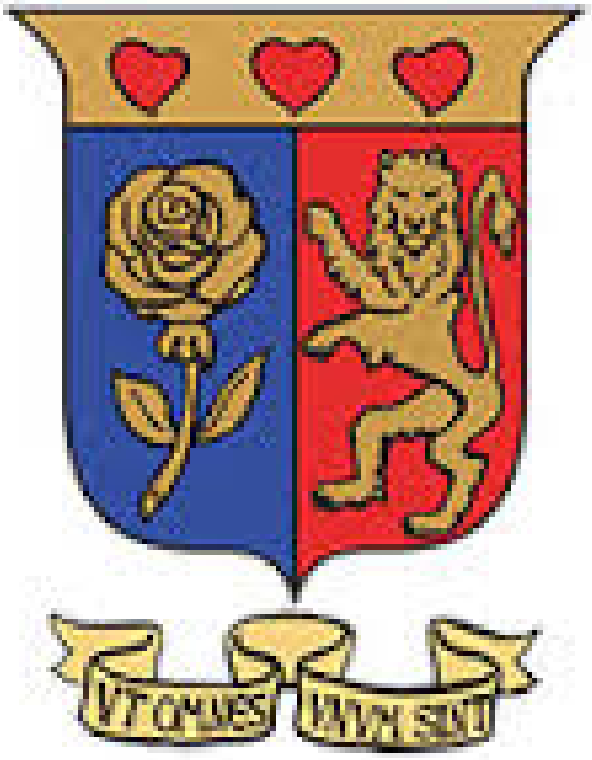


Strathmore
UNIVERSITY

Application of Machine Learning and Stochastic Control in Modeling Credit Default

Elizabeth Nduku Mutavi
Adm No. 147184

October 2024

Declaration and Recommendation

Declaration

I declare that this thesis is my original work and has not been submitted or approved for the award of a degree by Strathmore or any other University. To the best of my knowledge, the thesis contains no material previously published or written by another person except where due reference is made.

Elizabeth, Nduku Mutavi

Signature: 

Date 10th October 2024

Recommendation

This proposal has been submitted for assessment with our approval as supervisors according to Strathmore University regulations.

Name: Dr.Fred Mayambala

Signature 

Date 11th October 2024

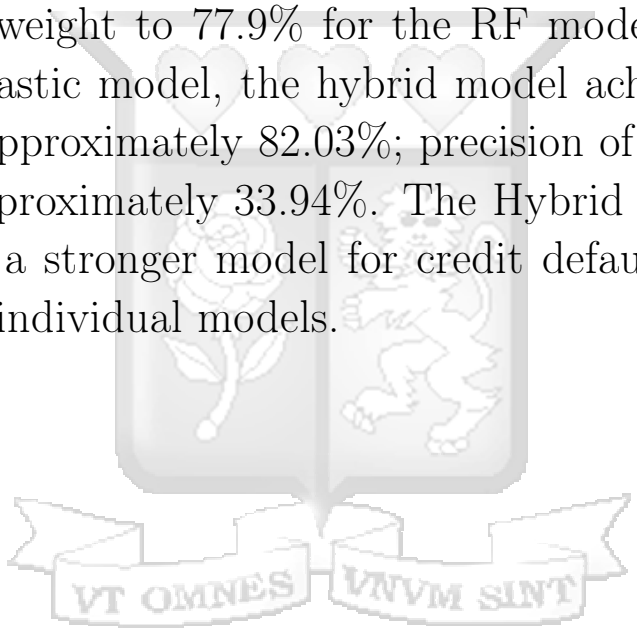
Name: Ferdinand Othieno

Signature 

Date 11th October 2024

Abstract

This study sought to develop a credit risk model that combines machine learning algorithms with stochastic control techniques to enhance the prediction of rare events, specifically credit defaults. The study was anchored on Stochastic control theory is a mathematical-based model for decision making in forecasting and probability. The study was underpinned by exploratory research design and used qualitative methods to gather and analyze data about the research topic. Data for the study was collected through searching secondary sources online. Data was reprocessed to remove missing values. The data was then split into test and train data sets for the machine learning model, the model was trained and then used to make predictions. The model was a combination of Machine learning and stochastic control models in a ration that was determined based on the option that led to optimal efficiency. The findings showed that combining the Random Forest and Monte Carlo models, demonstrates a balanced and effective approach to credit default prediction. By optimizing the weight to 77.9% for the RF model and 22.1% for the Stochastic model, the hybrid model achieved a high accuracy of approximately 82.03%; precision of 68.96%; and a recall of approximately 33.94%. The Hybrid model therefore presents a stronger model for credit default prediction compared to individual models.



Contents

List of Abbreviations

BSM Black-Scholes-Merton

ML Machine Learning

DGHLN Deep Genetic Hierarchical Learner Network

XGBOOST Extreme Gradient Boosting

ANN Artificial Neural Networks

SDE Stochastic Differential Equations

MDP Markov Decision Processes

HJB Hamilton-Jacobi-Bellman

LQG Linear-Quadratic-Gaussian

1 Introduction

1.1 Background

The financial world has gone through significant changes in recent years marked by the increasing complexity of financial instruments and the growing inter-connectedness of global markets,(Balling and Gnan 2013). Amidst this evolving financial ecosystem, one critical concern that has garnered significant attention is the modeling and prediction of default events (Adede 2012),specifically concerning the ability of borrowers to fulfill their financial obligations, particularly in repaying borrowed money. To address this critical issue, financial institutions have been actively investing in research to identify effective strategies for credit risk management (Musyoki and Kadubo 2012).Accurate modeling of default events is a fundamental component of risk assessment and mitigation strategies in the financial industry. This research is propelled by the growing recognition of the paramount role that credit risk plays in the stability and profitability of financial institutions.

Traditional methods for predicting defaults, such as credit scoring and statistical models, have long been relied upon. However, these methods often struggle to capture the dynamic and stochastic nature of financial markets(Sudhakar and Naganjaneyulu 2023), leading to inaccurate risk assessments and potential financial loss upon default. A credit score, often represented as numerical data, is derived from an individual's credit history (Aduda, Magutu, and Wangu 2012). It takes into account a range of factors, including financial assets, employment history, existing debts, and any past instances of defaults. Essentially, the credit score serves as a pivotal determinant of the likelihood that an individ-

ual will honor their loan commitments. In the backdrop of these developments, the financial landscape has been continually transformed by technological advancements. Traditional financial institutions have integrated technology into their business models, paving the way for network-based financial solutions. Internet finance products have emerged as a prominent facet of this technological integration (Lee 2009), encompassing innovative financial institutions that rely on the Internet for various financial operations, such as mobile financing and third-party payment platforms. However, these internet finance products are not without their challenges and associated risks.

Financial institutions offering internet-based financial products encounter a range of risks, including the possibility of substantial losses caused by borrowers failing to make timely payments, fraudulent activities by operators, and the potential collapse of peer-to-peer (P2P) platforms,(Mutembete 2022). These risks underscore the pressing need for establishing an effective finance risk model, one that can adeptly navigate the intricacies of credit risk within the digital realm. Traditionally, the finance industry has relied on historical credit data to compute credit ratings (Aduda, Magutu, and Wangu 2012). However, the traditional financial model is limited in its capacity to accurately process the complex historical data that characterizes Internet finance. In response to this challenge, this research paper endeavors to harness the potential of machine learning and stochastic models to predict and mitigate risk effectively. Machine learning can process large and intricate data sets, uncovering hidden patterns and connections that traditional methods might miss(Zhao, Li, and Yu 2017). In contrast, stochastic control furnishes a sturdy framework for decision-making in uncertain and ever

changing situations, making it well-suited for modeling rare events of default. The machine learning model proposed in this study will be rooted in the mining and collection of internet data, allowing it to adapt to the unique dynamics of the online financial landscape. These advancements in the field of machine learning and stochastic modeling hold the potential to play a trans formative role in enhancing the management of credit risk for Internet financial products. This research aims to contribute to the growing body of knowledge and practical applications in the domain of machine learning algorithms for financial risk management. It strives to navigate the complexities of the modern financial landscape and offer valuable insights that can influence the future of managing risks and maintaining financial stability.

1.2 Scoring Models in Use currently

The traditional credit scoring model has a substantial historical foundation, dating back to the 1950s (Fan, Shen, and Peng 2020). It introduced the concept of a credit scoring system, allowing financial institutions to base their credit decisions on a framework known as the "5C credit discrimination strategy(Aduda, Magutu, and Wangu 2012)." This strategy evaluates various factors, including capital, lender's Character, environmental Conditions, capacity, and collateral, to forecast the performance of borrowers. While effective for its time, this traditional model faces limitations in handling complex and dynamic data, particularly in the modern context of internet finance (Ordabayeva, Moldagulova, and Riza 2023).It however provides a foundational approach to assessing credit risk and are often used as a basis for more advanced credit risk modeling techniques, especially in the modeling of default events.

Machine Learning and Financial Control- With the advancement of technology, financial institutions have gained access to substantial amounts of data and sophisticated data mining techniques, especially over the Internet (Kshetri 2014). This progress gives rise to the application of machine learning in the domain of financial risk control. Machine learning algorithms have emerged as a promising tool for predicting borrowers' default risk (Mhlanga 2021), offering notable advantages in terms of reduced prediction errors when compared to the traditional model.

Machine learning models are strong in handling large data sets and uncovering non linear relationships. machine Learning models learn the data and are therefore more adaptable. This improves the prediction of defaults in a highly dynamic financial environment.(Zhao, Li, and Yu 2017; Mhlanga 2021). However one major weakness of machine learning models is over reliance of large balanced data sets and struggle with imbalanced data like rare events. They tend to under perform when it comes to accurately predicting such rare events, as seen in classification problems involving imbalanced data (Sudhakar and Naganjaneyulu 2023). Additionally, Machine learning models often act as black boxes, lacking interpret-ability, making it difficult for financial institutions to understand the reasoning behind specific predictions (Mhlanga 2021).

Stochastic models are grounded in probability theory and dynamic decision-making under uncertainty. They are better suited to handle rare events and capture the randomness in financial systems like credit default.(Pun 2018; Óskarsdóttir et al. 2019).The weaknesses of stochastic models lie in their need for complex mathematical formulations and assumptions, such as constant volatility or normal distribu-

tions, which may not hold in real-world data. They may also be less flexible in adapting to new patterns in data because they rely heavily on predefined assumptions about the system dynamics.(Ghevariya and Thakkar 2019)

The hybrid model, which combines the strengths of both machine learning and stochastic models, addresses these weaknesses. By combining Machine learning ability to handle large datasets and discover complex patterns with the stochastic model's strength in predicting rare events, the hybrid approach balances accuracy, precision, and recall(Ampountolas, Nyarko Nde, Constantinescu, et al. 2021).In particular, the stochastic model compensates for the machine learning model's inability to predict rare events effectively, while machine learning compensates for the stochastic model's rigidity by adapting to new data more easily. (Joseph 2013)

1.3 Problem Statement

Financial institutions rely heavily on credit scoring models to evaluate creditworthiness of borrowers (Aduda, Magutu, and Wangu 2012).However, this faces challenges due to the complexity of the data and also noting that credit defaults are rare events where there exists data imbalance. The changes in economic environment is a great challenge in modelling such complexities.

Previous researchers have explored the application of various machine learning algorithms in credit risk modeling and stochastic control models that have demonstrated an improvement in predictions compared to traditional models.(Mhlanga 2021; Mutembete 2022).

There exists a gap in literature where machine learning algorithms and stochastic control have been integrated to model credit default as a rare event.The primary problem

to address is how to enhance credit risk modeling to better predict and mitigate risks in the context of Internet finance, given that default is a rare event. The study will evaluate whether machine learning and stochastic models can significantly improve the accuracy and adaptability of credit risk assessment in this evolving landscape when applied together.

By doing so, it aims to bridge the existing gap in the traditional credit risk assessment process and contribute to the development of more accurate, adaptable, and efficient credit risk management strategies.

1.4 Research Objective

1.4.1 Main Objective

The primary objectives of this research is to develop a credit risk model that combines machine learning algorithms with stochastic control techniques to enhance the prediction of rare events, specifically credit defaults.

1.4.2 Specific objective

- The model the probability of default using machine learning.
- The model probability of default using stochastic control model.
- To model probability of default using the hybrid model.

1.5 Research Questions

- Can machine learning algorithms effectively predict rare events, such as loan defaults, when default cases are significantly lower in number compared to non-default cases in credit risk modeling?

- Is a combination of Machine learning algorithm and Stochastic control model better in predicting credit default?

1.6 Significance of Study

This research aims to provide insights into improving credit risk modeling. The research will contribute to how machine learning and stochastic control methods can improve the accuracy and adaptability of credit risk models. Financial institutions can use this paper to develop frameworks that can optimize credit risk mitigation strategies. Additionally, policymakers and financial institutions can use the paper to guide the practical implementation of machine learning and stochastic control techniques in credit risk modeling.

1.7 Scope and Limitations of the Study

1.7.1 Scope

The study will examine the application of machine learning and stochastic control in credit risk modeling. The study will examine the effectiveness of integrating machine learning techniques with traditional credit risk models to improve credit risk prediction and management accuracy and efficiency. Additionally, the research explores the potential implications and benefits of combining machine learning and stochastic control for optimizing credit risk mitigation strategies. The study primarily utilizes secondary data to develop and assess credit risk models. The findings may help contribute to the field of credit risk modeling and offer practical insights for financial institutions and policymakers.

1.7.2 Limitations

The research may fail to address all aspects of credit risk management as the scope is limited to the integration of ma-

chine learning and stochastic control in credit risk modeling specifically default as a rare event. Secondly, the research relies on secondary data. Therefore, the quality of research findings will depend on the quality of secondary data. Additionally, some secondary data regarding the research scope may require additional permissions from authors and publishers

2 Literature Review

2.1 Credit Risk Modelling Overview

Research regarding finance risk modeling has gained a lot of attention lately. Financial institutions are faced with credit risk regarding the ability of borrowers to meet their obligation of paying back borrowed money. Financial institutions fund research to help identify effective strategies for credit risk management. (Mutembete 2022) state that many financial institutions use credit scores to evaluate a borrower's creditworthiness. The credit score is numerical data based on the borrower's credit history. The credit score numeric data is a representation of financial assets, employment history, existing debt, and default history of a borrower. The credit score is used to determine the probability of an individual to repay their loan.Various articles highlighted how the advancement of technology has enabled the integration of tech into business models. The traditional finance industry is integrated with technology, enabling network-based financial solutions. Internet finance products include e-commerce innovation-based financial institutions that use the Internet for financial operations, such as mobile financing and payments via third-party platforms. However, internet financing faces various challenges and risks. Financial institutions with Internet finance products may be subjected to huge losses

based on overdue payments by borrowers, operator fraud, and the collapse of Peer to peer(P2P) platforms (Ghevariya and Thakkar 2019). The credit risk leads to the urgency of establishing an effective finance risk model. The credit score used by the traditional finance industry relies on large historical credit data to obtain the relevant credit rating (Ghevariya and Thakkar 2019). Additionally, the traditional financial model is limited to processing complex historical data accurately.

2.2 Traditional Approaches to credit modelling

The traditional scoring model revolutionized the credit assessment process for financial institutions(Fan, Shen, and Peng 2020).The model was established in the 1950s and formed the foundation of credit scoring systems used by lenders to make informed credit decisions. The system operates on the principles of the 5C credit discrimination strategy, which assesses borrowers based on capital, lender's role, environment, capacity, and collateral to forecast their performance and creditworthiness (Fan, Shen, and Peng 2020) Financial institutions have used the credit scoring model to analyze quantifiable characteristics of borrowers, especially historical credit data. This data is crucial in calculating a credit score, which serves as a numerical representation of an individual's creditworthiness. The credit score plays a key role in determining whether a loan or credit will be granted to the borrower. However, despite being an advanced tool in its time, the traditional scoring model has its limitations. The old-fashioned model struggles to effectively handle complex data, impacting its accuracy and efficiency in evaluating credit risk in today's sophisticated financial landscape. As a result, financial institutions increasingly recognize the need

for advancements and improvements in credit scoring models to better adapt to the evolving and intricate nature of modern financial data and risk assessment.

2.3 Machine Learning techniques for modeling credit Risk

Machine learning-based systems are increasingly popular across various research disciplines. In this field, significant insights into decision-making have been derived from data, particularly through decision tree-based ensemble techniques that are effective for supervised classification problems. This section reviews the existing literature on the machine learning techniques used for credit scoring evaluation. While these techniques are widely recognized, the literature on their application offers valuable insights for a deeper understanding of the subject.

The advancement of technology based on big data and data mining strategies has significantly influenced research in machine learning (ML) associated with credit risk prediction and evaluation globally. The primary objective in credit management for financial institutions is to minimize risks as well as optimize business performance. There is a need to have credit risk models based on effective guidelines to help financial institutions make proper credit-related decisions.

Credit scoring systems relied widely on clustering algorithms in the past. Huang and William developed techniques for insurance risk identification by merging the K-means clustering technique with supervision methods. Additionally, (Yeo et al. 2001) utilized predicted risks in the automobile industry via hierarchical clustering. (Yeo et al. 2001).model helped identify different customer risk levels ideal for financial institutions to make operational decisions regarding credit

limits.

Currently, the availability of data at a large scale has enabled the development of complex models regarding credit risk management. For instance, the vast available data based on customer transactions and credit management agencies was used by researchers Khandhani and Kim to model a credit risk management strategy by employing statistical and ML algorithms (Khandani, Kim, and Lo 2010). The model was able to predict borrower default risk. The research concluded that the ML-based model was better than traditional models based on linear regression methods as it reduced prediction errors by 6%. (Fan, Shen, and Peng 2020). Additionally, research by Joseph and Chakraborty concluded that models based on ML outperformed models based on traditional logistic regression methods. Their research was based on an ML model trained to predict financial distress, improving about 10% when compared with statistical models regarding discrimination based on operational characteristics. Ticknor furthered the research via a model based on neural network algorithms used to predict the behavior of the financial market. The model improved accuracy prediction without relying on data processing.

An ML model by Agrapetidou and Gogas was based on vector machines. The model made predictions by analyzing data based on financial statements retrieved from banks. The vector machine-based model achieved a prediction accuracy of about 99.2% in relation to bankruptcies of financial institutions in the US between 2007 and 2013 (Fan, Shen, and Peng 2020). Research by Enneya and Rtayli utilized the vector machine feature to model an algorithm based on a roster classifier to predict fraud-related risks. The model improved the identification of credit card risks, outperforming other

algorithms based on big data. (Pławiak et al. 2020) modeled an algorithm known as (DGHLN), a deep genetic hierarchical learner network. The DGHLN algorithm proved to be an effective learner-training method based on genetic hierarchical training principles. The algorithm used cross-validation to predict 21% of Germany's credit rate (Fan, Shen, and Peng 2020).

The model by (Khandani, Kim, and Lo 2010) relied on ML algorithms that were able to reduce losses by 19% as it was able to increase and improve prediction accuracy regarding defaults of credit card holders (Ampountolas, Nyarko Nde, Constantinescu, et al. 2021). The Machine Learning model was trained via a dataset based on bank customers' transactions and usage of credit. The study provided a foundation to examine whether systematic risk prediction may be improved via credit risk analytics.

A Machine Learning model by (Yap, Ong, and Husain 2011) utilized a dataset based on payment data club members of a recreational club. The model predicted the credit score of defaulters regarding the club subscription. In contrast to many ML models highlighted in this literature, (Yap, Ong, and Husain 2011) study concluded the old-fashioned credit scorecard model performed with an accuracy almost similar to that of ML and other models. (Zhao, Li, and Yu 2017) Machine learning model relied on a credit dataset from Germany. The model was trained, and its accuracy concerning MLP, multilayer perceptron, and neural network was examined. The results highlighted an improvement in the accuracy of the MLP model compared to statistical and regression models. A machine learning model by (Addo, Guegan, and Hassani 2018) relied on machine deep learning methods. The model was based on Binary Classifiers to predict

the probability of loan defaulters. The results showed that the neural network model was outperformed by old-fashioned credit models such as the random forest, logistic regression, and gradient boosting. A dataset retrieved from Greek loan data was examined by (Petropoulos et al. 2019). The authors utilized Machine Learning models based on classifications to make predictions for about 10 years (Ampountolas, Nyarko Nde, Constantinescu, et al. 2021). The results demonstrated the Machine Learning model had better accuracy in its predictions than the traditional models based on logistic regression methods. The Machine Learning model was able to make better predictions via the financial credit ratings. (Provenzano et al. 2020) Machine Learning model utilized a dataset based on bankruptcies, balance sheets, and macroeconomic variables. The results showed the Machine Learning models outperformed traditional models in areas of credit rating as well as reducing bankruptcy probability. These previous studies on the application of Machine Learning in credit risk modeling demonstrate the evolving landscape of Machine Learning. Most of the studies highlighted in this section demonstrated the potential of Machine Learning models to improve prediction accuracy. The various Machine learning algorithms identified help provide valuable insights for credit risk assessment. The Machine Learning algorithms are further highlighted below.

2.4 Machine Learning Algorithms

This section highlights the various machine Learning algorithms that outperformed traditional models.

2.4.1 AdaBoost

AdaBoost, an acronym for Adaptive boosting, demonstrated improved accuracy based on the enhanced performance of decision tree-based classifiers. (Freund and Schapire 1997) used the algorithm to train an Machine Learning model via the merging of weak classifiers to create an improved classifier. The improvement attributes made the boosting algorithm popular among researchers from diverse disciplines (Schapire 2013).

2.4.2 XGBoost

The extreme gradient boosting (XGBoost) proposed by Chen and Guestrin uses a scalable tree-boosting approach (Chen and Guestrin 2016). The tree-boosting approach employs Machine Learning methods to enhance the outcome of a prediction regarding credit risk ratings. The Machine Learning model is trained via historical data, and the gradient boosting method is applied to improve the outcomes of the model's predictions.

2.4.3 Decision Tree Classifier

(Panigrahi et al. 2018) describe the decision tree classifier as an algorithm with a tree flow that resembles a progress diagram. The algorithm's tree structure is based on a training set of objects associated with one of the various classes. The dataset used to train the model helps in decision-making based on the generated decision tree (Panigrahi et al. 2018).

2.4.4 Extra Tree Classifier

The Extra Trees Classifier is an Machine Learning method based on extremely randomized trees. The algorithm utilizes a top-down approach, relying on a collection of decision

trees. The Extra Trees Classifier shares similarities with the random forest classifier. However, the extra trees classifier constructs the decision trees differently. Each decision tree is built from the initial training dataset. The decision tree structure is based on a random selection of elements and cut-points when dividing nodes. The randomness feature in cut-point selection sets it apart from other tree-based algorithms. The algorithm utilizes the entire training dataset rather than bootstrap replicas that fail to prevent biased decisions (Ampomah, Qin, and Nyame 2020). Each decision selects the optimal feature to split the data according to specific criteria. The algorithm's key parameters include the total number of decision trees, the number of randomly chosen features at each node, and the minimum number of instances needed to split a node.

2.4.5 Random Forest Classifier

(Geurts, Ernst, and Wehenkel 2006) describe the random tree classifier as a tree-based algorithm with multiple classifiers with tree structure. The multiple tree-structured classifiers increase the accuracy of the algorithm by decreasing overfitting. Therefore, the accuracy of the random tree classifier depends on the strength of the multiple tree-structured classifiers.For this research, Random forest classifier will be used as the machine learning model to predict credit default.

2.4.6 K-Nearest Neighbors Classifier,K-NN

K-nearest neighbors, bases its algorithm on the notion that similar things exist nearby (Abu Alfeilat et al. 2019). The algorithm is trained by a large dataset with data points characterized by various variables. Distance metric is used to determine the similarity, and efficiency depends on the used

distance metric.

2.4.7 Neural Network

Artificial Neural Networks,ANNs, are significant non-linear models. The algorithms are effectively used in forecasting as well as demand prediction across multiple disciplines, such as finance and electricity. ANN models' empirical evidence demonstrates good forecasting performance. The advancement in Machine Learning has contributed to the development of different ANN models, especially in forecasting scenarios. ANN models estimate edge weights to enhance forecasting accuracy (Dornaika et al. 2017).

2.5 Stochastic Control Application in Credit Risk modelling

Stochastic control theory is a mathematical-based model for decision-making in forecasting and probability. Regarding credit risk modeling, it provides a mathematical framework to improve the accuracy of credit-based decisions, reducing the implications of credit risks. (Pun 2018) states credit risk is impacted by various factors such as market trends, economic conditions, and borrower behavior that are subject to changes at any given time. (Óskarsdóttir et al. 2019) argues stochastic control addresses the uncertainty caused by these factors by employing strategies with randomness methods to credit risk models. The stochastic control-based models have a more realistic approach to the dynamic credit environment. Stochastic control models help financial institutions in risk management via optimal credit-related decisions. Stochastic control theory facilitates dynamic portfolio management as well. Application of the theory to credit risk modeling help financial institutions provide credit to different borrow-

ers based on changing risk preferences and market condition. Integrating machine learning with stochastic control improves credit risk modeling via the usage of large datasets to identify patterns regarding credit behavior. The identified patterns and trends allow the real-time adaptability of credit risk models. Additionally, stochastic control supports the development of risk mitigation strategies. Stochastic control helps financial institutions maximize returns via the identification of borrowers with good credit ratings and minimizing potential losses by identifying potential defaulters. The integration of stochastic control and machine learning provides improvement opportunities in credit risk management (Dornaika et al. 2017). Below are some of the stochastic models that have been previously used to predict credit default.

2.5.1 Hamilton-Jacobi-Bellman (HJB) Equation

HJB is a partial differential equation used to model optimal policy in stochastic control for continuous time.Studies have demonstrated that HJB-based models can significantly improve the prediction of defaults in volatile markets by optimizing decisions dynamically rather than relying on fixed strategies(Oksendal 2013)

2.5.2 Linear-Quadratic-Gaussian (LQG) Control

Linear-Quadratic-Gaussian (control is used when system dynamics is linear and noise is Gaussian.Commonly applied in credit risk models to predict borrower behavior under the assumption that economic shocks affecting creditworthiness follow a Gaussian distribution.Studies have shown that LQG control models improve the accuracy of default predictions compared to basic models, especially in situations where market conditions can be approximated by Gaussian

processes(Anderson and Moore 2007)

2.5.3 Stochastic Differential Equations (SDE)

SDE are used to model the evolution of variables over time with both deterministic and random components. In credit risk modeling, SDEs are employed to model the dynamic behavior of credit markets and borrower creditworthiness, which are affected by random shocks like changes in interest rates or economic crises.(Karatzas and Shreve 2014)showed that SDE-based models are more effective at capturing the complex interplay between borrower behavior and macroeconomic factors compared to static models, leading to improved risk management strategies.

2.5.4 Markov Decision Processes (MDP)

MDP models guide decision-making over discrete time periods. Each decision influences future states, making MDPs ideal for modeling credit risk, where loan repayment behaviors evolve over time. MDPs can evaluate the probability of a borrower defaulting based on their current state and provide optimal strategies for granting loans(Puterman 2014)demonstrated that MDPs outperform static credit models by providing more adaptive lending strategies that reduce the likelihood of defaults and increase profitability over time.

2.5.5 Monte Carlo Model

Widely used to predict risk in uncertain environments.It estimates the likelihood of defaults by simulating various future states of economic variables and borrower behaviors.It is powerful for dealing with rare events, such as credit defaults, because they can explore a wide variety of potential outcomes based on different assumptions about risk factors like

interest rates, economic downturns, or changes in borrower creditworthiness.Research by (Glasserman 2004) showed that Monte Carlo simulations offer greater accuracy in estimating default probabilities, especially in portfolios with complex risk factors and correlated defaults.This research will adopt Monte Carlo as the preferred stochastic model that will be optimised together with Random forest model to come up with an optimal hybrid model for credit default prediction. The below equations will be run in R code.

$$PD = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(L_i > I_i) \tag{1}$$

(1)

where: PD is the probability of default,

N is the number of simulations for this research, N=1000

$$1(L_i > I_i) \tag{2}$$

is an indicator function that equals 1 if the borrower defaults in the i-th simulation, and 0 otherwise

2.6 Benefits of Credit Risk Models

Various articles highlighted credit risk models as essential risk assessment and management tools. (Wang 2009) argued credit risk models help financial institutions manage credit risks via structured frameworks. The frameworks improve the ability of financial institutions to identify, measure, and effectively manage risks. (Kollár and Gondžárová 2015) stated credit risk models can provide a more accurate prediction of credit risk. The models consider a borrower's portfolio mainly based on historical data. The model-based methods provide a better reflection of concentration risk compared. (Joseph 2013) highlighted that credit models can adapt and

respond to changes. The adaptability feature of the models makes them more responsive tools for risk management. The implementation of credit risk models makes financial institutions improve their systems based on the need to increase data collection efforts. Financial institutions that use credit risk model make more accurate risk assessment and adopts better pricing strategies. An increase in accuracy leads to improved transparency regarding the decision-making process. (Saeed and Zahid 2016) stated models based on machine learning and stochastic control demonstrate flexibility in adapting to changes. The adaptability promotes a more stable and resilient financial system.

3 Methodology

According to (Silverman and Patterson 2021), research methodology refers to the way researchers collect data to find new facts and solutions to existing research problems. This section will highlight all the activities and steps the researcher will do to collect and analyze data into useful information. Research methodology is very important for most scientific studies and helps us tell the difference between facts and opinions (Taylor, Bogdan, and DeVault 2015).

3.1 Research Design and Method

(Hammersley 2017) stated the three main ways researchers use for scientific investigations: qualitative, quantitative, and mixed methods. This study plans to use qualitative methods to gather and analyze data about the research topic. The researcher will use secondary data based on historical credit information, economic indicators, or financial market data, which government agencies, financial institutions, or other research groups have already collected and shared with the

public to answer the research questions. (Lewis 2015) argued that the research plan should make it easy to collect, process, and analyze data to answer the research questions and address the research problem. The researcher will use an exploratory research methodology to define how to collect the data.

3.2 Research Setting

(Creswell 2014) describes the research setting as where the data for a particular study are collected. However, with improved internet access in the modern world, the research setting could be online, allowing researchers to get results from anywhere in the world with internet connectivity. This study's research setting will be online, where peer-reviewed articles will be searched and selected.

For this research, the data will be run on R for the Random Forest model for the Machine learning, the monte carlo model for the stochastic model and the hybrid model that combines the two models with optimal weight. The analysis is conducted in a systematic way to ensure the models not only predict but also generalize well on unseen data. Below is an in-depth look at how these models will work during the analysis.

3.3 Random Forest Model

The Random Forest algorithm operates by constructing a multiple decision trees during the training process. Each tree is built using a subset of the training data, created through bootstrap aggregation where random samples of the data are selected, with replacement, to train individual trees. For each tree:

1. Random Feature Selection: At each node, a random sub-

set of features is chosen, and the best feature from this subset is used to split the data. This ensures that the trees are diverse and reduces the chances of over fitting.

2. Decision Tree Growth: The trees are grown to their maximum depth, meaning they learn the training data very well, which is then averaged out in the forest to generalize predictions.

3. Voting Mechanism: Each tree in the forest votes for a class (default or no default), and the final prediction is made based on the majority vote. This ensemble technique reduces variance and increases the model's robustness, especially in predicting the majority class.

The model is trained on the training dataset, and performance is evaluated using the test dataset. Various metrics such as accuracy, precision, and recall are calculated to assess how well the model identifies both defaults and non-defaults.

3.4 Monte Carlo Model

The Monte Carlo model focuses on simulating the probability distribution of potential outcomes by running repeated random sampling with a goal to model the uncertainty and randomness in borrowers behavior.

1. Data Sampling: The Monte Carlo model generates numerous simulated data points based on the known statistical properties of the dataset, such as the probability distributions of key variables used as income, loan amount, past defaults.

2. Scenario Simulation: Each simulation involves the creation of a "future state" for each borrower, which accounts for variations in economic conditions (interest rates, macroeconomic shocks) and borrower behavior. For each

simulation, the likelihood of a borrower defaulting is computed based on these inputs.

3. Aggregation of Results: The final probability of default is determined by aggregating the outcomes from all the simulations.

Both models are then integrated into the analysis by balancing the Random Forest's strength in handling structured datasets with the Monte Carlo's power in simulating rare events, making the hybrid model a robust tool for predicting credit defaults.

This section will be divided into the research design, research method, and analysis section (Taylor, Bogdan, and DeVault 2015).

3.5 Data Collection

Data for the study was collected through searching secondary sources online. The main terms used to search for secondary sources included machine learning, the Black scores calculus model, pricing prediction models, stochastic control, and credit risk modeling. The study used secondary sources to gather current knowledge on using machine learning and stochastic control in credit risk modeling. The search engines that were used to gather the data included Google Scholar.

3.6 Data Analysis

The goal of this research was to develop a credit risk model that combines machine learning algorithms with stochastic control techniques to enhance the prediction credit default as a rare event. The study used the selected secondary data to explore the current state of knowledge in using machine learning and stochastic control in credit risk modeling and analyzed their effectiveness in predicting credit risk. The R

software was used for this analysis. The researcher explored the data to gain insights on the structure and characteristics. After identifying the credit default distribution and the rarity of the credit defaults, data was reprocessed to remove missing values. The data was then split into test and train data sets for the machine learning model, The model was trained and then used to make predictions. The model was a combination of Machine learning and stochastic control models in a ration that was determined based on the option that led to optimal efficiency. Below in the Hybrid equation was used:

Below in the Hybrid equation that was used:

$$P_{\mathrm{h}}X = \omega \cdot P_{\mathrm{ML}}X + (1-\omega) \cdot P_{\mathrm{St}}X] \tag{3}$$

where;

The probability of default predicted by the model that combines Machine Learning and Stochastic models with optimized weights is given by $P_{\mathrm{h}}(X)$.

$P_{\mathrm{ML}}X$: probability of default predicted by the machine learning model.

$P_{\mathrm{St}}X$: Probability of default predicted by the stochastic model.

ω : contribution of machine learning into the model where $0 \leq \omega \leq 1$

$(1-\omega)$: is the contribution of the stochastic in the model

The value of $(\omega)$was established through optimization.

4 Research Findings

4.1 Introduction

This Chapter discusses the results obtained from the developed model of loan default prediction. The results are analyzed with respect to the research objectives: to determine probability of default using machine learning; to determine probability of default using stochastic model; and to model probability of default using the hybrid model.In this session, the study analyzed the Credit Card Clients dataset from the UCI Machine Learning Repository to predict whether a person will default on their online credit, using Random Forest and Monte Carlo models.The dataset consists of 24 variables and 30,000 individual instances. Key variables include: Default Payment Status (0 = No, 1 = Yes), Credit Amount, Gender (1 = Male, 2 = Female), Education (1 = Graduate School, 2 = University, 3 = High School, 4 = Other, 5 = Unknown), Marital Status (1 = Married, 2 = Single, 3 = Others), Age, Payment History from April 2005 to September 2005 (where -2 = no consumption, -1 = paid on time, 0 = revolving credit used, 1-9 = months of delayed payment), Bill Statement Amounts from April to September 2005, and Previous Payments from April to September 2005. There is a separate variable for each past payment, bill statement, and previous payment from April to September. The the process: loading required packages; model establishment for stochastic control; model optimization; data analysis: evaluation; interpretation of results as well as conclusion is as shared in the appendix. Additionally, the results revealed only 11 duplicate entries in the dataset, so the data frame was filtered to remove these duplicates. The output showed that 4 variables were converted from numerical to names: Sex (Male,

Female); Education (Graduate School, College, High school, others); Marriage (Married, Single, Unknown, Others); Default (1=Yes, 0=No).

4.2 Data Distribution

Next, bar plots and distribution tables were generated to examine the proportion of the variables. This was done to assess whether the data followed a normal distribution. If the data was not normally distributed, it was useful to observe the direction and extent of skewness.

4.2.1 Figure 1: Gender

Based on the output, it is clear that Female made up more than half (60.4%) of the sample data while Male counterparts were only 39.6%. This suggest that Female were more involved in online credit than men. The following output presented the frequency distribution of customers by education levels

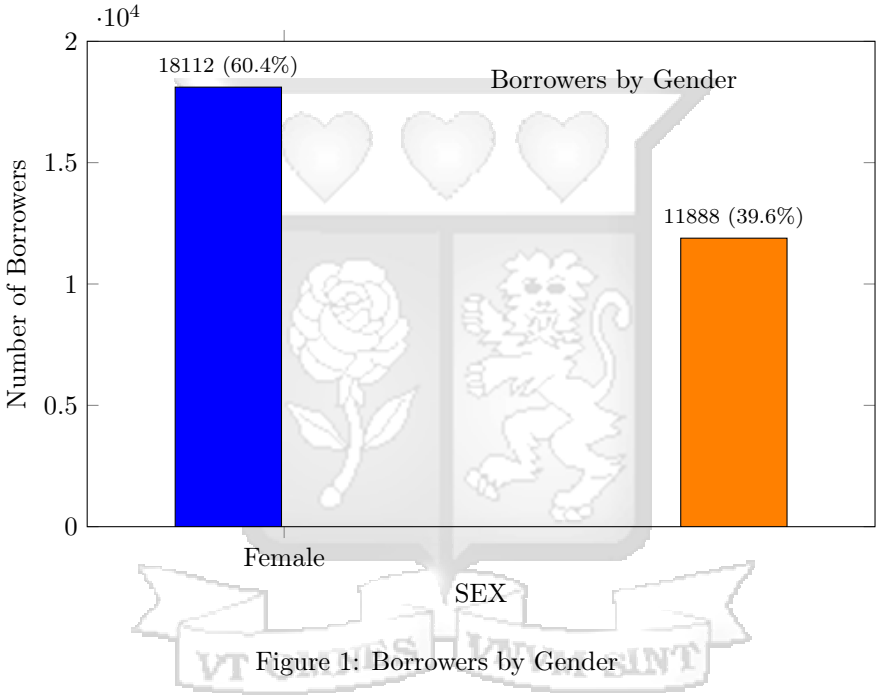


Figure 1: Borrowers by Gender

4.2.2 Figure 2: Education Level

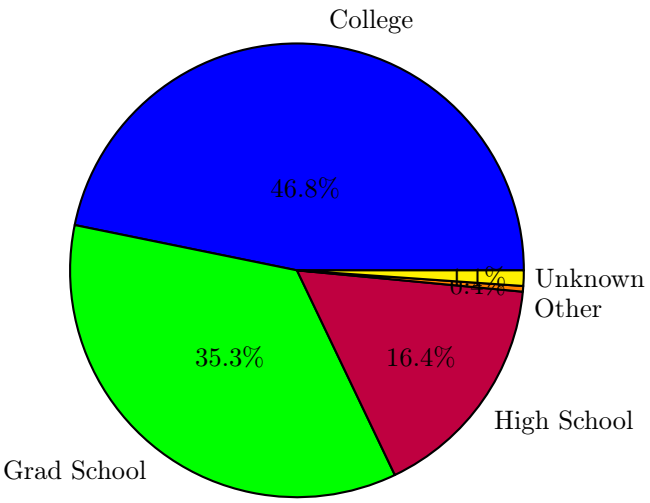


Figure 2: Distribution of Education Levels

The results showed that majority of the customers in the sample data were College graduates (46.8%) followed by Graduate school (35.3%) while high school came third at 16.4%, 1.1% were unknown while the least (0.4%) other categories were different from the defined ones.

4.2.3 Figure 3: Marital Status

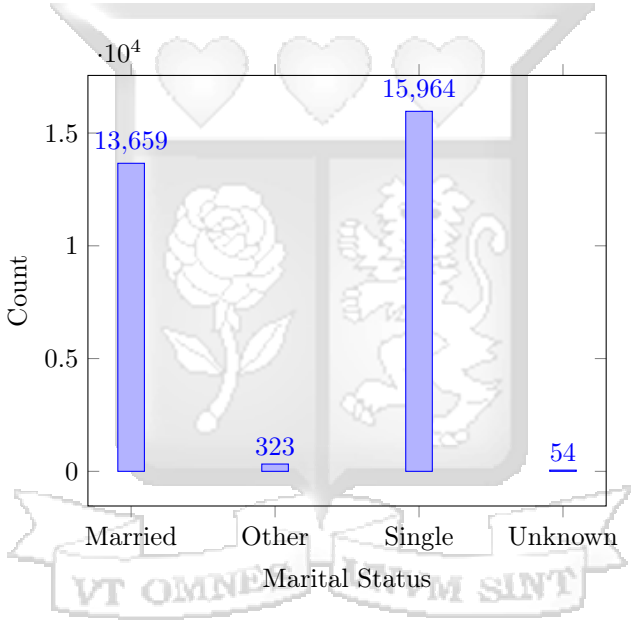


Figure 3: Distribution by Marital Status

The results in Figure 3 shows that majority (53.2%) of the customers in the sample data were single followed by married couples (45.5%) while others at 1.1%. Only 0.2% were unknown.

4.2.4 Figure 4: Default Status

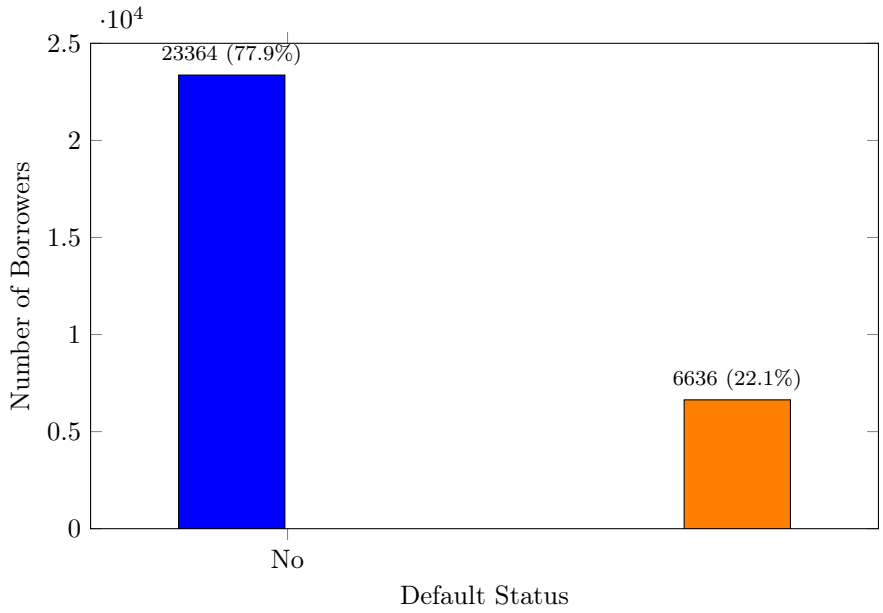


Figure 4: Borrowers by Default Status

The output showed that majority of the customers (77.9%) in the sampled data did not default while 22.1% defaulted.

4.2.5 Default Status by Gender

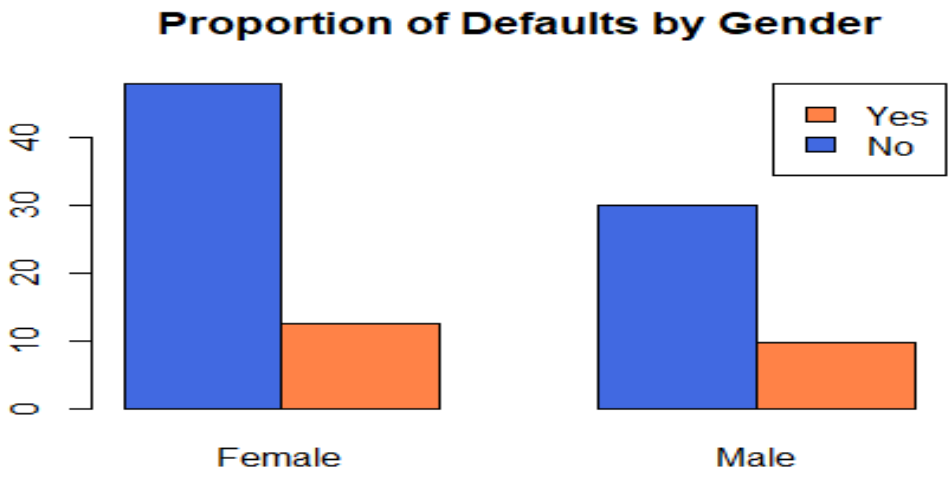


Figure 5: Default status by Gender

The output showed that the highest number of Female did not default followed by the men.

4.2.6 Skewedness of Age

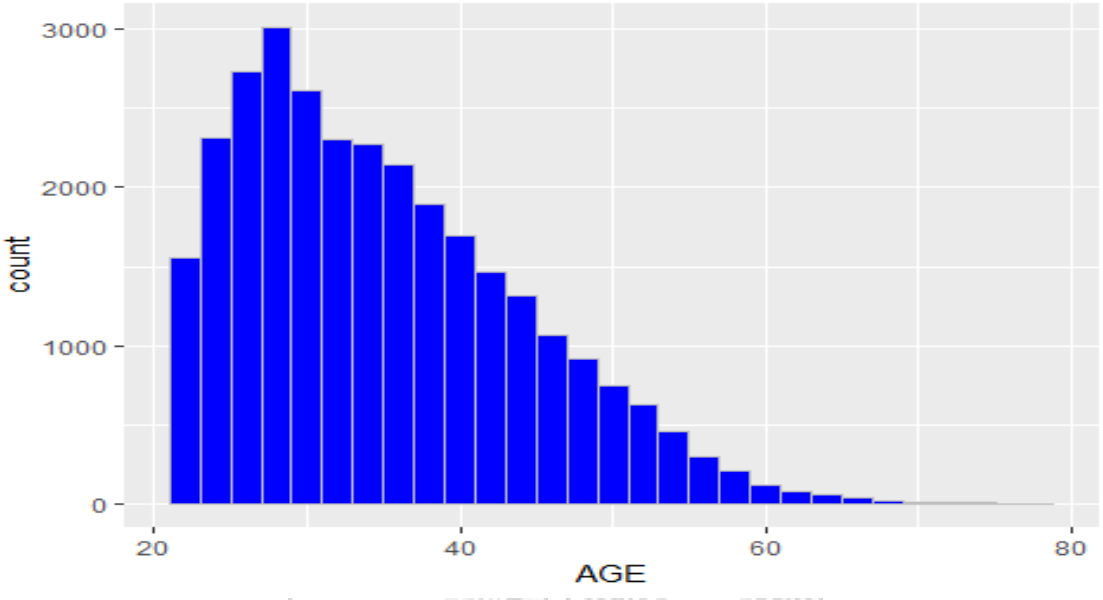


Figure 6: Skewness by Age

Figure 6 indicated that the age distribution was skewed to the right, suggesting that the dataset was predominantly composed of younger participants. The results revealed that the median age was 35.5 years.

4.3 Train and Test data Sets

Before creating the prediction models, the dataset was split into training and testing sets. The training data was used to build the model, while the testing data was reserved for making predictions. After fitting the model on the training data, identifying errors, and minimizing them, the Random Forest model was applied to predict outcomes on the unseen test data.

4.4 Random Forest

By constructing a "forest" of decision trees, the classification model aims to identify the best model by running multiple decision trees and selecting the majority vote to determine the final classification.(Geurts, Ernst, and Wehenkel 2006). To accomplish this, the Random Forest utilized out-of-bag (OOB) sampling. The OOB error was employed to estimate the internal error rate of the Random Forest. The results are as follows: listings

```
predictions_rf   No  Yes
             No  5525 1037
             Yes  316  619

#Predictions in Percentage Format

predictions_rf        No       Yes
              No  73.696145 13.832200
              Yes  4.215019  8.256636

# Calculate accuracy of the RF model
accuracy_rf <- sum(diag(conf_matrix_rf)) / sum(
   conf_matrix_rf)

```

```
## [1] "RF Model Accuracy: 0.81952781112445"

# Print Precision, Recall, and F1 Score

## [1] "Precision: 0.662032085561497"

## [1] "Recall: 0.373792270531401"

## [1] "F1 Score: 0.477807796217677"
```

Examining the confusion matrix, the Random Forest model demonstrated strong performance in predicting "No" for credit default, achieving a class error rate of 5.4%. However, it faced challenges in predicting "Yes" for credit default, resulting in a class error rate of 62.57%.Overall, the random forest model had 81.95% accuracy and Precision was at 66.20% indicating a good proportion of predicted defaults are correct.

4.5 Monte-Carlo

The following results were obtained for the predictions for the stochastic model.

listings

```
# Set number of Monte Carlo simulations
N <- 1000  # Number of simulations
"Average Confusion Matrix (Stochastic):"
          No      Yes
No  5598.773 1102.851
Yes  242.700  552.676
"Average Confusion Matrix Percentage (Stochastic):"

           No       Yes
No  74.680179 14.710564
Yes  3.237295  7.371962

"Average Accuracy (Stochastic): 0.820521408563426"

"Average Precision_st: 0.695728469711039"

"Average Recall_st: 0.336955297672593"

"Average F1 Score_st: 0.453944279372914"
```

Analyzing the confusion matrix, the Monte Carlo model

effectively predicted "No" for credit default, with 74.68% of predictions accurately classified as "No" and 7.37% correctly identified as "Yes." The model achieved an overall accuracy of 82.05%. In comparison, the Monte Carlo model outperformed the Random Forest model, which had an accuracy of approximately 81.95% in predicting credit defaults. The Monte Carlo Model has a slightly higher precision 69.57% compared to the Random Forest model 66.20%.The Monte Carlo Model has a lower recall 33.69% compared to the Random Forest model 37.38% implying that the Monte Carlo model is less effective at identifying all customers who will default. Overall, the Decision Tree model has a marginally better accuracy and precision, but it sacrifices recall and F1 Score. This implies that while the Decision Tree is better at correctly identifying defaults when it does predict them, it misses more actual defaults compared to the Random Forest model.

4.6 Hybrid Model

This section now combines Machine Learning Random Forest Model and the Stochastic Monte carlo Model to form an Hybrid model that is used to predict default. The optimal weight is determined through optimization.

listings

```
1
2     # Define the Hybrid Model
3 hybrid_model <- function(w, rf_probs, st_probs) {
4   w * rf_probs + (1 - w) * st_probs
5 }
6
7 # Optimize the Weight w
8 actual_probs <- ifelse(test_set$default.payment.next.month
     == "Yes", 1, 0)
9 loss_function <- function(w) {
10   hybrid_probs <- hybrid_model(w,
     predicted_probabilities_rf[,2],
     predicted_probabilities_stochastic)
```

```
  mean((hybrid_probs - actual_probs)^2)  # Mean squared
   error
}

opt_result <- optim(par = 0.5, fn = loss_function, method =
    "L-BFGS-B", lower = 0, upper = 1)
optimal_w <- opt_result$par

[1] "Optimal Weight (w): 0.779072209908921"
predictions_hybrid   No  Yes
                  No  5588 1094
                  Yes  253  562

## [1] "Precision: 0.689570552147239"
)
## [1] "Recall: 0.339371980676329"

## [1] "F1 Score: 0.454876568191016"

#Comparing different values of w
 Accuracy Recall Precision
0.1          0.8205949          0.3339372          0.6955975
0.2          0.8205949          0.3339372          0.6955975
0.3          0.8205949          0.3339372          0.6955975
0.4          0.8207283          0.3339372          0.6964736
0.5          0.8207283          0.3321256          0.6979695
0.6          0.8207283          0.3297101          0.7000000
0.7          0.8201947          0.3309179          0.6954315
0.8          0.8208617          0.3442029          0.6892382
0.9          0.8204615          0.3617150          0.6745495
```

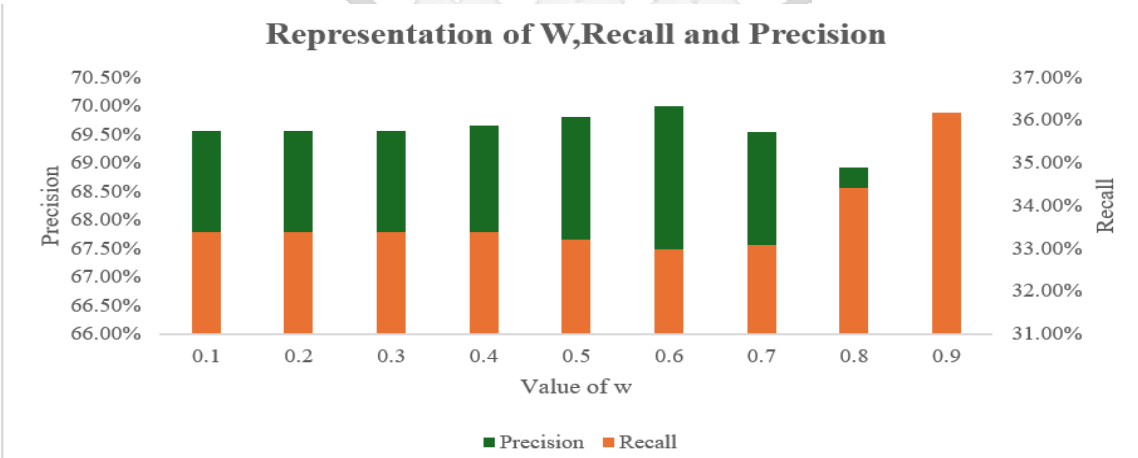


Figure 7: Recall and Precision representation

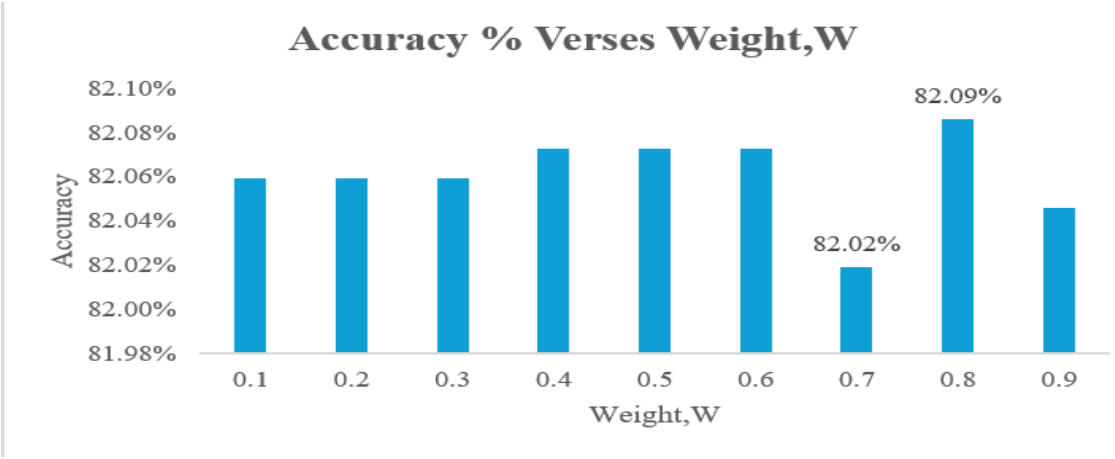


Figure 8: Accuracy Verses values of Weight

The Optimal Weight for combining the Random Forest (RF) and Stochastic (Monte Carlo) models is calculated using optimization. The Optimal Weight for the data used (w) is 0.779. This means that the final hybrid model gives approximately 77.9% weight to the Random Forest model's probabilities and 22.1% weight to the Stochastic model's probabilities.The accuracy of the hybrid model is 82.03% is very close to that of the individual models random forest at 81.95%, and Monte carlo at 82.06%. This indicates that combining the models does not significantly degrade performance and maintains the benefits of both models. The precision of the hybrid model is slightly lower than the Monte Carlo model 69.96% but higher than the Random forest model 66.20%.The recall is at 33.94% which is slightly higher than the Monte Carlo model,33.39% and lower that that of the random forest model 37.38%. The F1 Score of the hybrid model is 45.49% which is better than the Monte Carlo Model 45.12% but lower than the random forest model 47.78%. The F1 Score balances precision and recall, indicating a more balanced performance overall.

5 Discussion of Results

The hybrid model, combining the Random Forest and Monte Carlo models, demonstrates a balanced and effective approach to credit default prediction. By optimizing the weight to 77.9% for the RF model and 22.1% for the Stochastic model, the hybrid model achieves the following results:

Accuracy: The hybrid model maintains a high accuracy of approximately 82.03%, which is comparable to the individual models 81.95% for the random forest and 82.06% for the Monte Carlo. This indicates the model's robustness and reliability in predicting defaults correctly.

Precision: The precision of the hybrid model is around 68.96%, which is an improvement over the random model's precision 66.20% and slightly lower than the Monte Carlo's model precision of 69.96%. This means that the hybrid model effectively reduces false positives while maintaining a reasonable level of true positive predictions.

Recall: The recall is approximately 33.94%, which is higher than the Monte Carlo's model recall 33.39% and lower than to the random forest model's recall 37.38%. This shows that the hybrid model improves the ability to identify actual defaults compared to the Stochastic model. The Hybrid model therefore presents a stronger model for credit default prediction compared to individual models. It brings about:

Balanced Performance: The hybrid model effectively combines the strengths of both random forest and Stochastic models, leading to a well-rounded performance in terms of accuracy, precision, and recall.

Robustness: By leveraging on strengths of both models, the hybrid approach offers robustness in predictions, maintaining high accuracy and reasonable precision while slightly

improving recall over the Stochastic model.

Optimal Weighting: The optimized weight distribution 77.9% random forest and 22.1% Monte Carlo ensures that the hybrid model capitalizes on the superior performance of the random forest model while incorporating valuable insights from the Monte Carlo model. Overall, the hybrid model presents a strong and balanced solution for credit default prediction, making it a valuable tool for financial institutions seeking to improve their prediction accuracy and reliability.

6 Conclusion and Recommendations

The integration of machine learning with stochastic control, specifically using a hybrid model combining Random Forest and Monte Carlo techniques, offers a promising approach to predicting credit defaults.The combination effectively addressed the weaknesses of individual models by improving precision and maintaining accuracy, making it suitable for credit risk assessments where rare events such as defaults are involved.

Financial institutions can leverage the hybrid model to enhance their credit risk management strategies especially in areas where traditional credit scoring models fall short.By implementing models that blend machine learning's capacity to handle large datasets with stochastic control's ability to manage uncertainty in financial markets, institutions can better predict and mitigate the risk of default.

In future researchers can explore alternative machine learning algorithms and stochastic control techniques to further optimize credit risk prediction models. They should also explore the interpret-ability of these models to make them more transparent for financial institutions, ensuring their application in real-world credit risk scenarios.

A Appendix

A.1 R Code used

listings

```
1
2  \begin{verbatim}
3   required_packages <- c("readxl", "rpart", "rpart.plot", "
       caret", "pROC")
4  for (pkg in required_packages) {
5    if (!require(pkg, character.only = TRUE)) {
6      install.packages(pkg, dependencies = TRUE)
7      library(pkg, character.only = TRUE)
8    }
9  }
10
11 library(readxl)
12 file_path<- "C:/Users/lizmu/Desktop/Thesis/Credit Default
       Data.xlsx"
13 Credit_default_data <- read_excel(file_path, sheet = "
       Credit Default Data")
14 print(Credit_default_data)
15
16 #Inspecting data frame classes
17 sapply(`Credit_default_data`, class)
18
19 #Fixing missing Columns
20 sum(is.na(`Credit_default_data`))
21
22 #Checking for duplicates in the data
23 duplicated(Credit_default_data)
24
25 #Removing duplicated elements
26 unique(Credit_default_data)
27
28 #Data Preprocessing
29
30 #Renaming factor variables
31 Credit_default_data$EDUCATION[Credit_default_data$EDUCATION
       %in% "1"] = "Grad School"
32 Credit_default_data$EDUCATION[Credit_default_data$EDUCATION
       %in% "2"] = "College"
33 Credit_default_data$EDUCATION[Credit_default_data$EDUCATION
       %in% "3"] = "High School"
34 Credit_default_data$EDUCATION[Credit_default_data$EDUCATION
       %in% "4"] = "Other"
35 Credit_default_data$EDUCATION[Credit_default_data$EDUCATION
       %in% "5"] = "Unknown"
36 Credit_default_data$EDUCATION[Credit_default_data$EDUCATION
```

```
    %in% "0"] = "Unknown"
Credit_default_data$EDUCATION[Credit_default_data$EDUCATION
    %in% "6"] = "Unknown"

Credit_default_data$default.payment.next.month[
   Credit_default_data$default.payment.next.month %in% "0"]
    = "No"
Credit_default_data$default.payment.next.month[
   Credit_default_data$default.payment.next.month %in% "1"]
    = "Yes"

Credit_default_data$SEX[Credit_default_data$SEX %in% "1"] =
    "Male"
Credit_default_data$SEX[Credit_default_data$SEX %in% "2"] =
    "Female"

Credit_default_data$MARRIAGE[Credit_default_data$ MARRIAGE
   %in% "0"] = "Unknown"
Credit_default_data$MARRIAGE[Credit_default_data$ MARRIAGE
   %in% "1"] = "Married"
Credit_default_data$MARRIAGE[Credit_default_data$ MARRIAGE
   %in% "2"] = "Single"
Credit_default_data$MARRIAGE[Credit_default_data$ MARRIAGE
   %in% "3"] = "Other"

#View the changes
head(Credit_default_data)

#Data Distribution
#View the bar plots
counts_SEX <- table(Credit_default_data$SEX)
percentages_SEX <- round(100 * counts_SEX / sum(counts_SEX)
   , 1)
barplot_heights <- barplot(counts_SEX, col = c("blue", "
   orange"), ylim = c(0, max(counts_SEX) * 1.2))
text(barplot_heights, counts_SEX, labels = paste(counts_SEX
   , "(", percentages_SEX, "%)", sep = ""), pos = 3, cex =
   0.8, col = "black")
title(xlab = "SEX", ylab = "Counts", main = "Counts and
   Percentages by SEX")

#frequency distribution of customers by education levels
counts_EDUCATION <- table(Credit_default_data$EDUCATION)
percentages_EDUCATION <- round(100 * counts_EDUCATION / sum
   (counts_EDUCATION), 1)
labels <- paste(names(counts_EDUCATION), ": ",
   counts_EDUCATION, " (", percentages_EDUCATION, "%)", sep
    = "")
colors <- c("red3", "green4", "purple3", "orange", "blue4")
pie(counts_EDUCATION, labels = NA, col = colors, main = "
```

```
   Counts and Percentages by EDUCATION")
plot.new()
legend("bottomright", legend = labels, fill = colors, title
    = "Education Levels")

#Distribution by marital status
counts_MARRIAGE <- table(Credit_default_data$MARRIAGE)
percentages_MARRIAGE <- round(100 * counts_MARRIAGE / sum(
   counts_MARRIAGE), 1)
labels <- paste(counts_MARRIAGE, " (", percentages_MARRIAGE
   , "%)", sep = "")
barplot_heights <- barplot(counts_MARRIAGE, col = c("
   magenta2", "cyan3", "goldenrod"), ylim = c(0, max(
   counts_MARRIAGE) * 1.2), main = "Distribution by Marital
    Status", xlab = "Marital Status", ylab = "Counts")
text(barplot_heights, counts_MARRIAGE, labels = labels, pos
    = 3, cex = 0.8, col = "black")

#Frequency Distribution of Customers by Default Status
counts_default.payment.next.month <- table(
   Credit_default_data$default.payment.next.month)
percentages_default.payment.next.month <- round(100 *
   counts_default.payment.next.month / sum(counts_default.
   payment.next.month), 1)
labels <- paste(counts_default.payment.next.month, " (",
   percentages_default.payment.next.month, "%)", sep = "")
barplot_heights <- barplot(counts_default.payment.next.
   month, col = c("turquoise2", "sienna1"),ylim = c(0, max(
   counts_default.payment.next.month) * 1.2), main = "
   Percentage per Category",xlab = "default.payment.next.
   month", ylab = "Counts")
text(barplot_heights, counts_default.payment.next.month,
   labels = labels, pos = 3, cex = 0.8, col = "black")

#Proportion of defaults by Gender
table.default.payment.next.month_gender <- table(
   Credit_default_data$default.payment.next.month,
   Credit_default_data$SEX)
category_percentages <- prop.table(table.default.payment.
   next.month_gender) * 100

barplot(category_percentages,
        col = c("royalblue", "sienna1"),
        beside = TRUE,
        names.arg = c("Female", "Male"),
        main = "Proportion of Defaults by Gender")
legend("topright", legend = c("Yes", "No"), fill = c("
   sienna1", "royalblue"))


```

```
#Distribution of sample by age
library(ggplot2)
ggplot(data = Credit_default_data, aes(x = AGE)) +
   geom_histogram(fill = "Blue", col = "Grey", bins = 30)


#Train and test data sets
#Initializing number generator.
set.seed(123)
#New sample created for the training and testing data sets.
    The data is split with 75% in training and 25% in
   testing.
sample <- sample(c(TRUE, FALSE), nrow(Credit_default_data),
    replace = TRUE, prob = c(0.75, 0.25))
train_set <- Credit_default_data[sample, ]
test_set <- Credit_default_data[!sample, ]

#Machine Learning Model-Random Forest
library(randomForest)
library(rpart)
library(rpart.plot)

set.seed(123)
#Random Forest for variables. mtry = 5 since there are 24
   variables (square root of 24 is close to 5).
fit_rf <- randomForest(factor(default.payment.next.month)
   ~., mtry = 5, data = train_set)
print(fit_rf)
#Making predictions on the test data set
predictions_rf <- predict(fit_rf, newdata = test_set)
predicted_probabilities_rf <- predict(fit_rf, newdata =
   test_set, type = "prob")

# Comparing predicted labels with actual labels
conf_matrix_rf <- table(predictions_rf, test_set$default.
   payment.next.month)
print(conf_matrix_rf)
# Calculate percentages for each cell
conf_matrix_rf_percentage <- prop.table(conf_matrix_rf) *
   100

# Print confusion matrix with percentages
print(conf_matrix_rf_percentage)

# Calculate accuracy of the RF model
accuracy_rf <- sum(diag(conf_matrix_rf)) / sum(
   conf_matrix_rf)
print(paste("RF Model Accuracy:", accuracy_rf))

# Calculate Precision, Recall, and F1 Score for the RF
```

```
   model
TP_rf <- conf_matrix_rf["Yes", "Yes"]
FP_rf <- conf_matrix_rf["Yes", "No"]
FN_rf <- conf_matrix_rf["No", "Yes"]

precision_rf <- TP_rf / (TP_rf + FP_rf)
recall_rf <- TP_rf / (TP_rf + FN_rf)
F1_rf <- 2 * (precision_rf * recall_rf) / (precision_rf +
   recall_rf)

# Print Precision, Recall, and F1 Score
print(paste("Precision:", precision_rf))
print(paste("Recall:", recall_rf))
print(paste("F1 Score:", F1_rf))

#Stochastic Model
#Monte carlo Simulation
library(caret)
library(rpart)
library(rpart.plot)

# Set number of Monte Carlo simulations
N <- 1000  # Number of simulations

# Initialize matrices and variables to store results
predicted_probabilities_stochastic <- matrix(0, nrow = nrow
   (test_set), ncol = N)  # Store probabilities across
   simulations
conf_matrix_sum_st <- matrix(0, nrow=2, ncol=2, dimnames=
   list(c("No", "Yes"), c("No", "Yes")))  # Initialize
   aggregate confusion matrix
precision_sum_st <- 0  # Initialize sum of precision
recall_sum_st <- 0  # Initialize sum of recall
f1_sum_st <- 0  # Initialize sum of F1 score
accuracy_sum_st <- 0  # Initialize sum of accuracy

# Run Monte Carlo simulations
for (i in 1:N) {
  # Randomly sample from test_set for each simulation
  test_set_sample <- test_set[sample(nrow(test_set),
   replace = TRUE), ]

  # Train a Decision Tree on the training set
  fit_tree <- rpart(factor(default.payment.next.month) ~ .,
    data = train_set, method = "class")

  # Predict probabilities on the sampled test set
  predictions_st <- predict(fit_tree, newdata =
   test_set_sample, type = "prob")

```

```
  # Store the probability of default (class 'Yes') for this simulation
  predicted_probabilities_stochastic[, i] <- predictions_st[, 2]

  # Confusion Matrix for each simulation
  conf_matrix_st <- table(predictions_st[, 2] > 0.5, test_set_sample$default.payment.next.month)

  # Update the aggregate confusion matrix
  conf_matrix_sum_st <- conf_matrix_sum_st + conf_matrix_st

  # Calculate accuracy for each simulation
  accuracy_st <- sum(diag(conf_matrix_st)) / sum(conf_matrix_st)
  accuracy_sum_st <- accuracy_sum_st + accuracy_st
}

# Average the predicted probabilities across simulations
predicted_probabilities_stochastic <- rowMeans(predicted_probabilities_stochastic)

# Calculate average metrics across all simulations
average_conf_matrix_st <- conf_matrix_sum_st / N  # Average confusion matrix
average_accuracy_st <- accuracy_sum_st / N  # Average accuracy

# Print results
print("Average Confusion Matrix (Stochastic):")
print(average_conf_matrix_st)

# Calculate percentages for each cell for the stochastic model
conf_matrix_percentage_st <- prop.table(average_conf_matrix_st) * 100
print("Average Confusion Matrix Percentage (Stochastic):")
print(conf_matrix_percentage_st)

print(paste("Average Accuracy (Stochastic):", average_accuracy_st))
print(paste("Average Precision_st:", average_precision_st))
print(paste("Average Recall_st:", average_recall_st))
print(paste("Average F1 Score_st:", average_f1_st))



# Load necessary libraries
library(randomForest)
library(dplyr)
```

```
# Hybrid Model
# Number of simulations
num_simulations <- 1000

# Initialize a matrix to store predicted probabilities across simulations
predicted_probabilities_stochastic_all <- matrix(0, nrow = nrow(test_set), ncol = num_simulations)

# Run simulations
for (i in 1:num_simulations) {
  # Fit the decision tree model
  fit_tree <- rpart(factor(default.payment.next.month) ~ ., data = train_set, method = "class")

  # Predict on the test set and get predicted probabilities
  predicted_probabilities_stochastic_all[, i] <- predict(fit_tree, newdata = test_set, type = "prob")[, 2]  # Probabilities of "Yes"
}

# Average the predicted probabilities across all simulations
predicted_probabilities_stochastic <- rowMeans(predicted_probabilities_stochastic_all)

# Now use the averaged probabilities in the hybrid model
actual_probs <- ifelse(test_set$default.payment.next.month == "Yes", 1, 0)

# Define the Hybrid Model
hybrid_model <- function(w, rf_probs, st_probs) {
  w * rf_probs + (1 - w) * st_probs
}

# Loss function and optimization
loss_function <- function(w) {
  hybrid_probs <- hybrid_model(w,
   predicted_probabilities_rf[, 2],
   predicted_probabilities_stochastic)  # Use averaged probabilities
  mean((hybrid_probs - actual_probs)^2)  # Mean squared error
}

# Optimization
opt_result <- optim(par = 0.5, fn = loss_function, method = "L-BFGS-B", lower = 0, upper = 1)
optimal_w <- opt_result$par

```

```
# Compute the Hybrid Model Probabilities with Optimal Weight
final_hybrid_probs <- hybrid_model(optimal_w, predicted_probabilities_rf[, 2], predicted_probabilities_stochastic)

# Print results
print(paste("Optimal Weight (w):", optimal_w))
print("Final Hybrid Probabilities:")

# Performance Evaluation of the Hybrid Model
predictions_hybrid <- ifelse(final_hybrid_probs > 0.5, "Yes", "No")
conf_matrix_hybrid <- table(predictions_hybrid, test_set$default.payment.next.month)
print(conf_matrix_hybrid)

# Calculate accuracy of the Hybrid model
accuracy_hybrid <- sum(diag(conf_matrix_hybrid)) / sum(conf_matrix_hybrid)
print(paste("Hybrid Model Accuracy:", accuracy_hybrid))

# Calculate Precision, Recall, and F1 Score for the Hybrid model
TP_hybrid <- conf_matrix_hybrid["Yes", "Yes"]
FP_hybrid <- conf_matrix_hybrid["Yes", "No"]
FN_hybrid <- conf_matrix_hybrid["No", "Yes"]

precision_hybrid <- TP_hybrid / (TP_hybrid + FP_hybrid)
recall_hybrid <- TP_hybrid / (TP_hybrid + FN_hybrid)
F1_hybrid <- 2 * (precision_hybrid * recall_hybrid) / (precision_hybrid + recall_hybrid)

# Print Precision, Recall, and F1 Score
print(paste("Precision:", precision_hybrid))
print(paste("Recall:", recall_hybrid))
print(paste("F1 Score:", F1_hybrid))

# Compare different values of w
# Initialize variables to store results
results <- data.frame(w = numeric(), accuracy = numeric(), recall = numeric(), precision = numeric())

for (w in seq(0.1, 0.9, by = 0.1)) {
  # Calculate Hybrid Model Probabilities with the current weight (w)
  final_hybrid_probs <- hybrid_model(w, predicted_probabilities_rf[, 2], predicted_probabilities_stochastic)

```

```
  # Obtain predictions based on Hybrid Model Probabilities
  predictions_hybrid <- ifelse(final_hybrid_probs > 0.5, "
   Yes", "No")

  # Confusion Matrix for the Hybrid Model
  conf_matrix_hybrid <- table(predictions_hybrid,
   test_set$default.payment.next.month)

  # Calculate Accuracy, Recall, and Precision for the
   Hybrid Model
  TP_hybrid <- conf_matrix_hybrid["Yes", "Yes"]
  FP_hybrid <- conf_matrix_hybrid["Yes", "No"]
  FN_hybrid <- conf_matrix_hybrid["No", "Yes"]

  accuracy_hybrid <- sum(diag(conf_matrix_hybrid)) / sum(
   conf_matrix_hybrid)
  precision_hybrid <- TP_hybrid / (TP_hybrid + FP_hybrid)
  recall_hybrid <- TP_hybrid / (TP_hybrid + FN_hybrid)

  # Append results to the data frame
  results <- rbind(results, c(w, accuracy_hybrid,
   recall_hybrid, precision_hybrid))
}

# Print the results
print(results * 100)

library(ggplot2)

\end{verbatim}
```

A.2 Credit Data used

https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

References

Abu Alfeilat, Haneen Arafat et al. (2019). "Effects of distance measure choice on k-nearest neighbor classifier performance: a review". In: *Big data* 7.4, pp. 221–248.

Addo, Peter Martey, Dominique Guegan, and Bertrand Hassani (2018). "Credit risk analysis using machine and deep learning models". In: *Risks* 6.2, p. 38.

Adede, Chrisgone O (2012). "Model for predicting the probability of event occurrence using logistic regression: case for a Kenyan commercial bank". PhD thesis. University of Nairobi.

Aduda, Josiah, Peterson O Magutu, and Githinji Mary Wangu (2012). "The relationship between credit scoring practices by commercial banks and access to credit by small and medium enterprises in Kenya". In.

Ampomah, Ernest Kwame, Zhiguang Qin, and Gabriel Nyame (2020). "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement". In: *Information* 11.6, p. 332.

Ampountolas, Apostolos, Titus Nyarko Nde, Corina Constantinescu, et al. (2021). "A machine learning approach for micro-credit scoring". In: *Risks* 9.3, p. 50.

Anderson, Brian DO and John B Moore (2007). *Optimal control: linear quadratic methods*. Courier Corporation.

Balling, Morten and Ernest Gnan (2013). "The development of financial markets and financial theory: 50 years of interaction". In: *M. Balling & E. Gnan (Eds.)* 50, pp. 157–194.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Creswell, John W (2014). *A concise introduction to mixed methods research*. SAGE publications.

Dornaika, Fadi et al. (2017). "Object categorization using adaptive graph-based semi-supervised learning". In: *Handbook of Neural Computation*. Elsevier, pp. 167–179.

Fan, Shuangshuang, Yanbo Shen, and Shengnan Peng (2020). "Improved ML-based technique for credit card scoring in internet financial risk control". In: *Complexity* 2020, pp. 1–14.

Freund, Yoav and Robert E Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1, pp. 119–139.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel (2006). "Extremely randomized trees". In: *Machine learning* 63, pp. 3–42.

Ghevariya, SJ and DR Thakkar (2019). "BSM model for the Generalized ML-Payoff". In: *Journal of Applied Mathematics and Computational Mechanics* 18.4.

Glasserman, Paul (2004). *Monte Carlo methods in financial engineering*. Vol. 53. Springer.

Hammersley, Martyn (2017). "Deconstructing the qualitative-quantitative divide 1". In: *Mixing methods: Qualitative and quantitative research*. Routledge, pp. 39–55.

Joseph, Ciby (2013). *Advanced credit risk analysis and management*. John Wiley & Sons.

Karatzas, Ioannis and Steven Shreve (2014). *Brownian motion and stochastic calculus*. Vol. 113. springer.

Khandani, Amir E, Adlar J Kim, and Andrew W Lo (2010). "Consumer credit-risk models via machine-learning algorithms". In: *Journal of Banking & Finance* 34.11, pp. 2767–2787.

Kollár, Boris and Barbora Gondžárová (2015). "Comparison of current credit risk models". In: *Procedia economics and finance* 23, pp. 341–347.

Kshetri, Nir (2014). "The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns". In: *Big Data & Society* 1.2, p. 2053951714564227.

Lee, Ming-Chi (2009). "Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit". In: *Electronic commerce research and applications* 8.3, pp. 130–141.

Lewis, Sarah (2015). "Qualitative inquiry and research design: Choosing among five approaches". In: *Health promotion practice* 16.4, pp. 473–475.

Mhlanga, David (2021). "Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment". In: *International journal of financial studies* 9.3, p. 39.

Musyoki, Danson and Adano Salad Kadubo (2012). "The impact of credit risk management on the financial performance of banks in Kenya for the period". In: *International Journal of Business and Public Management* 2.2, pp. 72–80.

Mutembete, M. B. (2022). "Credit Risk Assessment Model Using Machine Learning". Doctoral dissertation. KCA University.

Oksendal, Bernt (2013). *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media.

Ordabayeva, Zhanna, Aiman Moldagulova, and Ibrahim Riza (2023). "Building a Credit Scoring Model Based on the Type of Target Variable". In: *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*. IEEE, pp. 31–36.

Óskarsdóttir, María et al. (2019). "The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics". In: *Applied Soft Computing* 74, pp. 26–39.

Panigrahi, Ranjit et al. (2018). "Classification and analysis of facebook metrics dataset using supervised classifiers". In: *Social Network Analytics: Computational Research Methods and Techniques* 1.

Petropoulos, Anastasios et al. (2019). "A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting". In: *IFC Bulletins chapters* 49.

Pławiak, Paweł et al. (2020). "DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring". In: *Information Sciences* 516, pp. 401–418.

Provenzano, Angela Rita et al. (2020). "Machine learning approach for credit scoring". In: *arXiv preprint arXiv:2008.01687*.

Pun, Chi Seng (2018). "Robust time-inconsistent stochastic control problems". In: *Automatica* 94, pp. 249–257.

Puterman, Martin L (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Saeed, MS and N Zahid (2016). "The impact of credit risk on profitability of the commercial banks". In: *Journal of Business & Financial Affairs* 5.2, pp. 2167–0234.

Schapire, Robert E (2013). "Explaining adaboost". In: *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Springer, pp. 37–52.

Silverman, Robert Mark and Kelly Patterson (2021). *Qualitative research methods for community development*. Routledge.

Sudhakar, Kalva and Satuluri Naganjaneyulu (2023). "Enhancing stock market forecasting using sequential training network empowered by tunicate swarm optimization". In: *Multimedia Tools and Applications*, pp. 1–24.

Taylor, Steven J, Robert Bogdan, and Marjorie DeVault (2015). *Introduction to qualitative research methods: A guidebook and resource*. John Wiley & Sons.

Wang, Yu (2009). "Structural credit risk modeling: Merton and beyond". In: *Risk Management* 16.2, pp. 30–33.

Yap, Bee Wah, Seng Huat Ong, and Nor Huselina Mohamed Husain (2011). "Using data mining to improve assessment of credit worthiness via credit scoring models". In: *Expert Systems with Applications* 38.10, pp. 13274–13283.

Yeo, Ai Cheo et al. (2001). "Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry". In: *Intelligent Systems in Accounting, Finance & Management* 10.1, pp. 39–50.

Zhao, Yang, Jianping Li, and Lean Yu (2017). "A deep learning ensemble approach for crude oil price forecasting". In: *Energy Economics* 66, pp. 9–16.