering recommendations and guidelines, the research will assist banks in effectively leveraging machine learning technologies to improve their credit risk management practices.

**1.5 Scope of the study**

* Exploring the application of machine learning techniques in credit risk management within Tanzanian banks.
* Investigating how machine learning algorithms can analyze historical credit data and develop models for predicting credit default probabilities.
* Assessing the integration of economic factors and alternative data sources (e.g., social media data) to improve credit risk assessment accuracy.
* Enhancing model interpretability and ethical considerations to ensure transparency and fairness.
* Measuring the potential impact of enhanced credit risk management on the financial stability of Tanzanian banks.
* Providing recommendations for the adoption and implementation of machine learning-based credit risk management practices

**CHAPTER TWO**

**LITERATURE REVIEW**

Credit risk management plays a pivotal role in the banking sector, as financial institutions face the challenge of assessing the creditworthiness of borrowers to make informed lending decisions. Traditionally, credit risk assessment has relied on manual processes and subjective judgment, which are often time-consuming and prone to human biases. Techniques such as weight‐of‐evidence measure, regression analysis, discriminant analysis, probit analysis, logistic regression, linear programming, Cox’s proportional hazard model, support vector machines, decision trees, neural networks, K‐nearest neighbor (K‐NN), genetic algorithms, and genetic programming are some of statistical credit scoring methods used by researchers, credit analysts, and lenders until recent (Abdou & Pointon, 2011). However, recent advancements in machine learning (ML) techniques have opened up new avenues for improving credit risk management by enabling the development of predictive modeling approaches. This literature review explores the existing research and studies that focus on enhancing credit risk management in banking using machine learning, with a specific emphasis on predictive modeling techniques.

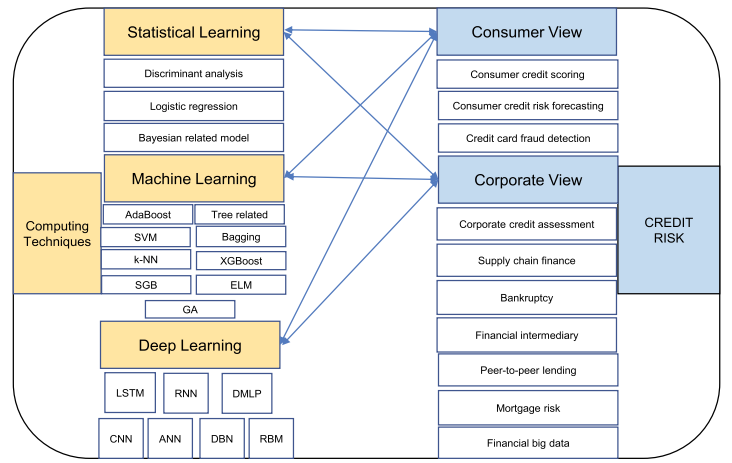


Figure 1. Taxonomy of approaches and algorithms (Shi et al., 2022)

1. Credit scoring

Credit scoring is a crucial aspect of credit risk management in the banking industry. It involves classifying clients into different risk categories, typically categorized as "good" and "bad" borrowers. Traditionally, credit scoring has been approached as a classification problem, aiming to accurately predict the creditworthiness of borrowers based on various factors. One technique that has been applied to credit scoring is the use of genetic algorithms. Genetic algorithms utilize a fitness function to evaluate the predictive performance of different solutions or individuals in

the population. In the context of credit scoring, each individual represents a genetically encoded solution to the classification problem, and its fitness score reflects its ability to provide an accurate credit risk assessment. The genetic algorithm operates by evolving an initial population of solutions into a new population. This evolutionary process involves genetic-inspired operations, such as mutation and crossover, to generate new individuals with potentially improved fitness scores. The goal is to find the most optimal combination of features or variables that can effectively classify borrowers into the appropriate risk categories (Vanve & Patil, n.d.). Furthermore, credit scoring can also be approached from a profit-based perspective. In this approach, the classification of customers is associated with potential profit or loss. Correctly classifying borrowers as good or bad can lead to profit, while incorrect classification may result in potential profit loss. This profit-based approach considers the financial implications of credit risk assessment, emphasizing the importance of accurately categorizing borrowers to maximize profitability.

1. The Evolution of Credit Risk Management:

Credit risk management has undergone significant transformations over the years, from traditional rule-based approaches to more sophisticated models that leverage statistical and ML techniques. This section provides an overview of the evolution of credit risk management practices, highlighting the challenges faced and the need for improved predictive modeling techniques.

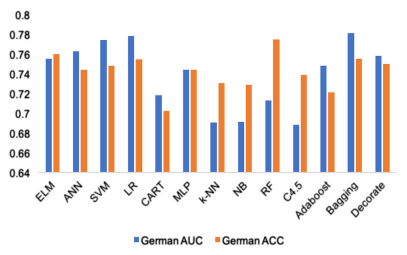
1. Machine Learning in Credit Risk Management: This section delves into the applications of machine learning in credit risk management. It examines the various ML algorithms and methodologies that have been employed to predict credit risk, including logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks (Shi et al., 2022).

Figure 2. The accuracy for Germany credit data

The strengths and weaknesses of each approach are discussed, along with their effectiveness in improving credit risk assessment.

3. Utilization of Alternative Data Sources:

To enhance credit risk management, researchers and practitioners have explored the use of alternative data sources beyond traditional financial information. This section reviews studies that have incorporated non-traditional data, such as social media activity, transactional data, and mobile application data, into credit risk modeling. The potential benefits and challenges associated with leveraging alternative data sources are examined, along with the impact of such data on credit risk prediction accuracy (Simão, n.d.).

4. Integration of Economic Effects:

Credit risk is not solely influenced by individual borrower characteristics but is also affected by broader economic factors. This section explores research that focuses on integrating economic effects, such as unemployment rates, inflation, and housing market indicators, into credit risk modeling. The findings highlight the importance of considering macroeconomic factors in credit risk assessment and demonstrate how their inclusion can improve the accuracy and robustness of predictive models (Simão, n.d.).

5. Model Explainability and Interpretability:

The interpretability of credit risk models is crucial for regulatory compliance and transparency. This section examines research that addresses the challenge of model explainability in machine learning-based credit risk management. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and fuzzy logic are explored for their potential to provide transparent and understandable explanations for credit risk predictions.

**RELATED WORKS**

The paper (Simão, n.d.) has been a closely related work that builds up a major contribution for the research I am conducting in the following aspect

Historical Development: The concept of credit scoring dates back to the 1940s, and Linear Discriminant Analysis was one of the early techniques used for assessing credit risk. Since then, various developments have been proposed in the literature, including Logistic Regression, Survival Analysis, and machine learning methods.

Machine Learning in Credit Scoring: Many studies have explored the application of machine learning techniques in predicting loan default and credit risk. These techniques include Support Vector Machines (SVM), Genetic Algorithms, Neural Networks, Random Forests, Decision Trees, AdaBoost, and ensemble methods. The performance of these models has been evaluated using metrics such as accuracy, AUC, Brier Score, and Type I and Type II errors.

Comparative Studies: Several comparative studies have been conducted to assess the performance of different models. These studies have shown that ensemble methods, such as stacking, bagging, and boosting, often outperform individual classifiers. Tree-based models, such as Random Forest and XGBoost, have demonstrated good performance and stability. It has been found that machine learning models, such as boosted regression trees, can provide better predictive performance for mortgage credit risk compared to logistic regression.

Important Variables: Common variables found to be important across multiple studies include income, loan amount, loan purpose, employment status, and home ownership. These variables have been consistently used to predict loan default and assess creditworthiness.

Performance Evaluation: Various performance evaluation criteria have been used to compare different models, such as accuracy, AUC, Brier Score, H-measure, and Kolmogorov-Smirnov statistic. Heterogeneous ensemble classifiers have been found to perform well, outperforming logistic regression in many cases.

**RESEARCH GAP**

Despite the advancements made in credit scoring modeling, there exist several limitations and areas for future research. One significant research gap is the exploration of alternative data sources and the integration of economic effects to improve the accuracy and interpretability of credit scoring models (Teng & Lee, 2019).

Currently, the availability of sensitive and confidential data from financial institutions poses a challenge, leading to limited access for researchers (Simão, n.d.). Therefore, this research focus on identifying and utilizing new data sources, such as the vast amount of digital information recorded on social networks and mobile applications. This data, when analyzed from a behavioral perspective, may provide valuable insights for consumer credit risk research.

Additionally, while personal characteristics play a crucial role in determining borrowers' creditworthiness, the external economic environment also impacts loan performance. Factors such as unemployment rates and house prices can influence borrowers' ability to repay loans. Hence, this research investigates the integration of detailed economic information into credit scoring models, aiming to improve their accuracy and interpretability by considering the broader economic context.

Another aspect that requires attention is the model explainability. The decisions made based on credit scoring models should be transparent and operate within equal opportunity laws. Although variable-importance scores provide some insight into the predictors' significance, a deeper exploration of model explainability is necessary. Techniques like LIME (Local Interpretable Model-Agnostic Explanations) and fuzzy logic have shown potential in enhancing the explanatory capability of predictive models. Exploring these methods, along with other interpretability frameworks, can contribute to developing credit scoring models that are more transparent and explainable.

## CHAPTER THREE

## RESEARCH METHODOLOGY

## In this chapter, we will discuss the methodology used in our research.

**3.1 Research Approach and Design**

The research approach for this study will be primarily quantitative, utilizing a combination of retrospective analysis of historical credit data and predictive modeling. A longitudinal design will be employed to track credit risk over a specific time period, with a focus on assessing the impact of machine learning on risk assessment. Additionally, a comparative design will be used to evaluate the effectiveness of machine learning-based models in comparison to traditional credit risk assessment methods.

**3.2 Research Methods**

The primary research methods will include data collection, data preprocessing, algorithm selection and implementation, model training, model validation, and performance evaluation. Machine learning algorithms such as logistic regression, decision trees, random forests, and neural networks will be employed to develop predictive credit risk models. The study will also involve the use of statistical analyses to compare the performance of different models and assess the significance of economic indicators in the models.

**3.3 Study Area**

The study will focus on the Tanzanian banking sector. Specifically, it will target a diverse set of banks within Tanzania to ensure a representative sample and comprehensive understanding of credit risk management practices in the country.

**3.4 Sample Size**

The sample size will be determined based on the availability of historical credit data from Tanzanian banks. The aim is to collect a sufficiently large and diverse dataset to ensure the robustness of the developed machine learning models.

**3.5 Materials**

The primary materials for this study will include historical credit data from Tanzanian banks, economic indicators (such as unemployment rates, inflation, GDP growth), alternative data sources (potentially social media data), machine learning software/tools for model development (Python, R, or relevant libraries), and computing resources for model training and analysis.

**3.6 Data Analysis**

Data analysis will involve various steps, including exploratory data analysis (EDA) to understand the characteristics of the credit data, feature engineering to extract relevant features for the models, model training and validation, and statistical analysis to evaluate the performance of the machine learning models. Advanced techniques such as cross-validation and AUC-ROC analysis will be used for model evaluation.

**3.7 Ethical Practices/Procedures**

Ethical considerations will be given utmost importance in this research. The study will adhere to ethical guidelines for data privacy and security, ensuring that all sensitive information is anonymized and protected. Consent will be obtained from the banks for the use of their historical credit data. Additionally, the research will be conducted in accordance with relevant regulatory standards and will aim to ensure fairness, transparency, and unbiased decision-making in credit risk assessment.

## CHAPTER FOUR

## REFERENCES

Abdou, H. A., & Pointon, J. (2011). CREDIT SCORING, STATISTICAL TECHNIQUES AND EVALUATION CRITERIA: A REVIEW OF THE LITERATURE: CREDIT SCORING, TECHNIQUES & EVALUATION CRITERIA: A LITERATURE REVIEW. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3), 59–88. https://doi.org/10.1002/isaf.325

*Monetary Policy Statement—2023.pdf*. (n.d.).

Shi, S., Tse, R., Luo, W., D’Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications*, *34*(17), 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Simão, S. B. S. (n.d.). *Data Science and Advanced Analytics*.

Teng, H.-W., & Lee, M. (2019). Estimation Procedures of Using Five Alternative Machine Learning Methods for Predicting Credit Card Default. *Review of Pacific Basin Financial Markets and Policies*, *22*(03), 1950021. https://doi.org/10.1142/S0219091519500218

Vanve, R., & Patil, P. (n.d.). *A survey on: Predictive Analytics for Credit Risk Assessment*. *03*(06).

## CHAPTER FIVE

**OTHER RELEVANT INFORMATION**

## 5.1 Financial Arrangement

**5.1.1 Sponsorship**

The research project will be self-funded.

**5.1.2 Proposed Budget**

The following are the items and their costs respectively need to accomplish this research.

|  |  |
| --- | --- |
| **ITEM** | **COST(TSH)** |
| Data Collection and Preparation | 500,000/= |
| Software and tools | 300,000/= |
| Research materials | 300,000/= |
| Computing resources | 2,500,000/= |
| Travel | 200,000/= |
| Miscellaneous expenses | 500,000/= |
|  | **4,300,000/=** |

**5.2 Duration of the Study**

Below is the Gantt chart that shows the duration of study in months from August 2023 to July 2024.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **2023** | | | |  | **2024** | | | | | | |
| **Aug** | **Sept** | **Oct** | **Nov** | **Dec** | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** | **Jul** |
| Literature Review and Refinement |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Collection and Preprocessing |  |  |  |  |  |  |  |  |  |  |  |  |
| Exploratory Data Analysis and Algorithm Selection |  |  |  |  |  |  |  |  |  |  |  |  |
| Model Development and Integration of Economic Factors |  |  |  |  |  |  |  |  |  |  |  |  |
| Incorporation of Alternative Data Sources and Model Interpretability |  |  |  |  |  |  |  |  |  |  |  |  |
| Model Fine-Tuning and Optimization |  |  |  |  |  |  |  |  |  |  |  |  |
| Ethical Considerations and Model Explainability |  |  |  |  |  |  |  |  |  |  |  |  |
| Model Validation and Results Analysis |  |  |  |  |  |  |  |  |  |  |  |  |
| Impact Assessment and Recommendations |  |  |  |  |  |  |  |  |  |  |  |  |
| Research Paper Finalization |  |  |  |  |  |  |  |  |  |  |  |  |
| Submission and Presentation |  |  |  |  |  |  |  |  |  |  |  |  |

Candidate’s Name**:** ………………………………………… Candidate’s Signature**:** …………

Date…………………………...

**Supervisors’(s) & Principal’s Comments**

Comment by Supervisor

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

Name**:** ………………………………………………………… Signature: …………………..

Date**:** ………………………

SUPERVISOR

Principal’s comment

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

Name**:** ……………………………………………………… Signature**:** ……………………

Date**:** ………………………

PRINCIPAL