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 of Commercial Banks in Tanzania Using Predictive Modeling Techniques.

INTRODUCTION

1.0 General Introduction

Financial institutions play a crucial role in the functioning of any economy, akin to the vital function of blood arteries in the human body. Much like how blood vessels transport essential nutrients throughout the body, financial institutions channel financial resources for economic growth from depositories to areas where they are needed (Shanmugan & Bourke, 1990). Among these financial institutions, commercial banks stand out as key contributors, serving as significant providers of financial information to the broader economy.They assume a particularly vital function in developing economies, especially in situations where borrowers lack access to capital markets (Greuning et al., 2003). There is empirical support indicating that efficiently operating commercial banks contribute to the acceleration of economic growth, whereas poorly functioning ones hinder economic advancement and worsen poverty levels (Barth et al., n.d.).

Commercial banks (CBs) encounter various risks that can be classified into three main categories: financial risks (where credit risk, or CR, is a significant component), operational risks, and strategic risks (Saunders & Cornett, 2018). These risks exert varying impacts on the performance of CBs, with credit risk, in particular, having severe consequences that can lead to bank failures (Chijoriga, 1997). In both matured and emerging economies, there has been a noticeable rise in significant banking issues over the years. Researchers have delved into the causes of these problems, identifying numerous factors (Chijoriga, 1997; Santomero & Babbel, 1997 ; BrownBridge & Harvey, 1998 ;Kimei, 1998 ; Basel, 2004).

Credit-related challenges, particularly weaknesses in credit risk management (CRM), emerge as major contributors to banking difficulties. Given that loans often represent a substantial portion of credit risk, typically surpassing the equity of a bank by 10-15 times (Kitua, 1996), the banking sector becomes vulnerable to challenges when there is even a slight deterioration in loan quality. The origins of poor loan quality can be traced back to the information processing mechanism. (Brownbridge, 1998) noted that these issues are particularly acute in developing countries, often manifesting at the loan application stage (Liukisila, 1996) and escalating during the loan approval, monitoring, and control stage, especially when there are gaps or weaknesses in CRM guidelines related to policy and strategies/procedures for credit processing.

In response to these challenges, the utilization of predictive modeling techniques emerges as a critical tool in enhancing credit risk management for commercial banks.Predictive modeling offers a data driven approach to risk assessment, allowing for more accurate predictions and proactive management strategies,By leveraging advanced analytics and modeling methodologies, commercial banks can strengthen their ability to identify and mitigate credit risks, ultimately contributing to the overall stability and sustainability of the banking sector in Tanzania.

1.1 Statement of the Problem

Effectively managing credit risk stands as a paramount concern for commercial banks in Tanzania, given its profound implications on financial stability and profitability. The conventional approaches to credit risk management, predominantly reliant on manual processes and simplistic statistical models, exhibit shortcomings in capturing the intricate dynamics inherent in credit portfolios. The inadequacies of traditional credit scoring models necessitate a paradigm shift towards more sophisticated methodologies.This research identifies the critical need to address the limitations of existing credit scoring models through the integration of advanced predictive modeling techniques, specifically leveraging machine learning. The inadequacies of conventional models become evident in their limited capacity to adapt to the complexity of modern financial landscapes. As financial institutions navigate through intricate credit portfolios, the study aims to enhance credit risk management practices by exploring alternative data sources, incorporating economic effects, and prioritizing model explainability.

1.3 Objectives

1.3.1 Main Objective

The main Objective is to significantly enhance credit risk management practices within commercial banks operating in Tanzania through the strategic application of advanced Predictive modeling techniques.

1.3.2 Specific Objectives

1. To build a comprehensive dataset for model development by collecting historical credit data from Commercial banks in Tanzania, including loan performance, borrower information, and relevant financial indicators.
2. To develop predictive models for credit risk management using machine learning techniques.
3. To evaluate the performance of the predictive models using metrics such as accuracy.

1.4 Research Questions

To meet the specific research objectives, the study will aim to answer the following questions:

1. What are the key components and variables needed in constructing a comprehensive dataset for credit risk management of commercial banks ?

2. How can machine learning techniques be effectively applied to develop predictive models for credit risk management?

3. What criteria and metrics are appropriate for assessing the performance of predictive models in the context of credit risk management?

1.5 Significance of the Study

The significance of this study is multifaceted, addressing critical issues within the Tanzanian banking sector and contributing to the broader field of credit risk management.

Advancement of Credit Risk Management Practices: This research holds paramount importance in enhancing credit risk management practices within commercial banks in Tanzania. By leveraging advanced predictive modeling techniques, the study aims to introduce a paradigm shift from traditional approaches, which have exhibited limitations in adapting to the intricate dynamics of credit portfolios. The strategic application of machine learning promises to bring about a significant improvement in the accuracy, efficiency, and adaptability of credit risk assessment.

Risk Mitigation and Financial Stability: The utilization of predictive modeling techniques serves as a proactive measure for identifying and mitigating credit risks. By developing robust models, commercial banks can navigate the challenges posed by credit-related difficulties more effectively, ultimately contributing to the overall stability and sustainability of the Tanzanian banking sector.

Contribution to Academic Research: The study contributes to the academic field by addressing gaps in existing credit scoring models and proposing innovative solutions. By exploring the integration of machine learning techniques, alternative data sources, and economic factors, the research expands the knowledge base in credit risk management, potentially paving the way for further advancements in the field.

1.6 Scope of the Study

This study will focus on advancing credit risk management practices within the Tanzanian banking sector through the strategic application of predictive modeling techniques. The scope of the study includes the following key components: The primary geographical focus is on commercial banks operating within Tanzania. The study will delve into the unique characteristics and challenges present in the Tanzanian banking sector . The study will analyze historical credit data, emphasizing loan performance, borrower information, and relevant financial indicators. The temporal scope will be determined by the availability of data, with an emphasis on capturing patterns and trends relevant to contemporary credit risk management practices.And the predictive modeling techniques that will be used in this study are, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Long short-Term Memory.

LITERATURE REVIEW

2.1 Introduction

Any research builds on the existing knowledge in a particular field. This chapter reviews various prior literatures on Credit Risk Management of Commercial Banks in Tanzania using Predictive modeling Techniques. 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. Finally, relevant and related empirical works will be reviewed and the knowledge gap will be identified.

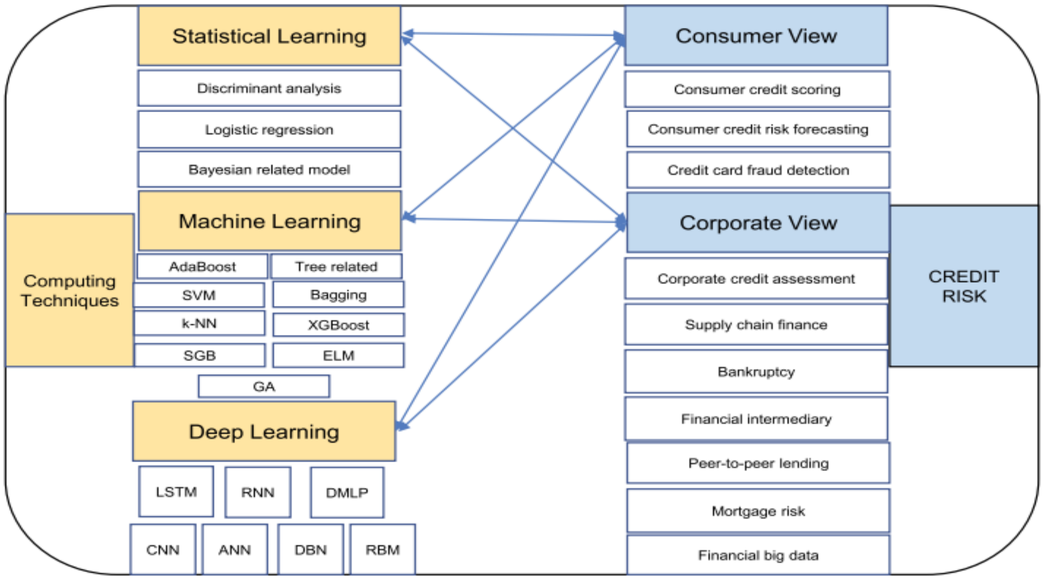

Figure 1. Taxonomy of approaches and algorithms (Shi et al., 2022)

2.2 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.

2.3 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.

2.4 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).

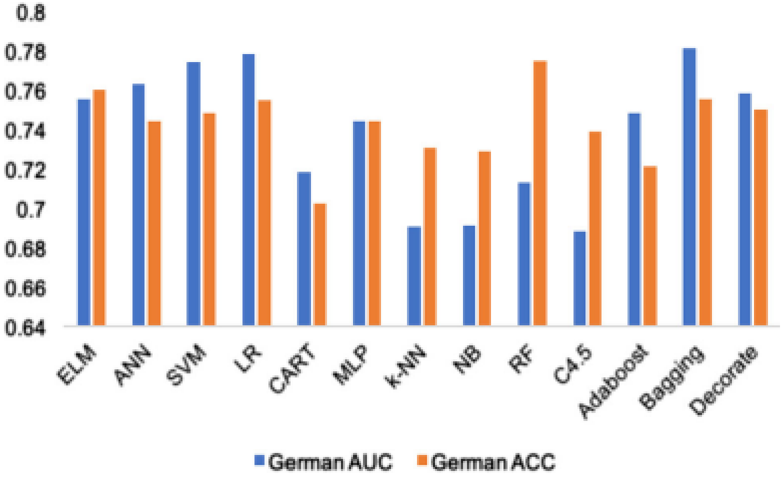

Figure 2. The accuracy for Germany credit data

2.5 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.).

2.6 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.).

2.7 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.

2.8 Related work

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 statistics. Heterogeneous ensemble classifiers have been found to perform well, outperforming logistic regression in many cases.

2.9 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 focuses 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.

RESEARCH METHODOLOGY

3.1 Research Design

This study involves experimentation, wherein specific groups of variables remain consistent, while another group of variables is observed as the primary focus of the investigation.The study involves conducting experiments using various predictive modeling techniques for credit risk of commercial banks in Tanzania. The objective is to apply different Predictive modeling Techniques for Credit Risk. The performance of each trained Predictive modeling algorithm is then evaluated to determine the best predictive model.

3.2 Research Approach

3.2.1 Variables

The study will consider a range of variables, including but not limited to: Loan performance metrics, Borrower demographics, Financial indicators (e.g., liquidity ratios, leverage ratios), Economic indicators, Alternative data sources.

These variables will be instrumental in constructing a comprehensive dataset for the development of predictive models.

3.2.2 Predictive Modeling Techniques

The Study will employ several machine learning algorithms to develop predictive models for credit risk management. The chosen algorithms include Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Long Short-Term Memory. The selection of these techniques is based on their suitability for handling different aspects of credit risk and their relevance to the Tanzanian banking context.

3.2.3 Model Development

The predictive models will be developed using a two-step process. In the first step, a comprehensive dataset will be prepared by cleaning, preprocessing, and transforming the collected data. In the second step, the dataset will be used to train and validate the selected machine learning algorithms. The models will be fine-tuned to ensure optimal performance in predicting credit risk.

3.2.4 Model Evaluation

The performance of the developed models will be assessed using relevant metrics such as accuracy, precision, recall, and the area under the ROC curve. The evaluation process will provide insights into the effectiveness of the predictive models in comparison to traditional credit scoring methods.

3.3 Study Area

This research is centered on the United Republic of Tanzania, a nation in East Africa known for its diverse economic landscape and a thriving financial sector. The Study will narrow its focus to three commercial banks operating in Tanzania, namely Stanbic Bank Tanzania, CRDB Bank, and NMB Bank. These banks were selected based on their significant market share, operational scale, and diverse credit portfolios. By concentrating on these institutions, the study aims to provide comprehensive insights into credit risk management practices across different banking models.

3.4 Tools and Materials

This study will utilize a variety of materials, including but not limited to: -

S/No	Material	Tool description
1	portable computer	for computation
2	software	JetBrains PyCharm for the analysis
3	modem	for Internet access to retrieve various literature materials
4	interview	for capturing quantitative data
5	Programming Languages	Python version 3.10.4, JavaScript ECMAScript 2023
6	Data formats	JavaScript Object Notation (JSON), Comma Separated Value (CSV), Excel (xls, xlsx)
7	Virtual Environment	Pipenv
8	Web Development Frameworks	Django 3.1.6, StreamLit
9	Version control	Git
10	Libraries	NumPy, Pandas, JupyterLab, Scikit-learn, Gensim, Keras matplotlib

Table 3.1 tools and materials to be used in this study

3.5 Data Collection

The data for this study will be gathered from the three commercial banks in Tanzania, namely Stanbic Bank Tanzania, CRDB Bank, and NMB Bank.The selection criteria are based on the market share,operational scale, and diversity of credit portfolios of the three selected banks.This ensure that the dataset encompasses a broad spectrum of banking activities, contributing to a thorough analysis of credit risk management. The data spans from 2020 to 2023.

This study will employ a combination of primary and secondary data collection methods. Historical credit data will be obtained, including information on loan performance, borrower characteristics, and relevant financial indicators.

Secondary data will be acquired from various sources such as financial reports, credit histories, industry publications, and academic journals will be consulted to gather supplementary information. And the source will be from the official records of Stanbic Bank Tanzania, CRDB Bank, and NMB Bank. These documents will serve as the primary source for evaluating credit risk management strategies and outcomes.

3.5.1 Document Review

As part of the secondary data collection, this study will review multiple documents from diverse sources including research papers, academic articles, journals, open-data platforms, and relevant websites. The purpose of this document review is to gain insights into different techniques used for analyzing and predicting credit risk for commercial banks. It aims to gather information about the approaches adopted by other researchers, their methodologies, and the findings they obtained. Additionally, the review will aid in identifying existing Predictive modeling Techniques and selecting suitable models that can be customized to align with the specific context of this research.

3.6 Data Analysis

In this research, the application of Exploratory Data Analysis (EDA) will be employed to conduct initial investigations on the collected data. The purpose is to uncover patterns, derive meaningful insights, identify anomalies, gain a comprehensive understanding of the data, and provide summary statistics and graphical representations. The initial data analysis will specifically focus on the datasets containing customer information from Commercial banks in Tanzania during the period of 2020 to 2023. To facilitate this process, a Python virtual environment will be established. Python scripts will be developed to handle the large JSON files associated with each month, enabling them to be split into smaller chunks. These chunks will be saved in the Comma Separated Value (CSV) format, ensuring ease of use and effective data manipulation.

3.7 Ethical Consideration

To address the ethical considerations associated with this study, research clearance letters will be obtained from UDSM and also the research clearance letters will be provided to three commercial banks in Tanzania, which will be involved in data collection: Stanbic bank Tanzania, CRDB, and NMB Banks . The potential respondents will be informed about the benefits of the study and will be requested to participate voluntarily throughout the data collection process. They will be fully informed and assured that this research is a necessary component of academic requirements and will be utilized to develop a model for predicting Credit risk of Commercial Banks in Tanzania.

REFERENCES

Barth, J. R., Caprio, G., & Levine, R. (n.d.). *Bank Regulation and Supervision: What Works Best?*

Brownbridge, M. (1998). Financial Distress in Local Banks in Kenya, Nigeria, Uganda and Zambia: Causes and Implications for Regulatory Policy. *Development Policy Review*, *16*(2), 173–188. https://doi.org/10.1111/1467-7679.00057

BrownBridge, M., & Harvey, C. (1998). *Banking in Africa, James Currey*. Oxford.

Chijoriga, M. M. (1997). *An application of credit scoring and financial distress prediction models to commercial bank lending: The case of Tanzania*. na.

Greuning, H. van, Greuning, H. van, & Brajovic Bratanovic, S. (2003). *Analyzing and managing banking risk: A framework for assessing corporate governance and financial risk* (2nd ed). World Bank.

Kimei, C. S. (1998). Sound banking management and macroeconomic stability in Africa. *Seminar for Chief Executives on Monetary Policy Stance in Tanzania*.

Kitua, G. D. Y. (1996). *Application of multiple discriminant analysis in developing a commercial Banks loan classification model and assessment of significance of contributing variables: A case of National Bank of Commerce.* University of Dar es Salaam.

Liukisila, C. (1996). IMF Survey, Healthy Banks are Vital for a Strong Economy, Finance and Development. *Washington, DC*.

Santomero, A. M., & Babbel, D. F. (1997). Financial Risk Management by Insurers: An Analysis of the Process. *The Journal of Risk and Insurance*, *64*(2), 231. https://doi.org/10.2307/253730

Saunders, A., & Cornett, M. M. (2018). *Financial institutions management: A risk management approach* (Ninth Edition). McGraw-Hill Education.

Shanmugan, B., & Bourke, P. (1990). *The Management of financial institutions: Selected readings*. Addison-Wesley Publishing, Reading, MA.

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

OTHER RELEVANT INFORMATION

5.1 Financial Arrangement

5.1.2 Sponsorship

The research project will be self-funded.

5.1.3 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/=
Total	**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						
	May	June	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	March	April	May	June	July
Literature Review	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Research Proposal Writing and Submission	■	■	■	■	■	■	■	■							
Data Collection							■	■	■	■					
Data Preprocessing, Cleaning, and Analysis									■	■	■	■	■		
Model development									■	■	■	■	■	■	
Results and Interpretation										■	■	■	■	■	
Report Writing												■	■	■	■
Report Submission													■	■	■

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