CREDIT RISK ASSESSMENT MODEL USING MACHINE LEARNING

BY

MUTEMBETE B MATHIAS

REG NO: 20/00498

A DISSERTATION SUBMITTED TO THE SCHOOL OF TECHNOLOGY IN PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN DATA ANALYTICS, KCA UNIVERSITY

19TH SEPTEMBER 2022

DECLARATION

I declare that this Dissertation is my original work and has not previously been published or submitted for a degree elsewhere. I also declare that this does not include any material written or published by others, unless a proper reference is made, and the author is properly acknowledged.

Student Name: **MUTEMBETE B MATHIAS** Reg No**: Reg: 20/00498**

Sign……..…… **Date**……..……

I hereby confirm that I have reviewed MUTEMBETE B MATHIAS's Master's Dissertation and approved it for examination.

Sign: ……. Date: ……..

Name: **Dr. Simon Mwendia**

ABSTRACT

An individual's creditworthiness is quantified by their credit score, which is based on their credit history. Credit ratings are used by financial companies to distinguish between debtors who will fulfill their obligations and those who won't. This system has not yet been digitized or used in Kenya's banking sector. The study's objective is to develop a reliable credit scoring model that will give organizations like these a reliable reference score to rely on when verifying a client. The data used included customer details including age, loan amount, marital status, and sex, among other factors. It was collected from Kenyan commercial banks between 2016 and 2021. It was possible to create an optimized model with an accuracy of 93%. The model was built using the Gradient Boosting method (GBM) and is based on classification and Regression Trees (CART). In addition, a new hybrid model with a two-step architecture is proposed. The first uses distributed random forests, where each decision tree's output is fed into a deep neural network (DNN) that has been trained to outperform the random forest method on its own. Giving a rating without adequate rationale is unethical because a person's creditworthiness is a sensitive matter. An examination of the model's interpretability was conducted, and the results created visual representations of the factors that influence the model's output and the data required for a successful client analysis. The outcomes were reliable and accurately replicated the process of appraising an individual. The proposed model could be put into practice in order to give failure evaluation and prediction a real-world foundation. Simultaneously provide a thorough explanation of the outcomes at the same time. This could considerably aid financial organizations in preventing the loss of millions of dollars from non-performing loans.

Keywords: Machine learning, Credit Score, Machine Learning algorithms, Probability of default, Gradient Boosting, Classification and Regression Trees.

ACKNOWLEDGEMENT

Over the past few years, I've had the honor of working and getting to know many incredible people. I'd like to express my heartfelt gratitude to the people listed below.

First and foremost, I would like to thank my supervisor, Dr. Simon Mwendia, for his patience, encouragement, expertise, and trust. Dr. Mwendia was excellent at communicating his ideas and comprehending my mumbled, half-baked explanations. His friendly demeanor and upbeat outlook turned a seemingly technical research discipline into an engaging and interesting topic. He was an excellent mentor and motivator. Under his guidance, I learned how to develop the focus required for scientific research. His never-ending enthusiasm was exemplified by his draft paper comments, late-night/early-morning emails, and unwavering support.

I'd like to thank members of the Computing and Information Management department, particularly Dr. Kibuku Rachael, Dr. Dennis Kamau, and Dr. Lawrence Nderu, for their assistance and guidance over the last few years. I'd like to thank Prof. Joshua Gisemba Bagaka in particular for having such a positive influence on my formal education.

My parents have undoubtedly made tremendous sacrifices to ensure that their children received a good education and were raised to be honest and fair to others. Thank you also to my siblings for their encouragement and gentle ribbing.

ACRONYMS AND ABBREVIATIONS

CART- Classification and Regression Trees

GBM- Gradient Boosting method

DNN- Deep Neural Network

P2P- Peer-to-peer lending

AI- Artificial Intelligence.

CBK- Central Bank of Kenya

NPL- Nonperforming Loan

CRB- Credit Reference Bureau.

WCDR- Worst-Case Default Rate.

PD- Probability of Default.

EAD- Exposure at Default.

LGD- Loss given Default.

CPD- Cumulative Probability Distribution

RMSE- Root Mean Squared Error

AUC- Area Under the Curve

ROC- Receiver Operating Characteristic Curve.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background of the Study

The risk associated with credit risk is the borrower's inability to meet the obligation. It is a major risk in financial markets. As a result, credit risk modeling has gotten a lot of attention in the financial industry, particularly since the 2008 financial crisis and the expanded concerns raised by Basel II and III. Financial institutions are working in a credit-constrained environment while raising interest rates to mitigate risk. There has also been an increase in peer-to-peer lending (P2P) and the emergence of microfinance institutions, where effective credit risk management is critical in ensuring profitability in a competitive environment. The ability of a company to manage its default process is critical to its long-term success and performance.

Credit risk is an area of concern to researchers. All transactions and contracts are based on information and take place during financial intermediation. Several issues may arise, such as the issue of all participants not being fully informed, or certain transactions containing additional information that may not be available to both parties. This results in an asymmetric flow of information, making financial agreements difficult to enter and potentially leading to inefficient intermediation. Adverse selection models are characterized by one side not having the information while performing the transaction, whereas moral hazard models have the issue arise after the transaction, such as lender’s inability to monitor borrower's actions which directly impacts obligor’s default probability.

Credit scoring and rationing are widely used by financial institutions to reduce risk and maximize profits. Credit scoring is the application of statistical models to transform a set of relevant data into a numeric transform that can guide credit decisions for a financial institution. Using characteristics such as income, age, and marital status, these models classify applicants as either good or bad. Credit scoring minimizes credit cost by reducing the chances of default by assessing a customer's creditworthiness and, in some cases, detecting fraud. It can also monitor existing loan accounts and thus prioritize repayment collection. Almost all credit institutions now use some form of credit scoring before extending a line of credit or making a loan to individuals or businesses.

A credit score is a numerical representation of individual's creditworthiness based on their credit history. In contrast, credit scoring uses statistical analysis to evaluate a borrower's creditworthiness. This method was developed in the 1950s (Tonester524, 2019). Credit scoring is commonly used in consumer lending to collect and assess an obligor's credit history in order

to determine the ability to meet stipulated facility conditions. Data collected may include, among other things, default history, employment history, financial assets, existing debt, and the type of bank accounts one owns. The optimal set of variables to predict delinquency or default, as well as the appropriate weights to give each variable, are determined using regression analysis. When the correlations between the elements are looked at, it becomes evident that some of the initial factors used by the model creator have minimal effect since, in comparison to the other variables, they don't add much value to the model. A renowned provider of scoring models, Fair, Isaac and Company, Inc., states that although 50 or 60 variables may be originally taken into account, the best predictive combination may wind up being 8 to 12 in the final scorecard (Tonester524, 2019). First Data Resources' Anthony Saunders claims that 48 criteria are used to determine the probability of credit card delinquency. Higher credit scores often signal lesser risk and based on the amount of risk they are willing to accept, lenders set a cutoff score.

Recent methodologies for estimating default probabilities include AI algorithms and models based on options pricing theory. Using experience, neural networks are AI algorithms that detect existing relationships between obligor characteristics and default risk. They then identify significant characteristics that influence the likelihood of default. Because no assumptions need to be made about the distributions of the model's variables or errors, or the functional form of the relationship between characteristics and default likelihood, this method is more adaptable and superior to standard statistical credit scoring methods.

Currently, most large and small banks utilize credit scoring for loans under $100,000. There is no universal scoring model. Due to the non-homogeneous nature of business loans, most lenders take longer to model them than credit cards or other types of consumer facilities. Furthermore, the volume of business loans is low, so there is insufficient data to train the model. Credit scoring models have been adopted by companies other than lending institutions, such as insurance companies and landlords. Alternative data sources are also used by digital finance companies, such as online lenders, to determine borrowers' creditworthiness.

Assessing the credit risk of loan debtors is the main goal of this study. There have been a rising number of repossessed or charged-off loans. Financial organizations and banks suffer significant losses as transactions are suspended and assets are frozen. According to the 2021 CBK Bank Supervision Report, there has been a significant increase in nonperforming facilities from 2010 to date. For example, in 2010, the sector's NPL rate was 6.3 percent, but by 2016, it had risen to 9.2% percent. The trend continued until 2021, when the NPL ratio reached 14.2%. According to several analysts, Kenya's current state will hinder corporate growth and prevent

the implementation of numerous plans to increase employment for the general population (CBK Survey, 2021). This suggests that default loans have a detrimental effect on financial institutions as well as the nation's economy. Filtering out apps that are more likely to fail would be a trustworthy solution to this issue. This can be accomplished using a pattern recognition technique, at which machine learning excels. The ability to spot basic patterns in a field and train on them to gradually get better. Numerous research have proven the efficacy of using machine learning algorithms for credit risk assessment. The effectiveness of utilizing machine learning algorithms for credit risk assessment has been demonstrated by numerous studies.

The feature selection methods used in Du, G.'s (2021) neural network and genetic algorithm model for credit risk assessment included forward selection, information gain, gain ratio, and Gini index. They arrived at the conclusion that a neural network and genetic algorithm combination, in addition to the genetic algorithm, was the optimum choice for their data set. *K*-folds cross validation was used instead of train-test splitting to produce a more reliable result. With promising results for credit risk prediction, supervised learning on credit risk datasets was carried out utilizing ensemble techniques, tree-based models, and neural network models in a number of additional publications, such as (Huang, X., 2018), (P. Addo, 2018), and (Khemakhem, S., 2018).

In this study, a number of supervised algorithms and feature selection methods were used to evaluate the prediction of a representative score based on an individual profile. The applicant's credit history and personal data were included in the data set. The underlying patterns in previous borrowers' credit scores were discovered using gradient boosted regression, extreme gradient boosting, random forest, and linear regression regressors, as well as a proposed hybrid model made up of a Deep Neural Network and a Random Forest Regressor. Supervised learning was used to generate scores for unseen data. By instantiating log and sq rt of numeric columns, correlations and feature engineering was used to select the best features. The best hyper-parameters for the selected models was then picked using grid search and 4-fold cross validation randomization. Each model was contrasted in order to determine which one is the most useful for determining credit risk. Inference would make use of LIME (P. Ferrando, 2018) for model interpretation. The model's operation was explained, and considerable data for potential human examination was also provided.

1.2 Problem Statement

Customer credit management is a critical issue for every commercial bank; therefore, banks exercise extreme caution when dealing with customer loans to avoid making mistakes that could result in lost opportunities or financial losses. Manual creditworthiness estimation has become both time- and resource-intensive. Furthermore, a manual approach is subjective (reliant on the bank employee who provides this estimate), which is why developing and implementing programming models that provide loan estimates is the only way to eliminate the 'human factor' in this problem. This model should make recommendations to the bank about whether or not to make a loan, or it should provide a probability about whether the loan will be returned.

A number of models have been developed in recent years, but there is no ideal classifier among these models because each gives some percentage of incorrect outputs; this is an important consideration because each percent of incorrect answers can result in millions of dollars of losses for large banks. However, the LR remains the industry standard tool for developing credit-scoring models. To that end, an investigation is conducted on the combination of the most efficient classifiers in the credits coring scope in an attempt to produce a classifier that outperforms each of its classifiers or components. The study also intends to make sure that the model's output isn't seen as coming from a "black box" with no underlying reasoning. Credit worthiness is a critical factor in determining an individual's future, thus the study explains why the model was given a specific score and which factors most significantly influenced the conclusion.

1.3 Objective

1.3.1 Main Objective

The study's objective is to develop a reliable credit scoring model that will give institutions a precise reference score to rely on when evaluating the credit risk of clients.

1.3.2 Specific Objective

The following research objectives guided the study:

i. To identify the critical variables that influence individual’s credit score.
ii. To create a model for predicting individual’s credit score.
iii. To assess and validate the developed model's ability to predict individual’s credit score.

1.4 Research Questions

The research aims to answer the following questions to meet the study objectives:

i. Which variables have a significant impact on individual's credit score?
ii. How efficient is the model at modeling individual's credit score?
iii. How accurate is the developed model in predicting individual's credit score?

1.5 Significance of the Study

In this work, several methods have been developed and enhanced to significantly enhance the performance of classifiers and combiners. The study's main contribution is that it provides a credit scoring model capable of classifying a credit as risky or not, backed up by an architecture that improves the development and maintenance process of exploring the available data.

The findings of the study add to the literature in this field while also informing the various stakeholders by understanding and improving the way consumer credit risk modeling is done for the Kenyan market. Kenya's credit market is expanding, particularly in the microfinance and digital lending sectors. The study's findings will assist institutions in lowering credit cost and maximize on profits through sound lending.

Banks, on the other hand, are much more heavily regulated, and they are required to have internal rating models to help manage credit risk under the Basel II and Solvency II frameworks, in addition to the new IFRS 9 regulation. Banks need a stronger credit scoring system to evaluate consumers and decrease defaults due to the rise in non-performing loans, especially when interest rate earnings are dropping as a result of fierce competition from microfinance and digital lenders. The results of this study will be helpful in this situation because even a little percentage increase in the models' predictability can result in significant savings for banks, which will boost their profitability. The study will also provide as a starting point for other studies in the same area, which will expand on some of the topics this one did not address.

1.6 Motivation

In Kenya, there is growing concern about the rate of growth of NPLs. According to the 2021 CBK Bank Supervision Report, there has been a significant increase in nonperforming facilities from 2011 to date. For example, in 2010, the sector's NPL rate was 6.3 percent, but by 2016, it had risen to 9.2 percent. Due to increased credit risk, there was a significant fluctuation in NPLs between 2010 and 2016. The trend continued until 2021, when the NPL ratio reached 14.2%.

Doubtful and loss loans and advances rose by 6.7 percent and 13.1 percent, respectively, as of December 2021 compared to the position in 2020. The elevated levels of these categories

over the entire loan book serve as more evidence of this. In comparison to 3.2 percent, 8.3 percent, and 3.1 percent in 2020, the substandard, doubtful, and loss categories made up 2.9 percent, 8.1 percent, and 3.2 percent of the loan book in 2021, respectively. As a result of the COVID-19 epidemic, declining asset quality, enhanced loan categorization and provisioning, business issues, and higher default on digital loans, there was a surge in non-performing loans in the doubtful and loss categories.

TABLE 1:

Risk Classification of Loans and Advances

	2020		2021			
	Amount Ksh' Million	% of Total	Amount Ksh' Million	% of Total	Change Ksh' Million	% Change
	A		B		C=B-A	D=C/A
Normal	2,254,006	75	2,401,620	73.8	147,614	6.5
Watch	316,031	10.5	393,801	12.1	77,770	24.6
Substandard	95,721	3.2	94,670	2.9	(1,051)	(1.1)
Doubtful	248,689	8.3	262,716	8.1	14,027	5.6
Loss	91,657	3	102,623	3.2	10,966	12.0
Total	**3,006,104**	**100**	**3,255,430**	**100**	**249,326**	**8.3**

Source: CBK, Bank supervision Annual report, 2021 pg 21

These alarming statistics necessitates development of a model that will minimize the default rate while maximizing the granted loans.

1.7 The Study Scope

The study relied on historical loan data on Kenyan market with the aim of developing a reliable credit scoring model that will give institutions a precise reference score to rely on when evaluating the credit risk of clients.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

When assessing capital adequacy, the new Basel Capital Accord holds banks accountable for having sound internal credit risk management practices (Wagacha & Othieno, 2015). CBK suggests that banks have a good enough assessment procedure in place to ensure that they can assess the risk profile of the borrower, and that they look at the credit rating report obtained from any licensed credit bureau (CBK, 2021). Non-banking financial institutions are also in the lending business, and as such, their risk must be managed, especially as mobile loans and digital lending become more popular.

The growing use of digital loans is due to the quick access to funds and the lack of a collateral requirement, as well as the use of alternative credit scoring models that use information from mobile money transactions when determining a borrower's eligibility. Use of such alternative data as well as data from the Credit Reference Bureau (CRB) can really fit well with machine learning models that are robust in examining the relationships between variables. Because of the increased competition, the industry has become a margins' industry, where a small increase in efficiency in their processes will increase their chances of success. To compete with banks, their credit scoring mechanism must be efficient and automatic, lowering the cost of credit analysis.

Market developments have also resulted in an increase in transactional data, which has led to the development of computational and power computers for data mining and analysis. In order to find patterns and trends in data, a process known as data mining is used. It is based on mathematical analysis. The widespread use of machine learning algorithms in predictive analytics, including risk assessment, identification, mitigation, and prediction, is supported by a number of academic studies. Thus, using ensemble classifiers, a machine learning predictive model for assessing credit risk can be developed.

This chapter describes concepts that underlie the study, a review of empirical literature, Conceptual Framework and how variables are operationalized.

2.2 Theoretical Review

This section explores credit scoring, how it affects bank lending decisions, the advantages, and disadvantages of credit modeling, and more. The next step is to describe the most common variables in credit scoring. After that, a review of the modeling strategies employed in other studies with the same research approach follows.

2.2.1 Banks' lending decision

The performance of consumer loans impacts the financial institutions' profitability and stability, and screening loan applications is a crucial step in reducing credit risk. Credit analysis should be performed as part of the screening process before making any credit decisions. Credit analysis, which includes valuing the applicant's financial history and financial documents, attempts to assess the borrower's likelihood of payback, ascertain the borrower's financial stability, and reduce the risk of non-payment to a manageable level. A loan would be approved for good borrowers with low credit risk but denied for borrowers with high risk. The appraisal of all relevant aspects of an applicant concurrently and the objective assessment of each applicant are the two main challenges in credit analysis. Loan applicators are judged using quantitative and qualitative characteristics.

According to Pang (2021), there are two basic techniques for determining a borrower's creditworthiness: the subjective judgment of the loan officer, sometimes known as the judging methodology, and the credit score technique. An applicant's creditworthiness is assessed based on whether or not they possess the qualities necessary to qualify for a loan. If a person does not have good credit, they will not be eligible for the loan (Pang, 2021). The six Cs—Character, Capacity, Cash, Collateral, Conditions, and Control—are used to evaluate an applicant's creditworthiness (Munguti, 2020). According to Pang (2021), the judgmental approach to credit is ineffective, illogical, incompatible, and non-standardized. Traditional methods of loan decision-making rely on past decisions' experience and human judgment of the danger of default.

However, the rise of new computer technologies, more economic competitiveness, and the increased need for credit have all contributed to the development of credit scoring methodology. In compared to subjective judging judgments by loan officers, classifications produced by credit scoring algorithms are more precise. The advantages of the credit scoring method over the judgmental approach are discussed by Siddiqi (2017). For instance, credit scoring is more effective because it enables loan officers to concentrate solely on ambiguous instances and because it helps lenders regularly assess the creditworthiness of their customers. As a result, credit scoring models are a preferred method for assessing credit risk.

Altman was the first to utilize a statistical model to estimate a borrower's likelihood of defaulting (1968). His goal was to better accurately assess the credit risk of the borrower. As a result, numerous statistical credit scoring models, including logistic regression, neural networks, smoothing nonparametric, and expert systems, were created. These models are now frequently employed to evaluate credit risk (Xia et al., 2018). Following this, a variety of statistical credit scoring models, including logistic regression, discriminant analysis, the linear probability technique, the probit model, and neural networks, were created and are now commonly used to evaluate credit risk (Siddiqi, 2017).

2.2.2 What is credit scoring?

In the 1950s, American retailers and mail-order businesses used credit rating for the first time along with the early implementation of investment portfolios to manage and diversify borrowers default risks (Siddiqi, 2017). Credit scoring models are currently one of the banking and financial industries' most successful modeling methods. Credit scoring, which is based on statistical analysis, separates the impact of different application factors on delinquencies and defaults by using historical data and credit characteristics. Banks can use credit scoring algorithms to help with lending choices. Since credit scoring analyzes applicants' credit risk far more precisely and swiftly than can subjective assessment, it complements and even occasionally replaces it. Although mortgage lending has just started to adopt this form of credit risk measurement, consumer loans, particularly credit cards, still make up the majority of its applications. Furthermore, this scoring approach is now applicable to complicated commercial loans thanks to advances in computer technology that make data more readily available to businesses.

In order to make applicants' default risk more predictable, several banks evaluate loan applications using credit scoring algorithms. A credit scoring system is a computerized procedure that generates a score based on a variety of pertinent borrower data, including income, occupation, age, wealth, past loans, etc. The ultimate score is calculated by adding each borrower's individual score. If the score is higher than a predefined bank's "cut-off-level," credit will be approved; if not, credit will not be granted. The foundation of credit scoring is statistical probabilities, or more precisely, the combinations of the borrower's qualities that distinguish good borrowers from poor. In this manner, a score is created to serve as an estimate of the level of risk associated with each new. According to Pang (2021), the objective of credit scoring is to forecast risk rather than to explain it. It is not necessary for the prediction model

to also provide an explanation for why some borrowers miss their loan payments while others do not.

Credit scoring electronically examines a borrower's credit history as well as additional variables relating to repayment capacity that are typically provided by the borrower. Credit models might forecast the default risk of any loan provided based on prior experience with borrowers with comparable loan profiles. High scores should be given to borrowers whose loans would perform well, and low ratings should be given to borrowers whose loans would not perform well, in accordance with a good credit model. Reviewing the credit worthiness of the borrower on a regular basis is a vital step in creating a superior credit scoring model, as changes in the economic climate may have an impact on loan performance. There are generally no top credit scoring models. It is likely that some problematic borrowers may receive loans and excellent scores, and vice versa. According to Mao et al (2018), if credit scoring were used, about 8% of applications would be allowed when they were actually for bad loans and 18% would be refused when they were for good loans.

In each credit market, borrowers have the choice to default. Those who default are not prohibited from borrowing in the future, thus lenders and borrowers may enter the market at any time, and lenders are not permitted to work together to punish defaulters. Instead, a lender derives information about a borrower's personality from his or her borrowing and repayment habits and summarizes this information in a credit score. Due to a lack of knowledge regarding the borrower's actions and income realizations, lenders may only offer small amounts of credit or credit at higher interest rates.

2.2.3 Pros and cons of credit scoring

Credit scoring is increasingly used in loan appraisal because it offers some clear advantages over judgmental procedures for both lenders and borrowers. First off, clear binary decisions can be handled instantly by credit scoring models. Credit officers now have more time to focus on the more complex instances that the models do not handle well. The time it takes to provide a loan to a consumer is decreased from weeks to days or hours thanks to credit scoring, which is another effective technique to save time. In comparison to the conventional loan assessment process, the credit officer can handle more loan applications. Time savings translate into cost savings for the bank and advantages for the customer. Customers simply need to submit the data needed for the scoring system, which makes applications less time-consuming. Additionally, credit scoring has the important benefit of lowering the likelihood of prejudice. This can be explained by the fact that credit scoring is a common method of loan approval.

Officers from the bank employ the model's constructed criteria and assess applicants in light of them. The scoring algorithms take into account the traits of both good and bad borrowers in contrast to judging techniques, which are typically negatively biased towards poor borrowers. The model helps lenders make sure that all borrowers have been subjected to the same underwriting standards, regardless of their gender, nationality, or any other characteristics that are illegal under commercial law to consider when evaluating credit.

In terms of how loans are categorized in relation to credit risk, credit analysts can reasonably assess each borrower's riskiness, and the cut-off level can be changed in accordance with the risk of each loan portfolio. As a result, credit scoring gives lenders the capacity to manage risk.

Additionally, this computerized process enables bank employees to periodically assess the borrowers' creditworthiness. As a result, it is simple to monitor the risk associated with each loan.

Not to mention, the fact that credit scoring is founded on historical data is a benefit. This shows that because of the database and credit history stored in the system, lenders can more accurately forecast whether the next application will perform well or default.

Despite the fact that credit scoring lowers costs and improves the efficiency of the loan-granting process, its drawbacks should not be overlooked.

First of all, since credit history scoring models are simply dependent on information contained in credit reporting agency files, they have the advantage of being affordable and trustworthy quick screening tools that can be used to monitor just about any prospective consumer. However, they likely have the drawback of being less accurate than models based on a wider collection of data because they are based on less information than that frequently utilized in consumer credit screening. Siddiqi (2017) highlight the practical challenges involved in developing credit scoring models that incorporate situational data as well as the possible implications of not incorporating situational data into consumer credit evaluations. These difficulties result from the inherent limits of the databases utilized by credit reporting agencies to create numerous scoring models.

The use of credit scoring also requires careful consideration of accuracy. If the models are inaccurate, loans that are performing poorly could have a detrimental impact on the cost reductions and other benefits of credit scoring. The complexity of credit scoring algorithms means that they can only be effective with complete and reliable data. Otherwise, the model will produce findings that are not precise. The information utilized in credit assessment should contain a sample of both loans with good performance and those with poor performance.

Regular data reviews and periodic model re-estimations are required to guarantee that any changes in the link between prospective factors and loan performance are captured.

The fact that the borrower's characteristics are strongly connected with their likelihood of repayment and defaults is another crucial component of scoring models. Although credit models try to predict the likelihood of a loan defaulting, understanding the borrower is a need. Because credit scoring cannot replace loan officer choices, which are based on informal qualitative information, credit scoring algorithms frequently include human mistake. In order to cope with clients who have not had impeccable credit, banks must consider their history in addition to credit rating when making credit decisions.

Making predictions when the economy is either recovering or in a recession is a crucial component of an accurate model. Therefore, the model's data should include both eras of strong and weak economic growth.

Credit scoring algorithms are used to forecast the likelihood of default, as was already mentioned. The models, however, often only employ a sample of accepted applications. This bias in selection may result in inaccurate credit score model estimation. According to Zeng, (2018), banks should use the credit scoring model on loans that have already been conditionally approved by credit officers in order to prevent bias in the model. For instance, the First National Bank of Chicago used credit modeling to reject around 25% of small company loan applications; however, the credit officers eventually approved the same applications. Finally, the bank's credit evaluation can be a combination of traditional lending through credit scoring models in order to eliminate bias in credit scoring.

2.2.4 Variables commonly used in credit scoring

When determining a borrower's creditworthiness, the model must take into account all of the borrower's traits and information that are readily available and have a clear relationship to default risk. If the variables are sequentially added or removed, the model's predictive accuracy is at its highest (Siddiqi, 2017). For choosing a variable, there are two key factors. First, the factors must be substantial and contribute to the variance of the dependent variable's explanation. Second, there should be a strong association between the variables and the other variables. According to Mao (2018), there is no requirement that each variable be justified. Use it if it makes the predictions better. The primary variables in credit scoring models, however, are the borrower's personal characteristics, such as their income, age, gender, education, occupation, region, length of time at their current address, residential status, and marital status, followed by their banking relationship, such as the value of their collateral, the length of their

loan, how long they have been with their bank, how many loans they have, and their current account balance (Peprah et al., 2017).

Income is the borrower's yearly income, which is a frequent indicator of the borrower's financial well-being and repayment capacity (Peprah et al., 2017). Because a higher income is associated with a lower chance of default, income and the default rate of borrowers are positively connected (Peprah et al., 2017).

Regarding occupation, this is a factor in credit scoring that is strongly correlated with income.

Education improves borrowers' capacity to repay loans. Borrowers can be categorized based on their level of education because those with greater levels of education are thought to have more stable jobs with higher incomes and a reduced likelihood of defaulting.

Employer designates the sort of business a borrower works for, including state-owned, foreign, joint-stock companies, etc. This factor is crucial since the borrower's company type may serve as a good proxy for their income level and financial stability. Since the highest default rates are displayed by borrowers who do not respond to this question, missing values for this variable are likewise quite instructive.

The borrowers' length of employment with their present employer is indicated by the variable "time with employer." It demonstrates how content the borrower is with their current employment. Borrowers' ability to repay loans will be better and their employment will be more stable if their job satisfaction is higher (Peprah et al., 2017). It should be made clear that a woman may be discriminated against based on how long she spends working for a company because of pregnancy and childbirth.

The borrower's age is expressed in years. It is confirmed by Peprah et al. (2017) that elder borrowers are less risk-averse and less prone to default. As a result, banks are less likely to lend to younger, riskier borrowers.

There is ample evidence that women default on loans less frequently than males, presumably because women are more risk averse, despite the fact that gender is often regarded to be prejudiced in industrialized nations due to statistical differences between men and women (Francis et al., 2022).

The borrower's region represents where in the nation they reside. The postal code is a popular stand-in for this variable. Because those with comparable wealth are more likely to reside in the same area, the geographic criterion can indicate the financial wealth of a borrower.

Some suburbs might draw wealthier people, which might cause real estate property values to rise. Additionally, this affects the collateral value and default likelihood.

Whether borrowers own their home, rent it, or remain with their parents is indicated by their residential status. If the borrower owns a home, this variable can represent their financial prosperity. Additionally, a borrower's home status denotes the financial strain that insurance premiums, taxes, or utility prices may place on their income. According to Francis et al., (2022), debtors who live with their parents had a lower default rate.

The number of years the borrowers have resided at their current location is referred to as "time at present address." According to Francis et al., (2022), the time spent at the current address and the default risk have a negative correlation, showing that the latter could be a proxy for the maturity, stability, or risk aversion of the borrower. Changing your address could mean your financial situation is strong or getting better quickly.

The borrower's level of maturity, dependability, or accountability is influenced by their marital status. According to statistics, married borrowers experience higher default rates than single borrowers. According to research by Bonga et al. (2019), the number of dependents and marital status are both related to the financial strain that a borrower is under and their capacity to repay a loan.

A sort of guarantee used to reduce the risk of default for borrowers is collateral. Demanding collateral for retail loans may be an indicator of danger. For instance, the likelihood of default is quite low if the loans are secured by real estate. This is because the borrowers are risk averse and anxious about losing their homes. The greater the value of the collateral, the greater the motivation for the borrowers to repay the loan because they would keep the collateral. Given that the collateral value has a strong positive correlation with the borrowers' income, it may also serve as a proxy for their financial wealth (Bonga et al., 2019).

The number of months a loan will take to mature is its loan duration. This variable is a result of the discussion between the borrower and the bank. If the strain on the borrower's income is reduced, there is a chance that the borrower will be accepted for a longer loan of the same size but rejected for a shorter one. Loan term reflects the intention, risk aversion, or self-evaluation of payback capacity of the borrower.

The number of years the customer has been a bank customer is expressed as "Time with the bank." One can presume that the longer a borrower stays with the bank, the more information the bank has about them, and the less likely they are to fail. However, because the condition of the borrowers is subject to unforeseen changes, this variable should be updated frequently.

The total amount of loans a borrower has gotten from the bank over the course of their relationship is counted. Many borrowers have a history of loans, and many times they have multiple loans from the same bank. This proxy provides information about the borrower's default risk because a borrower who has not repaid a loan that has already been given will find it challenging to obtain a new loan. Therefore, this variable captures how challenging it is for a defaulted borrower to obtain new loans from the same bank.

Whether the borrower has a current account with the bank is indicated by the binary variable current account. This variable is significant and provides some insight into the financial well-being of borrowers as well as their interactions with banks. The risk of default is reduced for borrowers who have current accounts with their banks.

2.2.5 Credit-Scoring Evaluation Techniques

Banks and other financial institutions don't just hand out loans to anyone who asks for them; instead, they undertake an assessment to gauge the applicants' level of risk before deciding whether or not to extend credit. In other words, whether or not a loan can be granted depends on whether the features of current applicants are comparable to those of previous applicants (whether they defaulted or not) and on the basis of their past performance. In general, there are two methods or approaches that can be used to achieve the results, or scores produced by the scoring systems, namely the statistical approach and the judging approach (Siddiqi, 2017).

a. Judgmental Scoring Systems

Prior to the use of numerical scoring systems, the traditional technique of deciding whether to award consumers credit in banking systems was based on the credit officer's judgmental and subjective opinion (Basu, 2017). Additionally, these arbitrary choices incorporate standards and other credit regulations established by bank policy (Hussain, 2019). Basu (2017) claims that in judgmental scoring systems, the borrower is awarded points or weights based on specific characteristics; they are then weighted and converted into a score, which determines whether or not to provide a loan. The credit officer makes the ultimate decisions based on his experience, common sense, and straightforward numerical backing. According to Peprah (2017), the well-known 5Cs are useful in determining a borrower's creditworthiness and include (1) Character (the borrower's background and reputation); (2) Capital (the borrower's contribution to the investment); (3) Collateral (guarantees to back-up the loan in case of default); (4) Capacity (the borrower's financial ability to pay the loan); and (5) Condition (the overall economy of the borrower).

The efficacy and dependability of judging techniques are frequently contested. According to Basu (2017), on the one hand, the value of judgmental evaluation has come under fire because of some drawbacks, including the potential for human error, high training costs, and inconsistent application of credit policies among credit officers. As a result, lenders are looking for more computerized ways to carry out credit evaluation and decision-making. The outcomes of the judgmental approach are ineffective, inconsistent, and lacking in uniformity, according to Siddiqi (2017).

Hussain et al (2019) claims that, in contrast, loan choices are still made using evaluative procedures that are based on sparse or unstructured data as well as personal experience. Additionally, Munguti (2020) cites three possible reasons why lending institutions have resisted deploying credit-scoring systems: a desire to retain highly experienced credit officers, potential model flaws, and a lack of quantitative expertise in credit management. Additionally, they think that the credit-granting process falls short of statistical system use. However, because of the consumer credit sector's rapid expansion and the volume of data it generates, banks and other financial institutions frequently use objective, quick, uniform, and fast processes in instead of or in addition to judgmental approaches (Munguti, 2020)). These strategies are constructed utilizing statistical methodologies that have been continually improved throughout time.

b. Credit-Scoring Systems

As previously mentioned, the rapid expansion of consumer lending and technology in recent decades has compelled banks and financial institutions to upgrade their credit strategies to a level that can handle this growth, with modifications made to the way credit is assessed and increased engagement in more sophisticated methods that can replace and overcome the shortcomings of their approaches.

In his article, Dastile et al (2020) highlighted the value of employing statistical credit-scoring. He demonstrated how lenders are becoming more aware of how statistical credit-scoring can be a better alternative to judgmental approaches when considering objectivity, consistency, and reliability. Quantitative and statistical techniques were not widely used until the early 1960s, when the required computer technology was developed. This was made worse by economic pressures, which subsided as a result of the creation of the credit-scoring system, an objective framework for making credit choices (Dastile et al, 2020).

According to Dastile et al (2020), today's credit-scoring is based on statistical or operational techniques including neural networks, discriminant analysis, logistic regression, and decision trees. Hussain et al (2019) claims that banks experiment with credit scoring in an

effort to lower default rates and gain better control over their loan practices. They discovered that credit scoring had several benefits, including "reduced processing costs, more effective credit control, racial and ethnic lending nondiscrimination, ease of altering credit standards, and speedier loan approval decisions." Additionally, banks saw a growth in their clientele without a commensurate rise in delinquency rates. Objective approaches may be a useful tool for comprehending the credit evaluation process up to the decision-making stage, according to Peprah, et al. (2017).

On the other hand, despite the expansion of consumer financing, it might be claimed that using statistical methods in industries with a small clientele can be expensive. Thomas et al (2017), argued in favor of the employment of judgmental evaluation in circumstances where the statistical scoring system is unable to do so, such as when a highly valued customer who can generate significant profit is involved. According to Thomas et al (2017), utilizing statistical scoring has advantages and disadvantages that are restricted when compared to using judgmental evaluation while making credit choices. The next parts go into great length into credit-scoring methods (statistical and machine-learning), along with their empirical research.

Quantitative Credit-Scoring

When creating a credit-scoring model, the primary goal is to identify the best classification methods that can distinguish between good and bad credit and, in turn, predict new loan applications. In the field of finance, and specifically in the field of banking, credit-scoring models are used extensively. To develop credit score, a variety of classification techniques can be utilized, ranging from statistical (such as LDA, LR, and NB) to machine-learning (such as NN, SVM). Fisher (1936) was the first to utilize statistical methods to address a classification problem, and the Fair Isaac Company proposed using them to develop credit scoring in the late 1960s (Thomas et al, 2017). Up until the advent of machine-learning or artificial intelligence (AI) approaches, which was sparked by the development of computer technology, statistical techniques had been used in the development of credit-scoring procedures. However, these methods are thought to perform better than statistical methods (Siddiqi, 2017). The primary distinction between statistical techniques and machine-learning techniques is that statistical techniques concentrate on analyzing already-existing data and studying their relationships while making assumptions in order to predict an outcome, whereas machine-learning techniques concentrate on building systems that can learn directly from the data that is already present (Siddiqi, 2017).

Despite the wide range of variations in the approaches and methods employed, the basic goal is to develop a model that can forecast borrower loan applications and accurately categorize and measure borrowers' payback behavior (Thomas et al, 2017). This part provides an overview of the most popular modern statistical and machine-learning categorization approaches that are pertinent to this study and are used to create credit-scoring models.

2.2.6 General Supervised Algorithms

Linear Regression

A machine learning algorithm called linear regression carries out a regression job following supervised learning. In order to forecast and understand the link between variables, the model looks for a prediction value based on the independent variables. The following describes the linear regression hypothesis function:

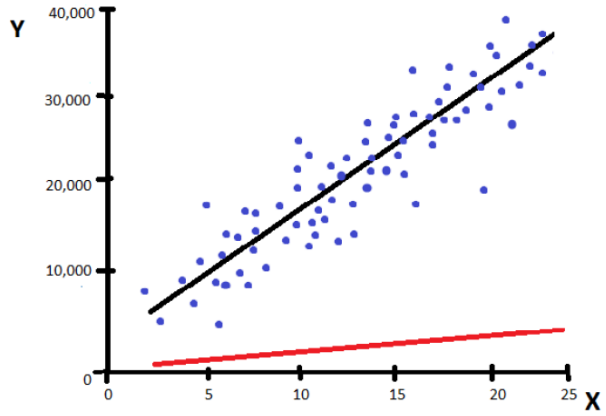
$$Y = ax\theta 1 + \theta 2 \quad (2.11)$$

When the model is being trained, the *x* and *y* variables are provided.

As the model is being trained, a best fit line is generated that forecasts the label (*y* values) for a certain input value (*x* values). Finding the ideal values of $\theta 1$ and $\theta 2$ results in the best regression fit line.

FIGURE 1:

Linear Regression Intuition



Training & loss

Cost Function (J) : The model's objective is to predict y values with a minimum error between the projected value (y) and the actual value (x). To achieve the least amount of inaccuracy, it is therefore required to update the values of $\theta 1$ and $\theta 2$.

$$minimize = (\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2) \quad (2.12)$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \quad (2.13)$$

The Mean Squared Error (MSE) between the expected and true value, then, serves as the cost function.

Gradient Descent:

By adjusting the values of θ1 and θ2, gradient descent is utilized to obtain the best fit line. As a result, the MSE value is minimized, and the cost function is reduced. The model changes the values by iterating in order to arrive at the minimal cost by starting with random values for θ1 and θ2. The objective is to identify the best θ1 and θ2 parameters.

K-Nearest Neighbor (KNN)

A non-parametric pattern recognition technique called KNN is employed for both classification and regression applications. Finding the k-closest training examples in the feature space is the underlying idea, and the result is dependent on the characteristics of those k-closest neighbors. Because it is categorized based on the majority vote of its neighbors. It is given the average value of its neighbors for regression.

In pattern recognition, the KNN algorithm is a method for classifying objects based on the nearby training examples in the feature space. When utilizing KNN, only local approximations of the function are employed, and all calculations are postponed until classification (Zhang, S.,2017).

When there is little to no prior knowledge about the distribution of the data, the *KNN* is the most basic and elementary classification technique (Shah, K.,2020). This rule maintains the integrity of the whole training set throughout learning and categorizes each query according to the majority label of its *K*-nearest neighbors in the training set. The most straightforward application of *KNN* is the Nearest Neighbor rule (*NN*) when $K = 1$.

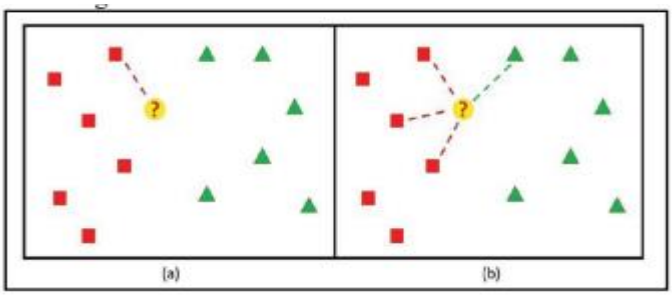
Each sample is categorized using this method in the same manner as the samples around it. This makes it possible to anticipate unknown samples using the samples around them. Given a training set and an unknown sample, one can determine the separations between each sample in the training set and the unknown sample. For the sample, there is the smallest difference between it and the sample from the training set. Consequently, the classification of the unknown sample's nearest neighbors can be used (Zhang, S.,2018).

The *KNN* decision rule is depicted in *Figure 2* for two classes of samples with K=1 and K=4. *Figure 2*(b) illustrates how to categorize an unknown sample using a range of known

samples, in contrast to *Figure 2*(a), which shows how to do so using just one known sample. The parameter *K* is set to 4 in the final scenario, which means that the four examples that are closest to the unknown sample are utilized to classify it. Three of them are from the same class, whereas just one of them is from the other class. The unidentified sample is classified as being a member of the class on the left in both cases. The *KNN* method is displayed in *Figure 3*.

FIGURE 2.

(a) The 1-NN decision rule: the point is assigned to the class on the left; (b) the KNN decision rule, with K= 4: the point is assigned to the class on the left as well



Source: Mohammad B (2013).

FIGURE 3.

The KNN Algorithm.

```
for all the unknown samples
UnSample(i)
    for all the known samples Sample(j)
            compute the distance
between
UnSamples(i) and Sample(j)
       end for
       find the k smallest distances
        locate the corresponding samples
Sample(j1),..,Sample(jk)
       assign UnSample(i) to the class
which
appears more frequently
end for
```

Source: Mohammad B (2013).

The value of *K* and the distance measure used have the most effects on a *KNN* classifier's performance (Zhang, S.,2018). The estimate is sensitive to the choice of neighborhood size *K* since the radius of the local region is determined by the distance from the Kth nearest neighbor to the query. Furthermore, different *K* values result in different conditional class probabilities. Due to data scarcity, noisy, ambiguous, or mislabeled points, a

very bad local estimate is expected if *K* is very small. We can further smooth the estimate by increasing *K* and considering a large area around the question. Unfortunately, a high value of K easily results in an estimate that is excessively smooth, and the effectiveness of classification suffers when outliers from other classes are included. Related research has been done to enhance the *KNN*'s classification performance in order to overcome the problem.

The Best Neighborhood Size to Choose *K* is a key variable that has a big impact on how well *KNN* classification works. In the case of *KNN*, a small training sample size can have a significant impact on choosing the appropriate neighborhood size *K*, and the KNN's classification performance may suffer the effects of the sensitivity with which *K* is selected. Due to data scarcity, the classification results can be significantly impacted by noisy, ambiguous, or mislabeled points if K is too little or by a large number of outliers from neighboring classes if *K* is too large. Theoretically, the calculation of the query's conditional class probabilities in a particular region of the data space, which is determined by the distance of the query's *Kth* nearest neighbor, determines the classification performance of *KNN*. The *K* value used has a considerable effect on classification performance as a result. Furthermore, the simplest majority voting of combining the class labels for *KNN* can be problematic if the distances between the closest neighbors varies noticeably and the closer ones better represent the class of the query item. The sensitivity issue with various neighborhood size *K* options has been addressed using a few weighted voting algorithms for *KNN*.

It has been shown that finding the value of *K* gets challenging when the points are not spread evenly. Larger values of *K* typically have smoother class boundaries and are more resistant to the noise that is given. As a result, it becomes almost difficult to choose the same (optimal) *K* for several applications.

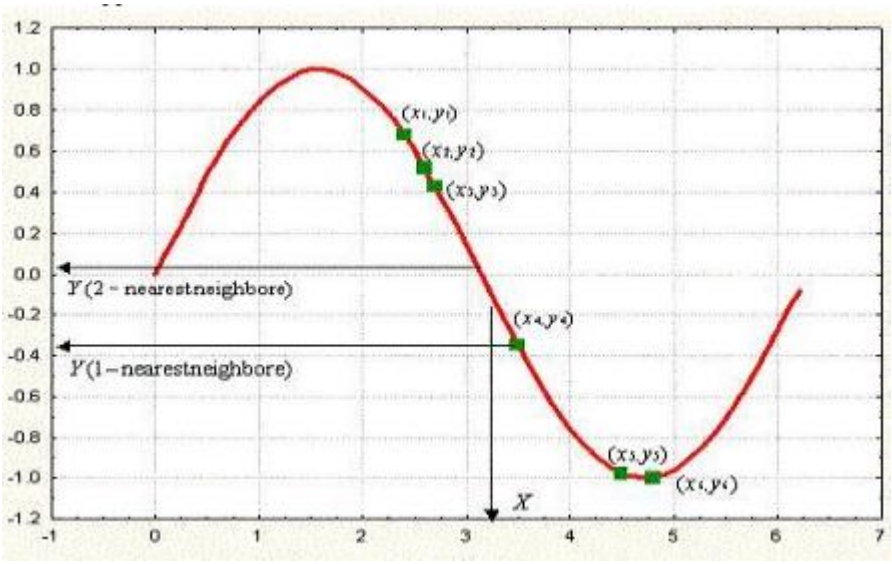
KNN for Regression

I. Theory

Regression can be performed using the same way by simply averaging the values of the object's *K* Nearest neighbors' properties. It can be useful to weight neighbor contributions so that closer neighbors give more on average than distant neighbors.

FIGURE 4.

The KNN Decision Rule for Regression



Source: Mohammad B (2013).

Each sample is categorized using this method in the same manner as the samples around it. This makes it possible to anticipate unknown samples using the samples around them. Given a training set and an unknown sample, one can calculate the separations between each sample in the training set and the unknown sample. For the sample, there is the smallest difference between it and the sample from the training set. Consequently, the classification of the unknown sample's nearest neighbors can be used (Zhang, S.,2018). It so happens that in this situation, this is x_4. Thus, it is presumed that the answer to the query of X (i.e., Y) is the result of x_4 (i.e., y_4) . As a result, Y can be expressed as Y 4 for the nearest neighbor.

The following stage is to take into account the 2-nearest neighbor algorithm. The first two positions closest to X in this situation are y_3 and y_4, which are found. The answer to Y is as follows when averaging their outcomes:

$$Y = \frac{y_3+y_4}{2} \qquad (2.14)$$

Any quantity *K* of nearest neighbors can be added to the previous discussion. A *KNN* technique, in essence, presupposes that the result Y of the query point X is the average of the outcomes of its *K* nearest neighbors.

II. Distance Metric

KNN bases its forecasts, as was already mentioned, on the outcomes of the K sites closest to the target location. The distance between cases from the examples sample and the query point must be measured in order to apply KNN for prediction. The term "Euclidean," or simply

"Euclidean," refers to one of the most widely used methods of calculating this distance. Euclidean squared, City-block, and Chebychev are other metrics.

$$(x,p) = \begin{cases} \sqrt{(x-p)^2} & Eucledian \\ (x-p)^2 & Eucledian\ Squared \\ |x-p| & Cityblock \\ Max(|x-p|) & Chebychev \end{cases}$$

where x and p, respectively, represent the query point and a case from the examples sample.

III. K-Nearest Neighbor Predictions

Following your selection of K's value, you can use the *KNN* examples to inform your predictions. The average outcome of the *K* nearest neighbors represents *KNN* regression prediction:

$$y = \frac{1}{k}\sum_{i=1}^{k} y_i \qquad (2.15)$$

where y is the anticipated value of the query point and y_i is the ith case of the examples sample (outcome). In contrast to regression, the winner of a vote process is utilized to label the query in *KNN* predictions for classification issues.

KNN analysis haven't been discussed yet without taking into account how close the *K* earliest samples are to the query point. To put it another way, *K* neighbors are enabled to influence forecasts equally regardless of how close or far they are to the query location. An alternate strategy is to give examples nearest to the query point greater weight by using arbitrarily high values of *K* (if not the full prototype sample). "Distance weighting" technique is used to achieve this.

Distance Weighting

It becomes sense to differentiate between the K nearest neighbors while creating forecasts in order to give the query, point additional effect from the K nearest neighbors that are closest to it. KNN predictions are based on the idea that similar objects should be comparable. Adding a set of weights W, one for each nearest neighbor, based on how close they are to the query point is one way to accomplish this. Thus:

$$W(x,p_i) = \frac{\exp\left(-D(x,p_i)\right)}{\sum_{i=1}^{k}\exp\left(-D(x,p_i)\right)} \qquad (2.16)$$

where $D(x,p_i)$ is the separation between the query point x and the *ith* case p_i of the example sample. It is clear that the weights listed above will correspond to:

$$\sum_{i=1}^{k} W(X_0, X_i) = 1 \quad (2.17)$$

As a result, for regression problems, we have:

$$y = \sum_{i=1}^{k} W(X_0, X_i) y_i \quad (2.18)$$

The maximum of the aforementioned equation is applied to each class variable in classification tasks. The argument presented above clearly shows that when K>1, the standard deviation for predictions in regression tasks can be defined naturally by using,

$$\text{err bar} = \mp \sqrt{\frac{1}{K-1} \sum_{i=1}^{k} (y - y_i)^2} \quad (2.19)$$

The study's algorithm operates as follows:

- Determine the Euclidean distance between the training instances and the labeled ones.
- Sort the labeled examples in ascending order of distance.
- Using Root Mean Squared Error (RMSE), determine the heuristically ideal number of nearest neighbors, *K*.
- Determine the uniform weighted average using the k-nearest multivariate neighbors.

2.2.7 Ensemble Models

Decision Trees and Random Forests

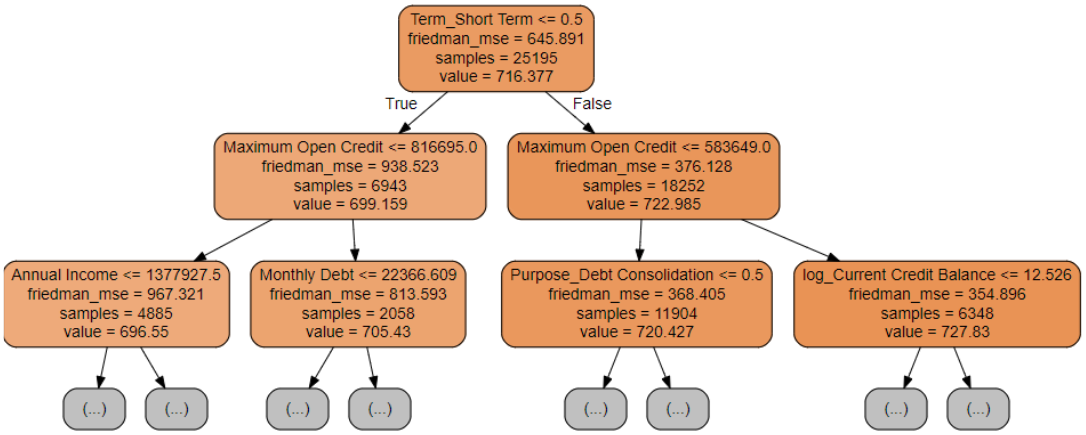
a) Decision Trees

A decision tree is a technique that uses a tree structure resembling a flowchart, a collection of decisions, and every conceivable outcome, including input cost and utility. A supervised learning approach called decision-tree is effective for both continuous (regression) and categorical (classification) output variables.

The primary characteristic of decision trees is their capacity to recursively subset a target field of data based on the values of related input fields or predictors to create partitions and associated descendent data subsets (referred to as leaves or nodes), which contain target values that are progressively similar within leaves (or within nodes) and progressively different between leaves (or between nodes) at any given level of the tree (Patel, H. H., 2018).

FIGURE 5:

Decision Tree Path



An illustration of how decision trees use binary splits (Yes or No) on conditions to produce the best split between the sample data is shown in Example 3.2 from our analysis. All the input samples are contained in the root node ($x_1, x_2, \ldots\ldots, x_n$). The samples are then divided based on specific characteristics to maximize information gain.

- Each internal node of the tree represents an attribute or feature.
- An associated label or prediction is found at each leaf node.

J. R. Quinlan's ID3 algorithm, which implements a top-down, greedy search through the space of potential branches without backtracking, is the fundamental algorithm for creating decision trees [2]. Regression decision trees can be created using the ID3 algorithm by swapping Information Gain for Standard Deviation Reduction.

Top-down construction of a decision tree involves dividing the data into sets of samples with related values (homogenous). When determining the homogeneity of a sample instance, standard deviation is used. When a sample's standard deviation is zero, it is said to be entirely homogeneous.

$$(\text{StandardDeviation})\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \text{ where } \textit{(Mean)}\ \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{Coefficient Of Variation(CV)} = \left(\sigma/\mu\right) \times 100$$

Standard deviation for two variables (Target, Feature):

$$S(T; X) = \sum_{c \ni X} P(c)S(c)$$

The decision node is assigned to the characteristic that has the greatest standard deviation decrease.

The standard deviation is decreased as part of the standard deviation reduction process after splitting a data set by an attribute. Find the characteristic that reduces standard deviation (SDR) the most in order to design a decision tree (i.e., the most homogeneous branches). The decision node makes use of the feature with the highest standard deviation reduction (SDR).

$$SDR(T,X) = S(T) - S(T,X) \quad (2.20)$$

Based on the values of the selected attribute, the data set is partitioned. Up till every sample in the dataset has been analyzed, the procedure is repeated on the non-leaf branches.

A common recursion criterion is the coefficient of variation (CV). We halt the splitting process and assign the average value at the leaf node to that subgroup if the CV for a specific branch is less than a predefined threshold, say 10%.

b) Random Forests

A novel approach to ensemble supervised machine learning is Random Forest. In data mining, machine learning techniques are used. The two main categories of data mining are descriptive and predictive. The primary goals of descriptive data mining are to describe, categorize, and summarize the data. Predictive data mining looks at historical data to find patterns or make inferences about the future. The process of developing statistical models traditionally is where predictive data mining got its start. Feature analysis of predictor variables is the basis for the creation of predictive models. One or more qualities may theoretically function as predictors. The result, which is a function of the predictors, is represented by the hypothesis. To determine if the developed hypotheses are true or not, they are put to the test. This model's accuracy is evaluated using a number of error estimation techniques. While descriptive data mining often employs unsupervised machine learning techniques, predictive data mining frequently makes use of supervised machine learning techniques.

Labeled data samples are used in supervised machine learning to group samples into different categories. The training dataset is where the predictive model learns. The model's accuracy is calculated using the test dataset. One well-liked supervised machine learning method is the decision tree. The underlying classifier used by Random Forest (Paul, A., 2018) is a decision tree. Numerous decision trees are produced by Random Forest, and two distinct randomization processes take place: (1) random data sampling for bootstrap samples (similar to bagging), and (2) random selection of input traits for creating unique base decision trees. The generalization error of a Random Forest classifier depends on the power of each individual

decision tree classifier and the correlation between base trees (Paul, A., 2018). It has been discovered that the Random Forest classifier's accuracy is on par with that of contemporary ensemble techniques like bagging and boosting. Breiman (2001) claims that Random Forest has the following advantages: it can handle tens of thousands of input variables without deleting any variables; it provides estimates of significant variables; it generates an internal, unbiased estimate of generalization error as the forest expands; it has a useful method for estimating missing data and maintaining accuracy when a significant portion of the data is missing; and it has techniques for balancing class error. Due to Random Forest's inherent parallelism, parallel implementations have utilized multithreading, multi-core, and parallel architectures. Numerous classification and prediction systems now in use make use of Random Forest due to the attributes listed above.

Random forests frequently outperform other constituent trees because of the vast number of uncorrelated trees that work together as a committee to achieve superior results. Low correlation in random forest models or trees is a benefit since low correlations cluster and bind together to produce predictions that are more accurate than the sum of their individual forecasts. This argument is supported by the effectiveness of the trees' mutual defense against one another's flaws. The following model has a better chance of picking the appropriate path because there are many trees that could be right or wrong (Tonester524, 2019).

As a result, in order for random forest to function perfectly, the following conditions must be met:

- For the model to perform better at random guessing, features must have an actual signal, which implies they must have some predictive potential.
- There shouldn't be much of a correlation between specific tree flaws and expected outcomes. The features and hyper-parameters selected will have an impact on the correlations even though the program tries to build these correlations for us using feature randomness.

Gradient Boosted Regression

It is a common challenge in many machine learning applications to build non-parametric regression or classification models from data. Starting with theory and adjusting the model's parameters based on seen data is one method for developing models in domains that are specific to that theory. Unfortunately, most real-world scenarios do not have access to such models. The researcher typically has little access to even preliminary expert-driven hypotheses about potential correlations between input variables. The lack of a model can be avoided by using

non-parametric machine learning methods to create a model directly from the data, such as neural networks, support vector machines, or any other algorithm of one's choice. These models must have the necessary target variables ready in advance because they are supervised models.

Making a single, potent prediction model is the most typical method to data-driven modeling. An alternative technique would be to create a group of models, or an ensemble of models, for a certain learning assignment. Consider creating a series of "strong" models, such as neural networks, which can be combined to produce a more precise prediction. However, in practice, the ensemble technique relies on combining a number of weak fundamental models to provide a more reliable ensemble prediction. Neural network ensembles (Li, W.,2021) and random forests (Paul, A., 2018) are two well-known examples of machine learning ensemble approaches that have found many successful applications in a variety of disciplines (Rokach, L. 2019).

Simple model averaging is a common ensemble technique used in random forests and other techniques. The foundation of the boosting method family is an original, beneficial approach to ensemble building. Adding more models to the ensemble incrementally is the fundamental concept underpinning boosting. Every iteration, a new weak, base-learner model is taught in reference to the mistake of the previously learned entire ensemble. Due to the fact that the earliest well-known boosting approaches were totally algorithm-driven, it was challenging to thoroughly analyze their characteristics and effectiveness (Li, T. R., 2019). Many ideas regarding why these algorithms either performed better than every other approach or, conversely, were inapplicable because of significant overfitting were spawned by this (Grover, 2019).

A formulation of gradient-descent boosting methods was created in order to link boosting techniques to the statistical framework (Freund and Schapire, 1997; Friedman et al., 2000; Friedman, 2001). After the development of boosting techniques and the accompanying models, the gradient boosting machines were given their name. By supplying the explanations for the model hyperparameters, this technique also lay the groundwork for future gradient boosting model development.

Gradient boosting machines, or GBMs, sequentially fit new models as part of the learning process to provide a more precise estimate of the response variable. The basic idea behind this method is to construct new base-learners that are maximally correlated with the negative gradient of the ensemble associated loss function. To be explicit, if the error function is the conventional squared-error loss, the learning process will lead to sequential error-fitting.

You can pick loss functions at random. The loss function is often chosen by the researcher from a large pool of previously derived functions or from a task-specific loss that can be applied.

GBMs are highly adaptable and can be used for any data-driven job. It adds a great deal of flexibility to the model design, making it difficult to predict which loss function will work best. However, because boosting techniques are very easy to implement, several model designs can be tested. Additionally, GBMs have shown substantial effectiveness in a variety of machine learning and data mining problems in addition to practical applications (Grover, 2019).

Ensemble models are a useful practical tool for a variety of predictive tasks in neurorobotics since they routinely outperform standard single strong machine learning models in terms of accuracy. To detect and track human movement, for instance, ensemble models may successfully translate EMG and EEG sensor signals. However, these simulations of memory and brain development can also provide insightful information. The base-learners act as the memory medium in boosted ensembles, creating the captured patterns sequentially and gradually enhancing the level of pattern detail. Unlike boosted ensembles, artificial neural networks do not maintain learned patterns in the connections of artificial neurons. The field of brain simulation can benefit from improvements in boosted ensembles since network development strategies and ensemble formation models can be merged. Base-learners can be built into ensembles with different graph features and topologies, such as the small-world networks seen in biological neural networks, if they are viewed as network nodes, which in the context of the connectome will indicate neurons. It is required to first describe the approach and algorithmic basis for boosted ensemble models in order to enable sophisticated neurorobotics applications.

Since this is an ensemble method, several predictors are used in place of only one to create the prediction. GBR employs the Boosting technique, which generates the predictors consecutively, as opposed to Random Forests, which construct the decision tree predictors independently. This method makes use of the technique whereby the succeeding forecasters pick up on the errors of the preceding predictors. The samples are picked based on the mistakes made by the preceding predictors rather than using bootstrapping. Faster convergence and a tighter match to real predictions are the results of this, however stopping conditions must be carefully specified to avoid overfitting the training data. GBR reduces bias and variance (Grover, 2019).

The approach reduces the loss iteratively using gradient descent after creating a loss function.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (2.21)$$

$$\hat{Y}_i = \hat{Y}_i + \alpha \times \delta \sum\left(Y_i - \hat{Y}_i\right)^2 / \delta \times \hat{Y}_i \quad (2.22)$$

$$\hat{Y}_i = \hat{Y}_i - \alpha \times 2 \sum\left(Y_i - \hat{Y}_i\right)^2 \quad (2.23)$$

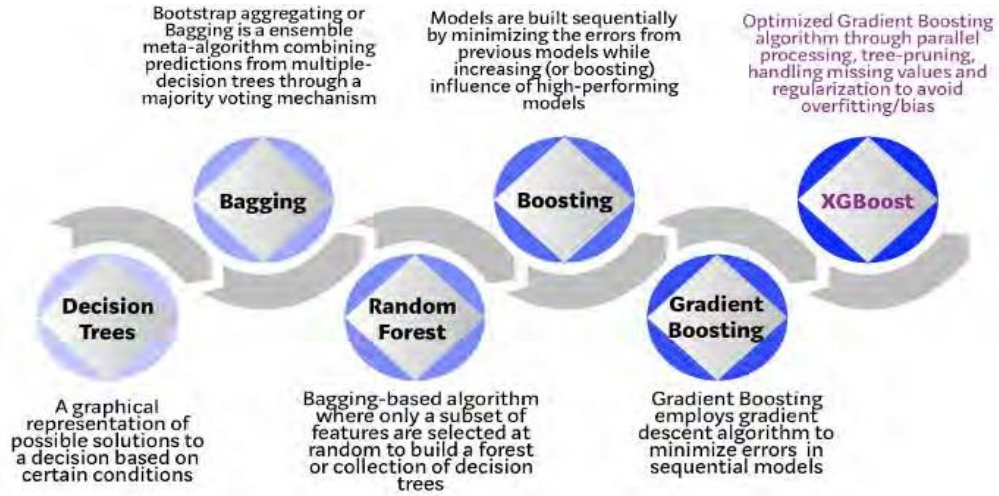
Where, α is the learning rate, and $\sum\left(Y_i - \hat{Y}_i\right)^2$ is the sum of residuals.

XGBoost

XGBoost is a decision tree-based machine learning method. It makes use of the gradient boost framework. XGBoost has been used to successfully address a variety of difficulties, including classification, regression, ranking, and user-defined predictions. This method is portable since it works with almost all widely used programming languages and runs on Windows, Linux, and OS X. XG Boost also supports cloud connectivity with AWS, Azure, and Yarn Clusters. The diagram below depicts the evolution of XG boost from decision trees:

FIGURE 6:

XGBoost Overview



Source: Vishalmorde, 2019

Machines that boost gradients efficiently and accurately can use XGBoost. It has the capacity to exceed the computational capacity of boosted tree algorithms. This algorithm's development objective was to increase model performance and computing speed by making the most of all available hardware and memory. The block structure of XGBoost allows for concurrent tree construction, handling of missing values, and excellent fit and boost on newly added data to the training model. According to Tianqi Chen, the creator of the system, the

method uses a more regularized model formalization to control overfitting and improve performance.

2.3 Empirical Review

The evaluation of credit risk is a crucial subject in banking and finance. Since its inception, statistics and human judgement have been closely connected. On the other hand, credit risk assessment utilizing pattern recognition and machine learning has greatly increased interest in the academic community as a result of recent advancements in data science and machine learning. In this area of study, a number of significant research papers have been published that have drawn a lot of attention. These papers often employed artificial neural networks. Information-processing patterns called artificial neural networks, or ANNs, are based on the biological nervous system. The output of Jiang, Q (2022)'s RBF multilayer feed forward network was compared to that of a conventional logistic regression model. They got to the conclusion that the Logistic Regression model was superior for positive classification, while the Neural Network model was superior for negative classification. G. V. Attigeri (2017) used the chi-square test to grade the defaulters in a comparable comparison. The logistic regression model outperformed the neural network when trained with 1,000 examples under supervision. However, ensemble techniques have received widespread praise from the scientific community.

The most efficient neural network models were chosen via de-correlation maximization by Golzadeh, M. (2018), who employed neural networks to build an ensemble agent. Multiple ensemble approaches, including mean, median, max, min, and product, were used to integrate the model results. Single-based agents (such Logistic Regression, Support Vector Machines, and ANN), hybrid agents (like Neuro-fuzzy, Fuzzy SVM), and a voting-based reliability ensemble approach were all employed to compare the ensemble methodology. Compared to the other models, the reliability-based neural network agent marginally outperformed them. Additionally, it has been highlighted that several applications of distributed Random Forests, the Gradient Boosting Method, and generalized linear modeling have yielded notable results. R. G. Lopes used GBM on a dataset from a Brazilian bank (2016). The study also made use of the Random Forest and the Generalized Linear Model. The models were 70% trained using more than 20,000 cases. With an AUC Score of 99, GBM outperformed the other techniques significantly. To determine credit risk, Classification and Regression Trees (CART) or Decision Trees are widely utilized. In [34], the authors compared and contrasted neural network-based methods with those based on tree-based models. A comparison of the LogR,

GBM, Random Forests, and Neural Network models was provided in this proposal. Computing became more accurate as cost fell. Elastic net was used to train the lambda and alpha hyperparameters of LogR in order to prevent "over-regularization." There were 120 trees in the GBM and Random Forest. The next step was to create four neural network models using various regularization strategies and hidden layer counts. The drop out ratio, activation functions, layers, and regularization functions were all optimized using grid search. As evaluation measures, the RMSE and AUC Score were applied. The outcomes demonstrated that neural network and LogR models were outperformed by tree-based methods such as Random Forests and GBM. A comparison of decision tree, artificial neural network, naive bayes, k-nearest neighbor, and linear discriminant analysis models was also carried out in 2010 by B. Twala. Ensemble models were also made using these classifiers. The naïve bayes classifier model and the decision tree produced the best outcomes. The authors claim that it was challenging to determine the ideal network topology for the neural network model and the ideal value of k for the knn-based model.

The significant barriers and challenges linked to credit scoring were clearly shown by earlier studies on the issue. According to Guotai, C. (2017), it was found that metric models, such as Logistic Regression and LDAs, were superior to neural network approaches, such as Multi-layer Perceptron (MLP), at predicting an accurate overall score. However, Mixture-of-Experts (MOE), a modification of the MLP that divides the credit scoring task into smaller parts and assigns local experts to learn specific parts of the problem, showed accuracy predicting bad credits comparable to Logistic Regression. In the MOE architecture, the weights that are restricted to the expert networks have an effect on back propagation, which is the main source of this. Unlike other neural network models, RBFs (radial basis function networks) have a unique advantage.

In their study on loan delinquency prediction, Reddy, K. L. et al. 2020 intended to develop a method to help lenders determine whether to approve loans or reject them using machine learning. The 2019 Analytics ML Hackathon's data was used (B. A. Kumar et al., 2020). Each of the 116059 records in the collection has 28 attributes. Three machine learning approaches were used to the data sets: first, logistic regression; second, fitting into decision trees to create a model; and third, random forest to improve the predictions. Confusion data were used to test the research findings, and all of the models examined in the paper had accuracy rates of more than 80%.

The study also looked at A. Lawi (2017), which modified the logistic regression model by using the Generalized Linear Model algorithm. Logistic regression, in essence a

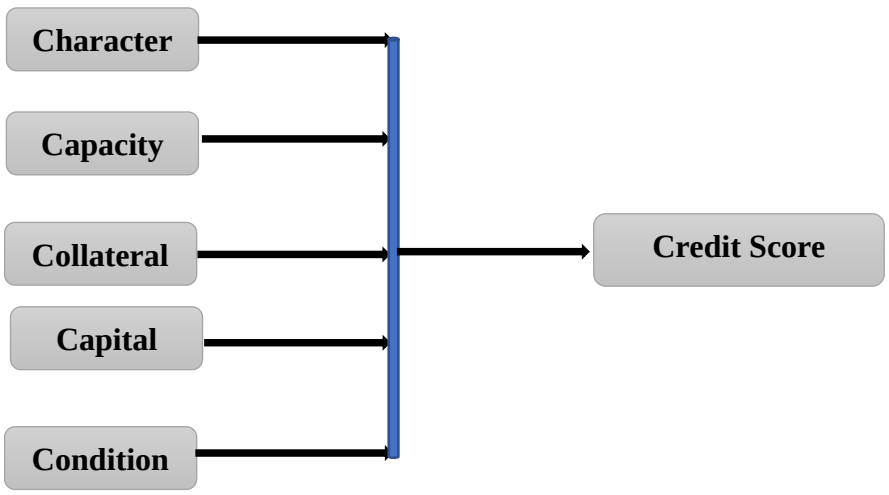
classification technique, produces a binary response from a collection of independent factors. GLM delivers a confidence bound showing the likelihood of a good outcome, which is an enhancement to this method. Their model has a projected confidence of 97.337 percent and an overall accuracy of more than 98 percent. On German and Australian datasets accessible in the UCI machine learning repository, J. Nalic (2018) showed additional gains utilizing Ensemble Logistic Regression improved by GradientBoost, reaching accuracy rates of 81 and 88.4 percent, respectively. This study decided to use an ensemble method to analyze the data after taking everything into consideration.

2.4 Conceptual Framework

The accompanying conceptual framework was acknowledged in accordance with the literature review. It demonstrates the relationship between the examination's independent and dependent variables. The independent elements in this analysis are clients' characteristics which can be broadly classified into the five Cs (Character, Capacity, Collateral, Capital, and Condition), which quantify the credit worthiness of the client. The conceptual foundation for the variables' operationalization is shown in *Figure 7.*

FIGURE 7:

Conceptual framework



a) Character

Character is a method in determining an average that is based on a number of loan application attributes. The overall weighted score evaluates the customer's creditworthiness (Myers and Forgy, 2005). Four groups can be made up of customer-related factors: social, economic, personal, and cultural. A customer's lifestyle, or how he lives, his social circle (or reference group), consumption patterns, and entertainment trends are frequently used to gauge

his social component. Because it is known to have a substantial impact on a customer's credit worthiness, credit institutions are particularly interested in learning who a customer's reference group is (Moti et al., 2012). Economic factors included details about the customer's or company's property ownership as well as relative ownership within the reference group. Personal characteristics include things like your age, profession, personality, financial situation, and family situation. For instance, families with younger children may struggle more frequently due to outside expenses, whereas more established families with older children are more likely to have steady collateral on their assets.

b) Capacity

When assessing a borrower's ability to pay, lending institutions look at their cash flow as a consumer or firm, how frequently they pay back credit, and how well they've previously repaid loans. Orlando (1990) asserts that a customer's financial ratios can assist lending organizations in determining if the borrower is able to pay both the current interest charges accruing from the loan and any potential future credit advancement costs.

c) Collateral

Collateral refers to the assets that a borrower will use as a backup source of payment for a loan. The majority of collateral will consist of physical assets like real estate, farm and industrial machinery, office supplies, and so on. Lending institutions often only accept collateral with a lifecycle value that is equal to the term of the loan.

d) Capital

This is the sum of money that a borrower has put up, whether they are a person or a business. The amount of risk that a borrower will incur in the event that a firm fails is often defined by capital.

e) Condition

This indicator assesses how susceptible a borrower is to outside factors including interest rates (static or dynamic), inflation, economic cycles, and market pressures. Even though most of the factors listed here are outside the customer's control, a customer's state frequently indicates their vulnerability, which could affect credit risk.

2.5 Operationalization of Variables

TABLE 2:

Operationalization of Variables

Variables	Indicator	Values
Character	Age	Age group
	Financial Standing	Net worth
	Occupation	employment status
	Gender	Male or Female
	Marital Status	Married, Single, divorced
Capacity	Customer financial ratio	Capital debt repayment capacity (CDRC) %
Collateral	Security held	Value of Security held
Capital	Business investment amount	Share of wallet
Condition	Purpose of the loan	

CHAPTER 3
RESEARCH DESIGN AND METHODOLOGY

3.1 Introduction

The research methodology defines the path the study will take to achieve its goals. It defines the processes involved in problem formulation, objectives, and findings from the collected dataset. The chapter also discusses how the study's results will be obtained in accordance with the study's objectives**.** As a result, from the research plan through the result dissemination, all methods and processes used during the study are covered.

3.2 Dataset

The data used for this study's analysis was obtained from Kenya commercial banks for the years 2016-2021. There were 65,079 applications for consumer loans in total who had taken out loans from various banks and were tracked over time to see if they defaulted or not. An applicant's personal history and credit history are included in our data set.

3.3 Research Design

3.3.1 Project Workflow

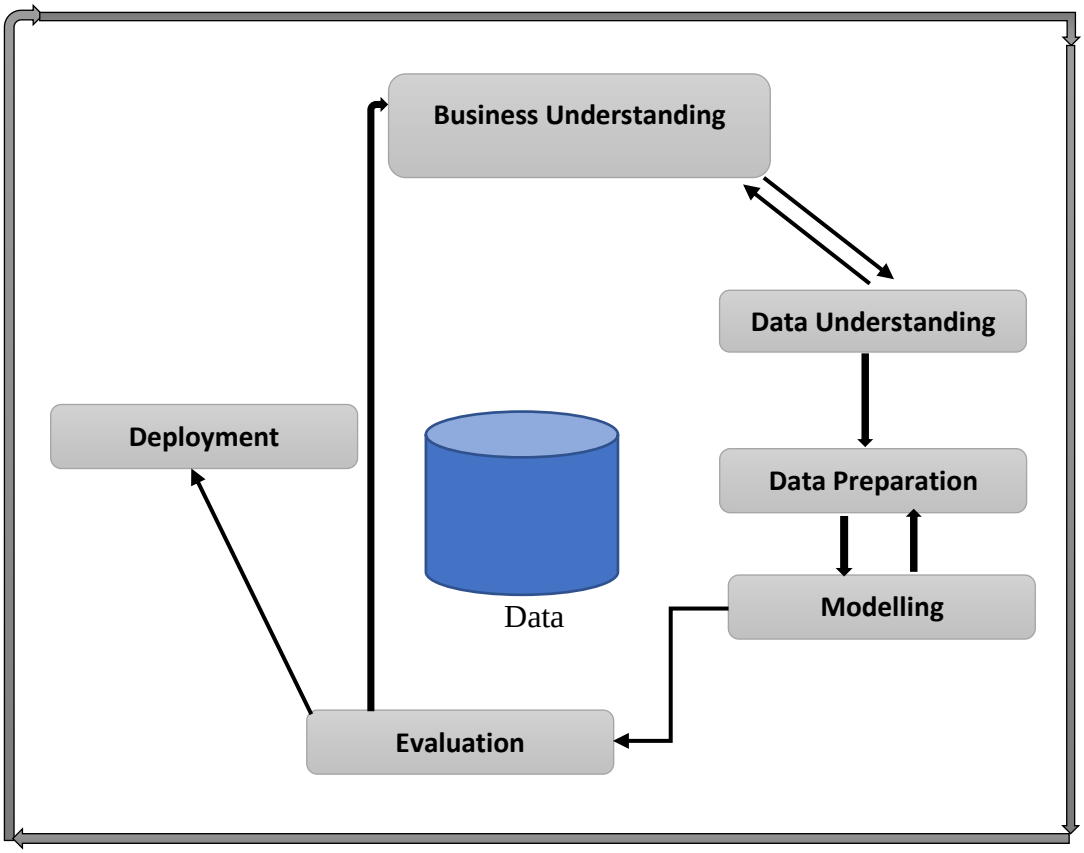
The workflow entails developing a model that can predict a person's credit score using the provided bank loan data, then examining the results to identify the variables that are most indicative of the score. Given a set of data (x) and targets (y), this supervised, regression machine learning challenge tries to train a model that can learn to map features (commonly referred to as explanatory variables) to the target (in this case, the credit score).

- Supervised problem: both the features and the target are provided
- The target variable in a regression analysis is a continuous one (credit score is a number between 0-800)

Despite some variations in implementation specifics, a machine learning project's overall structure often remains consistent. The study adopted CRISP-DM Phases, adapted from CRISP-DM (2000) as shown in *Figure 8.*

FIGURE 8:

Overall Design and Flow process of the research



Source: CRISP-DM Phases, adapted from CRISP-DM (2000)

Business Understanding

The step focuses on developing a data mining problem definition and an initial project plan from the project's business objectives and requirements. This process is broken down into four parts: identifying business objectives, evaluating the environment, identifying DM objectives, and creating a project strategy, as was described in Chapter 1.

Data Understanding.

This phase starts with gathering the data and goes on to include any actions that can help users familiarize themselves with the information. The information was gathered from an existing database, and all variables were clearly defined. To ensure data quality, data exploration was carried out. Python 3.1 was used for data exploration, with the Pandas library performing data preprocessing and Matplotlib and Seaborn for visualization. The Scikit-learn library was used to create machine learning models.

Data Preparation

A key factor in improving a model's generalization performance is the quality of the data. This is mostly determined by the adequacy of the data to be utilized in relation to the sample size, the significance of the features employed in the analysis, and the presence of outliers in the dataset. As a result, data pre-processing became a crucial stage in issues with credit-scoring classification (Thomas et al., 2017). In general, datasets can be created in a variety of ways and from a variety of sources.

On the other hand, datasets gathered from the real world may entirely consist of unclean, untransformed, or altered raw data. Accuracy, completeness, and consistency are three key factors that can be used to gauge the quality of data. Contrarily, real-world datasets are not like this since they are more susceptible to noise, outliers, missing attribute values, and inconsistent data (Garcia et al., 2015). Before doing any additional analysis or procedures, it is crucial to confirm the data representation and its quality. If the dataset contains any samples or attributes that are irrelevant, redundant, noisy, or untrustworthy, this could cause issues with model training since it makes knowledge mining and discovery challenging (Garcia et al., 2015). Data pre-processing thus becomes a crucial step in verifying the accuracy of the data, thereby enhancing and soothing the models' knowledge finding process. The stage of developing a model that deals with raw datasets is known as data pre-processing, and it includes numerous techniques such data imputation, normalization, feature selection, and data-filtering or instance selection. After the data has been processed, a new training dataset is prepared for additional analysis (Garcia et al., 2015).

This stage required data preparation tasks to be completed. These tasks were among them.

a) Data integration is the process of combining various data elements into a single data set.
b) Data cleaning to reduce noise, fill missing values, and resolve inconsistencies.
c) Identifying appropriate data sources, analytics libraries/algorithms, and relevant variable data.
d) Data transformation to make data ready for predictive analytics.

This stage was performed in python by first Selecting and install relevant libraries for predictive analytics, installation of various libraries and lastly dataset & data samples selection.

Modelling

Train-Test Split

The target (Y="Credit Score") and the features (X) were divided, and the train test was divided containing 30% of the test set and 70% of the training set. The training set would be used to

train our model. Additionally, the trained model's accuracy would be assessed using the test set, which contains unknown data. Using the median of the training set, a baseline prediction was created: The starting estimate is 716.28. Test set baseline performance: The mean absolute error is 17.6026.

Model Implementation and Optimization
Workflow Overview

To create the best model feasible, it is essential to select the proper model space given the training set. Therefore, the study's primary goals were to reduce actual error on the test set and prevent overfitting the data on the training set. The workflow is summarized in the section below.

a) A clean-up and processing are performed on the raw data set. Feature scaling, outlier detection, and filling in missing values are some examples of this.

b) Engineering for the chosen features is finished. Correlation and the significance of the feature are used as a guide when selecting important features.

c) Divide the test set into the train set in the proportion of 70:30, saving the test set for assessment in the final model.

d) Randomized Search is carried out over selected models' hyper-parameters to further filter the models and parameters for Grid Search.

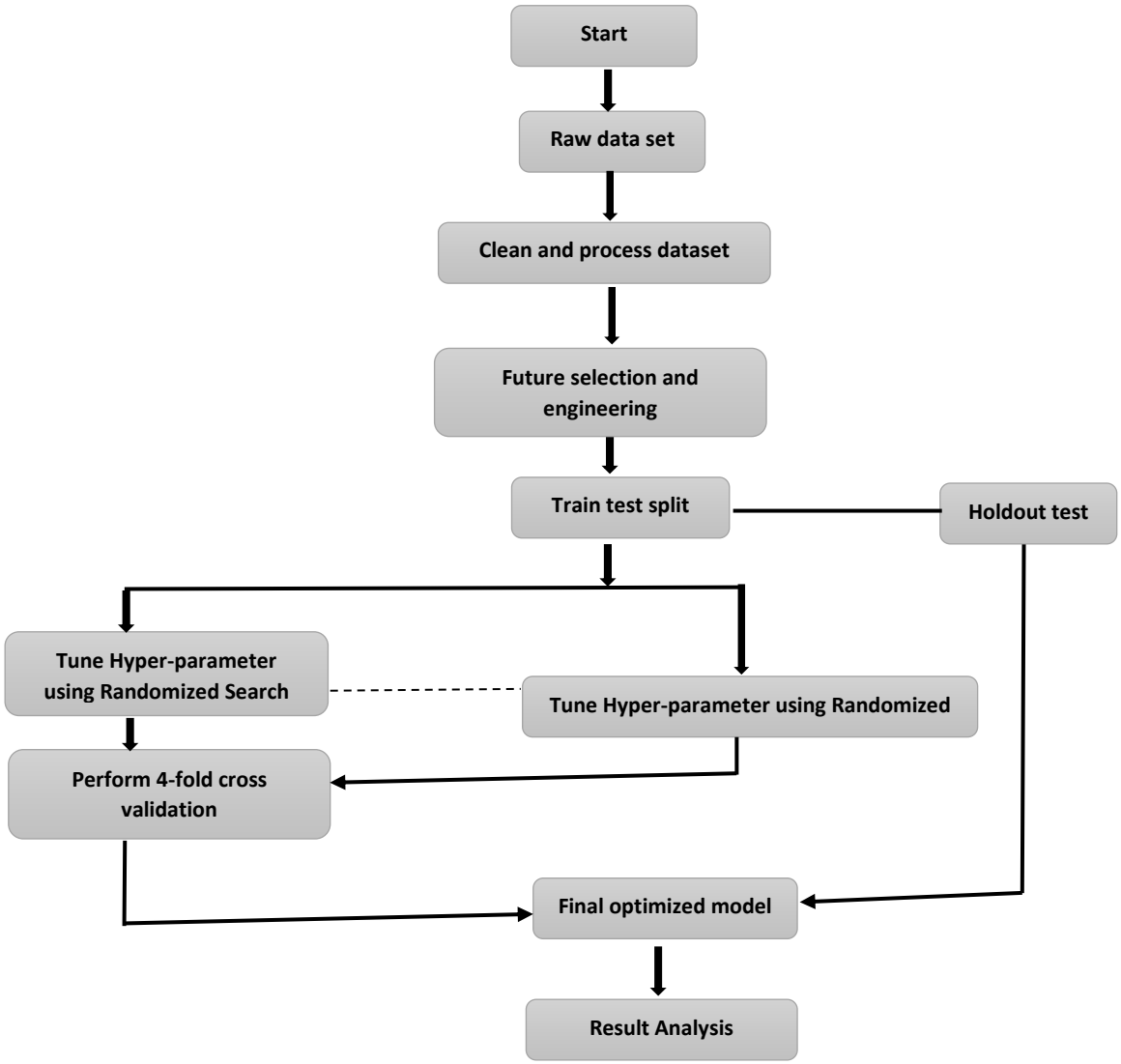
e) Cross-validation, as well as Randomized and Grid Search, are used to optimize models in the hyper-parameter space.

f) The final model was tested on a holdout test set. The results were obtained and analyzed.

A comparison analysis was carried out between the models and the final model.

FIGURE 9:

Proposed Workflow



Features Selection

As was already established, both the raw data and the data that are used to develop classification models are each connected with variables or features. Although there may be a large number of features available, it is frequently desired for a classification model to be trained on a small set of characteristics in order to simplify the model and decrease the amount of data it needs (Li et al, 2017). The following advantages can be attained by creating a classifier with specific features: 1) makes data simple to visualize and comprehend; 2) decreases data storage needs; 3) shortens training time; and 4) reduces dimensionality to enhance prediction performance (Li et al, 2017). The dataset is prepared for additional further processing once the missing variables are replaced and the datasets are normalized with new entries.

In general, datasets are made up of several properties or features that differ from one dataset to the next. However, datasets could contain duplicate and irrelevant features that complicate the training of models, resulting in models with poor performance and accuracy. As a result, in order to improve the performance of the model's prediction, analyzing characteristics and determining their significance has become a crucial and important activity for data pre-processing in data mining in general and credit-scoring in particular (Cai et al, 2018). In other words, feature selection is the process of choosing a subset of representative features that can affect a model's performance. It is a crucial step in choosing the most pertinent and appropriate features and subsequently discarding the unnecessary ones.

Data-filtering (Instance Selection)

Studies on credit scoring have paid minimal attention to feature selection or elimination during the data pre-processing stage, whilst data-filtering or instance selection was given very little consideration prior to training the data. The goal of data-filtering or instance selection is to construct a representative training dataset while reducing the size of the original dataset and maintaining the integrity of it (Nikpey et al, 2020). The performance of a model may be significantly impacted by noisy data or outliers, just as it may be by duplicated and pointless features.

Nikpey et al, 2020 et al (2020) found that in some circumstances, eliminating outliers can improve the performance and accuracy of classifiers by smooothing the decision boundaries between data points or feature space. Outliers in a dataset, which can include unusual data, data without a prior class, or data that are incorrectly labeled, are samples of the dataset that appear inconsistent within other samples in the same dataset. If all of this is present in a dataset, then these outliers must be removed by filtering those samples that contain such characteristics that could interfere with the training process, as their presence can result in ineffective data training for classifiers (Nikpey et al, 2020).

Data Splitting and Partitioning Techniques

The data is prepared to be divided into training and testing sets, which are to be used for developing and analyzing the model, respectively, once any missing variables in the data have been replaced and normalized. However, the training set can be further processed by using feature and instance selection once the dataset has been partitioned. It is acknowledged that data-splitting, partitioning, or resampling is a crucial stage in the development, testing, and validation of models. Datasets must be divided into training and testing portions for two

reasons: first, the model must be trained on the seen data portion of the dataset, and then it must be validated and applied to the unseen data portion of the dataset, which will reveal how well the model performed and how it would perform in real-world future cases.

Data-related issues include how many data should be saved for training and testing; the more data in the training set, the better the model fits the data; however, the more data in the testing set, the more accurate the model's estimates of accuracy are; for example, the model is more confident that it will have good accuracy on 1,000 testing data than it will on 100 testing data.

The size of the available data and the number of data samples associated with each prior class can also be a problem in this situation, and because of the different dataset sizes and the distribution of the data class sizes, the use of a particular splitting technique can have a significant impact on the model performance. In order to ensure that data from various classes are trained effectively and have acceptable model generalization over the testing set, a fair distribution of these data between the training and testing sets is also crucial. But in the world of credit rating, various data separation methods have been applied.

a. Holdout Technique

This method is based on dividing the dataset into two parts: one for learning and training the model, and the other for testing and validating it. This approach is fairly straightforward and has received a lot of attention in the literature. The typical approach entails randomly preserving 70% of the dataset for training and 30% for testing. The holdout technique's accuracy results, however, may be skewed because data can be misused and the training and testing sets may not be representative (the testing set might contain simple or difficult data, for example) (Li et al, 2017).

Even if the training and testing sets may overlap and may not be ideal, this problem can be avoided by repeatedly using the holdout technique to have randomly picked training and testing sets data each time. This reduces the likelihood of receiving a lucky testing set.

Performance Evaluation Measurement

The most crucial phase of the modeling development process is thought to be the model performance evaluation. The constructed model is put to the test over the gathered datasets during this stage, and the performance evaluation metrics will indicate how well-learned the model is and whether the outcomes are solid and trustworthy so that it is prepared to predict fresh real-world data. Three different sorts of indicator measures, each addressing different aspects of how the generated model interprets the outcomes, need be taken into account in order

to draw a valid conclusion about how well it worked (Lessmann et al., 2015): first, metrics that evaluate the model's predictive power (e.g., differentiating between good and bad loans); second, measures that evaluate the model's discrimination power; and third, measures that evaluate the accuracy of the model's predictions probability. Thus, taking into account these metrics offers a thorough assessment of the produced model's performance. Additionally, it is suggested that using many performance evaluation measures would help ensure the model's value by enabling the collection of all of its key characteristics (Japkowicz & Shah, 2011; Lessmann et al., 2015).

As a result, accuracy and mean absolute error were used in this study to evaluate our model and draw a solid and dependable conclusion about its prediction abilities.

Deployment

The goal of this step is to organize and present the newly discovered knowledge in an understandable manner. This section is fully covered results analysis section.

3.4 Mapping of Activities and Methods to objectives

TABLE 3:

Mapping of Activities and Methods to objectives

Objective	Activities	Methods
Objective 1	Data Preparation	Data Pre-Processing
	Data Understanding.	Feature Selection & Engineering
Objective 2	Modeling.	Model Implementation and Optimization
Objective 3	Post Processing Stage/ Evaluation & Deployment	Performance Measures

CHAPTER FOUR
DATA ANALYSIS, FINDINGS AND DISCUSSION

4.1 Introduction

The main findings of the study are presented in this chapter. It begins by displaying the summary descriptive statistics results. Following that, the results of mathematical and statistical procedures are used to evaluate both the customer's creditworthiness and the likelihood of default. The outcomes of different validation techniques are shown. Effectiveness of relevant variables is also discussed.

4.2 Descriptive Statistics

There were 65,079 applications for consumer loans in total. It's important to note that there were no missing values in the sample of data. The sample includes personal data such as gender, age, marital status, type of residence, nationality, and postal code as well as characteristics connected to banks such as other loans, other banking transactions, loan term, and underwriter (annual income). The study's time frame is from 2016 to 2021.

4.3 Research Findings

4.2.1 Objective 1 Results

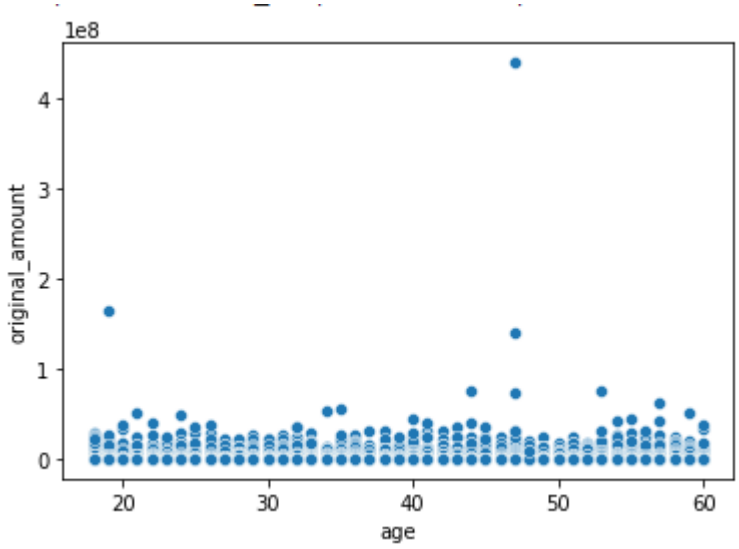
a) Data Pre-Processing

The initial concern was how to handle the dataset's missing values, however there were none.

Outliers were detected using Scatter plots if any were present in the dataset. *Figure 10* shows an example of an outlier in borrowed funds.

FIGURE 10:

Scatterplot of Age Original borrowed amount against Age



The entry on the far right of the plot is an example of an outlier that was removed due to its large deviation from the median.

To gain insight into the dataset, the skewness of column distributions was observed. To normalize the numerical data, feature scaling was used. The min-max strategy was employed. Which the formula denotes

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.1}$$

b) Feature Selection & Engineering

To ascertain how categorical values relate to one another, credit was used. The study developed density graphs for each categorical feature. The plots illustrated how the credit score was impacted by various categorical feature values and how the density distribution varied as a result. Features that had no effect on how credit was distributed were found so they may be given less weight in the final model.

The distribution of the credit score was influenced by all of the category features other than "Years in Service." Since changes in its value had no impact on how the Credit Score was distributed, the absence of "Years in Service" posed no issues in the final model. Furthermore, the distributions were bimodal. A continuous probability distribution having two distinct modes is referred to as a bimodal distribution in statistics. These show up as different peaks in the probability density function (local maxima).

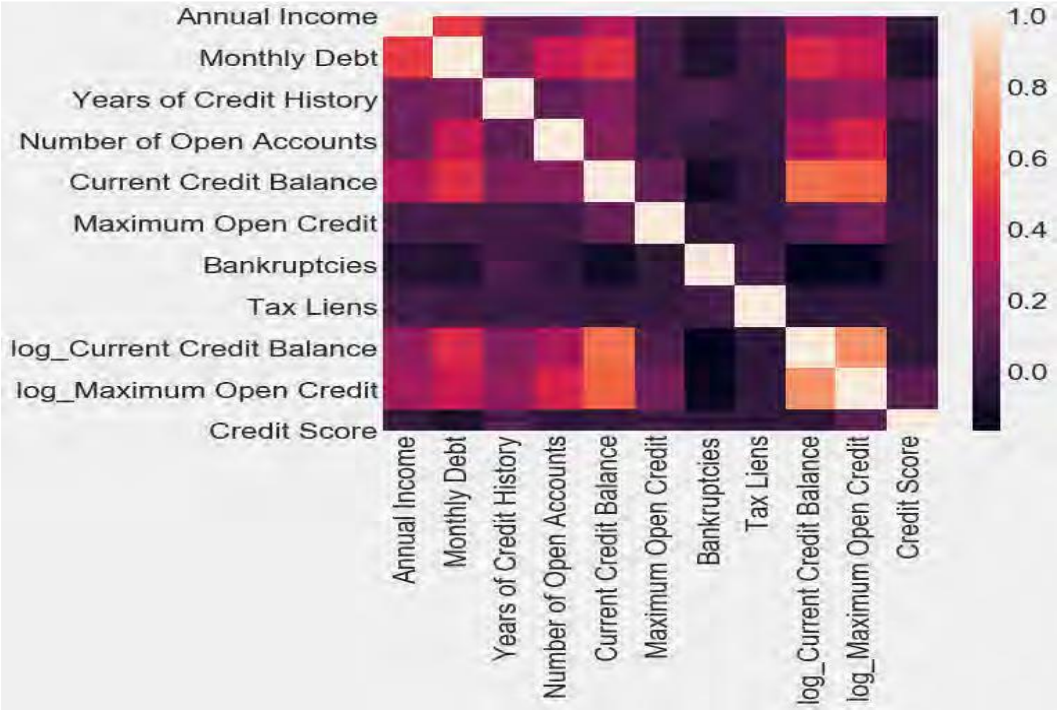
- Feature selection is the process of choosing the data features that are most relevant based on a number of different considerations. It is possible to classify a characteristic as very relevant if it has the highest correlation or variance with the target. In order to improve the model's ability to generalize and understand the new data, less significant features are thus removed.
- Feature engineering is the process of selecting raw data and extracting features that enables machine learning models to take in a mapping of features to targets. Commonly used variable transformations include logarithms and square roots. On the other hand, categorical variables of one-shot encoding are applicable. As a result, feature engineering can be summed up as the process of extracting more pertinent features from raw data.

To succeed, iterative procedures like feature engineering and selection were necessary. The study routinely went back and revise feature selection using modeling results, such as the feature importance from a random forest, or may later find relationships that call for the construction of additional variables, necessitating feature engineering. Additionally, these procedures frequently combined statistical data quality with domain expertise.

After completing Feature Engineering, it was necessary to eliminate multi-linearity by incorporating the log and square roots of numeric columns. It was unnecessary to identify features that were highly col-linear due to an underlying commonality and preserve them. Collinearity between characteristics that exceeded a predetermined cutoff (0.65) was eliminated. A heat-map displaying the relationships between the features was made in order to identify multi-linearity or collinearity, *Figure 11.*

FIGURE 11:

Heat-map of Numeric Features



Following feature selection and engineering, below feature set was arrived:

TABLE 4:

Final Set of Features selected

Current Loan Amount	float64	Categorical values are one hot encoded
Annual Income	float64	Loan Status
Monthly Debt	float64	Term
Years of Credit History	float64	Home Ownership
Number of Open Account	float64	Years in current job
Number of Credit Problem	float64	Purpose
Current Credit Balance	float64	
Maximum Open Credit	float64	
Bankruptcies	float64	
log MonthlyDebt	float64	
log MaximumOpenCredit	float64	
sqrt TaxLiens	float64	

4.2.2 Objective 2 Results

Proposed Hybrid-Stacked Model (RfDNN)

Overview

A feature detector serves as the input for a neural network in this paradigm. This approach was inspired by Yunchuan Kong and Tianwei Yu's implementation of a supervised feature detector on top of DNN architecture in their publication (Y. Kong, 2018). Random forests (RF), as compared to other machine learning algorithms, have been a standout performer in learning feature representations due to their strong classification's capabilities and easily comprehensible learning mechanism.

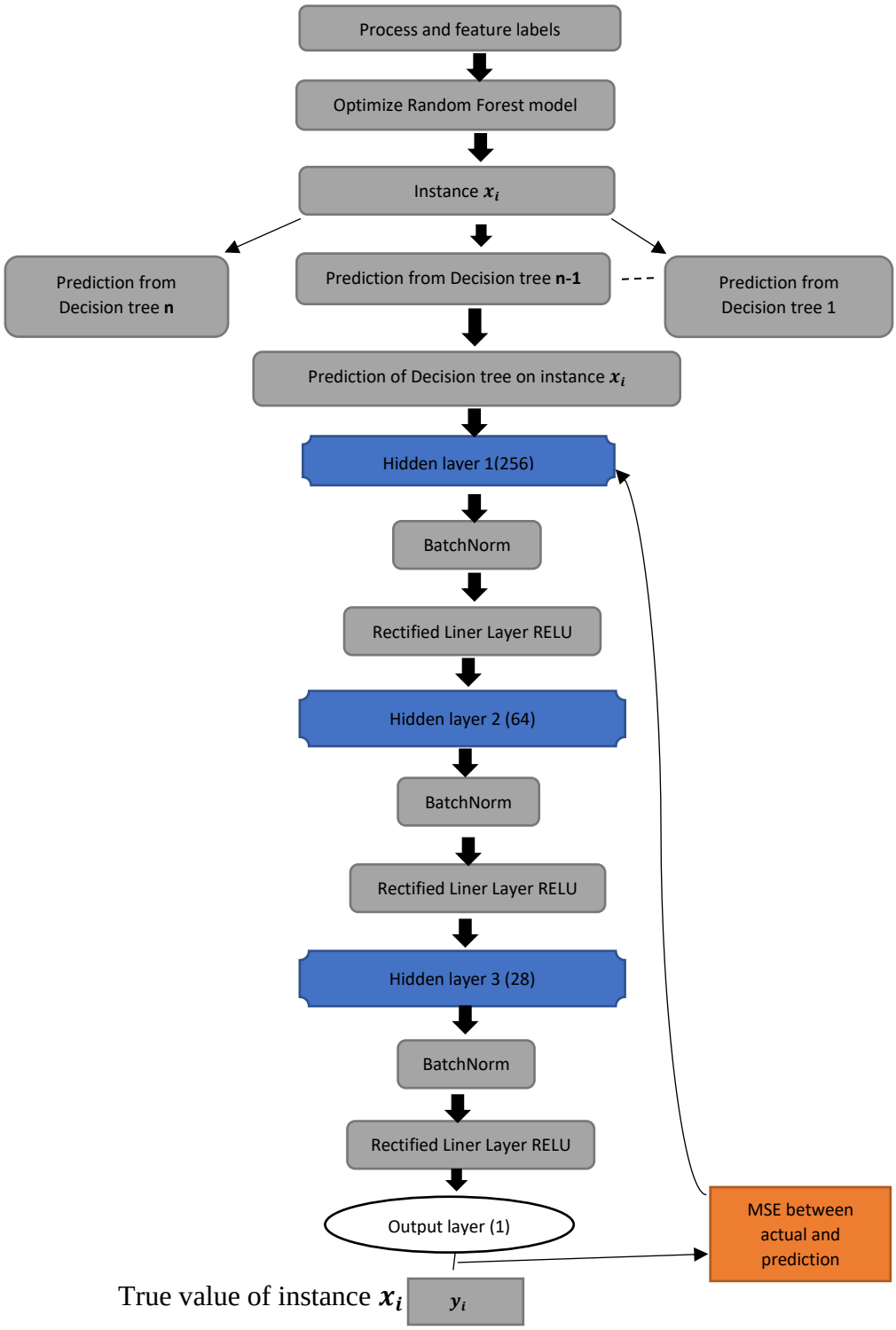
The modified Random Forest model was chosen by the study to serve as a "feature detector" and input to the subsequent neural network model. Gradient Boosted Regression (GBR) trains the Decision trees sequentially while reducing the error caused by the succeeding trees, in contrast to Random Forest, which trains the Decision trees independently from bootstrapping. Because of this, Random Forest was chosen for this methodology as opposed to GBR. As a result, each decision tree's output from Random Forest offers an impartial feature representation that can be applied to the training of a neural network.

The proposed model is based on the Forest Deep Neural Network (fDNN) architecture developed in Yunchuan Kong and Tianwei Yu's study (Y. Kong, 2018).

The flowchart below provides a summary of the architecture of our RfDNN model.

FIGURE 12:

RfDNN model Architecture



Training and Architecture details

It takes two steps to train the RfDNN regressor. The Optimized Random Forest Model was fitted using selected and processed features in the first stage. The second part of the process involved recording and individually inputting predictions from each decision tree in the forest into the fully connected DNN for training. The fitted forest and DNN used the entire model to determine the prediction given a testing instance following two stages of training-on-training data.

The Deep Neural Network architecture consisted of an output layer that outputs one value, the credit score, and three hidden layers with sizes of 256, 64, and 28. The Mean Squared Error between this prediction's Y value and the actual value Y was determined for instance X_i in order to train the model. Rectified Linear Units are the name given to the activation function utilized in the model (RELU).

$$\sigma ReLU(x) = \max(x, 0)$$

This activation function has an advantage over Sigmoid and Tanh activation in that it addresses the problem of missing gradients during model backpropagation.

In order to reduce covariate shift in the concealed unit values, the study used batch normalization between the layers as a regularizer. Because it has a slight regularization effect, it also protects against over-fitting.

The most widely used Gradient Descent Algorithms version in Deep Learning research, the Adam optimizer (D. P. Kingma, 2014), was used to select the model's optimizer. A small subset of the samples were randomly chosen by the optimizer for each iteration of the mini-batch training approach, which was also employed (I. Goodfellow,2016). Specific hyper-parameters must be taken into consideration in this two-step model. Randomized and Grid Search have already been used to modify the Random Forest model's hyperparameters. The beta and gamma values of the Batch Normalization layer, the number of training epochs, and the optimizer's learning rate were among the hyperparameters of the neural network that were tweaked using Randomized Search throughout a preset range. The Python code for the model made use of the Scikit-learn and Pytorch tools. The suggested model outperformed the baseline model, which consisted of a single random forest prediction, in terms of Mean Absolute Error (MAE).

Model Interpretability

They improved the ensemble models and tuned the hyperparameters to produce an outstanding level of accuracy on the test dataset (93 percent). Although this might be adequate for some issue areas, it is insufficient for the process of evaluating a person's credit. Using the model as

a "black box" with no logic guiding the outcomes is unethical and unreliable. The architecture of these algorithms demonstrates that there is no simple way to comprehend why or how the output was created. The focus today is on better understanding these "black-box" models. Recent research has made significant progress in providing correct information on dynamics and the connection between input, output, and intermediates.

The standard dimension reduction and Principal Component Analysis (PCA) methodologies are being improved, according to the paper (M. T. Ribeiro, 2016). To provide better visual artifacts for a deeper comprehension, more study is being done. Additionally, various models might use various analytical techniques. Despite the observations, they might not always lead to accurate conclusions (A. Vellido,2012). This study discusses the importance and downsides of model interpretation as well as the various strategies currently used to adequately express the fundamental concepts of machine learning models.

Z. C. Lipton (2016) illustrates the benefits of interpreting an ML model utilizing a model-agnostic approach, where an interpretable technique is produced from the predictions of the "black box model." It can therefore comprehend more sophisticated models, such Deep Neural Networks, because it is not model-dependent. By avoiding relying entirely on conventional interpretable models like linear or logistic regression, practitioners are able to be more flexible.

As a result, the study opted to show the model output using two Model-Agnostic techniques. Agnostic model techniques are ones that work well independent of the methodology and are not model-dependent.

- Local Interpretable Model-Agnostic Explanations (LIME)
- Feature Importance
- Single Decision Tree interpretation

4.2.3 Objective 3 Results

Evaluating and Comparing Machine Learning Models

Finding the model with the most potential for advancement is the goal (such as hyper parameter tuning). The mean absolute error was used in the study to compare the models. The baseline model's prediction for the median score was 17.6 points off.

a) Scaling Features

This is necessary because the units in which features are measured vary, and the study intended to normalize the characteristics to prevent the algorithm from being affected by the units. Since support vector machines and k-nearest neighbors do not take the Euclidean distance between

the data into account, feature scaling is required for linear regression and random forest. As a result, it is advised to scale characteristics while contrasting various approaches (Z. Zhao,2015).

There are two methods for scaling features:

- Divide each value by the feature's standard deviation after deducting the feature's mean. Each feature has a mean of 0 and a standard deviation of 1 after being subjected to standardization.
- The minimum value of the feature should be deducted from each value after dividing each value by the maximum minus the minimum for the feature (the range). Scaling to a range or normalizing refers to ensuring that all values for a feature are between 0 and 1.

The test and training sets both underwent scaling and normalization. Five distinct machine learning models were trained and assessed using the Scikit-Learn toolkit.

1) Linear Regression
2) Support Vector Machine Regression
3) Random Forest Regression
4) Gradient Boosting Regression
5) K-Nearest Neighbors Regression

The default models' mean absolute error on the test set was calculated after they were trained on the training set.

TABLE 5:

Baseline model MAE on Test Set

Model	Mean Absolute Error
K-Nearest Neighbour	17.1126
Linear Regression	16.8247
Gradient boost	15.2336
XGBRegressor	15.3622
Random Forest	12.6155

Although all of the default parameters were used to build the models, making this comparison was biased, the errors showed that the issue was learnable because all of the models outperformed the baseline MAE of 17.62.

Model Optimization

In machine learning, determining the best set of hyperparameters for a particular problem is known as optimizing a model. Model parameters and model hyperparameters are distinguished in the following ways, according to I. Goodfellow (2016):

- Construct a range of alternatives, then randomly choose combinations to try. This process is called as random search, and it is how the hyper parameters is determined for evaluate. Grid search, in contrast, analyzes each and every combination that is specified. Utilize grid search with a more constrained set of possibilities to adjust particular model hyperparameters with better precision when unsure of the appropriate model hyperparameters and need to narrow the options (H. Ma,2018).
- The method utilized to evaluate the performance of the hyperparameters is cross validation, as indicated in *Table 6*. Instead of dividing the training set into separate training and validation sets, which would limit the amount of training data the study could utilize, the study employed K-Fold Cross Validation. Divide the training data into K folds, train iteratively on fold K-1, and then evaluate performance on fold k to achieve this. The study repeated this procedure K times, with the crucial distinction that it was testing on data that it did not train on in each iteration, until it has tested on every example in the training set. Before training the model with all of the training data at once, it applied K-fold cross validation and used the average error on each of the K iterations as its final performance indicator. The recorded performance was then used to compare various hyper-parameter combinations (H. Ma,2018).

TABLE 6:

Cross-Validation Overview

Iteration 1	**Test**	**Train**	**Train**	**Train**	**Train**
Iteration 2	**Train**	**Test**	**Train**	**Train**	**Train**
Iteration 3	**Train**	**Train**	**Test**	**Train**	**Train**
Iteration 4	**Train**	**Train**	**Train**	**Test**	**Train**
Iteration 5	**Train**	**Train**	**Train**	**Train**	**Test**

Source: H. Ma, (2018).

The top two options from the default model test were taken into consideration for model optimization.

- Random Forest Regression
- Gradient Boosting Regression

Based on the information entropy of the features, both of these algorithms generated regression and decision trees. Comparison of the differences between them (Z. C. Lipton,2016).

- Weak students are the foundation for boosting (high bias, low variance). Weak learners are shallow trees in the context of decision trees. Boosting effectively decreases error by reducing bias, or the error for the erroneous assumptions made when developing the learning algorithm. It is the main reason why the model was underfit. Since boosting uses sequential processing, parallel processing capability cannot be utilized, increasing run-time.
- However, Random Forest makes use of fully evolved decision trees (low bias, high variance). It employs the opposite strategy to mistake reduction by minimizing variance. The approach cannot minimize bias because the trees are uncorrelated to optimize variance reduction (which is slightly higher than the bias of an individual tree in the forest). Therefore, in order to keep the bias as low as possible from the beginning, big, unpruned trees are needed. Runtime is decreased by Random Forests' simultaneous tree generation.

Figures 13 and *14* show that, after performing a randomized search of the following hyper-parameters on both approaches, Gradient Boosting Regression surpassed Random Forests in terms of mean absolute error.

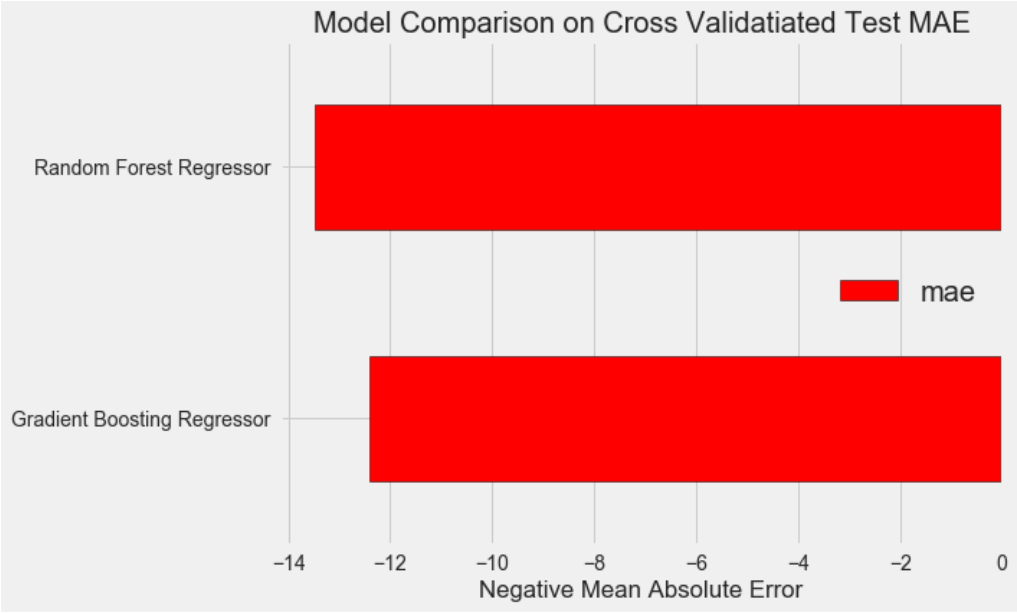
FIGURE 13:

List of Hyper-parameters tuned

```
# Loss function to be optimized
loss = ['ls', 'lad', 'huber']

# Number of trees used in the boosting process
n_estimators = [100, 500, 900, 1100, 1500]

# Maximum depth of each tree
max_depth = [2, 3, 5, 10, 15]

# Minimum number of samples per leaf
min_samples_leaf = [1, 2, 4, 6, 8]

# Minimum number of samples to split a node
min_samples_split = [2, 4, 6, 10]

# Maximum number of features to consider for making splits
max_features = ['auto', 'sqrt', 'log2', None]

# Define the grid of hyperparameters to search
hyperparameter_grid = {'loss': loss,
                       'n_estimators': n_estimators,
                       'max_depth': max_depth,
                       'min_samples_leaf': min_samples_leaf,
                       'min_samples_split': min_samples_split,
                       'max_features': max_features}
```

FIGURE 14:

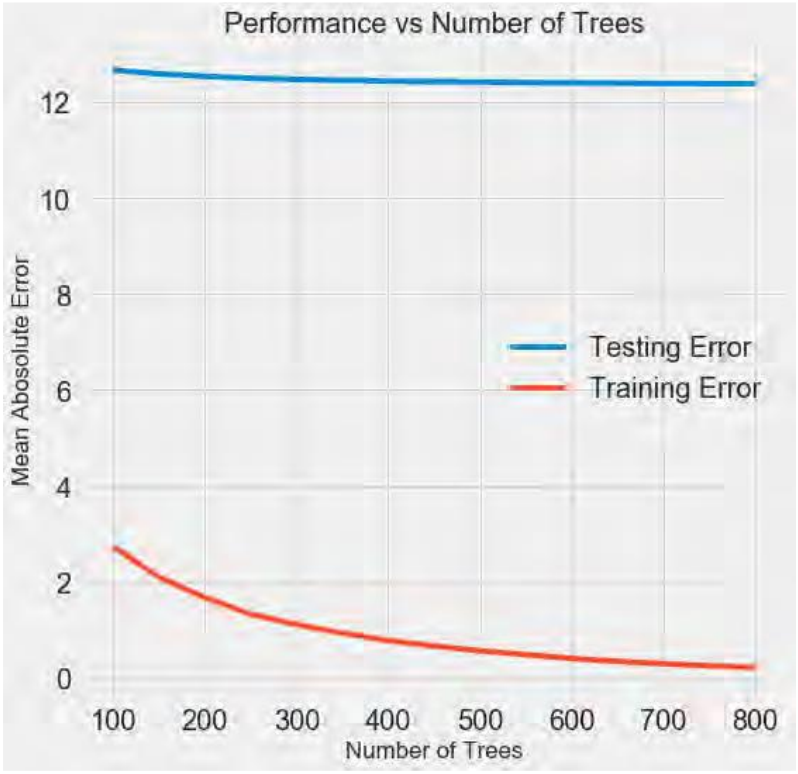
Error Comparison of best estimators predicted after Randomized Search and Cross Validation



The number of trees on the best estimate of GBR were determined using a grid search, which makes use of all characteristics rather than just a handful at random. *Figure 15* illustrates the relationship between training and test error and the number of trees used in the model.

Figure: 15

Effect of Number of trees on train and test error.



Once there were 200 trees, it was discovered that adding more trees reduced training error but had no effect on test error. As a result, the model overfits on the training data as the number of trees used increases. As a result, the study decided to continue fitting the model using 200 trees.

4.4 Discussion of Results

Feature Importance

Calculating feature relevance is based on a straightforward yet effective concept. It argues that the strength of a given feature in a dataset is inversely related to the growth in the model's prediction error when a feature's values have been permuted or shuffled, which destroys the relationship between the feature and the actual outcome.

In order to assess the significance of a feature, the study looked at the rise in the model's prediction error after permuting it. The model in this case relied on the feature for prediction since a feature is only considered "important" if changing its values increases the model error. A feature is unimportant, on the other hand, if altering its values has little to no impact on how well the model predicts the future. The shuffling feature significance measurement for random forests was first presented by L. Breiman (2001). A. Gelzinis (2014) developed model reliance,

a model-independent variation of feature importance, based on this idea. The trained model f, the feature matrix X, the goal vector y, and the error measure L are all considered inputs (y,f).

The algorithm used is as follows (C. Molnar, 2019):

1. Calculate eorig = L(y, f(X)) the original model error (e.g. mean squared error)
2. For each feature j = 1,...,p, do the following:
    - By shu ing feature j in the data X, generate feature matrix Xperm. This breaks the link between feature j and the true outcome y.
    - Estimate the error eperm = L(Y,f(Xperm)) based on the permuted data predictions.
    - Calculate the importance of permutation features FIj= eperm/eorig. Alternatively, the distinction can be used: FIj = eorig - eperm
3. Sort features by descending FI.

On the valid set predictions, the study tested the feature importance of the optimized models and ranked the top eight features in order of importance.

The annual income, maximum open credit, current credit balance, monthly debt, and current loan amount were found to have a substantial impact on the models' forecasts for both the Random Forest and Gradient Boosted Regression models. It is also reasonable to suppose that these aspects will be important when deciding whether to submit a candidate for manual evaluation. As a result, when generating a forecast, the models seemed to give the right qualities priority, just like a human evaluator would. The study further examined both model predictions to support our assertion.

Local Interpretable Model-Agnostic Explanations

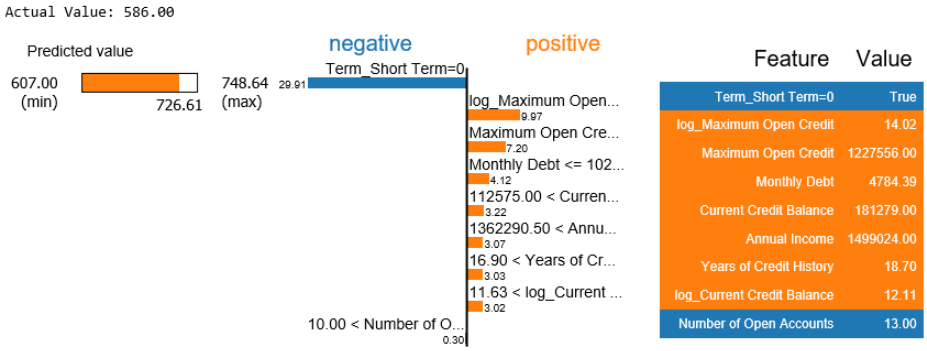
The study that follows provides a thorough explanation of a specific prediction made by an ML model put out by P. Ferrando, (2018). The study believes it satisfies all three fundamental standards for model interpretability.

- Agnostic Model: It is not model-specific. Only changes the input and predicts behavior based on how the prediction changes.
- Interpretability: Explanations need to be clear, which can be challenging even for linear models with a lot of features and a complicated feature space. The explanations provided by LIME use an alternative data representation (known as an interpret-able representation) from the original feature space.
- Locality. By utilizing an interpretable model to roughly mimic the black-box model in the vicinity of the instance that need to be explained, LIME generates

an explanation (for example, a linear model with a few non-zero coefficients).

FIGURE 16:

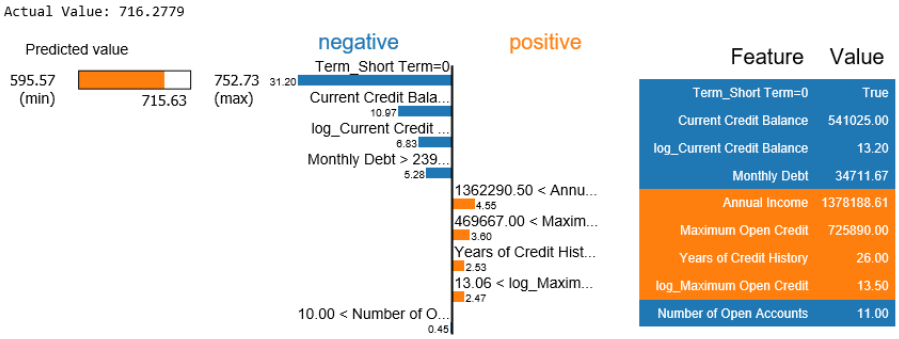
Wrong Prediction Interpretation



Maximum Open Credit appeared to have the greatest influence on the incorrect prediction, followed by loan term or duration. Considerations including monthly debt, credit card debt, and annual income are crucial.

FIGURE 17:

Right Prediction Interpretation



The model's accurate prediction reveals that the factors that have a negative impact on the prediction—high current credit balances and monthly debt values, which are regarded as warning signs when considering a loan request—are related to these factors. Long-term loans are more likely to default, so it makes sense that the duration has a negative impact on the forecast. In contrast, an acceptable wage and a long credit/loan history benefit the prediction conclusion. Based on these interpretations, their is more faith that the model will be able to pick the relevant characteristics to produce positive and negative weights during a forecast.

Single Decision Tree Interpretation

The chosen interpretation was finally model-specific. Because decision trees are used as individual predictors in both Random Forest and Gradient Boosted Trees, it was necessary to

dig deeper into a single decision tree in order to comprehend the model and how and where the splits were produced while producing a forecast.

FIGURE 18:

Tree Interpretation of RF

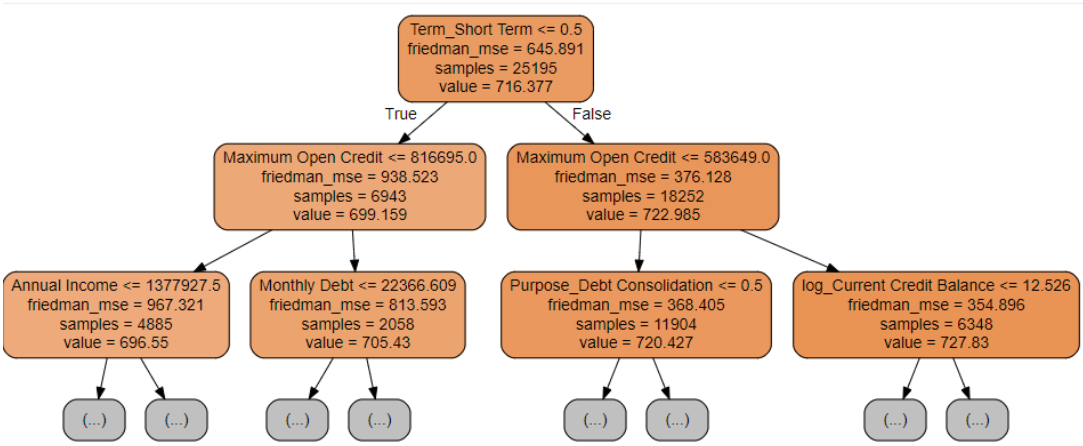
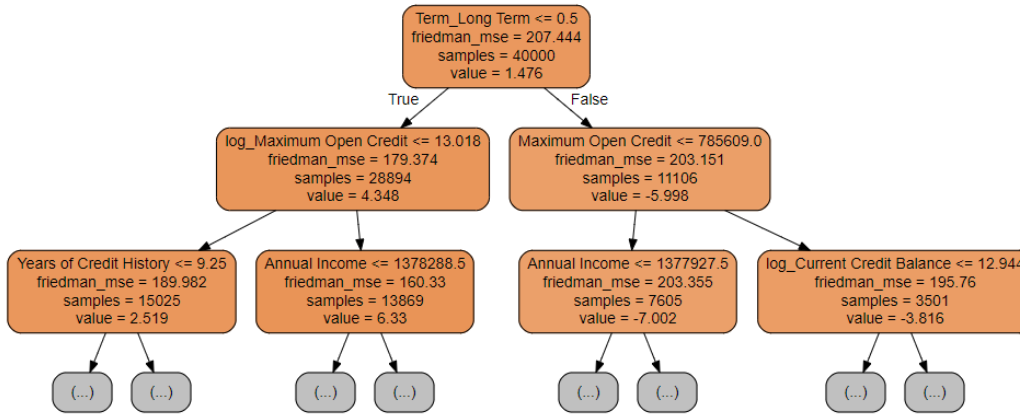


FIGURE 19:

Tree Interpretation of GBR



If-then analysis can be used to interpret the trees. *Figures 18* and *19* were limited to three layers in order to provide an overview of the branches. This is because the tree nodes were a little too long for all of the variables under consideration. **If** the short term ,in Figure 18, is = 0.5, the maximum open credit is =816695, and the annual income is 1377927.5, **then** credit score(value) is 696.55. Other branch nodes are interpreted using the same methodology.

4.1 Comparative Analysis of Supervised Models

The study began by fitting the training data to a few well-known existing models from the literature.

- Linear Regression

- Random Forest Regression
- Gradient Boosting Regression
- K-Nearest Neighbors Regression
- Extreme Gradient Boosting Regression

The following performance was observed on the validation set, *Table 7.*

TABLE 7:

Comparison of Initial models

Model	Mean Absolute Error	Accuracy
Linear Regression	18.83	88.65%
Random Forest Regression	17.61	89.10%
Gradient Boosting Regression	18.36	88.93%
K-Nearest Neighbors Regression	19.11	88.48%
Extreme Gradient Boosting Regression	18.36	88.93%

After applying Random Search to each model's parameters, the outcomes improved by the margins depicted below.

Table 8:

Accuracy of different models after initial tuning

Model	**Mean Absolute Error**	**Accuracy**
Random Forest	13.47	91.87%
Extreme Gradient Boosting	15.36	90.75%
Gradient Boosting Regresser	15.36	90.73%
Linear Regression	16.82	89.85%
K-Nearest Neighbors	17.11	89.68%

Because it would be impossible to tune all of the models given the required time and computational complexity. On the top two most accurate models, a decision was made to perform comprehensive analysis and hyper-parameter tuning. Both Gradient Boosting Regression and Random Forests are Regression Decision Tree-based techniques. The Random Forest and GBR models underwent 4-fold Cross Validation and Randomized Search over the hyperparameters.

Only the best model (GBR) was chosen after additional filtering was done after the Random-CV Search. The optimized GBR model was utilized to obtain a maximum accuracy of (93%) on the test set utilizing Grid Search across the hyperparameters and 5-fold Cross Validation.

4.2 Final Model Evaluation

The final model with optimized hyperparameters was fitted to the training set also examined using the test set. As demonstrated in *Table 9*, the model performed significantly better than all earlier models as well as the default untuned Gradient Boosting Regressor Model in terms of Mean Absolute Error.

TABLE 9:

Accuracy of all the Models

Model	Mean Absolute Error	Accuracy
GridCV-Gradient Boosting Regressor	11.571	93.023%
RandomCV-Gradient Boosting Regressor	12.416	92.521%
RandomCV-Random Forest	12.623	92.404%
Default Random Forest	13.475	91.882%
XGBRegressor	15.36	90.763%
Default Gradient Boosting Regressor	15.362	90.742%
Linear Regression	16.822	89.865%
K-Nearest Neighbors	17.114	89.696%

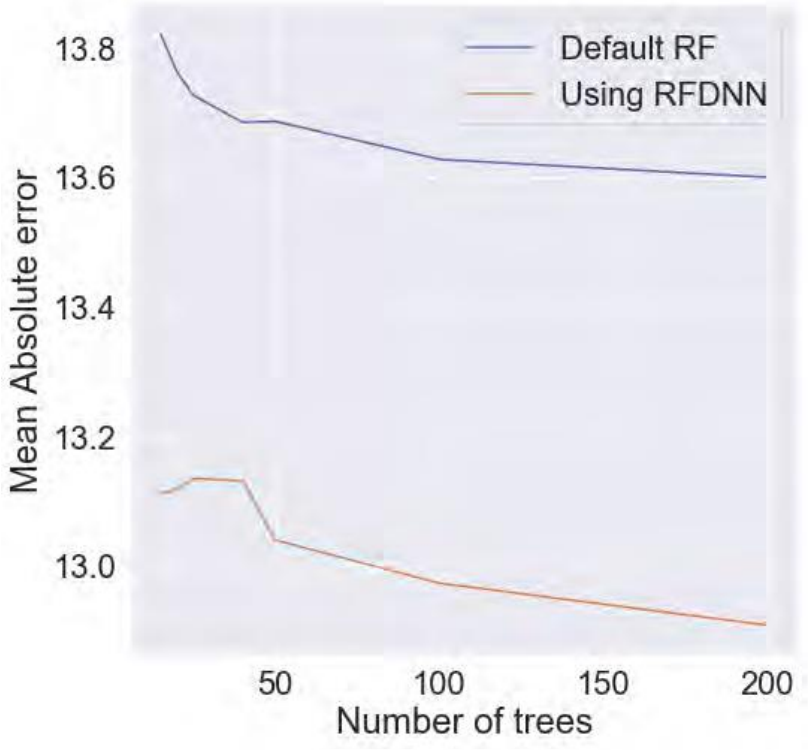
The model's predictions and the actual values had a similar distribution. Credit score values below the qualifying criterion were accurately predicted by the model (i.e 710). The model did a good job of predicting defaulters as a result. The bimodal distribution of credit ratings over the eligibility criterion was not very well predicted by it. That's not a major concern because credit scoring's objective is to correctly anticipate defaulters and ineligible individuals.

4.3 RfDNN Result Analysis

The Random Forest model could only be utilized to run the RfDNN model since it generates separate base learners or decision trees that can be used as input downstream of the DNN architecture. Both the suggested Stacked RfDNN model and the default Random Forest Regressor were tested using a range of tree values. The suggested model outperformed all other models after being verified on a test set of 25000 instances.

FIGURE 20:

Comparison between RF and RF-DNN for different number of trees used



The table below provides a summary of the experiment's findings:

TABLE 10:

RFDNN and RF results

Trees	RFtrainLoss	Rfvalid MAE	RFDNN Valid MAE
15	10.523	13.821	13.111
20	10.481	13.762	13.117
25	10.389	13.726	13.124
40	10.37	13.684	13.13
100	10.325	13.627	12.972
200	10.319	13.599	12.907

Trial and error was used to choose the RfDNN model's hyper-parameters. The study was unable to grid search properly across all possible combinations of different hidden layers, activation functions, regularization layers, Batch Norm(,), learning rate, learning rate annealing, and other parameters.

CHAPTER 5
CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This section provides conclusion derived from the study and descriptions of how the study extended previous studies, models, or frameworks. The section is concluded by providing recommendations for future research.

5.2 Conclusion

To sum up, using financial and profile data, the study was able to successfully find the best supervised regressor to forecast a reference score. The research shows that tree-based models are more effective at identifying patterns in tabular data with a predominance of numeric values. It was found that 4-fold Cross Validation along with Randomized and Grid Search over the hyper-parameter space is a solid method for supervised model optimization.

Both the Random Forest and Gradient Boosted Regression models found that annual income, maximum open credit, current credit balance, monthly debt, and current loan amount had a significant impact on model forecasts. It is also reasonable to assume that these factors will be considered when deciding whether or not to submit a candidate for manual evaluation. As a result, when generating a forecast, the models appear to prioritize the right qualities, just as a human evaluator would.

The accurate prediction of the model reveals that the factors that have a negative impact on the prediction—high current credit balances and monthly debt values, which are regarded as red flags when considering a loan request—are related to these factors. Because long-term loans are more likely to default, the duration has a negative impact on the forecast. An acceptable wage and a long credit/loan history, on the other hand, help the prediction conclusion. Based on these interpretations, there is more confidence in the model's ability to select the relevant characteristics to produce positive and negative weights during a forecast.

The distribution of the model's predictions and the actual values was similar. The model correctly predicted credit score values below the qualifying criterion (i.e 710). As a result, the model did a good job of predicting defaulters. It did not predict the bimodal distribution of credit ratings over the eligibility criterion very well. This is not a major concern because the goal of credit scoring is to correctly predict defaulters and ineligible individuals.

As a result, the study demonstrates an intriguing approach to predicting an individual's credit score. In today's ever-changing economy, implementing such a system can yield remarkable results, which can then be used to assess borrowers' credit risk and allow all financial institutions to continue operating in a transparent and profitable manner.

5.3 Contributions of the study

In this work, several methods have been developed and enhanced to significantly enhance the performance of classifiers and combiners. The study's main contribution is that it provides a credit scoring model capable of classifying a credit as risky or not, backed up by an architecture that improves the development and maintenance process of exploring the available data.

5.4 Recommendation for Future Research

This study suggests a number of recommendations based on the limitations. Analysis should be performed on data with many more features, such as monthly transactional details of customers' accounts, to enable feature engineering, which may improve performance by having a more optimal subset. Furthermore, given the growth of mobile loans in Kenya, the use of alternate data, which is a trending issue in digital lending, would be an interesting area to investigate. Future works should consider going a step further and modeling the default experience after classification into examining different window lengths to capture the customer's month-to-month changes. Future work should consider examining how RfDNN can perform grid search properly across all possible combinations of different hidden layers, activation functions, regularization layers, Batch Norm(,), learning rate, learning rate annealing, and other parameters.

REFERENCE:

A. Gelzinis, A. Verikas, E. Vaiciukynas, M. Bacauskiene, J. Minelga, M. Hallander, V. Uloza, and E. Padervinskis, \Exploring sustained phonation recorded with acoustic and contact microphones to screen for laryngeal disorders", Dec. 2014, pp. 125{132. doi: 10.1109/CICARE.2014.7007844.

A. Lawi, F. Aziz, and S. Syarif, (2017). "*Ensemble gradientboost for increasing classification accuracy of credit scoring*", in 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), Aug. 2017, pp. 1-4. DOI: 10.1109/CAIPT.2017.8320700.

Arabameri, A., Chandra Pal, S., Rezaie, F., Chakrabortty, R., Saha, A., Blaschke, T., ... & Thi Ngo, P. T. (2021). Decision tree based ensemble machine learning approaches for landslide susceptibility mapping. *Geocarto International*, 1-35.

Ardabili, S., Mosavi, A., & Várkonyi-Kóczy, A. R. (2019, September). Advances in machine learning modeling reviewing hybrid and ensemble methods. In *International Conference on Global Research and Education* (pp. 215-227). Springer, Cham.

A. Vellido, J. D. Mart in-Guerrero, and P. J. Lisboa, (2012). "Making machine learning models interpretable.", vol. 12, pp. 163-172.

Basu, A. (2017). Use of Scoring Approach in Credit Decision.

Bonga, W. G., Chirenje, G., & Mugayi, P. (2019). Analysis of Credit Culture in the Zimbabwean Banking Sector. *DRJ-Journal of Economics & Finance (2019)*, *4*(2), 45-55.

Brieman L, (2001). Random Forests, Machine Learning, 45, 5-32.

B. Twala, "Multiple classifier application to credit risk assessment", (2010). *Expert Systems with Applications*, vol. 37, no. 4, pp. 3326-3336.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70-79.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.

C. Molnar, *Interpretable machine learning*, Sep. 2019. [Online]. Available: https://christophm. github. io/interpretable- ml- book/feature - importance.html.

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263.

D. P. Kingma and J. Ba, (2014). "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*.

Du, G., Liu, Z., & Lu, H. (2021). Application of innovative risk early warning mode under big data technology in Internet credit financial risk assessment. *Journal of Computational and Applied Mathematics*, *386*, 113260.

Francis, A., Caleb, T., & Eton, M. (2022). Credit Risk Management Practices and Loan Performance of Commercial Banks in Uganda.

Gareth, James; Witten, Daniela; Hastie, Trevor & Tibshirani, Robert. 2021. *An Introduction to Statistical Learning: with applications in* R. New york: Springer.

Guotai, C., Abedin, M. Z., & Moula, F. E. (2017). Modeling credit approval data with neural networks: an experimental investigation and optimization. *Journal of Business Economics and Management*, *18*(2), 224-240.

G. V. Attigeri, M. Pai, and R. M. Pai, (2017). "Credit risk assessment using machine learning algorithms", *Advanced Science Letters*, vol. 23, no. 4, pp. 3649-3653.

Golzadeh, M., Hadavandi, E., & Chelgani, S. C. (2018). A new Ensemble based multi-agent system for prediction problems: Case study of modeling coal free swelling index. *Applied Soft Computing*, *64*, 109-125.

H. Ma, X. Yang, J. Mao, and H. Zheng, (2018). "The energy efficiency prediction method based on gradient boosting regression tree", *in 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2)*, IEEE, 2018,pp. 1-9.

Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. Cognitive Systems Research, 52, 317-324.

Hussain, A., Khan, M., Rehman, S. U., & Khattak, A. (2019). Credit scoring model for retail banking sector in Pakistan. *Journal of Managerial Sciences*, *14*(4), 153-161.

I. Goodfellow, Y. Bengio, and A. Courville, (2016). *Deep learning*. MIT press.

Jiang, Q., Zhu, L., Shu, C., & Sekar, V. (2022). An efficient multilayer RBF neural network and its application to regression problems. *Neural Computing and Applications*, *34*(6), 4133-4150.

J. Nalic and A. _Svraka, (2018). "*Using data mining approaches to build credit scoring model: Case study|implementation of credit scoring model in microfinance institution*", in INFOTEH-JAHORINA (INFOTEH), 2018 17th International Symposium, IEEE, 2018, pp. 1-5.

Khemakhem, S., Said, F. B., & Boujelbene, Y. (2018). Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines. *Journal of Modelling in Management*.

Kokate, & Chetty, M. S. R. (2021). Credit Risk Assessment of Loan Defaulters in Commercial Banks Using Voting Classifier Ensemble Learner Machine Learning Model. International Journal of Safety and Security Engineering, 11(5), 565–572.

Kumari, R., & Srivastava, S. K. (2017). Machine learning: A review on binary classification. *International Journal of Computer Applications*, *160*(7).

Kuncheva L, (2005). Diversity in Multiple Classifier Systems, Information Fusion, Vol 6, Issue 1, 3-4.

L. Breiman, (2001). "Random forests", Machine learning, vol. 45, no. 1, pp. 5-32.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.

Li, T. R., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R., & Fu, F. (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *Frontiers in Physics*, *7*, 98.

Li, W., Paffenroth, R. C., & Berthiaume, D. (2021). Neural Network Ensembles: Theory, Training, and the Importance of Explicit Diversity. *arXiv preprint arXiv:2109.14117*.

Mao, D., Wang, F., Hao, Z., & Li, H. (2018). Credit evaluation system based on blockchain for multiple stakeholders in the food supply chain. *International journal of environmental research and public health*, *15*(8), 1627.

M. T. Ribeiro, S. Singh, and C. Guestrin,(2016). "Model-agnostic interpretability of machine learning", *arXiv preprint arXiv:1606.05386*.

Munguti, V. M., & Ngali, R. M. (2020). Evaluating Credit Worthiness of Small and Growing Technology Businesses.

Nikpey Somehsaraei, H., Ghosh, S., Maity, S., Pramanik, P., De, S., & Assadi, M. (2020). Automated data filtering approach for ANN modeling of distributed energy systems: Exploring the application of machine learning. *Energies*, *13*(14), 3750.

P. Addo, D. Guegan, and B. Hassani. (2018). "Credit risk analysis using machine and deep learning models", Risks, vol. 6, no. 2, p. 38.

Pang, S., Hou, X., & Xia, L. (2021). Borrowers' credit quality scoring model and applications, with default discriminant analysis based on the extreme learning machine. *Technological Forecasting and Social Change*, *165*, 120462.

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74-78.

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., & Kundu, S. (2018). Improved random forest for classification. *IEEE Transactions on Image Processing*, *27*(8), 4012-4024.

Peprah, W. K., Agyei, A., & Oteng, E. (2017). Ranking The 5C's of credit analysis: Evidence from Ghana banking industry. *International Journal of Innovative Research and Advanced Studies*, *9*, 78-80.

P. Ferrando, (2018). *Understanding how lime explains predictions*, Dec. 2018. [Online]. Available: https://towardsdatascience.com/understanding-how- lime-explains-predictions-d404e5d1829c.

P. Grover, *Gradient boosting from scratch*, Aug. 2019. [Online]. Available:https: //medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

R. G. Lopes, R. N. Carvalho, M. Ladeira, and R. S. Carvalho, (2016). "Predicting recovery of credit operations on a brazilian bank", in 2016 15th IEEE Inter-national Conference on Machine Learning and Applications (ICMLA), IEEE, 2016, pp. 780-784.

Rokach, L. (2019). *Ensemble learning: pattern classification using ensemble methods*.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 1-16.

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons.

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*. Society for industrial and Applied Mathematics.

Tonester524, Understanding random forest, Aug. 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

Vishalmorde, "Xgboost algorithm: Long may she reign!", Medium, Apr. 2019. [Online]. Available: https : / / towardsdatascience .com/https - medium- comvishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data. Procedia Computer Science, 174, 141-149.

Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, *93*, 182-199.

Y. Kong and T. Yu, (2018). "A deep neural network model using random forest to extract feature representation for gene expression data classification", *Scientific Reports*, vol. 8, no. 1, p. 16 477, 2018, issn: 2045-2322. doi: 10.1038/s41598-018-34833-6. [Online]. Available: https://doi.org/10.1038/s41598-018-34833-6.

Z. C. Lipton, (2016). "The mythos of model interpretability", *arXiv preprint arXiv:1606.03490*.

Zeng, B., Wei, X., Zhao, D., Singh, C., & Zhang, J. (2018). Hybrid probabilistic-possibilistic approach for capacity credit evaluation of demand response considering both exogenous and endogenous uncertainties. *Applied energy*, *229*, 186-200.

Zhang, S., Cheng, D., Deng, Z., Zong, M., & Deng, X. (2018). A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, *109*, 44-54.

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(3), 1-19.

Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, (2015). "Investigation and improvement of multi-layer perceptron neural networks for credit scoring", *Expert Systems with Applications*, vol. 42, no. 7, pp. 3508-3516.

APPENDIX

i. Research Schedule

Timeline Activity	Start Date	Target End Date	Completion date
Topic selection	1st September 2021	1st September 2021	
Design Problem Statement, Objectives & Research Questions	2nd September 2021	15th September 2021	
Obtain and completion ofLiterature review	16th September 2021	15th October 2021	31st October 2021
Decide on research methods	1st November 2021	15th November 2021	30th November 2021
Completion of Research Proposal		15th December 2021	31st December 2021
Addressing Panel Feedback/Corrections	26th February 2022	7th March 2022	7th March 2022
Proposal approved		7th March 2022	17th March 2022
Data Collection and Analysis	17th March 2022	15th April 2020	20th April 2020
Completion of First Draft		24th April 2022	
Addressing supervisor Feedback	30th April 2021	5th May 2022	30th May 2022
Submit revised draft		5th July 2022	23rd July 2022
Approval of Master's Research Dissertation		23rd July 2022	30th August 2022

ii. Resources and Budget

Budget Items	Expected Cost (Kenya Shillings)
Project development- a) internet costs	Kes. 40,000.00 (Kes 3000 per Month for 8 Months)
Software	The project will rely on Python Open-Source Package.
Hardware	Kes. 50,000.00 (Average price for Student Laptop)
Printing and Hard Cover Binding	Kes. 12,000.00
Miscellaneous	Kes. 10,000.00
TOTAL BUDGET	**KES 112,000**

iii. Sample of data used in the study

LINE_CONTRACT	original_amount	Exposure_Kes'000	Exposure_Bracket	log_amount	Limit_KES	Arrears_Kes	ACCT_DPD	CUST_DPD	Default_Status	CLASSIFICATION	loan_groups	CURRENCY	FCY_LCY	BUSINESS_UNIT	PRODUCT	product_name	INTEREST_RATE	IFRS_INTEREST_RATE
2	2910.95	2.91095	0-50	3.464034746	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	GOLD ALL IN ONE CURRENT	Overdrafts	12.89	12.89
2	6842.24	6.84224	0-50	3.835198304	0	0	0	0	0	NORMAL	A	GBP	FCY	Personal Banking	GOLD ALL IN ONE CURRENT	Overdrafts	10.99	10.99
4	4025643.8	4025.6438	>1000	6.604835345	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Personal Unsecured Scheme Loan	Personal Unsecured Loans	13	13
5	205881.45	205.88145	100-250	5.313617218	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - ASSET FINANCE	Asset Finance Loans	13	13
7	2884812.63	2884.81263	>1000	6.460117611	0	92508.34	21	21	1	NORMAL	A	KES	LCY	Personal Banking	LOAN - PERSONAL	Personal Unsecured Loans	13	13
8	21115	21.115	0-50	4.324591086	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
9	960	0.96	0-50	2.982271233	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	GOLD ALL IN ONE CURRENT	Overdrafts	12.89	12.89
10	44557.8	44.5578	0-50	4.64892374	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
11	2147.02	2.14702	0-50	3.33183609	0	0	0	0	0	NORMAL	A	USD	FCY	Personal Banking	PLATINUM PAY AS YOU GO	Overdrafts	11.99	11.99
11	344.18	0.34418	0-50	2.536785631	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	PLATINUM PAY AS YOU GO	Overdrafts	12.89	12.89
12	367.29	0.36729	0-50	2.565009104	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Platinum ALL IN ONE	Overdrafts	12.89	12.89
13	321.04	0.32104	0-50	2.506559147	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	PLATINUM FLEXI CURRENT	Overdrafts	12.89	12.89
13	1078118.76	1078.11876	>1000	6.032666603	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - PERSONAL	Personal Unsecured Loans	13	13
13	30192.3	30.1923	0-50	4.479896198	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
14	522.53	0.52253	0-50	2.71811123	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	PLATINUM FLEXI CURRENT	Overdrafts	12.89	12.89
16	251295.12	251.29512	250-500	5.400184055	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - PERSONAL	Personal Unsecured Loans	13	13
17	372545.44	372.54544	250-500	5.571179252	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - PERSONAL	Personal Secured Loans	13	13
18	420140.25	420.14025	250-500	5.62339429	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - ASSET FINANCE	Asset Finance Loans	13	13
22	14084.75	14.08475	0-50	4.148749143	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
23	6017.05	6.01705	0-50	3.77938362	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
25	9130941.31	9130.94131	>1000	6.960515551	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - MORTGAGE	Mortgage Loans	5	5
26	20551.1	20.5511	0-50	4.312835072	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
27	86827.04	86.82704	50-100	4.938654996	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - ASSET FINANCE	Asset Finance Loans	13	13
28	687.63	0.68763	0-50	2.837354816	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	GOLD ALL IN ONE CURRENT	Overdrafts	12.89	12.89
29	580948.65	580.94865	500-1000	5.764137747	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Personal Unsecured Non Scheme Loan	Personal Unsecured Loans	13	13
30	12317.44	12.31744	0-50	4.090520455	0	11275.42	30	30	1	NORMAL	A	USD	FCY	Personal Banking	GOLD ALL IN ONE CURRENT	Overdrafts	11.99	11.99
30	1652561.08	1652.56108	>1000	6.21815752	0	0	0	30	1	NORMAL	A	KES	LCY	Personal Banking	LOAN - PERSONAL	Personal Unsecured Loans	13	13
32	50373.95	50.37395	50-100	4.702206007	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
33	57507	57.507	50-100	4.759720712	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Individual IPF	Insurance Premium Finance Loans	0	13
33	20	0.02	0-50	1.301029996	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	PLATINUM PAY AS YOU GO	Overdrafts	12.89	12.89
34	420	0.42	0-50	2.62324929	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	PLATINUM FLEXI CURRENT	Overdrafts	12.89	12.89
35	1291228.45	1291.22845	>1000	6.111003086	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - PERSONAL	Personal Unsecured Loans	7	7
36	802.56	0.80256	0-50	2.90447751	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	GOLD ALL IN ONE CURRENT	Overdrafts	12.89	12.89
37	2727641.7	2727.6417	>1000	6.435787321	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Personal Unsecured Non Scheme Loan	Personal Unsecured Loans	13	13
37	5225004.66	5225.00466	>1000	6.718086682	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - MORTGAGE	Mortgage Loans	13	13
39	21506670.86	21506.67086	>1000	7.332573189	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - MORTGAGE	Mortgage Loans	5	5
40	1199979.05	1199.97905	>1000	6.079173664	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Personal Unsecured Non Scheme Loan	Personal Unsecured Loans	13	13
40	51018.15	51.01815	50-100	4.707724708	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
41	30267.1	30.2671	0-50	4.480970812	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
44	13004.6	13.0046	0-50	4.114096999	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	Mobile Loan	Digital Loans	13	13
46	829577.73	829.57773	500-1000	5.918857085	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	LOAN - PERSONAL	Personal Secured Loans	13	13
47	242.57	0.24257	0-50	2.384837088	0	0	0	0	0	NORMAL	A	KES	LCY	Personal Banking	SALARY CURRENT ACCOUNT	Overdrafts	12.89	12.89

Contract_Date	Maturity_Date	Report_Date	time_diff	INDUSTRY_CODE	INDUSTRY	CBK_SECTOR	age	age_bracket	status	Employment_Status	Year_in_Service	gender
1/11/2007	1/1/1900	9/30/2021	5,376.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	27	18-34	Married	Employed	2	Male
11/6/2012	1/1/1900	9/30/2021	3,250.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	27	18-34	Married	Employed	2	Male
9/13/2021	10/5/2025	9/30/2021	17	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	41	34-43	Divorced	Employed	16	Male
1/10/2019	1/5/2022	9/30/2021	994	1845	WOOD & WOOD PRODUCTS	MANUFACTURING	58	>53	Widowed	Employed	33	Male
4/18/2019	11/9/2024	9/30/2021	896	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	52	44-53	Married	Employed	27	Female
9/29/2021	10/29/2021	9/30/2021	1	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	21	18-34	Single	Not employed	2	Female
4/7/2004	1/1/1900	9/30/2021	6,385.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	33	18-34	Married	Employed	8	Male
9/13/2021	10/13/2021	9/30/2021	17	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	31	18-34	Single	Not employed	6	Female
1/11/2021	1/1/1900	9/30/2021	262	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	58	>53	Divorced	Not employed	33	Female
8/17/2005	1/1/1900	9/30/2021	5,888.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	58	>53	Divorced	Not employed	33	Female
7/28/2005	1/1/1900	9/30/2021	5,908.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	35	34-43	Divorced	Employed	10	Female
7/10/2006	1/1/1900	9/30/2021	5,561.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	51	44-53	Widowed	Employed	26	Male
11/9/2018	11/5/2022	9/30/2021	1,056.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	51	44-53	Widowed	Employed	26	Male
9/12/2021	10/13/2021	9/30/2021	18	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	51	44-53	Widowed	Employed	26	Male
7/3/2006	1/1/1900	9/30/2021	5,568.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	37	34-43	Divorced	Not employed	12	Male
3/11/2019	2/28/2023	9/30/2021	934	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	52	44-53	Widowed	Employed	27	Male
3/20/2020	2/28/2025	9/30/2021	559	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	37	34-43	Married	Employed	12	Male
9/4/2020	9/5/2024	9/30/2021	391	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	46	44-53	Married	Employed	21	Female
9/14/2021	10/14/2021	9/30/2021	16	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	44	44-53	Divorced	Not employed	19	Male
9/23/2021	10/23/2021	9/30/2021	7	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	48	44-53	Divorced	Employed	23	Male
10/9/2019	3/30/2041	9/30/2021	722	2410	RETAIL	PERSONAL HOUSEHOLD	29	18-34	Single	Not employed	4	Male
9/23/2021	10/24/2021	9/30/2021	7	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	40	34-43	Single	Not employed	15	Male
2/1/2018	1/28/2022	9/30/2021	1,337.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	40	34-43	Single	Not employed	15	Male
10/12/2006	1/1/1900	9/30/2021	5,467.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	27	18-34	Divorced	Employed	2	Female
6/30/2021	1/26/2022	9/30/2021	92	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	35	34-43	Married	Not employed	10	Male
6/2/2018	1/1/1900	9/30/2021	1,216.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	60	>53	Divorced	Not employed	35	Male
3/16/2019	4/5/2024	9/30/2021	929	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	60	>53	Divorced	Not employed	35	Male
9/10/2021	10/10/2021	9/30/2021	20	3478	OTHER BUSINESS SERVICES	TRADE	21	18-34	Single	Not employed	1	Male
7/8/2021	10/8/2021	9/30/2021	84	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	53	44-53	Single	Employed	28	Female
3/18/2005	1/1/1900	9/30/2021	6,040.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	53	44-53	Single	Employed	28	Female
5/4/2007	1/1/1900	9/30/2021	5,263.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	21	18-34	Divorced	Employed	1	Male
10/2/2020	9/24/2026	9/30/2021	363	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	46	44-53	Single	Employed	21	Male
4/20/2012	1/1/1900	9/30/2021	3,450.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	19	18-34	Divorced	Employed	1	Male
12/22/2020	1/3/2025	9/30/2021	282	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	29	18-34	Widowed	Not employed	4	Male
8/4/2017	8/4/2027	9/30/2021	1,518.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	29	18-34	Widowed	Not employed	4	Male
9/4/2020	9/5/2041	9/30/2021	391	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	19	18-34	Divorced	Employed	1	Male
4/12/2021	3/26/2026	9/30/2021	171	4440	INSURANCE COMPANIES	FINANCE & INSURANCE	41	34-43	Married	Employed	16	Female
9/30/2021	10/31/2021	9/30/2021	-	4440	INSURANCE COMPANIES	FINANCE & INSURANCE	41	34-43	Married	Employed	16	Female
9/6/2021	10/6/2021	9/30/2021	24	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	27	18-34	Single	Not employed	2	Male
9/30/2021	10/31/2021	9/30/2021	-	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	27	18-34	Married	Not employed	2	Male
4/25/2016	4/25/2022	9/30/2021	1,984.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	36	34-43	Widowed	Not employed	11	Female
7/26/2007	1/1/1900	9/30/2021	5,180.00	4260	PERSONAL SERVICES	PERSONAL HOUSEHOLD	23	18-34	Widowed	Employed	1	Male