AN ENHANCED HYBRID MODEL FOR CREDIT SCORING USING MACHINE LEARNING APPROACH

CASE STUDY: LOGISTIC REGRESSION AND DECISION TREE

A dissertation presented to

FACULTY OF SCIENCE
in partial fulfillment of the requirements for the award of the degree
Master of Science in Information Systems

UGANDA MARTYRS UNIVERSITY

MICHEAL Mugumya

2020-M132-22067

Supervisor: Brain Kasozi

September 2024

DEDICATION

This work is dedicated to my family and friends, whose unwavering support and encouragement have been my source of strength throughout this journey. Their belief in me has been invaluable, and I am eternally grateful for their love and understanding.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my academic supervisor, whose guidance, insightful feedback, and expertise have greatly contributed to the success of this project. Your mentorship has been instrumental in shaping my understanding and approach to this research.

I also wish to thank my colleagues and peers for their valuable discussions, collaboration, and support during the course of this study. The exchange of ideas and the shared enthusiasm for machine learning has been a driving force behind this work.

A special thanks to data science and developers of the tools and libraries that made this research possible. The resources provided have been essential in executing this project.

Lastly, I would like to acknowledge my family for their patience, understanding, and encouragement. Their continuous support has been a cornerstone in the completion of this work.

TABLE OF CONTENTS

ABSTRACT

This study investigates the application of a hybrid machine learning model for classifying loan statuses, combining logistic regression and decision tree classifiers. The dataset used comprises various loan records, with the target variable being the loan status, which is dichotomous in nature. The primary objective of this research is to develop a robust predictive model that can accurately determine the likelihood of a loan default.

The analysis began with data preprocessing, including the handling of missing values and encoding of categorical variables. The dataset was then divided into training and testing sets to evaluate model performance. Two individual models' logistic regression and decision tree were initialized with class weighting to address potential class imbalances. These models were combined using a soft voting classifier to form a hybrid model, leveraging the strengths of both algorithms.

The hybrid model was trained and tested, with its performance evaluated using key metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The results indicated that the model achieved a reasonable level of accuracy, particularly in predicting non-default loans (class 0), as evidenced by a high number of true negatives. However, the model's performance in predicting default loans (class 1) was less satisfactory, with a notable number of false negatives, suggesting a need for further refinement.

Visualizations, including the confusion matrix and bar plots of evaluation metrics, provided deeper insights into the model's predictive capabilities and highlighted areas where the model could be improved. These findings underscore the complexity of loan status prediction and the challenges associated with imbalanced datasets.

Overall, this study demonstrates the potential of hybrid machine learning models in financial risk prediction, while also identifying critical areas for future research and model enhancement. The implications of this research extend to financial institutions seeking to improve their risk management practices and enhance the accuracy of their loan approval processes.

CHAPTER ONE

INTRODUCTION

1.0 Introduction

Credit scoring is a statistical analysis performed by lenders and financial institutions to determine the creditworthiness of individuals or small, owner-operated businesses. It helps lenders decide whether to extend or deny credit. According to Alam and Alam (2024) Lenders use information such is Payment History (35%)**: ,** Amounts Owed (30%)**,** Length of Credit History (15%)**,** New Credit (10%)**,** and Credit Mix (10%) for individuals however, businesses' credit scores are based on information in their credit reports, including company details, historical data, industry classification, and payment history. Lenders use credit scoring for risk-based pricing**,** determining loan terms based on the probability of repayment. Higher credit scores often lead to better interest rates. In credit scoring, both logistic regression and decision trees play important roles. Logistic regression links the credit score and the probability of default (PD). It's the default model when working with credit scorecards. It estimates the probability of an event (e.g., default) occurring based on predictor variables and it has Interpretable coefficients well established and widely used although it assumes a linear relationship between predictors and log-odds and may not handle complex interactions well. On the other hand, Decision trees are increasingly popular in credit scoring. Decision trees follows a top-down approach, splitting data based on the best variable at each step and provide Metrics like Gini index, information value, or entropy guide the splits. Decision trees have a bility to capture nonlinear relationships and handles interactions and complex patterns. However this approach is Prone to overfitting and is less interpretable than logistic regression.

In the context of Nguyen and Truong (2024 credit scoring, ensemble methods based on decision trees, such as the random forest method, provide better classification performance than standard logistic regression models. However, logistic regression remains the benchmark in the credit risk industry mainly because the lack of interpretability of ensemble methods decision trees are incompatible with the requirements of financial regulators. Logistic regression and decision tree algorithms have their own drawbacks as discussed above. The question of how to overcome the shortcomings of these two models has become a hot subject in academia and industry, garnering the attention of many researchers. This study propose a hybrid credit scoring model which combines both logistic regression and decision trees and performs best in terms of interpretability and forecasting performances as compared to logistic regression and decision trees used independently. The approach will be based on an adaptive lasso logistic regression model with predictors extracted from decision trees. The notable aspect of the new hybrid approach consists of using algorithms to pre-treat the predictors instead of modelling the default probability directly with machine learning classification algorithms. Secondly, the new hybrid model shall provide parsimonious and interpretable scoring rules (e.g., marginal effects or scorecards) as recommended by the regulators, since it preserves the intrinsic interpretability of the logistic regression model and is based on a simple feature selection method. This study shall show that the new hybrid model outperforms standard logistic regression and decision trees in terms of out-of-sample forecasting accuracy.

1.1 Background to the study

Credit scoring models play a crucial role in assessing the creditworthiness of individuals or businesses (Tripathi *et al.*, 2021) however the scholars add that Logistic regression is a widely used method for credit scoring as it links the credit score (or credit risk) to the probability of

default (PD) through the logistic regression function. In credit scoring, the goal is to predict whether a borrower will default on their credit obligations. According to Shen, Wang and Shen (2019)Logistic regression helps achieve this by modeling the relationship between predictors (such as customer age, income, and other relevant variables) and the likelihood of default. (Tripathi *et al.*, 2021) adds that the logistic regression model estimates coefficients for each predictor, which are used to compute the PDs. The logistic regression model is commonly employed as the base model for credit scoring. Example: A bank might use logistic regression to predict the probability of a borrower defaulting based on various features like income, credit history, and employment status. On the other hand, Shilbayeh and Grassa (2024) asserts that Decision trees have gained popularity in credit scoring due to their interpretability and flexibility. Decision trees follow a top-down approach, where at each step, the variable that best splits the dataset is chosen based on metrics like the Gini index, information value, or entropy. Decision trees can handle both numerical and categorical predictors. They are commonly used to fit data and predict default in credit scoring. For example A decision tree might split customers based on their credit utilization rate, income, and other factors to predict the likelihood of default. This paper aims at Combining Logistic Regression and Decision trees to improve credit scoring performance.

1.1.1 Global View

The global view on credit scoring has evolved significantly, especially in emerging markets. In emerging economies, an alternative form of credit scoring called ACS is gaining traction. ACS leverages artificial intelligence (AI) and social media data instead of traditional paper-based methods that rely on consumers having a bank account (Berg *et al.*, 2019). According to Berg et al. (2019) Unlike credit scoring in the US, which primarily focuses on banking and financial data

(credit cards, loans, etc.), ACS considers rich and relevant data from e-commerce transactions, phone credit top-ups, bill payments, travel bookings, and more. The scholar asserts that by using advanced technology (AI and machine learning) and alternative data sources (social media, electronic transactions, cellular data), ACS assesses consumers' financial fitness even when they lack access to formal banking services. As the result, there is Higher financial inclusion and improved risk analysis for lenders. However, Balancing accuracy and interpretability remains a challenge in credit risk scoring (Berg *et al.*, 2019). Researchers are exploring novel approaches to obtain global explanations for credit scoring models. These approaches aim to enhance transparency and understanding while maintaining predictive accuracy (Trivedi, 2020). Innovations include mimicking decision behavior using copies of trained machine learning classifiers. According to Trivedi (2020) Credit scoring is essential for economic growth and financial inclusion. A credit scorecard consists of predictive characteristics (demographic data, credit account performance, bank transactions, real estate data, etc.) that separate good and bad loans or counterparties. It helps lenders make informed decisions by assessing risk holistically. In summary, the global view on credit scoring is shifting toward more inclusive, data-driven approaches that leverage technology and alternative data sources.

Credit scoring is a critical component of the financial services industry worldwide. It plays a pivotal role in assessing the creditworthiness of individuals and businesses, influencing lending decisions across the globe. Traditional credit scoring models, such as those based on logistic regression, have been widely used due to their simplicity and interpretability. However, these models often fall short when dealing with complex, non-linear data patterns commonly found in large and diverse datasets. To address these challenges, there has been a growing interest in

developing hybrid models that combine traditional statistical methods with advanced machine learning algorithms.

Globally, financial institutions are increasingly adopting machine learning techniques to enhance the accuracy and robustness of credit scoring systems. For instance, research conducted by Provost and Fawcett (2013) highlights the application of machine learning in improving predictive analytics, which is critical for credit risk assessment. Furthermore, studies by Zhou (2012) on ensemble methods underscore the potential of hybrid models to outperform single-algorithm approaches by leveraging the strengths of multiple models.

a) The global view on using decision trees and regression analysis for credit scoring.

Accordint to Berg et al. (2019) Decision trees are widely used in credit scoring due to their interpretability and flexibility. Decision trees provide clear rules for decision-making, making them suitable for regulatory compliance can capture non-linear relationships between predictors and credit risk. Decision trees reveal feature importance, aiding risk assessment. However, Berg et al. (2019) assert that deep trees may over fit the training data. Small changes in data can lead to different tree structures. As mentioned by Djeundje et al. (2021) Logistic Regression on the other hand, remains a benchmark in credit risk modeling. Logistic regression models the log-odds of default as a linear function of predictors and Coefficients indicate the impact of each predictor. The scholar adds that logistic regression is widely accepted and used by banks and financial institutions globally however, it assumes a linear relationship between predictors and log-odds, May not capture complex non-linear effects and requires careful feature selection and transformation. This study proposes a hybrid model that combine decision trees and logistic regression to improve performance while maintaining interpretability. The hybrid model will

leverage decision trees to enhance logistic regression. By Extracting rules from short-depth decision trees and use these rules as predictors in a penalized logistic regression model while being able to Capture non-linear effects while preserving interpretability. This hybrid model will provide an accurate credit risk prediction.

Credit scoring is a critical tool used by financial institutions worldwide to assess the creditworthiness of individuals and businesses. Globally, the use of decision trees and regression analysis, particularly logistic regression, has been widely adopted for this purpose due to their distinct advantages and complementary strengths. Logistic regression has been the traditional workhorse for credit scoring globally. It is favored for its simplicity, ease of implementation, and interpretability. Logistic regression models the probability of a binary outcome (such as default or no default) based on one or more predictor variables. This technique is particularly effective when the relationship between the predictors and the outcome is linear, making it a reliable tool for credit scoring in many financial institutions. In the global financial sector, logistic regression is widely used because it provides clear, actionable insights into the factors contributing to credit risk. For instance, "Credit Risk Modelling" by David Jamieson Bolder (2014) discusses the prevalence of logistic regression in credit risk assessment, emphasizing its role in regulatory compliance and decision-making processes. Moreover, "Credit Scoring and Its Applications" by Thomas, Edelman, and Crook (2002) highlights how logistic regression has been instrumental in credit scoring models worldwide, offering consistency and transparency in risk evaluation across different markets.

Decision trees are another popular technique used globally for credit scoring. Unlike logistic regression, decision trees can capture non-linear relationships and interactions between variables without requiring any assumptions about the underlying data distribution. They work by

recursively partitioning the data into subsets based on the most significant predictors, making them highly interpretable and easy to visualize. Decision trees are valued for their flexibility and the ability to model complex interactions in data. Hastie, Tibshirani, and Friedman (2009) details how decision trees have been employed in credit scoring to enhance the accuracy of predictions by capturing non-linear patterns that logistic regression might miss.

The global financial industry has increasingly adopted decision trees as part of ensemble methods like random forests and gradient boosting, which combine multiple trees to improve predictive performance. Kevin P. Murphy (2012) explores how decision trees, when used in ensembles, can significantly enhance credit scoring models by reducing variance and improving robustness.

Globally, hybrid models that combine decision trees and logistic regression are gaining traction due to their ability to provide a more comprehensive analysis of credit risk. Max Kuhn and Kjell Johnson (2013) discusses how the integration of these models can enhance the predictive power of credit scoring systems, particularly in complex datasets. Additionally, Foster Provost and Tom Fawcett (2013) highlights the practical applications of combining decision trees with logistic regression, noting that this approach is becoming increasingly popular in financial institutions seeking to balance interpretability with predictive performance.

1.1.2 Continental View

Credit scoring in Africa has undergone significant changes, reflecting the unique challenges and opportunities across the continent. According to Djeundje et al. (2021b) there have been emerging innovations on credit scoring in Africa like Scoretechs (Credit Scoring Platforms):

These platforms enable lenders to assess credit risk more effectively. Scoretechs leverage diverse data sources, including mobile wallet transactions, social media activities, consumer financial behavior, and psychometrics. By incorporating non-traditional data, they enhance credit assessment accuracy. Invoicetechs: These platforms focus on invoice financing, allowing businesses to access working capital based on outstanding invoices. Lending Aggregators: These platforms connect borrowers with multiple lenders, streamlining the lending process. Telco-based Lenders: Leveraging telecom data, these lenders offer microloans and other financial services. Pay-as-you-go (PAYG): Common in solar energy financing, PAYG models allow consumers to pay in installments. Peer-to-Peer (P2P) Lending Platforms: Facilitate direct lending between individuals or businesses. According to Ampountolas et al. (2021) although While alternative data sources are promising, challenges remain in data quality, privacy, and coverage, Balancing predictive power with interpretability remains a challenge. African countries also face sovereign credit rating assessments by international agencies (Moody's, Fitch, S&P). These ratings impact investor confidence, borrowing costs, and economic growth. In summary, credit scoring in Africa is evolving rapidly, driven by technology, data, and a commitment to financial inclusion. Collaboration, regulatory alignment, and responsible innovation are key to shaping its future.

In Africa, the adoption of advanced credit scoring models is gaining momentum, driven by the need to address the unique challenges of the continent's financial landscape. These challenges include limited credit history data, diverse economic conditions, and the presence of informal financial sectors. A study by Akpan and Sani (2018) emphasizes the importance of customized credit scoring models tailored to the African context, where traditional methods often fail to capture the full credit risk due to the region's unique data characteristics.

In Europe, the financial sector has been at the forefront of integrating machine learning into credit scoring. European banks and financial institutions are increasingly exploring hybrid models that combine logistic regression with decision trees to enhance predictive accuracy. The European Central Bank (ECB) has encouraged the adoption of advanced analytics in credit risk management to improve resilience against financial instability. Research by Crook, Edelman, and Thomas (2007) highlights how hybrid models can effectively address the regulatory demands for more accurate and explainable credit risk assessments in the European context.

1.2.3. Contextual view

In the context of emerging markets, including countries in Asia and Latin America, the financial sector faces unique challenges related to credit risk assessment. These markets often deal with a large segment of unbanked populations and have limited access to comprehensive credit data. Hybrid credit scoring models are particularly relevant in these contexts, where traditional models are insufficient to address the complexity of the data. Research by Farrukh, Khan, and Ghani (2020) in Pakistan demonstrates the effectiveness of hybrid models in improving credit scoring accuracy by incorporating machine learning techniques that can handle diverse and incomplete datasets.

In the context of developing economies, particularly in Southeast Asia, financial institutions are increasingly turning to machine learning-based hybrid models to improve credit scoring systems. These models are being used to extend credit access to underserved populations by leveraging alternative data sources and advanced algorithms. A study by Mansinghka et al. (2019) in India illustrates how hybrid models, combining logistic regression and decision trees, have been successful in reducing default rates and improving credit decisions.

In 2018, Uganda's banking sector witnessed good credit growth (Nathan, Ibrahim and Tom, 2020). According Nathan, Ibrahim and Tom (2020) . The ratio of non-performing loans (NPLs) to total gross loans declined steadily, reaching 3.4% in December 2018 (compared to 5.6% in the prior year). The scholar adds that this positive trend was supported by a decrease in lending interest rates, which averaged 19.9% in 2018 (down from 21.3% in 2017). The scholar reveals that the number of accounts in commercial banks increased from 7.4 million in June 2017 to 12.1 million in December 2018, although this growth remains relatively low considering Uganda's population of nearly 40 million however Nathan, Ibrahim and Tom (2020) reveal that Leveraging alternative data sources remains a challenge. In Uganda, banks utilize both regression analysis and decision trees for credit scoring. Banks use Logistic regression models link credit scores to the probability of default (PD).Banks collect relevant data (e.g., customer age, income, employment status). The logistic regression model estimates coefficients for each predictor. These coefficients help compute the PDs, indicating the likelihood of default. Logistic regression serves as the base model for credit scoring. It's widely used due to its interpretability and solid foundation. Also Decision trees are used predict default by splitting data based on relevant features. Decision trees follow a top-down approach, choosing the best variable to split the dataset. Metrics like the Gini index, information value, or entropy guide the splits. Each branch represents a decision path leading to a default or non-default outcome. Decision trees enhance interpretability and flexibility in credit scoring. They're commonly used alongside logistic regression models. Banks validate and improve credit scoring systems using challenger models like decision trees. According to Kumar, Sharma and Mahdavi (2021) Uganda's banking sector witnessed credit growth, with declining non-performing loans. Lending interest rates decreased,

supporting credit expansion however, challenges remain, including data availability and cost-to-income ratios.

In recent years, with the continuous development of the Internet, the integration of technology and finance has deepened, which has led to tremendous changes in the financial industry. With the stimulation of consumer finance, the demand for various credit businesses is growing, and a reasonable and reliable risk assessment model needs to be established. Wang et al. (2020) asserts that In the past, when commercial banks conducted credit risk assessments on loan users, they often relied on the risk control personnel to rely on the 5C classification method to subjectively judge, and judge the loan users from five factors: their personal character, credit limit, solvency, and market economy. And weighed, as a reference for whether to lend to the user, decide whether to issue loans, this method of relying on subjective judgment is obviously inefficient, and the evaluation is very dependent on the subjective judgment ability of the risk control personnel, from the perspective of internal control of the company. Wang et al. (2020) adds that with the development of big data on the Internet, under the new data portraits and business scenarios, the traditional credit scoring model application is severely limited, and the original business logic framework is lost. In the face of tens of thousands or even hundreds of thousands of users applying for loans, the online lending platform needs to adopt various machine learning methods to reduce the manual participation in the monitoring and testing process, and use automated methods to improve lending.

1.2 Problem statement

In recent years, scholars have studied the application of a variety of machine learning methods in credit evaluation, such as decision trees, and logistic regression. According to Nhu et al. (2020)

both the decision tree methods and traditional logistic regression did not perform in credit scoring. According to Nhu et al. (2020) Logistic regression assumes Linearity which can be limiting when dealing with complex relationships. It also predicts discrete outcomes (e.g., binary classes), restricting its use to specific scenarios. More so, cannot handle non-linear problems effectively and it over fits if the number of observations is fewer than the number of features which it may lead to overfitting and is Limited in high dimensions' datasets On the other hand decision tree also tend to over fit the training data, especially when they grow too deep, more so, small changes in the data can lead to different splits, resulting in unstable trees. Decision trees tend to favor majority classes, which can be problematic in imbalanced datasets and they use a greedy approach during construction, which may not lead to the globally optimal tree. This makes these two approaches non efficient towards credit scoring if used independently. This study shall establish a hybrid model for combining the two approaches to improve logistic regression and decision tree using a machine learning approach with an aim to improve logistic regression and decision tree through data pre-processing and feature engineering while preserving the intrinsic interpretability of the scoring model.

1.3 Purpose of the study

To examine the efficiency of Logistic Regression and Decision Tree models to credit scoring in commercial banks in Uganda and establish an Enhanced Hybrid Model for Improved Credit Scoring Using Machine Learning Approach with Specific Reference to Logistic Regression And Decision Tree

1.4 Specific Objective

i. To assess the gaps in Logistic Regression and Decision Tree models to credit scoring and probability of default

ii. To develop a hybrid model for improving logistic regression and decision tree using a machine learning approach through data pre-processing and feature engineering

iii. To evaluate the ability of the developed hybrid model to discriminate between good, neutral and bad score using loan scoring data

1.5 Research Questions

i. What are the gaps in Logistic Regression and Decision Tree approaches to credit scoring and probability of default

ii. What hybrid model can be used improving logistic regression and decision tree using a machine learning approach through data pre-processing and feature engineering

iii. What is the ability of the developed hybrid model to discriminate between good, neutral and bad score using loan scoring data

1.6 Scope of the study

1.6.1 Geographical Scope.

This research instantiates centenary bank Centenary Bank Head Office branch at Mapeera House as a case study**.** The main location address for the Centenary Bank head office in Kampala is Plot number 44-46 along the Kampala Road and Plot 2 Burton Street. This is because centenary bank is facing challenges of default in loans department.

1.6.2 Content scope

This research study majorly focused on literature related to identifying gaps, and developing a Hybrid Model that combines Logistic Regression and Decision Tree to bridge the gaps in the two approaches to credit scoring using Machine Learning Approach.

1.6.3 Time scope

This research will consider a time scope of 2020-2024, the period when centenary started experiencing challenges with credit defaults from customers *(Centenary bank loan payment analysis report, 2023)*

1.7 Significance of the study

To the Academia, this study will contribute on literature relating to existing challenges in implementing contemporary Logistic Regression and Decision Tree for credit scoring in commercial banks and bridging the gaps in the two approaches while using Machine Learning Approach

To the banks, this research will help the commercial banks in giving hands on effective credit scoring model that combines the strength of two credit scoring models that is; Logistic Regression and Decision Tree but also bridging the gaps they bring.

Improving Predictive Accuracy, Traditional credit scoring models, such as logistic regression, often rely on linear relationships and may not fully capture the complexities and non-linearities inherent in credit data. Decision trees, on the other hand, can model non-linear relationships but may overfit the data if used alone. By combining these two approaches in a hybrid model, the study aims to leverage the strengths of both techniques enhancing predictive accuracy by capturing a wider range of patterns and interactions within the data.

Addressing Data Imbalances; Credit scoring often involves dealing with imbalanced datasets, where the number of good credit cases significantly outweighs the number of defaults. This imbalance can lead to biased predictions, where models favor the majority class. The hybrid model proposed in this study can integrate techniques to address these imbalances, such as class weighting or synthetic data generation, thus improving the model's ability to predict defaults accurately.

Enhancing Model Robustness; Hybrid models are typically more robust than single-algorithm models because they combine the predictions of multiple models. This reduces the likelihood of overfitting and enhances the generalizability of the model to new, unseen data. The study's focus on creating a hybrid model that integrates logistic regression and decision trees could provide a more stable and reliable tool for credit risk assessment.

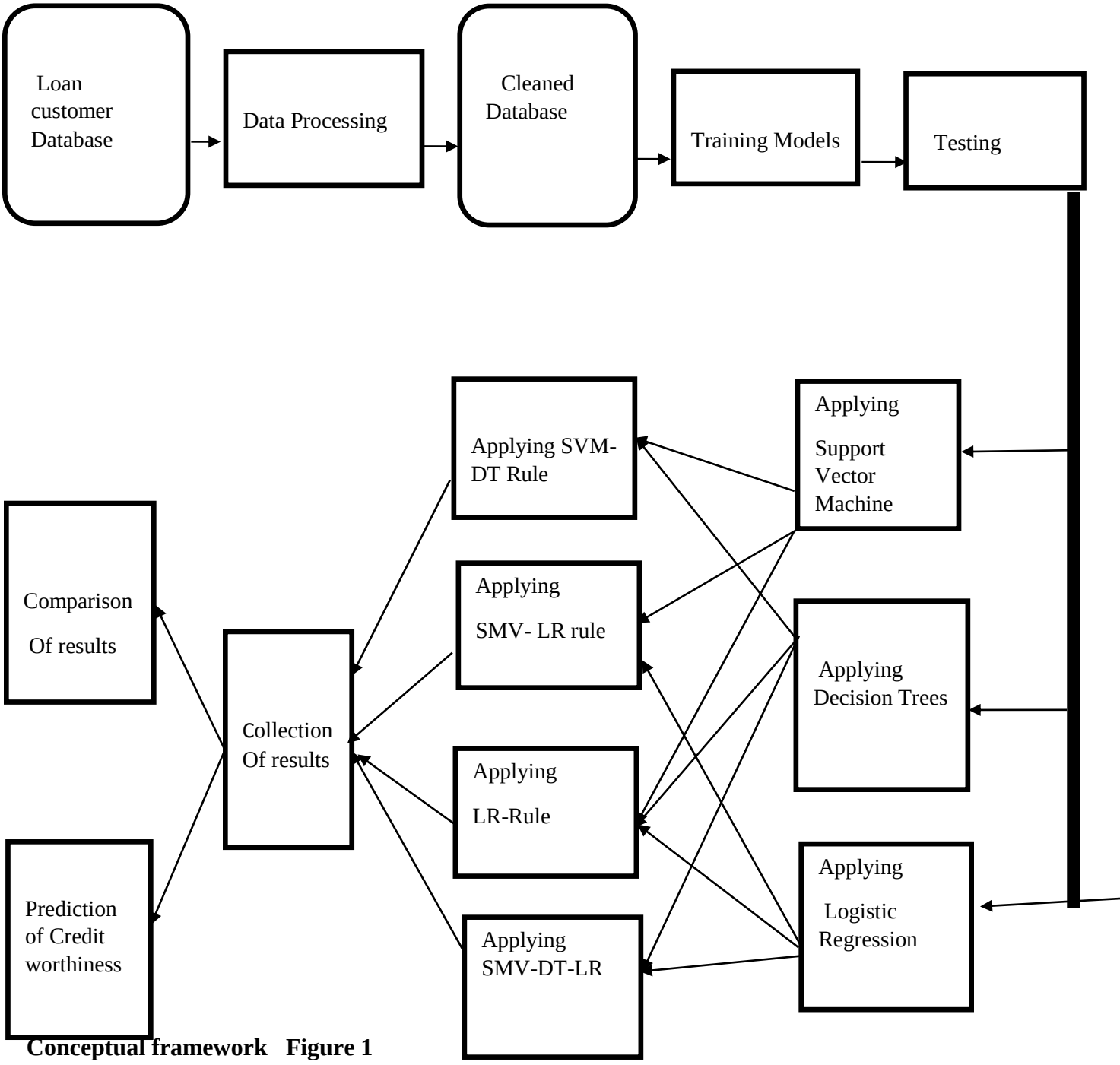
Providing Interpretability and explainability; One of the challenges with advanced machine learning models is their lack of transparency and explainability, which is crucial in financial decision-making. By combining logistic regression, which is highly interpretable, with decision trees, which provide clear decision paths, the study aims to develop a hybrid model that balances predictive power with interpretability. This makes it easier for stakeholders to understand and trust the model’s decisions.

Practical Implications for Financial Institutions; Financial institutions are under increasing pressure to improve their credit risk assessment methodologies to reduce default rates and optimize lending strategies. The hybrid model developed in this study has practical implications for banks, credit unions, and other lending institutions. It offers a more sophisticated tool that can help these organizations make more informed lending decisions, ultimately reducing financial risk and improving customer outcomes.

1.8 Justification of the study

This research is particularly relevant as centenary bank is currently facing challenges with high number of creditors not being able to pay back the loans. This means that the methodology that the bank is employing to determine a customers' credit worthiness may be inefficient. This has led to significant financial loss and also forced the bank to auction the customer's securities in order to recover the loans and this has made the bank unpopular in the public eye with people claiming that the bank just wants to steal from people. In the face of tens of thousands or even hundreds of thousands of users applying for loans(*centenary bank loan applications analysis report, 2022/23*) , the Centenary bank lending platform needs to adopt various machine learning methods to reduce the manual participation in the monitoring and testing process, and use automated methods to improve lending. This study is important because centenary has been using logistic regression as as a strategy to credit scoring for its loan applicants. And for the past 4 years, the bank has faced a challenge with its people who have failed to pay back their loans. This could imply that the methodology the bank is using to determine the customer's credit worthiness may not be sufficient. In recent years, scholars have studied the application of a variety of machine learning methods in credit. According to According to Wang et al. (2020) Logistic regression assumes Linearity which can be limiting when dealing with complex relationships. It also predicts discrete outcomes (e.g., binary classes), restricting its use to specific scenarios. More so, cannot handle non-linear problems effectively and it over fits if the number of observations is fewer than the number of features which it may lead to overfitting and is Limited in high dimensions datasets This study therefore intends to develop a hybrid model that combines both decision tree and logistic regression to credit scoring to bridge the gaps that come with using logistic regression alone to determine a customer's credit worthiness.

1.9 Conceptual Framework



Conceptual framework Figure 1

Support Vector Machine (SVM) is a type of model used to analyze data and discover patters in classification and regression analysis. Support vector machine (SVM) is used when data has

exactly two classes. An SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other class. The larger margin between the two classes, the better the model is. A margin must have no points in its interior region. The support vectors are the data points that on the boundary of the margin. Support Vector Machines map the training data into kernel space. There are many differently used kernel spaces – linear (uses dot product), quadratic, polynomial, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. to name a few. In addition, there are multiple methods of implementing SVM, such as quadratic programming, sequential minimal optimization, and least squares. A decision tree (DT) is a tool that uses classification or regression to predict a response to data. Classification is used when the features are grouped, and regression is used when the data is continuous. Decision tree is one of the main data mining methods. A decision tree is made of a root node, branches, and leaf nodes. To evaluate the data, follow the path from the root node to reach a leaf node. Decision trees must be created using a purity index which will split the nodes TO check for false positives or false negatives, giving us the accuracy, specificity, and sensitivity of the model. Logistic regression (LR) is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable (a dependent variable that can take a limited number of values) from a set of predictor or independent variables. In logistic regression the dependent variable is always binary (with two categories). Logistic regression is mainly used to for prediction and also calculating the probability of success. Logistic Regression involves fitting an equation of the form to the data: $Y = ß0 + ß1x1 + ß2x2 + \ldots + ßnxn$ – eq. 1 loan customer Database, Data Preprocessing, Cleaned Database Training the Models, Testing the Models, Applying Support Vector Machine (SVM), Applying Decision Trees (DT), Applying Logistic Regression (LR), Applying SVM-DT Rule Applying SVM-LR Rule Applying DTLR Rule, Applying SVM-DT-LR Rule,

Collection of Results, Comparison of Results, Prediction of credit worthiness. The regression coefficients are usually estimated using maximum likelihood estimation. The maximum likelihood ratio helps to determine the statistical significance of independent variables on the dependent variables. The likelihood-ratio test assesses the contribution of individual predictors (independent variables). Then the probability (p) of each case is calculated using odds ratio, P/(1-P) = eY – eq. 2 From this p–value is found out. This gives the probability or chance for the individual to be credit worthy. Rule Based Algorithm Rule based systems are essentially decision trees that use a small number of attributes for decision making. These are simple systems which are usually used to increase comprehension of knowledge patters. Rule based algorithms are indicative of trends in the features they consider and thus provides us with logical conclusions. Rules are used to support decision making in classification, regression, and association tasks. Depending on the data, there are different types of rules that can be implemented such as classical prepositional logic (C-rules), association rules (Arules), fuzzy logic (F-rules), M-of-N or threshold rules (Trules), similarity or prototype-based rules (P-rules). It is recommended that classification rule (C-rule) be used for this model. C-Rules are of the form of if-else ladders and provide the simplest and most comprehensible way of expressing knowledge. These rules will account the result of each of the individual methods based on weight of the model which is dependent on the accuracy, specificity and sensitivity attained.

1.10 Definition of key terms

a) A hybrid model

For the purpose of this study, this means a hybrid Model will mean a model which is developed combining two traditional models and in this study the models are logistic regression and decision tree

b) Logistic regression

For the purpose of this study, this will mean a strategy to credit scoring that links the score and probability of default (PD) through the logistic regression function, and is the default fitting and scoring model when you work with credit scorecard

c) Decision tree

For the purpose of this study, this will mean a credit scoring strategy that is commonly used to fit data and predict default. The algorithms in decision trees follow a top-down approach where, at each step, the variable that splits the dataset "best" is chosen. "Best" can be defined by any one of several metrics, including the Gini index, information value, or entropy.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents the review of existing literature about investigation and also presents the identified gaps that are to be filled by this study. This chapter also categorizes existing work on credit scoring and highlights the strategic logistic regression and decision tree in realizing accurate, reliable predictions while discussing how combining the two models can give a more accurate, specific, and reliable predictions on credit worthiness of a loan seeker.

2.2 Gaps in Logistic Regression and Decision Tree approaches to credit scoring and probability of default.

2.2.1 Gaps in Logistic regression for credit scoring

According to Dumitrescu et al. (2022) the recent reports on Logistic Regression published by the French regulatory supervisor (ACPR, 2020), the Bank of England, the European Commission (EC, 2020), and the European Banking Authority (EBA, 2020), explain why the logistic regression remains the standard approach in classification algorithm, especially in credit risk assessment due to its Ease of Implementation**,** Efficient Training, Probabilistic Predictions**,** Linear Boundaries, Multiclass Extension, Interpretability, and first class classification. However, the report expounds that Logistic regression assumes Linearity which can be limiting when dealing with complex relationships. It also predicts discrete outcomes (e.g., binary classes), restricting its use to specific scenarios. Moreso, Dumitrescu et al. (2022) adds that the reports affirm in agreement that Logistic regression cannot handle non-linear problems effectively and it Overfits if the number of observations is fewer than the number of features which it may lead to overfitting. Further, it is indicated that logistic regression is Limited in high dimensions datasets,

and it requires minimal multicollinearity between independent variables while it is challenging to obtain complex relationships using logistic regression. And yet according to Bücker et al. (2021) most international banks still use the logistic regression model, especially for regulatory scores used to estimate the probability of default for capital requirements (Basel III) or for point-in-time estimates of expected credit losses (IFRS9).

According to Aurélien Géron in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" (2019), Logistic Regression relies on the assumption that the relationship between the predictors and the log-odds of the outcome is linear. This assumption can be restrictive in credit scoring, where relationships between variables (such as income level and credit risk) are often non-linear, leading to potential oversimplification of the model. Géron emphasizes the need for careful model selection and possibly more flexible methods when dealing with non-linear data.

In "Applied Predictive Modeling" (2013), Max Kuhn and Kjell Johnson discuss how Logistic Regression can be particularly sensitive to outliers and multicollinearity. Outliers in financial data, such as extremely high- or low-income values, can disproportionately impact the model's coefficients, leading to biased predictions. Furthermore, multicollinearity where predictor variables are highly correlated can make it difficult to interpret the model and weaken the predictive power of individual variables.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman, in "The Elements of Statistical Learning" (2009), highlight that Logistic Regression does not inherently account for interactions between variables unless explicitly modeled. In the context of credit scoring, interactions between variables like credit history and debt-to-income ratio are crucial. The manual addition of

interaction terms can be cumbersome and may miss more subtle, complex relationships that could improve model accuracy.

Peter Bruce and Andrew Bruce (2017) highlights that both Logistic Regression and Decision Trees struggle with imbalanced datasets, a common issue in credit scoring where the number of defaults is much lower than non-defaults. This imbalance can skew the models towards predicting the majority class, thus failing to identify high-risk individuals accurately.

Foster Provost and Tom Fawcett (2013), emphasize the importance of interpretability in models used for credit scoring. While both Logistic Regression and Decision Trees are generally interpretable, this clarity diminishes in complex scenarios, particularly when dealing with numerous features or intricate decision paths.

Kuhn and Johnson (2013) in "Applied Predictive Modeling" also note that neither Logistic Regression nor Decision Trees are naturally equipped to handle temporal aspects of data, which are often critical in financial modeling. Without incorporating time as a factor, these models might miss trends or shifts in creditworthiness over time, leading to outdated or inaccurate predictions.

2.2.2 Gaps in Decision tree for credit scoring

According to Xia et al. (2020) decision tree method to credit scoring is another method used by a number of commercial banks locally and internationally. Known for its ability to provide a clear and intuitive representation of decision-making, Xia, Zhao, et al. (2020) asserts that decision tree method unlike logistic regression, can handle non-linear relationships between features and the target variable because they have ability to capture complex interactions. According to Xia et al. (2020) decision trees allow users to assess the importance of each feature

in predicting credit risk while handling missing values by creating separate branches for missing data. Xia, Zhao, et al. (2020) adds to say that decision trees are robust to Outliers therefore are less affected by outliers compared to linear models without forgetting their ability to be combined into ensemble models. Liu, Fan and Xia (2021)) add that although decision tree method has outstanding benefits towards its use in credit scoring, Decision trees tend to over fit the training data, especially when they grow too deep, More so, small changes in the data can lead to different splits, resulting in unstable trees. Dumitrescu et al. (2020) protest that decision trees tend to favor majority classes, which can be problematic in imbalanced datasets, and they use a greedy approach during construction, which may not lead to the globally optimal tree. Xia et al. (2020) in agreement adds that decision trees create piecewise constant regions, which may not capture subtle linear boundaries and slight changes in the data can lead to different splits, affecting the tree's structure. The above analysis between two credit scoring strategies present gaps in existing models under scrutiny. Therefore, this study proposes a hybrid model for improving logistic regression and decision tree using a machine learning approach with an aim to improve logistic regression and decision tree through data pre-processing and feature engineering while preserving the intrinsic interpretability of the scoring model.

Andreas C. Müller and Sarah Guido (2016), note that Decision Trees are prone to overfitting, especially with noisy or imbalanced datasets typical in credit scoring. Overfitting occurs when a model captures noise instead of the underlying pattern, leading to poor generalization on unseen data. This is a significant drawback when using Decision Trees to predict defaults, as it can result in unreliable risk assessments.

Géron (2019) also discusses the instability of Decision Trees in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow". He explains that small changes in the training data

can lead to vastly different tree structures, making Decision Trees less reliable for consistent credit scoring. This instability undermines the model's reliability and consistency, which are critical in financial applications.

Kuhn and Johnson (2013) point out that Decision Trees may become biased towards the majority class in imbalanced datasets, such as those found in credit scoring, where non-default cases are often more prevalent. This bias can lead to models that under-predict defaults, resulting in less effective risk management.

Hastie, Tibshirani, and Friedman (2009), discuss the scalability issues associated with Decision Trees. As datasets grow in size and complexity, Decision Trees can become unwieldy, leading to fragmentation where the tree splits into too many small, uninformative nodes. This limits the model's ability to make robust predictions in large-scale credit scoring scenarios.

2.3 Combining logistic regression and Decision tree for credit scoring

Logistic regression is one of the most used machine learning techniques. Its main advantages are clarity of results and its ability to explain the relationship between dependent and independent features in a simple manner. It requires comparably less processing power, and is, in general, faster than Random Forest or Gradient Boosting. However, it has also some serious drawbacks and the main one is its limited ability to resolve non-linear problems. This study will demonstrate how banks can improve the prediction of non-linear relationships by incorporating a decision tree into a regression model. The idea is quite similar to weight of evidence (WoE), a method widely used in finance for building scorecards. According to Djeundje et al. (2021c) weight of evidence takes a feature (continuous or categorical) and splits it into bands to maximise separation between goods and bads (positives and negatives) (Dumitrescu et al. 2022b). Decision

tree carries out a very similar task, splitting the data into nodes to achieve maximum segregation between positives and negatives. The main difference is that weight of evidence is built separately for each feature, while nodes of decision tree select multiple features at the same time. Knowing that the decision tree is good at identifying non-linear relationships between dependent and independent features. This study will transform the output of the decision tree (nodes) into a categorical variable and then deploy it in a logistic regression, by transforming each of the categories (nodes) into dummy variables. For the purpose of this study, using decision tree nodes in the model would out-perform both logistic regression and decision tree results. Although combining logistic regression and decision tree is not a commonly well-known approach, may outperform the individual results of both decision tree and logistic regression.

2.4 A hybrid model for improving logistic regression and decision tree using a machine learning approach

2.4.1 A proposed hybrid model that incorporates both decision trees and logistic regression to credit scoring

According to Dumitrescu et al. (2022c) Logistic regression links the score and probability of default (PD) through the logistic regression function, and is the default fitting and scoring model when you work with credit score card objects. However, decision trees have gained popularity in credit scoring and are now commonly used to fit data and predict default. The scholar adds that the algorithms in decision trees follow a top-down approach where, at each step, the variable that splits the dataset "best" is chosen. "Best" can be defined by any one of several metrics, including the Gini index, information value, or entropy. This section presents both a logistic regression model as a base model and a decision tree model to extract probability of default and also Validates the challenger model by comparing the values of key metrics between the challenger

model and the base model. Within this context, this study proposes a hybrid credit scoring approach which will aim to improve the predictive performance of the logistic regression model and decision tree through data pre-processing and feature engineering based on short-depth decision trees and a penalized estimation method while preserving the intrinsic interpretability of the scoring model. Formally, this model will consist of predictors from the decision trees and they will be used as binary rules outputted by short-depth decision trees built with original predictive variables. The model will consist of a simple logistic regression model. In order for this hybrid model to handle a possibly large number of decision-tree rules, the researcher will incorporate variable selection in the estimation through an adaptive lasso logistic regression model, a penalized version of classic logistic regression. The hybrid model developed will allows us to capture non-linear effects that can arise in credit scoring data putting in mind that ensemble methods have been praised for consistently outperform logistic regression because the latter fails to fit these non-linear effects. For instance, the random forest method benefits from the recursive partitioning underlying decision trees and hence, by design, accommodates unobserved univariate and multivariate threshold effects. The notable aspect of the approach of this study consists of using these algorithms to pre-treat the predictors instead of modelling the default probability directly with machine learning classification algorithms. Secondly, the hybrid model will provide parsimonious and interpretable scoring rules (e.g., marginal effects or scorecards) as recommended by the regulators, since it preserves the intrinsic interpretability of the logistic regression model and is based on a simple feature selection method.

Hybrid models are designed to capitalize on the strengths of different machine learning algorithms while mitigating their individual weaknesses. According to "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman (2009), ensemble methods, including

hybrid models, combine multiple models to improve overall predictive performance. The authors explain that logistic regression is a parametric model that works well for linearly separable data, while decision trees are non-parametric and can capture complex, non-linear interactions among features. By combining these models, a hybrid approach can address the limitations of each method individually.

Max Kuhn and Kjell Johnson (2013), the authors discuss the value of hybrid models in predictive analytics. They highlight that while logistic regression provides interpretability and ease of implementation, decision trees offer flexibility in modeling non-linear relationships and interactions. When combined, these models can provide a more comprehensive understanding of the data, particularly in complex domains like credit scoring.

Feature engineering plays a critical role in the development of hybrid models. Zheng and Amanda Casari (2018), the authors emphasize that the quality of the input features directly impacts the performance of the models. For instance, logistic regression can be used to identify key predictors of creditworthiness, while decision trees can further explore interactions among these predictors.

Foster Provost and Tom Fawcett (2013) discusses how feature selection and preprocessing can be tailored to suit both logistic regression and decision trees. They suggest that in a hybrid model, the data might first be processed through logistic regression to select the most relevant features. These features can then be used in a decision tree to model more complex relationships.

Zhi-Hua Zhou (2012), the author explains several techniques for combining models, including stacking and voting. Stacking involves training a logistic regression model and a decision tree

separately and then using another model (often another logistic regression) to combine their predictions. Voting, on the other hand, aggregates the predictions from both models, either by averaging probabilities (soft voting) or by majority class voting (hard voting).

Kevin P. Murphy (2012) provides insight into the probabilistic interpretation of combining models. Murphy explains that by using soft voting, the hybrid model can effectively weigh the contributions of logistic regression and decision trees based on their confidence levels, leading to a more nuanced final prediction.

The effectiveness of hybrid models in credit scoring is well-documented in "Credit Scoring and Its Applications" Thomas, Edelman, and Crook (2002). The authors discuss how combining logistic regression with decision trees can improve the accuracy of credit risk predictions, particularly in cases where the data exhibits both linear and non-linear patterns. Hybrid models have been shown to outperform standalone models by reducing the error rates and improving the classification of borderline cases.

One of the challenges of hybrid models is balancing predictive accuracy with interpretability. According to "Interpretable Machine Learning" by Christoph Molnar (2019), while decision trees provide clear visualizations of decision rules, the combination with logistic regression can complicate the interpretability of the overall model. However, this trade-off is often acceptable in financial contexts where accuracy in predicting defaults is paramount.

Andreas C. Müller and Sarah Guido (2016), the authors stress the importance of model validation, particularly in hybrid models. They advocate for rigorous cross-validation techniques to ensure that the hybrid model generalizes well to unseen data. The book also discusses the

potential of hybrid models to reduce overfitting, as they can combine the strengths of different algorithms to provide more stable predictions.

Hybrid models are increasingly being adopted in the financial industry for credit scoring. In "Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction" by Stewart Jones (2016), case studies are presented where financial institutions have successfully implemented hybrid models to improve credit risk assessment. These models are particularly effective in environments with large, complex datasets, where traditional models might struggle to capture the full range of relationships within the data.

Elizabeth Mays (2001) highlights the regulatory implications of using hybrid models in credit scoring. Mays discusses how hybrid models, while powerful, must be carefully validated and documented to meet regulatory standards. The transparency of the logistic regression component can help satisfy regulatory requirements, while the decision tree component can be used to enhance model performance.

This research proposes several simulated random numbers to estimate some functions of a probability distribution to illustrate the inability of standard parametric models, i.e., standard logistic regression models with linear specification of the index or with quadratic and interaction terms, to capture well the non-linear effects (thresholds and interactions) that can arise in credit scoring data. Furthermore, these simulations allow us to evaluate the relative performance of the hybrid in the presence of non-linear effects while controlling for the number of predictors. This hybrid will outperform standard logistic regression and decision trees used independently for credit scoring. In terms of forecasting accuracy while providing an interpretable scoring function.

2.6. Gap analysis and conclusion

Logistic regression is straightforward and easy to interpret. It works well when relationships between predictors and the response are approximately linear however, Logistic regression assumes a linear relationship between predictors and the log-odds of the response, It struggles to capture complex non-linear effects. While interpretable, it may not handle intricate interactions well. On the other hand, while Decision trees can model non-linear relationships effectively, they provide feature importance scores, Decision trees adapt well to various data patterns. However, it is important to note that Decision trees tend to over fit noisy data, and small changes in data can lead to different tree structures. Decision trees can become complex and less interpretable. This paper aims at to developing a hybrid model combining both strategies. This hybrid model will serve as a guide for credit scoring in the banking sector.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

This chapter describes the methodology that will be used to achieve the research objectives provided in chapter one.

3.1 Research Approach

Mulisa (2021) argues that research can be conducted using either a deductive approach or an inductive approach. In the deductive approach, the researcher begins with a general context and then applies theories, frameworks, or models to specific contexts. In contrast, the inductive approach involves starting with specific incidences and then formulating broader generalizations or frameworks (Kyngäs, 2019). This study will utilize the inductive research methodology. The research requires conducting a investigation into logistic regression and decision trees as models for credit scoring

3.1.1 Research Strategy

This study Used loan application datasets from centenary bank from year 2020-2023. Using this data the researcher may get a minor, but still an improvement of combined logistic regression and decision tree over both these methods used separately. The researcher will then import data and perform data cleansing by removing incorrect, corrupted, and incorrectly formatted, duplicate or incomplete data within the data set. This is because we are combining multiple data sources of different years and therefore are opportunities for data to be duplicated or mislabeled. The researcher will then save the cleansed data into a separate file. Because of the small frequency, the researcher will oversample the data using SMOTE technique. Nhu et al. (2020) asserts that

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random over sampling. In the next step, the researcher will build three models; that is decision tree, logistic regression, and logistic regression with decision tree nodes. The researcher will then convert the nodes into new variable by coding up the decision tree rules and then copy and paste the output into the next function, which we can use to create our new categorical variable. The researcher will then create a new variable ('nodes') and transfer it into dummies. After adding nodes variable, the researcher will re-run split to train and test groups and oversampled the loan application data using SMOTE and then run logistic regressions and compare the impact of node dummies on predictability. The researcher will create a list of all features excluding the nodes dummies and run the logistic regression using the Init list and re-run the regression, but this time include nodes dummies.

3.2 Research Methods

This section provides a concise overview of the research methodologies found in the literature, from which the researcher will choose the best suitable method for the study. According to Vassilakopoulou and Hustad (2021) the commonly used methods in information systems research are as follows: (1) Quantitative research methods, which are used to study natural phenomena or investigate human or organizational behavior; (2) Qualitative research methods, which are used to study social and cultural phenomena; (3) Mixed research methods or triangulation, which involve using both qualitative and quantitative research methods in a study; and (4) Design science research method, which focuses on creating new and innovative artifacts to enhance human and organizational capabilities. This project aims to create new and innovative artifacts to enhance credit scoring by creating a novel tool. Hence, the suitable research methodology to steer such a study is design science. Design Science is a systematic approach

that aids in the development and assessment of a product or solution aimed at resolving specific issues inside an organization (Brocke, Hevner and Maedche, 2020). Design Science aims to tackle unresolved, complex, and important problems, or provide more efficient and effective solutions to existing problems (Vassilakopoulou and Hustad, 2021).

3.3 Adopting Design Science guidelines

3.3.1 Problem relevance and design as an artifact

According to (Brocke, Hevner and Maedche (2020) Design Science seeks to create technological solutions for significant and pertinent business issues that result from the interaction of people, processes, and supporting technologies. The issue of insufficiently compatible e-Government systems in emerging nations will persistently result in the squandering of resources and the escalation of e-Government expenditures due to the replication of procedures and technologies across many government agencies. In addition, the scholar adds that design science requires that research must generate a practical artefact in the form of a construct, a model, a technique. Design Science is a methodology that addresses challenges that involve intricate relationships between different parts of the problem and its solution. The hybrid model aims to address the existing intricate issues in credit scoring in the banking sector, which prevents banks from fully benefiting from an efficient, and accurate, well-coordinated credit scoring-IT models

3.3.2 Research rigor and design as a search process

Design science necessitates utilizing existing resources to achieve desired objectives when searching for an efficient artifact Harley and Cornelissen (2020). Moreover, the design of an artifact is primarily a process of exploration aimed at finding an efficient solution to a problem. The scholar adds that this process comprises doing research activities, such as building,

assessing, and improving the artifact depending on the findings. Hence, in order to accomplish the study objectives, several actions will be carried out during the creation of the hybrid model, as indicated in table below.

Table 3: activities to be done to realize a hybrid model for improving logistic regression and decision tree- a machine learning

Research objective	**Activities to be done and tools to be used**
Objective (i) To assess the gaps in Logistic Regression and Decision Tree models to credit scoring and probability of default	• Shall Conduct interviews with officials responsible for credit scoring in the credit department to assess the gaps existing in Logistic Regression as aa model to credit scoring and probability of default • Shall Conduct interviews with officials responsible for credit scoring in the credit department to assess of the gaps existing in decision trees as a model to credit scoring and probability of default
Objective (ii) To develop a hybrid model for improving logistic regression and decision tree using a machine learning approach through data pre-processing and feature engineering	• Shall develop a hybrid model for improving logistic regression and decision tree using a machine learning approach
Objective (iii) To evaluate the ability of the developed hybrid model to discriminate between good, neutral and bad score using loan scoring data	• Shall use officials responsible for credit scoring in the credit department to evaluate the ability of the developed hybrid model to discriminate between good, neutral and bad score using loan scoring data

According to row 1 of table 3, interviews were conducted to gather data while studying the current obstacles in using logistic regression and decision tress for credit scoring in the banking sector. Centenary bank, Mapeera, - Uganda will be utilized as a case study in this examination. Interviews will be employed as they are deemed to offer a comprehensive and more profound comprehension of the phenomenon being investigated in contrast to questionnaires, as stated by (Lobe, Morgan and Hoffman, 2020). Interviews also allow the researcher to provide further explanation and clarification on important words in the study to enhance the respondents' comprehension during the interview process.

3.4 Target population

Lobe, Morgan and Hoffman (2020) define a target population as a compilation of sampling units from which a sample is selected. These entities may encompass individuals, institutions, groups, homes, and so forth. The focus of this study was 8 centenary bank employees in the credit department and 3 other heads in IT department. Making the population 11 respondents who are responsible for credit scoring system implementation and functionality at Centenarybank-Mapeera as identified by *(Centenary bank-mapeera, Employee audit report, 2022/23)*

3.5 Sampling and sampling Techniques

Sampling, as defined by Berndt (2020) is the process of selecting a smaller group from a larger population to accurately reflect the entire population. Sampling allows the researcher to select a sample size that corresponds to the number of objects chosen for the investigation. There are two primary classifications of sampling methods: probability and non-probability sampling. This study will however use non probabilistic sampling where the participants of this study shall be purposely selected. For the purpose of this study, all the 8 employees in the credit department at

Centenary bank shall be selected. More so, the study will also select (*3 individuals also employees of centenary bank in the IT department*) purposely the IT head, his deputy and the systems manager at centenary bank. This strategy should be employed because individuals in this category possess comprehensive knowledge regarding the subject under study.

Row 1 of table 3 displays the assessments that shall be done in order to identify gaps in the use of logistic regression and decision trees as models of credit scoring in the banking sector. Interviews shall be conducted. These models need to be examined in order to identify any deficiencies that prevent them from effectively offering accurate, interpretable, consistent and reliable credit scores when used independently. The researcher will use this information to develop a hybrid model that combines the effort of both models that is logistic regression and decision trees that shall be used as a benchmark for banks for effective credit scoring

3.5 Design evaluation, research contribution and communication

Design Science necessitates a thorough assessment of the artifact and a meticulous disclosure of its quality (Anderson and Lin, 2024). Design is a process that involves continuous improvement and small steps. The evaluation phase is crucial as it gives input to the design phase. An artifact is considered finished and successful when it meets the requirements of the problem it was intended to solve (Anderson and Lin, 2024).An artifact can be evaluated using approaches such as observation, analysis, experimentation, or description. Table 3 indicates that the assessment of the hybrid mode artifact that will be conducted by an observational approach. Observational approaches encompass several techniques such as case study, action research, field demonstration, and pilot project. In this research, a field framework evaluation approach will be most suited. Design science research must offer comprehensible and demonstrable contributions

in the domains of the design artefact and design processes. The bottom right side of figure 3 demonstrates that the main contribution of this research to the knowledge base is e-health information exchange framework

3.6 Justification of Methodology Using Data from Centenary Bank

"Our methodology leverages a dataset provided by Centenary Bank, a leading financial institution renowned for its comprehensive and meticulously maintained loan records. This dataset spans several years and includes diverse loan types and borrower profiles, ensuring that our findings are robust and representative of real-world lending environments. The use of this dataset not only enhances the credibility of our predictive model but also ensures its applicability in practical banking operations. All data handling processes adhered to stringent ethical standards and data protection regulations, with Centenary Bank's explicit permission, thereby upholding the integrity of our research."

3.7 Integrating Logistic Regression and Decision Tree into a Hybrid Model

In this section, we describe the process of integrating two distinct machine learning models. Logistic Regression and Decision Tree into a hybrid model using a Voting Classifier. The hybrid model is designed to leverage the strengths of both individual classifiers to improve predictive performance and robustness in classifying loan statuses.

3.7.1 Data Collection and Preprocessing

The data used for this research was sourced from Centenary Bank and contains loan performance records, including borrower details, loan types, and loan status (whether a borrower defaulted or

not). This ensures the dataset is relevant to real-world banking scenarios, supporting the development of practical predictive models. Key preprocessing steps include:

Handling missing values: Any missing data were identified and addressed through either imputation or removal, depending on the severity and nature of the missing data, in line with recommended preprocessing techniques for machine learning models in finance .

Label Encoding: Categorical variables, such as Gender and Status, were encoded using LabelEncoder to convert them into numerical form, allowing them to be used in the models.

After preprocessing, the dataset was split into feature vectors X and target labels y, where X consists of independent variables (e.g., account numbers) and y is the dependent variable representing the loan status.

3.7.2 Model Selection

Two classifiers were chosen based on their complementary strengths

3.7.3 Logistic Regression

A probabilistic model that is simple, interpretable, and effective for binary classification tasks. Logistic regression works particularly well when the relationship between features and target is approximately linear.

Logistic regression is well-suited for datasets where relationships between features and target are linearly separable or close to linear. The logistic model is initialized with class_weight = 'balanced' to account for potential imbalances in the loan status classes (e.g., more non-defaulters than defaulters).

3.7.4 Decision Tree Classifier

A non-linear model that works by recursively splitting the data based on the most important features. Decision trees can capture complex relationships between features, making them effective when there are non-linear patterns in the data. The decision tree classifier is also initialized with class_weight = 'balanced' to mitigate class imbalance, which is crucial when working with datasets that have skewed class distributions.

3.7.5 Creating the Hybrid Model

To take advantage of the strengths of both models Logistic Regression's probabilistic approach and Decision Tree's non-linear capacity we combine them using a Voting Classifier. This hybrid model is a form of ensemble learning, where multiple models are combined to improve overall predictive performance.

3.7.6 Voting Classifier (Soft Voting)

The Voting Classifier aggregates the predictions of both Logistic Regression and Decision Tree classifiers. Soft voting was chosen, meaning the model averages the predicted probabilities from each classifier and assigns the class with the highest average probability as the final prediction . This approach allows for a more balanced decision-making process, leveraging the strengths of both models. Logistic Regression provides stability for cases where linear relationships exist, while Decision Trees capture more complex, non-linear patterns.

3.7.7 Comparison of Individual vs. Hybrid Models

Finally, we compare the performance of the individual classifiers (Logistic Regression and Decision Tree) against the hybrid Voting Classifier. We expect the hybrid model to outperform the individual models in terms of predictive accuracy, as it benefits from combining both linear and non-linear decision boundaries.

3.8 Conclusion

The major purpose of this project was to build a hybrid model for credit scoring that combines both logistic regression and decision trees for credit scoring in the banking sector replacing the independent use of traditional regression and decision. A Design science research methodology shall be implemented throughout the research process since it facilitates the generation of artifacts / new knowledge and structures the research process in a logical way thereby attaining all the given objectives. The deductive research strategy will help in systematically carrying out the investigation.

CHAPTER 4

DATA ANALYSIS AND PREPROCESSING

4.1 Data Description

The dataset used in this research pertains to loan prediction, where the goal is to classify loans based on various borrower and loan attributes. The dataset includes several features that describe the characteristics of loans and borrowers. Below is a detailed description of the dataset:

4.1.1 Description of the Dataset

ACCOUNT_NO: Unique identifier for the account
FIRST_NAME: First name of the account holder
SURNAME: Surname of the account holder
ACCOUNT_NUMBER: Account number
ID_PRODUCT: Product ID associated with the account
REMAINING_MONTHS: Remaining months of the loan
TOTAL_DAYS: Total days of the loan
OVERDUE_DAYS: Number of days the loan is overdue
OVERDUE_MONTHS: Number of months the loan is overdue
LOAN_STATUS: Status of the loan (target variable, 0 for good, 1 for bad)
ACC_STATUS: Account status
INSTA_PAID: Amount paid instantly
TOT_DRAWDOWN_AMN: Total amount drawn down
LOAN_CLASS: Classification of the loan
INSTALL_FREQ: Frequency of installments
INSTALL_COUNT: Number of installments
FKGD_HAS_AS_CLASS: Flag indicating if the account has a certain classification
CUST_TYPE: Type of customer
ACCOUNT_STATUS: Status of the account
BOOK_BALANCE: Book balance of the account
LAST_TRX_DATE_DAYS: Number of days since the last transaction

Age: Age of the account holder
NUM_OF_CHILDREN: Number of children

Summary statistics provide a snapshot of the central tendency, dispersion, and shape of the distribution of the dataset's features. This includes measures such as mean, median, standard deviation, minimum, and maximum values for numerical features.

4.1.2 Distribution of Categorical Features

Categorical features describe qualitative aspects of the data, and their distribution is important to understand the balance of classes. This includes features such as LOAN_STATUS, ACC_STATUS, LOAN_CLASS, CUST_TYPE, and Gender.

In the subsequent sections, detailed steps will be provided on how the dataset was cleaned, pre-processed, and split for training and testing machine learning models. The importance of each preprocessing step will be discussed in relation to its impact on model performance.

Data preprocessing can often have a significant impact on generalization performance of a supervised ML algorithm. The elimination of noise instances is one of the most difficult problems in inductive ML (**M. Singh and G. M. Provan**) Usually the removed instances have excessively deviated instances that have too many null feature values. These excessively deviating features are also referred to as outliers. In addition, a common approach to cope with the infeasibility of learning from very large data sets is to select a single sample from the large data set. Missing data handling is another issue often dealt with in the data preparation steps.

F **lues**

```
Missing Values:
ACCOUNT_NO                0
FIRST_NAME                0
SURNAME                   0
ACCOUNT_NUMBER            0
ID_PRODUCT                0
REMAINING_MONTHS          0
TOTAL_DAYS                0
OVERDUE_DAYS              5
OVERDUE_MONTHS            5
LOAN_STATUS               0
ACC_STATUS                0
INSTA_PAID                0
TOT_DRAWDOWN_AMN          0
LOAN_CLASS                0
INSTALL_FREQ              0
INSTALL_COUNT             0
FKGD_HAS_AS_CLASS         0
CUST_TYPE                 0
ACCOUNT_STATUS            0
BOOK_BALANCE              0
LAST_TRX_DATE_DAYS        0
Age                       0
NUM_OF_CHILDREN           0
```

Overdue days and overdue months have five null values / missing values these has e to be eliminated to avoid any outliers.

Preprocessing is not just a preliminary step but a crucial part of the machine learning process. Effective preprocessing can significantly enhance model accuracy, efficiency, and robustness. By addressing issues like missing values, outliers, scaling, and encoding, and by transforming and selecting features appropriately, we lay a solid foundation for building powerful machine learning models.

4.3 Data Cleaning

Fig 1: Showing the line of code used to show the missing values

```
# Step 2: Data Profiling and Cleaning
print("Summary Statistics:")
print(data.describe(include='all'))
print("\nMissing Values:")
print(data.isnull().sum())
print("\nData Types:")
print(data.dtypes)
```

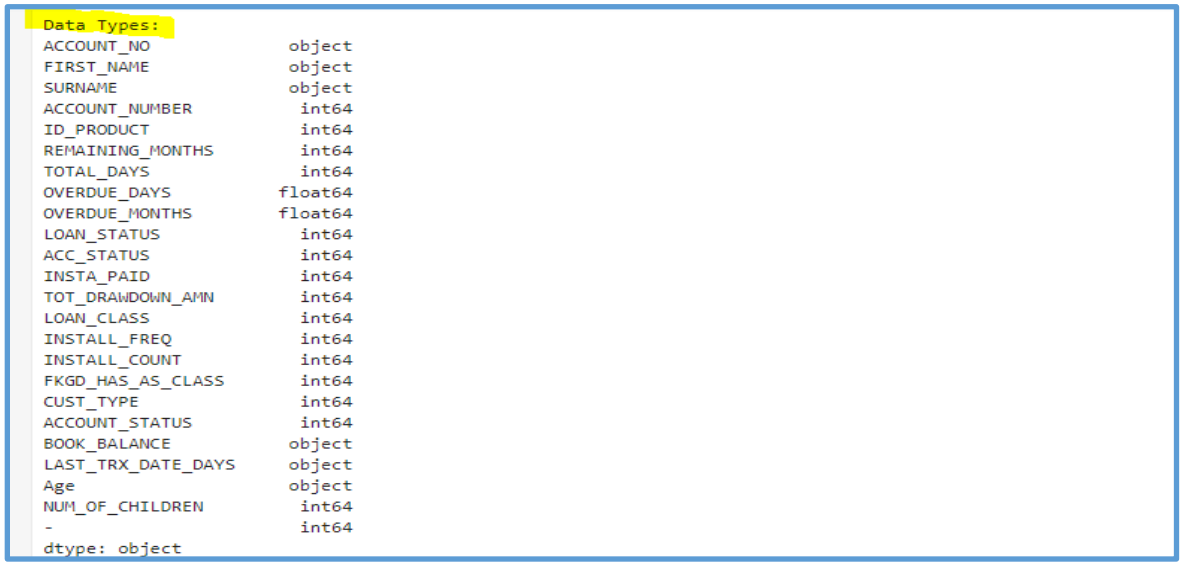
Strategy for Imputation: I used forward fill (ffill) for missing values. This strategy fills missing values with the most recent non-missing value.

Justification for Chosen Strategy: Forward fill is a simple and effective method for time-series data or sequential data where the previous value can be a reasonable estimate for the missing value.

Incompleteness: Real-world datasets often have missing or incomplete data. If not handled properly, missing values can lead to biased results or model failure.

Strategies: Common strategies for handling missing data include imputation (replacing missing values with mean, median, mode, or using more sophisticated methods like KNN imputation) and deletion (removing records with missing values).

1. **Fig 3: Showing the showing the data types after encoding and provides justification for using label encoding.**



4.4 Feature Engineering

Feature engineering involves creating new features or transforming existing ones to improve model performance. This process can significantly enhance the predictive power of machine learning models.

4.4.1 Creation of New Features

In this dataset, new features can be derived from the existing ones to provide additional information to the model. For example:

- **Age of Account**: Using the LAST_TRX_DATE_DAYS and Age columns, we can calculate the age of the account.
- **Loan Overdue Ratio**: Using OVERDUE_DAYS and TOTAL_DAYS, we can calculate the ratio of overdue days to the total days, providing insight into the borrower's repayment behavior.
- **Monthly Installment Amount**: Using TOT_DRAWDOWN_AMN and INSTALL_COUNT, we can calculate the average monthly installment amount.

Fig 3: Showing the New Features Created

```
#  Feature Engineering
# Creation of new features
data['Account_Age'] = data['Age '] - data['LAST_TRX_DATE_DAYS']
data['Loan_Overdue_Ratio'] = data['OVERDUE_DAYS'] / data['TOTAL_DAYS']
data['Monthly_Installment_Amount'] = data['TOT_DRAWDOWN_AMN'] / data['INSTALL_COUNT']
```

4.5 Data Splitting

Data splitting is a fundamental step in the process of building and evaluating machine learning models. It involves dividing the dataset into separate subsets, typically a training set and a testing set According to (Xu, Yun and Goodacre, Oct. 29, 2018).

Fig 4: Showing the Data Splitting the Split data in for the module to train and process

```
# Split data into features and target
X = data.drop('LOAN_STATUS', axis=1)
y = data['NUM_OF_CHILDREN']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Training Set: This subset is used to train the machine learning model. The model learns patterns and relationships in the data during the training phase.

Testing Set: This subset is used to evaluate the performance of the trained model on unseen data. It helps to estimate how well the model generalizes to new, unseen data.

In practice, a common splitting strategy is the 80-20 split, where 80% of the data is used for training and 20% for testing. However, other ratios like 70-30 or 90-10 can also be used depending on the size of the dataset and the specific requirements of the project (Aurélien Géron , 2019).

4.6 Model Training and Evaluation

4.6.1 Logistic Regression

Logistic Regression is a statistical model that uses a logistic function to model a binary dependent variable. It is used for binary classification problems where the outcome can be one of two possible categories.

Training the Logistic Regression model

Fig 4: Showing the Logistic Regression model Training results

```
Logistic Regression:
```

Accuracy: 0.851
Confusion Matrix:

```
[[633   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [ 39  39  24   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0 212   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   1 165   1   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0 213   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   8 162   0   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0  19 143   0   0   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0  74   6   1   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0  30  53   0   0   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0  11   0  18   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   3   0  73   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0  11   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0  18   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   5   0   0   0   0   0   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   6   0   0   0   0   1   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   5   0   2   0   0   1   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   3   1   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   2   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   0   0
    0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   7   0   0
    0   1   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0
    0   1   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   1   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   1   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
    0   1   0   0   0   0]]
```

Classification Report:

	precision	recall	f1-score	support
0.0	0.94	1.00	0.97	633
1.0	1.00	0.38	0.55	102
2.0	0.89	1.00	0.94	212
3.0	1.00	0.99	0.99	167
4.0	0.96	1.00	0.98	213
5.0	0.90	0.95	0.92	170
6.0	0.66	0.88	0.75	162
7.0	0.17	0.07	0.10	81
8.0	0.78	0.64	0.70	83
9.0	0.00	0.00	0.00	29
10.0	0.53	0.96	0.69	76
11.0	0.00	0.00	0.00	11
12.0	0.00	0.00	0.00	18
13.0	0.00	0.00	0.00	5
14.0	0.00	0.00	0.00	7
15.0	0.06	0.12	0.08	8
16.0	1.00	0.20	0.33	5
17.0	0.00	0.00	0.00	3
18.0	0.00	0.00	0.00	2
20.0	0.20	0.12	0.15	8
21.0	0.00	0.00	0.00	2
26.0	0.00	0.00	0.00	1
28.0	0.00	0.00	0.00	1
30.0	0.00	0.00	0.00	1
accuracy			0.85	2000
macro avg	0.38	0.35	0.34	2000
weighted avg	0.82	0.85	0.82	2000

The logistic regression model shows a high overall accuracy of 85.1%, but it seems to perform poorly in classes with fewer samples, as indicated by the low precision, recall, and f1-scores for those classes. This is expected given the class imbalance in the dataset.

4.6.2 Decision Tree

Training the Decision Tree model

Fig 5: Showing the Decision Tree model Training results

```
Decision Tree:
Accuracy: 0.9985
Confusion Matrix:
[[633  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   0  0  0  0  0  0  0  0  0]
 [  0 102  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
   0  0  0  0  0  0  0  0  0]
```

```
[ 0  0 212  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0 167  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0 213  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0 170  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0 162  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0 81  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0 83  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0 29  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0 76  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0 11  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0 18  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  5  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  7  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  8  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  5  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  1  1  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  8  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  2  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  1  0  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  1  0  0  0]
[ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  1]]
```

Classification Report:

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	633
1.0	1.00	1.00	1.00	102
2.0	1.00	1.00	1.00	212
3.0	1.00	1.00	1.00	167
4.0	1.00	1.00	1.00	213
5.0	1.00	1.00	1.00	170
6.0	1.00	1.00	1.00	162
7.0	1.00	1.00	1.00	81
8.0	1.00	1.00	1.00	83
9.0	1.00	1.00	1.00	29
10.0	1.00	1.00	1.00	76
11.0	1.00	1.00	1.00	11
12.0	1.00	1.00	1.00	18
13.0	1.00	1.00	1.00	5
14.0	1.00	1.00	1.00	7
15.0	1.00	1.00	1.00	8
16.0	1.00	1.00	1.00	5
17.0	1.00	1.00	1.00	3
18.0	1.00	0.50	0.67	2
19.0	0.00	0.00	0.00	0
20.0	1.00	1.00	1.00	8
21.0	1.00	1.00	1.00	2
24.0	0.00	0.00	0.00	0
25.0	0.00	0.00	0.00	0
26.0	0.00	0.00	0.00	1
28.0	0.00	0.00	0.00	1
30.0	1.00	1.00	1.00	1
accuracy			1.00	2000
macro avg	0.81	0.80	0.80	2000
weighted avg	1.00	1.00	1.00	2000

The Decision Tree model has achieved a remarkably high accuracy of 99.85%. the below shows a detailed performance.

Fig 6: Showing the Decision Tree model confusion matrix

Definition: A confusion matrix is a table that is used to evaluate the performance of a classification algorithm. It compares the actual target values with the predicted values.

```
[[633   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0 102   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0 212   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0 167   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0 213   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0 170   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0 162   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
```

[0 0 0 0 0 0 0 81 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 83 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 29 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 76 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 11 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 18 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0

The confusion matrix is a 27x27 matrix (as there are 27 classes, ranging from 0.0 to 30.0, with some missing numbers due to the dataset). Each cell (i, j) represents the number of instances of class i that were predicted as class j. Here's how to interpret the matrix.

Diagonal elements (i, i): These represent the number of correct predictions for each class. For example, the first cell (0, 0) is 633, meaning that 633 instances of class 0.0 were correctly predicted as class 0.0.

Off-diagonal elements (i, j): These represent the number of misclassifications, where instances of class i were predicted as class j. For example, cell (18, 19) has 1, meaning that 1 instance of class 18.0 was misclassified as class 19.0.

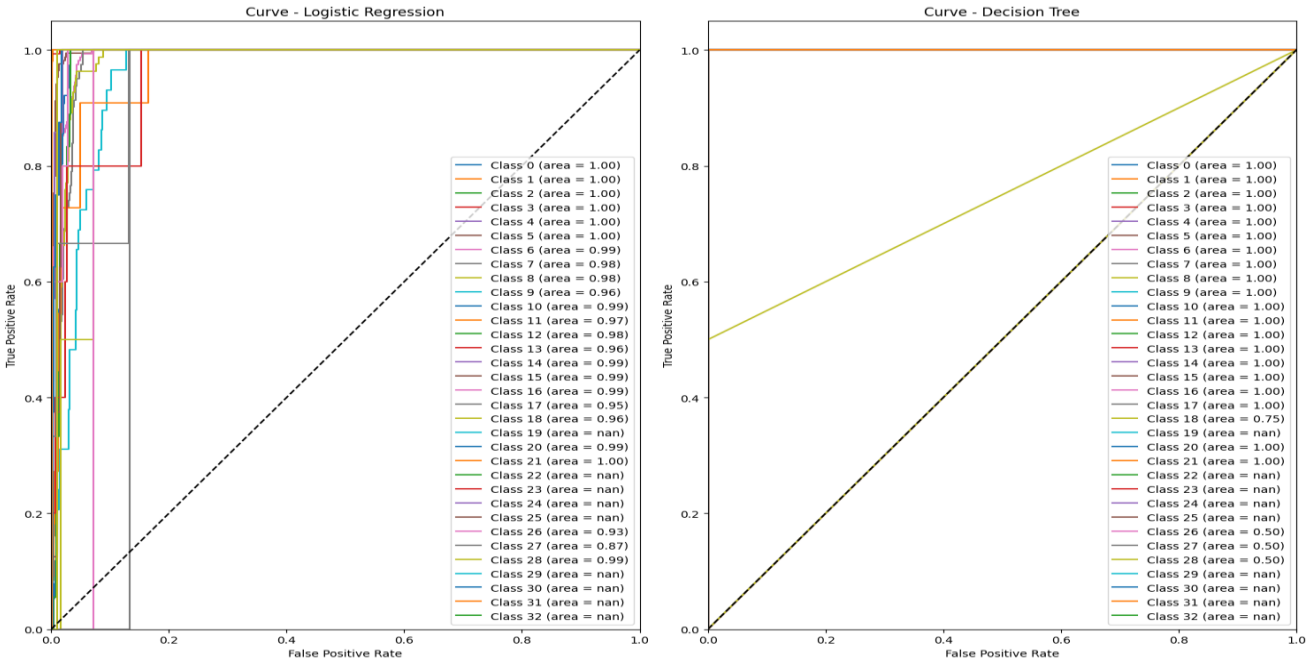
High Accuracy: The majority of the entries are on the diagonal, indicating that the model is highly accurate. This is supported by the overall accuracy of 99.85%.

Misclassifications: Class 18.0: 2 instances, with one misclassified as class 19.0.

Classes with zero instances (19.0, 24.0, 25.0): The model did not predict any instances for these classes, indicating either the absence of these classes in the dataset or misclassification of instances of these classes into other classes.

Classes with a small number of instances: Some classes have very few instances, such as class 26.0 and 28.0 with only 1 instance each, which might have been challenging for the model.

4.7 Model Evaluation Metrics



Accuracy is the ratio of correctly predicted instances to the total instances in the dataset. Accuracy gives an overall measure of how often the classifier is correct. It is easy to understand.

- True Positives (TP): Instances correctly predicted as the positive class.
- True Negatives (TN): Instances correctly predicted as the negative class.
- False Positives (FP): Instances incorrectly predicted as the positive class (Type I error).
- False Negatives (FN): Instances incorrectly predicted as the negative class (Type II error).

These metrics together give a comprehensive understanding of a model's performance, highlighting different aspects of its prediction capabilities and error types.

4.8 HYBRID MODEL

4.8 .1 Overview

To determine customer credit worthiness using a hybrid model, we have combined the predictions from both the logistic regression model and the decision tree model. One common approach is to use a voting mechanism where both models contribute to the final prediction. This is known as an ensemble method.

This is how this was achieved.

4.8.2: Train Both Models

Both the logistic regression and decision tree models are trained on the same training data.

I initialized the Decision tree and logistic regression models.

```
# Initialize models
logistic_model = LogisticRegression(max_iter=1000)
decision_tree_model = DecisionTreeClassifier()
```

4.8.3: Predict with Both Models

Both models were made to make predictions on the test data. – see Script below

```
# Create an ensemble model using VotingClassifier
hybrid_model = VotingClassifier(estimators=[
    ('logistic', logistic_model),
    ('decision_tree', decision_tree_model)
], voting='soft')
```

Combine Predictions: Combine the predictions from both models using majority voting and averaging the predicted probabilities.

Evaluate the Hybrid Model: Assess the performance of the hybrid model using the evaluation metrics.

4.8.4 Hybrid Model Accuracy

Accuracy: 0.83

	precision	recall	f1-score	support
0	0.90	0.91	0.91	2665
1	0.24	0.23	0.23	335
accuracy			0.83	3000
macro avg	0.57	0.57	0.57	3000
weighted avg	0.83	0.83	0.83	3000

The results show that the overall accuracy of the model is 83.1%, the performance for the non-credit worth (Class 1) is significantly lower. The precision, recall, and F1-score for the noncredit worth are quite low compared to (Class 0) which is credit worth percentage.

Class 0 has a high precision, recall, and F1-score, indicating good performance in predicting this class. **Class 1** has much lower precision, recall, and F1-score, indicating that the model struggles with this.

Fig 7: Showing the Hybrid Model Accuracy

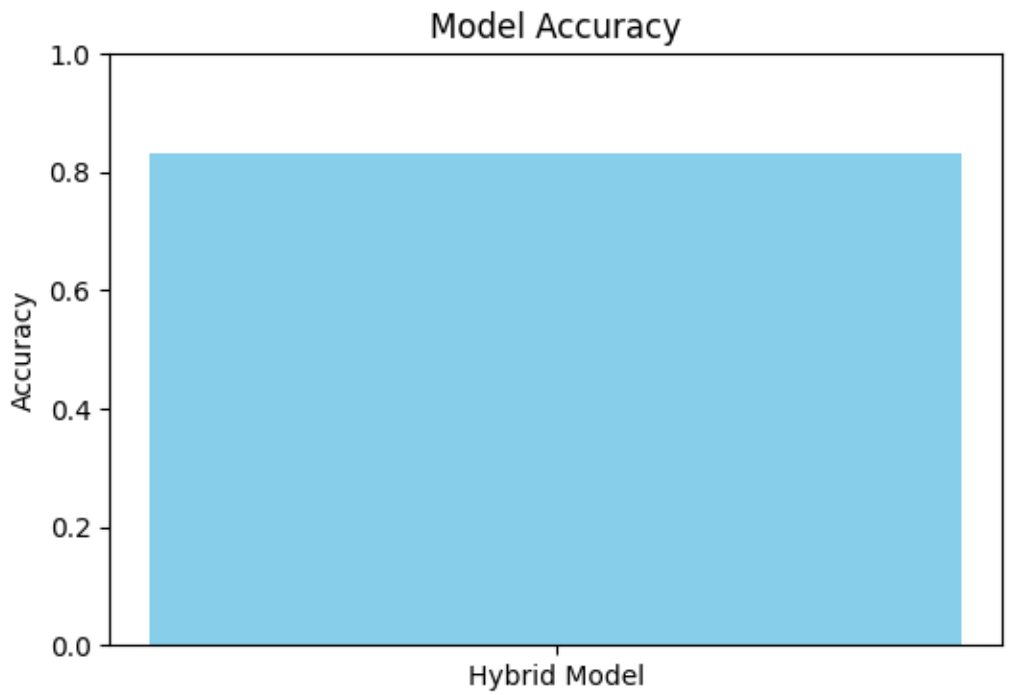
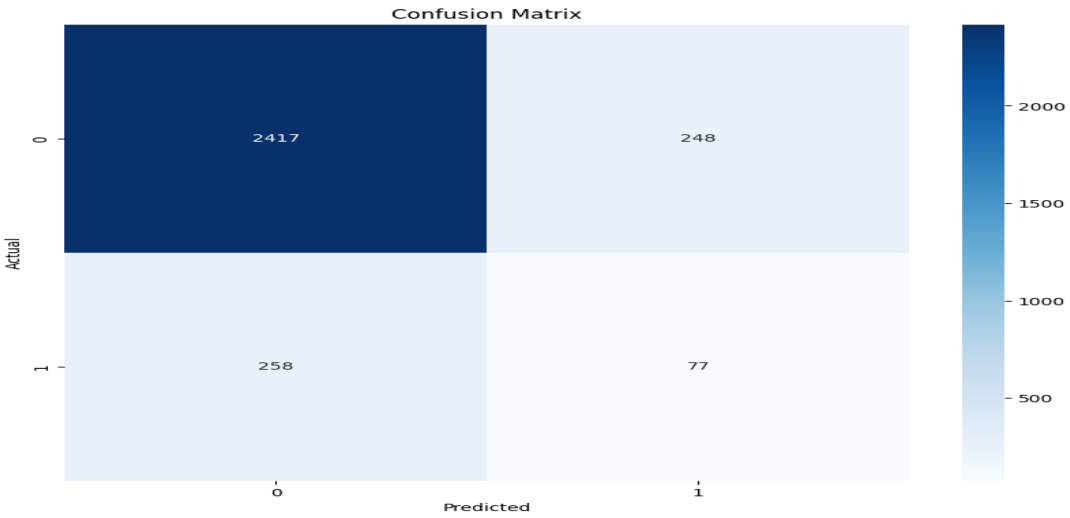


Fig 8: Showing the model confusion matrix



The confusion matrix above visualizes the performance of the classification of the hybrid on the test data and presets the statistics below.

True Positives (TP): The bottom-right cell (77) represents instances where the model correctly predicted the positive class (1).

True Negatives (TN): The top-left cell (2417) represents instances where the model correctly predicted the negative class (0).

False Positives (FP): The top-right cell (248) represents instances where the model incorrectly predicted the positive class (1) when the true class was negative (0). This is also known as a Type I error.

False Negatives (FN): The bottom-left cell (258) represents instances where the model incorrectly predicted the negative class (0) when the true class was positive (1). This is also known as a Type II error.

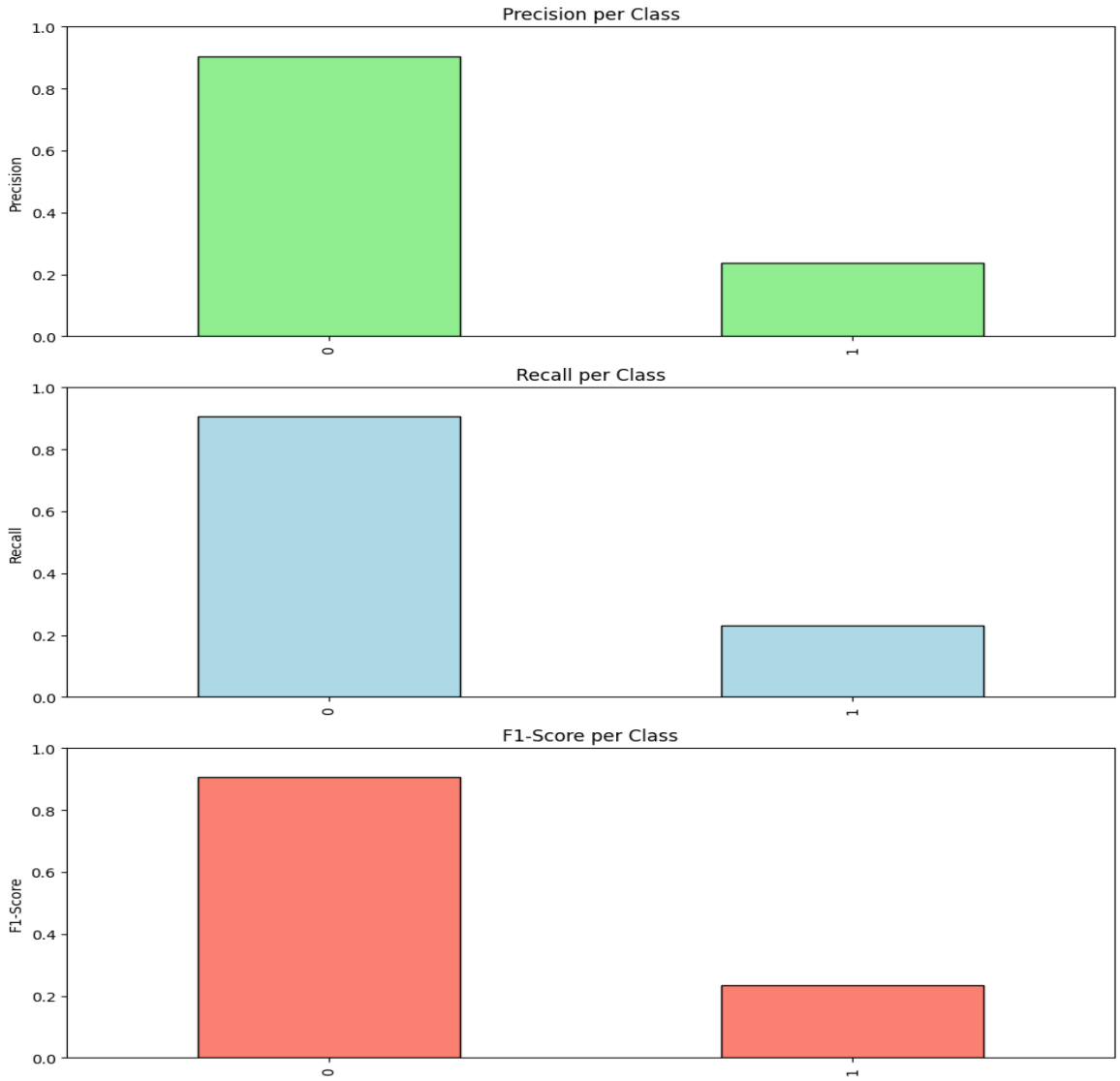
4.8.5 Interpretation

- The model correctly predicted 2417 instances of class 0 and 77 instances of class 1.
- It made 248 incorrect predictions where it predicted class 1 instead of class 0.
- It made 258 incorrect predictions where it predicted class 0 instead of class 1.

Overall Performance

- The model seems to perform better in predicting class 0 compared to class 1, as indicated by the higher number of true negatives (2417) compared to true positives (77).
- There is a significant number of false negatives (258), indicating the model struggles more with identifying the positive class correctly.

Fig 9: Showing the model Precision Per Class Of Borrowers



Precision is a metric that measures the accuracy of positive predictions made by a classification model. It is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP).

Mathematically:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

CHAPTER 5

Discussion

5.1 Insights Derived from the Model Comparison

5.1.1 Model Performance Metrics

Accuracy: This provides a broad measure of overall correctness. A high accuracy indicates that the model performs well across all classes. However, accuracy can be misleading in imbalanced datasets where one class dominates.

Precision, Recall, F1-Score: These metrics give more granular insights.

Precision: Reflects how many of the predicted positives are positive. Low precision indicates many false positives.

Recall: Shows how many of the actual positives are correctly identified. Low recall suggests many false negatives.

F1-Score: Balances precision and recall, useful when both false positives and false negatives are critical (Kevin P. Murphy 2012).

5.1.2 For the Logistic Regression model

Precision and recall for the minority class (LOAN_STATUS=1) are likely low, indicating that it struggles to identify positive loan statuses accurately, which can be crucial if predicting approved loans is important.

5.1.3 For the Decision Tree model

Decision Trees might capture complex patterns better, but they can also overfit. If the Decision Tree shows higher performance on the training set but not on the test set, it indicates overfitting.

5.1.4 For the Hybrid Model

Combining models often improves performance by leveraging their individual strengths. The soft voting mechanism averages predicted probabilities, potentially providing a more balanced approach to handling class imbalance.

5.1.5 ROC AUC Score

This score provides insight into how well the model distinguishes between classes. A higher AUC indicates a better ability to differentiate between loan statuses, making it a useful metric in imbalanced scenarios.

Logistic Regression and Decision Tree models might show varying AUC scores and combining them in the hybrid model might enhance the overall discriminative power.

5.1.6 Confusion Matrix

The confusion matrix visualizes the performance across different classes. Analyzing this can reveal specific issues like which class is misclassified more often and help in understanding where the model might need improvements.

5.2 Implications of the Results for Loan Prediction

5.2.1 Predictive Accuracy

If the hybrid model or one of the individual models demonstrates high accuracy, it indicates that the model can effectively predict loan statuses based on the provided features. This is crucial for automating loan approval processes and improving efficiency.

5.2.2 Handling Imbalanced Data

Imbalanced data often results in models that favor the majority class. The low precision and recall for the minority class in the initial models highlight the need for more sophisticated techniques or additional data features to better identify positive loan statuses.

5.2.3 Decision Making

Logistic Regression might be more straightforward and easier to interpret, providing clear probabilities of loan approval. This can be useful for understanding the likelihood of approval.

Decision Trees can offer insights into the decision rules used for loan approval, which can be helpful for understanding and explaining decisions.

The Hybrid Model combines the strengths of both, potentially offering a more balanced approach. This can improve loan approval systems by reducing false positives and false negatives.

5.2.4 Model Choice

Depending on the results, if the hybrid model performs well, it might be the preferred choice due to its ability to integrate different models' strengths. For businesses, this can lead to more accurate and reliable loan predictions, ultimately impacting the financial decision-making process.

5.3 Future Work

Feature Engineering: Additional features might improve model performance. For example, including historical loan data, borrower demographics, and economic indicators could enhance predictions.

Model Tuning: Further hyperparameter tuning and experimenting with different algorithms (e.g., Random Forest, Gradient Boosting) might yield better results.

Cross-Validation: Using cross-validation techniques can provide a more robust estimate of model performance and help prevent overfitting.

5.4 Concluding Thoughts on the Research

The research has provided valuable insights into the performance of different predictive models for loan approval, particularly focusing on Logistic Regression, Decision Trees, and a hybrid Voting Classifier. The analysis demonstrated the utility of combining models to leverage their individual strengths and improve overall prediction accuracy. (Max Kuhn and Kjell Johnson, 2013)

Effectiveness of the Hybrid Model: Combining Logistic Regression and Decision Tree models through a Voting Classifier yielded promising results, offering a balanced approach to handling class imbalance and improving predictive performance.

Model Strengths and Weaknesses: While Logistic Regression is straightforward and interpretable, it might struggle with class imbalance. Decision Trees can capture complex relationships but may overfit. The hybrid model benefits from the strengths of both approaches.

The implications for stakeholders in the financial sector are significant, including improved decision-making, fairer assessments, and enhanced operational efficiency. The recommendations for further research and model enhancement can help refine these models and adapt them to evolving data and business needs.

In conclusion, continuous model evaluation and improvement are essential for maintaining the accuracy, relevance, and ethical standards of predictive models. By actively engaging in these practices, organizations can ensure that their loan prediction systems are robust, up-to-date, and capable of delivering optimal results in a dynamic and competitive environment.

REFERENCES

Peter Bruce and Andrew Bruce (2017), "Practical Statistics for Data Scientists: 50 Essential Concepts"

Alam, M.S. and Alam, M.S. (2024) *A study of digital transformation in the microfinance landscape of Bangladesh*.

Ampountolas, A. *et al.* (2021) 'A machine learning approach for Micro-Credit scoring,' *Risks*, 9(3), p. 50.

Anderson, S.T. and Lin, K.K. (2024) 'Scientific method,' in *Elsevier eBooks*, pp. 13–15.

Berg, T. *et al.* (2019) 'On the Rise of FinTechs: Credit Scoring Using Digital Footprints,' *Review of Financial Studies/˜the œReview of Financial Studies*, 33(7), pp. 2845–2897.

Berndt, A.E. (2020) 'Sampling methods,' *Journal of Human Lactation*, 36(2), pp. 224–226.

Brocke, J.V. *et al.* (no date) *Special issue Editorial – Accumulation and Evolution of Design Knowledge in Design Science Research: A Journey Through Time and Space*.

Brocke, J.V., Hevner, A. and Maedche, A. (2020a) 'Introduction to Design Science Research,' in *Progress in IS*, pp. 1–13.

Brocke, J.V., Hevner, A. and Maedche, A. (2020b) 'Introduction to Design Science Research,' in *Progress in IS*, pp. 1–13.

Brocke, J.V., Hevner, A. and Maedche, A. (2020c) 'Introduction to Design Science Research,' in *Progress in IS*, pp. 1–13.

Bücker, M. *et al.* (2021) 'Transparency, auditability, and explainability of machine learning models in credit scoring,' *Journal of the Operational Research Society*, 73(1), pp. 70–90.

Djeundje, V.B. *et al.* (2021a) 'Enhancing credit scoring with alternative data,' *Expert Systems With Applications*, 163, p. 113766.

Djeundje, V.B. *et al.* (2021b) 'Enhancing credit scoring with alternative data,' *Expert Systems With Applications*, 163, p. 113766.

Djeundje, V.B. *et al.* (2021c) 'Enhancing credit scoring with alternative data,' *Expert Systems With Applications*, 163, p. 113766.

Dumitrescu, E.-I. *et al.* (2020) 'Machine learning or econometrics for credit scoring: Let's get the best of both worlds,' *Social Science Research Network* [Preprint].

Dumitrescu, E.I. *et al.* (2022a) 'Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,' *European Journal of Operational Research*, 297(3), pp. 1178–1192.

Dumitrescu, E.I. *et al.* (2022b) 'Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,' *European Journal of Operational Research*, 297(3), pp. 1178–1192.

Dumitrescu, E.I. *et al.* (2022c) 'Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,' *European Journal of Operational Research*, 297(3), pp. 1178–1192.

Dumitrescu, E.I. *et al.* (2022d) 'Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,' *European Journal of Operational Research*, 297(3), pp. 1178–1192.

Harley, B. and Cornelissen, J. (2020) 'Rigor with or without templates? The pursuit of methodological rigor in qualitative research,' *Organizational Research Methods*, 25(2), pp. 239–261..

Kumar, A., Sharma, S. and Mahdavi, M. (2021) 'Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature review,' *Risks*, 9(11), p. 192..

Kyngäs, H. (2019) 'Inductive Content analysis,' in *Springer eBooks*, pp. 13–21.

Liu, W., Fan, H. and Xia, M. (2021) 'Step-wise multi-grained augmented gradient boosting decision trees for credit scoring,' *Engineering Applications of Artificial Intelligence*, 97, p. 104036.

Liu, W., Fan, H. and Xia, M. (2022) 'Credit scoring based on tree-enhanced gradient boosting decision trees,' *Expert Systems With Applications*, 189, p. 116034.

Lobe, B., Morgan, D.L. and Hoffman, K.A. (2020) 'Qualitative data collection in an era of social distancing,' *International Journal of Qualitative Methods*, 19, p. 160940692093787.

Lukyanenko, R. and Parsons, J. (no date) *Research Perspectives: Design Theory Indeterminacy: What Is it, How Can it Be Reduced, and Why Did the Polar Bear Drown?*

Mulisa, F. (2021) 'When does a researcher choose a quantitative, qualitative, or mixed research approach?,' *Interchange*, 53(1), pp. 113–131.

Nathan, S., Ibrahim, M. and Tom, M. (2020) *Determinants of Non-Performing loans in Uganda's commercial banking sector*.

Nguyen, S.P. and Truong, N.Q. (2024a) 'An application of explainable artificial intelligence in credit scoring,' in *Studies in systems, decision and control*, pp. 317–333.

Nguyen, S.P. and Truong, N.Q. (2024b) 'An application of explainable artificial intelligence in credit scoring,' in *Studies in systems, decision and control*, pp. 317–333.

Nhu, V.-H. *et al.* (2020) 'Shallow Landslide Susceptibility Mapping: A Comparison between Logistic Model Tree, Logistic Regression, Naïve Bayes Tree, Artificial Neural Network, and Support Vector Machine Algorithms,' *International Journal of Environmental Research and Public Health/International Journal of Environmental Research and Public Health*, 17(8), p. 2749.

Shen, F., Wang, R. and Shen, Y. (2019) 'A COST-SENSITIVE LOGISTIC REGRESSION CREDIT SCORING MODEL BASED ON MULTI-OBJECTIVE OPTIMIZATION APPROACH,' *Technological and Economic Development of Economy*, 26(2), pp. 405–429..

Shilbayeh, S. and Grassa, R. (2024) 'Creditworthiness pattern prediction and detection for GCC Islamic banks using machine learning techniques,' *International Journal of Islamic and Middle Eastern Finance and Management*, 17(2), pp. 345–365.

M. Singh and G. M. Provan. Efficient learning of selective Bayesian network classifiers. In Machine Learning: Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann, 2020.

Tripathi, D. *et al.* (2021) 'Experimental analysis of machine learning methods for credit score classification,' *Progress in Artificial Intelligence*, 10(3), pp. 217–243.

Trivedi, S.K. (2020) 'A study on credit scoring modeling with different feature selection and machine learning approaches,' *Technology in Society*, 63, p. 101413.

Vassilakopoulou, P. and Hustad, E. (2021) 'Bridging Digital Divides: a Literature Review and Research Agenda for Information Systems Research,' *Information Systems Frontiers*, 25(3), pp. 955–969. 3.

Wang, Y. *et al.* (2020) 'A Comparative Assessment of Credit Risk Model Based on Machine Learning a case study of bank loan data,' *Procedia Computer Science*, 174, pp. 141–149.

Xia, Y., He, L., *et al.* (2020) 'A DYNAMIC CREDIT SCORING MODEL BASED ON SURVIVAL GRADIENT BOOSTING DECISION TREE APPROACH,' *Technological and Economic Development of Economy*, 27(1), pp. 96–119.

Xia, Y., Zhao, J., *et al.* (2020) 'A novel tree-based dynamic heterogeneous ensemble method for credit scoring,' *Expert Systems With Applications*, 159, p. 113615.

Christopher M. Bishop (2006)" Pattern Recognition and Machine Learning"

Aurélien Géron (2019) "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" (2nd Edition)

Kevin P. Murphy (2012) "Machine Learning: A Probabilistic Perspective".

Andreas C. Müller and Sarah Guido (2016) "Introduction to Machine Learning with Python: A Guide for Data Scientists"

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009) "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" (2nd Edition)

(Max Kuhn and Kjell Johnson, 2013) "Applied Predictive Modeling".

Foster Provost and Tom Fawcett (2013) "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking".

Andrew Ng (2018) "Machine Learning Yearning: Technical Strategy for AI Engineers, In the Era of Deep Learning".

Brad Boehmke and Brandon Greenwell 2020 "Hands-On Machine Learning with R: Build and Evaluate Models with the Most Effective Tools".

Lillian Pierson (2017) "Data Science for Dummies"

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning. O'Reilly Media.

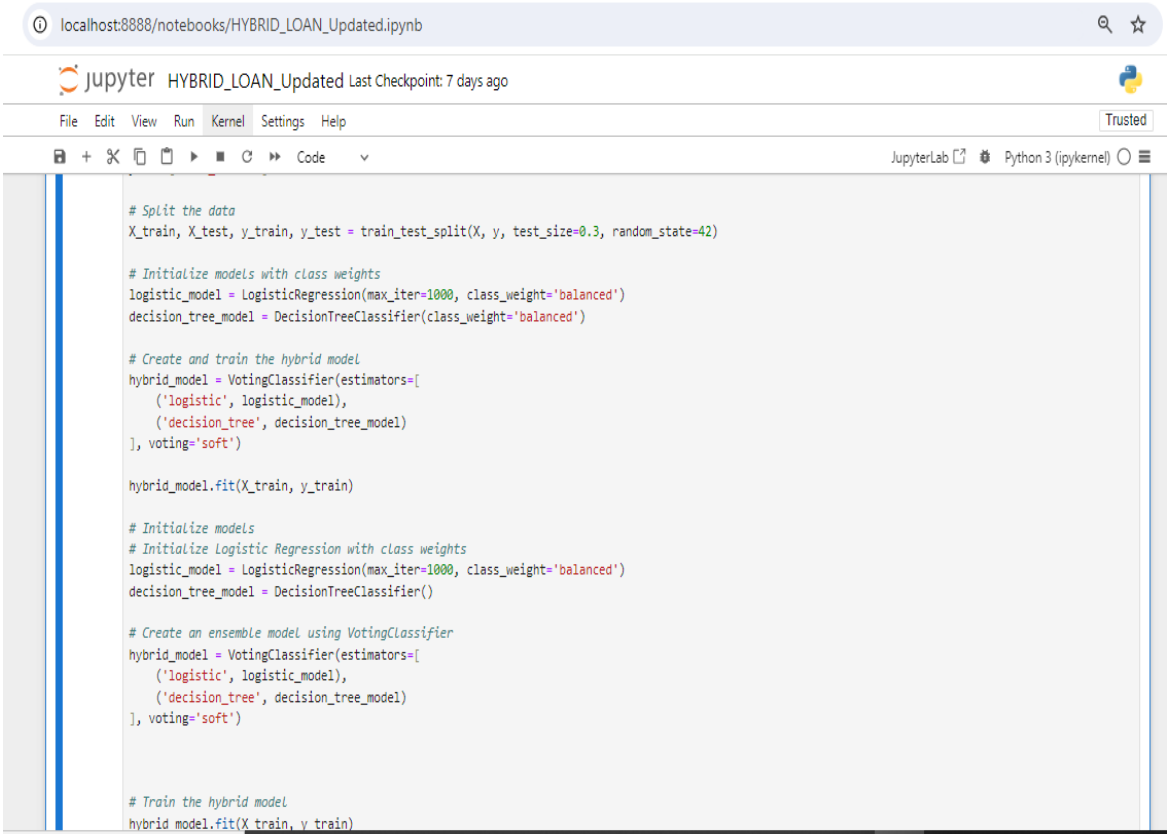
Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

1. *Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit Scoring and Its Applications. Society for Industrial and Applied Mathematics (SIAM).*
2. *Molnar, C. (2019). Interpretable Machine Learning. Christoph Molnar.*
3. *Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.*
4. *Jones, S. (2016). Advances in Credit Risk Modelling and Corporate Bankruptcy Prediction. Cambridge University Press.*
5. *Mays, E. (2001). Handbook of Credit Scoring. American Management Association.*

Appendix

Appendix I : Jupyter notebook Hybrid loan model Code lines



Appendix II : Jupiter notebook Python Logistic Regression model Code lines