2023

A Loan default prediction and loan amount recommendation tool for SACCOs in Nairobi: a case of Okoa Management SACCO.

Mwalozi, Purity Monje
School of Computing and Engineering Sciences
Strathmore University

A Loan Default Prediction and Loan Amount Recommendation Tool for Saccos in Nairobi: A Case of Okoa Management Sacco

By

Mwalozi Purity Monje

072959

Master of Science in Information Technology

2023

A Loan Default Prediction and Loan Amount Recommendation Tool for Saccos in Nairobi: A Case of Okoa Management Sacco

By

Mwalozi Purity Monje

072959

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University

School of Computing &Engineering Sciences

Strathmore University

Nairobi, Kenya

July, 2023

Declaration and Approval

Declaration

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

Student's Name: Mwalozi Purity Monje

Sign: _____ Date: _____

Approval

The thesis of Mwalozi Purity Monje was reviewed and approved for examination by the following:

Dr. Kennedy Ronoh
School of Computing & Engineering Sciences,
Strathmore University

Dr. Julius Butime,
Dean, School of Computing & Engineering Sciences,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

SACCOs loan delinquency is a severe danger to the organization's capacity to continue availing loans to loan applicants and to grow. SACCOs are unable to collect what they have lent out to loan beneficiaries as the default rate rises gradually. This research project aimed at using the analysis of the different factors that determine loan defaults in microfinance institutions, microlending institutions and SACCOs in Kenya with a focus on Okoa Management Ltd. and how the same factors can be used to predict the likelihood of a loan borrower to default in the repayment process by applying machine learning algorithms. Credit risk assessment precision is important to the functioning of lending institutions. Traditional and most existing credit score models are developed and designed using demographic characteristics, historical payment data, credit bureau data and application data, with most of them not suitable for developing countries such as Kenya which consider the employment type (casual, temporary, contractual or permanent) and the fact that we can lend up to 3 times as much as the borrower's savings. With these factors being constantly changing and dynamic, credit risk models based on machine learning algorithms provide a higher level of accuracy in predicting default as they can be continuously trained with new data sets should the variables that are used change. Risk management has been an increasing issue for credit lending institutions as the need to determine the likelihood of defaulting by borrowers is becoming more evident. By using machine learning, we can be able to reduce the uncertainty that comes with borrowing and even go further to recommending lower amounts for borrowers who we predict are likely to default in the repayment of the loan amount they have in mind. The research focused on three main algorithms: logistic regression, decision trees and tensor flow on the prediction. The algorithm that provided the best accuracy was the decision tree. The results of the research showed that people with little or no collateral (home-ownership/car ownership) were more likely to default and that there was a low correlation between months since last delinquent and the loan prediction default likelihood status.

Keywords: *loan default prediction, machine learning, credit lending*

Table of Contents

List of Figures

List of Tables

Abbreviations/Acronyms

AML	Amazon machine learning
AWS	Amazon Web Services
FOSA	Front Office Savings Account
FICO	Fair Isaac Corp
ML	Machine Learning
SACCO	Savings and Credit Co-operative Society

List of Appendices

Acknowledgments

I would like to take this opportunity to express my heartfelt gratitude to the individuals listed below, whose contributions have aided in the successful completion of this research project proposal. Dr. Kennedy Ronoh, my supervisor, without whose guidance and help this work would not have been feasible. Professor Ismail Ateya, for his advice on how to structure and compose this work, is also thanked by the researcher.

Strathmore University's university library has been quite helpful in allowing me to access various academic materials that were used in the analysis and compilation of this academic work. During the studies, all colleagues and friends who provided fascinating encouragement. Most importantly, for the gift of good health and the strength to complete my academic work effectively I am grateful to God.

Dedication

I wish to dedicate this work to my parents; both of whom gave me the foundation of education. And from them I have able to greatly appreciate the value of reading and learning and understand that it is a never-ending journey. I wish to dedicate it to my siblings as well; they encouraged me to push on even when things got tough (late nights, frustration when things did not go my way) and to work hard.

Definition of terms

FICO Score	A credit score used by lenders to evaluate and quantify a borrower's credit worthiness to determine whether to extend credit or not (Hayes, 2021).
FOSA	Front Office Service Activity. It is a transactional SACCO account that offers banking services like those offered by Commercial Banks such as loan repayment, standing order set up, etc. (Kimisitu Sacco, 2021).
SACCO	Savings and Credit Co-operative Societies. It is an organization in which a group of people stock their savings and offer loans to their own members (Gundaniya, n.d.).

Chapter 1: Introduction

1.1. Background Information

Customarily, advancing loans has been based on the establishment of believe and trust. In spite of the fact that there were credit report measuring tools like the Fair Isaac Corp (FICO) score from as early as 1989, in some areas the money loaning process is still a bit reasonably subjective, and potential borrowers are sometimes frequently judged by how trust-worthy their character appeared and by who the know and the connections they have. Nowadays, banks are able to utilize instruments like FICO Scores to measure how dependable potential borrowers are, minimizing arbitrariness. All of which is usually done for one reason: to decide how likely it is that a given borrower will default a credit.

Determining whether or not disbursing a loan will result in the making of a profit or loss is an important component of money lending that can be done by predicting default likelihood and default rates. Loans are normally profitable due to interest, but borrowers may default (Zhao, 2020), which is a breach of the moneylender's trust as well as a risk to the moneylender's business. As a result, it's critical for a lender to be able to assess the possibility of a borrower failing before issuing a loan to him or her.

Savings and Credit Co-operative Societies (SACCOs) are begun locally and have strong bases of little sparing accounts constituting a steady and moderately low-cost source of financing and more authoritative costs. Sacco Social Orders are not only the fastest-developing component of Kenya's Agreeable Development, but they are also the most significant in influencing individuals' capacity to how much they are allowed to borrow. (Olando, Mbewa, & Jagongo, 2013). They are as of now controlled through the Agreeable Act beneath SACCO Social orders Administrative Specialist in the SACCO Societies Regulatory Authority (SASRA) which is prudentially directing Front Office Savings Account (FOSA) working SACCOs. By focusing on the FOSAs, SACCOS can now operate like banking institutions, which means they can lend out money to members and borrow money from other finance institutions and that they are also likely to face the same challenges that banks face as well such as correctly disbursing loan amounts. According to a financial report by (Financial Access, 2013), SACCOs lost their market share in spite of the fact that their geographical spread in the country compared to other financial providers is greater and wider. This could be attributed to challenges that SACCOs faces due to the characteristics of the market segment it serves such as loan defaulting from its borrowers.

According to (Salaton et al, 2020) the SACCO movement is one of Kenya's most highly recognized strategies for increasing resource mobilization and utilization. However, studies on the elements that influence the performance of these cooperatives have yielded inconsistent results, indicating that more research is needed. Many individuals in Kenya and other poor countries benefit from savings and credit cooperative societies despite there being a large demand for the service compared to the supply (Marwa & Aziakpono, 2015). The inability of SACCOs to expand is hampered by a shortage of funding – and the failure of borrowers to repay their loans does not help with the situation. Saving and credit cooperative societies play an important role in financial intermediation, which is problematic when the demand for money exceeds the supply. The world Council of Credit Union suggested that SACCOs worldwide face similar challenges which brought about regulations that are remedial measures for future financial crisis (Olando & Mbewa, 2012).

SACCOs make money through profits. They earn these profits through interest repaid by members when they repay their loans (Opiyo, 2014). Without being able to correctly predict the member's ability to be able to repay their loans, SACCOs increase their risk to money loss. With SACCOs relying on this money for investment in different industries and for borrowing from larger financial institutions, it becomes difficult for them to increase their trustworthy score if most of their member default. If we find a way to allow member's loan repayment default likelihood to be predicted, then we reduce the risk of bad debts by the SACCO. To reduce the risk even further, the researcher aims at also providing recommended lower amounts for borrowers who have been predicted to default in the loan amount they have applied for.

According to (Gouda, A, Madivala, & R, 2021), we do need to keep updating and upgrading the tools, models and algorithms currently being used as they present challenges in the following areas:

i. Computing glitches. Computer glitches such as slow speeds, computational errors are rare but could occur. (Gouda, A, Madivala, & R, 2021) suggest that with ML languages being updated every so often, it is important for new the existing tools to be updated to ensure more accurate results by utilizing the patches, new release features, and so much more that comes with the language updates.
ii. Content errors. Since the existing models are made with the concept on one size fits all, the different features(factors) that affect defaulting in one country or organization cannot necessarily be the same for another. Other than the generic factors like age and levels of

income, different lenders have different needs, for instance like the SACCO to be used for this study has factors like availability of guarantors, temporary (casual and seasonal) employment as types of employment and this determine the loan amount you can borrow, etc. The aim of this research was to allow for the development of a tool that can allow the lenders to provide their own factors to use provided they have the matching data so that the model can be trained accordingly.

iii. Feature weighting. The researcher proposed to have a more dynamic weight adjustment by constantly updated the training set to ensure that the critical factors that will affect the rate of default keep changing based on the actual situation in the area of use.

iv. Aging. Since the tools were trained using old training data sets, it becomes difficult to get the best recommendations since the behavior keeps changing. The researcher's proposal aimed at making a tool that will keep re-learning, as any correct predictions and recommendations will continually be updated as part of the training set.

1.2. Problem Statement

Microfinance institutions and SACCOs use credit scoring models to evaluate loan default risks potential. The scores generated by these models translate to the likelihood of defaulting, making it easier to make lending decisions. However, these models are fixed and do not easily evolve with changing customer behavior to predict the likelihood of defaulting more accurately (Wanjohi et al. 2016). The authors argue that existing models use old data sets to train their models meaning that if new characteristics are introduced the model will not predict the output correctly. The accuracy of loan default prediction can be enhanced by machine learning approaches.

According to (Jemoek, 2013), loan defaulting has a direct correlation with the SACCOs and/or microfinances ability to borrow from other Bankers and has the effect of affecting the institution's liquidity. With better ways to predict the likelihood of loan repayment default, we can be able to be able to reduce the number of defaulters significantly. According to (Central Bank of Kenya, 2019), the number of defaulters increased by around 36% from 12% in 2016 to 48.7% in 2019.

With existing models not being able to correctly predict the likelihood of a borrower defaulting, SACCOs and microfinance institutions run a higher risk of losing revenues from the interests that would have been earned on the borrowed loans. Some of the existing models use variables that are custom to

their countries of use and don't include aspects that are specific to the Kenyan market e.g., casual laborers as a form of employment type, current savings amount, the three times savings amount as the loan limit, among others. Other than this, these models don't recommend a lower reduced amount to the borrowers if they end up having a bad credit score or having been predicted to be a likely defaulter. For instance, in the credit scoring method, the credit score just determines the rates that will be offered by the financial institution. The better the rate offered by the financial institution, the higher the credit score. (The Investopedia Team, 2023). The existing models for prediction use general variables that cut across all countries without having focus on the specific needs of the different countries. Being able to recommended lower loan amount to borrowers who are likely to default in the repayment of the amount they initially seek for will help in ensuring that there is still continuity of the business processes since interest will be earned. According to Felix Gichina, the correspondent from Okoa Ltd., they currently rely on only the previous borrowing history.

To ensure continuity of business through interest, if a borrower is likely to default on the amount they are seeking, we should have a way to recommend a lower amount for them based on their characteristics. This will help both the lender and the borrower as the lender will still earn some interest and the borrower will have some relief as they will get at least a partial amount of the amount they intended to borrow.

1.3. General Objective

To develop a machine learning model that can predict loan default likelihood and recommend lower loan amounts for borrowers who have been predicted to be likely to default.

1.4. Specific Objectives

i. To identify the leading characteristics and attributes analyzed when processing loans in loan default prediction models.
ii. To review existing techniques, models and algorithms applied in default prediction and amount recommendation.
iii. To develop a machine learning model that can predict loan defaulting and recommend lower amounts.
iv. To evaluate the model's performance in loan default prediction and amount recommendation.

1.5. Research Questions

i. What are the characteristics and attributes used for loan default prediction models?
ii. What are the currently existing models and algorithms used in loan default prediction and amount recommendation?
iii. How does one develop and train a machine learning model that can forecast default likelihood and recommend loan amounts?
iv. How will the model performance be evaluated?

1.6. Scope

The scope of this study was be a SACCO, as SACCOs have the highest rate of defaulters. (Central Bank of Kenya, 2019). The main reason am focusing on a SACCO as the case study target was to be able to have access to the data on previous loan defaults and the bio data of the borrowers excluding any information that may be considered to be personally identifiable information (PII). This will also provide a better sample as the SACCO under study had members of all ages and employment types and was as not as limiting as mobile loan applications as this favors the youth (Maina, 2021) and disadvantages the elderly from usage of such apps either because they are not as tech-savvy as the youth or because of lack of access to the internet. The main SACCO that will be used for the purposes of this research is Okoa Management Ltd. The SACCO was selected because of the fact that it caters to a wide and diverse clientele. It caters to lower, middle and upper middle earners which provide us with a broad data set to cater for all different income levels. According to Felix Gichina, my correspondent from the organization, they not only offer loans to high income earners, they also consider themselves a leading micro-lender in the country. The SACCO is located in Nairobi with their offices along Chepkorio Road, Industrial Area and has a membership of about 6000 members and a turnover of about Kshs. 10 million to Kshs. 20million a year. They are a good candidate as they currently have not yet adopted any automated system to predict bad loans and purely rely of the borrower's borrowing history with them yet there are much more factors that affect a borrower's ability to repay their loan. The main loan delinquency aspects that were the focus of study was the repayment time as this is what most SACCOs usually use to determine default. The researcher will use a late payment of a duration of one month for the borrower to be considered a defaulter as the target is small SMES and micro lending institutions as it is currently how Okoa determines a defaulting borrower. The one month is one month after the due date of the final payment.

1.7. Justification

As credit directly affects an institution's profitability, assessment of credit risk is crucial to the success of lending institutions. (Corporate Finance Institute, 2022) suggests that traditional procedures are inefficient and time-consuming, and don't provide the correct credit worthiness of the lenders as they rely mostly on the Cs of the credit (character, capacity, capital and collateral) which cannot be unbiasedly used to determine the borrower's likelihood to default in loan repayment.

According to (Central Bank of Kenya, 2019), the number of defaulters has increased by 36.7% between 2016(12%) and 2019(48.7%) which is among the top challenges facing credit facilities. (Business Daily, 2020) suggest that if there was a way to determine a borrower's ability to fully pay for the loan before issuing it, then the number of defaulters would possible reduce significantly. By recommending lower amounts for the borrowers who the model predicts to be likely to default, not only do we ensure revenue for the lender (through the interest earned on the lowered amount), we also provide the borrower with some money (even if it won't be the initial full amount they requested for), after all half a loaf is better than no loaf.

Chapter 2: Literature Review

2.1. Introduction

As of June 2022, the number of defaulters in banks and micro-finance institutions has increased by 30.6 billion Kenyan shillings (Mwaniki, 2022). Without a way of predicting the likelihood, this figure is predicted to keeping increasing year and year on. With no ways of finance SMEs being able to reduce the risk, it has reduced the amount they are willing to disburse into personal and development loans. With a lot of challenges being experienced in Kenya, when it comes to loan repayment, being able to predict the likelihood of the borrower defaulting will help SACCOs in reducing the risk of bad debt and defaulting.

With machine learning continuously evolving, it can not only be used in loan default prediction, but also in determining how much to lend and the lending terms. Machine learning can be utilized to automate the decision-making process for loan amount processing and terms negotiation. By utilizing information from past credit applications, machine learning models can learn to recognize designs that are predictive of loan default. These designs can at that point then be utilized to consequently decide loan amount and terms for new loan candidates (Kumar, 2022). In Kenya, a lot of machine-learning predictive models have been used to try predict loan defaults. Some of the models used include: extreme value regression models, logistic regression and linear analysis. All these models are parametric as they expect the returned response output feedback being considered, reviewed and analyzed takes a particular functional form which might not always be the case.

This chapter focused on reviewing existing literature on previous works that have been done in loan default prediction, the models and frameworks used, their limitations and how they work. It highlights the different architectures and algorithms as well as the limitations of existing systems. After analyzing all these, the researcher then used the gained knowledge to come up with a conceptual framework on how their proposed solution will work.

2.2. Theoretical Review and Empirical Review

2.2.1. Theoretical Review

The kind of information utilized in conventional credit scoring is chronicled information which incorporates bank conventional information such as past credit, credit bureau checks, records of late payment installments and commercial information such as financial statements and length of credit history. (World Bank, n.d.).

2.2.1.1 Using Logistic Regression Theory

According to (Zhao, 2020), we can establish key correlations between default rates and a few other variables by studying variables that describe loans and the financial situations of their borrowers. This method involves using predefined variables related to the historical information of the loan amount taken and the borrower's characteristics. It takes into account the following variables: loan amount taken, annual income of the borrower, interest charged on the loan, the term and the employment type of the borrower.

The output of logistic regression is always between (0, and 1), which is suitable for a binary classification task. The higher the value, the higher the probability that the current sample is classified as class=1, and vice versa as shown by equation 1 below.

$$h_\theta(X) = 1/(1 + e^{-\theta x}) \quad (1)$$

Where:

i. Θ is the specification to be learned or trained or optimized (such as the loan default)

ii. $\mathbf{X}$ is the input data (could be the different features to train the model such as age, income levels, level of education, among others)

iii. The output is the prediction value when the value is closer to 1, which means the instance is more likely to be a positive sample(y=1). If the value is closer to 0, this means the instance is more likely to be a negative sample(y=0). In the context of the research, 0 being defaulter, 1 non-defaulter

It is a widely used technique because of it being efficient and less consuming of the computational power of the resources (Arya, 2022). The main issue with using logistic regression is that it assumes that there is complete linearity (Rout, 2020) between the dependent (loan default prediction likelihood) and independent variables(characteristics of the loan or the borrower, e.g. income) as it predicts the probability of an event or class that is dependent on other factors (Arya, 2022).

2.2.1.2 Using Score Card Theory

Numerous credit decisioning frameworks are driven by scorecards, which are exceptionally simplistic rules-based frameworks. These are built by end-user organizations through industry information or through straightforward factual frameworks. (DataRobot, 2021)A few organizations go a step further and get scorecards from third parties which may not be customized for an individual's organization's book.

The disadvantage of this approach is that one cannot for sure know the parameters that were used (especially in the case of third-party score cards) when determining the borrower's score rate.

2.2.2. Empirical Review

Different models and frameworks provide different metrices for evaluating the success of the model built. While aspects such as accuracy, standard deviation, true positives, false negatives being important, the amount of data tested will also likely affect the results of the output. As the model will be focusing on both classification/prediction and recommendation algorithms, it's important to identify the different algorithms to use to accomplish both. The main reason to have different algorithms for the different functions of the model to prevent cyclic dependencies in the model. The researcher took a look at different models that were considered in this study, how they worked and their limitations.

2.2.2.1 Linear Regression

This is an elementary machine learning model that falls under the supervised class. It is considered the hello world of machine learning and thus forms a solid base to most machine learning classification problems. Linear regression can be employed to create a prediction model that is established on the relationship that exists between the dependent and independent variables, giving more emphasis on the independent (Geeks for Geeks, 2022).

This machine learning model attempts to predict a variable (dependent), with its basis being an input variable (independent). This can be presented by the simple formula prediction = O1 + O2X.

Where O1 represents the intercept and O2 represents the coefficient of the independent variable. In order to increase the model's accuracy, it is essential to fine tune the two variables. The two variables can be tuned using the cost function, which prioritizes the difference in the error.

Linear regression makes use of a straight line (regression line) to perform its prediction. The regression line can be calculated by a number of strategies. The most popular strategy being the least square method, that attempts to reduce the total sum of the deviations from the straight line. The regression line is the one which has the least sum (Least square method, 2018). For this study linear regression shall be used to create a prediction model that factors in the most important input (using correlation). The main limitation of this method is that it assumes there is normal distribution of the variables and that there is a complete linear relationship between dependent and independent variables (Rout, 2020) which might not always be the cause.

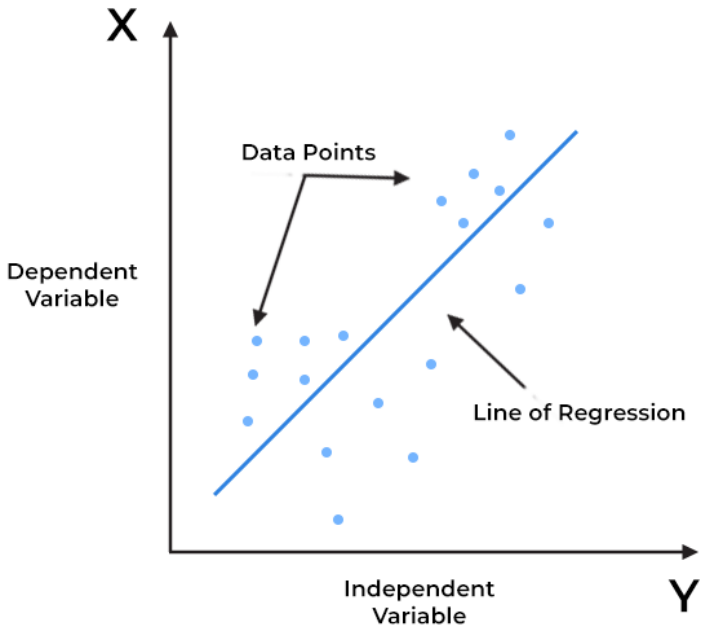


Figure 2.1: Linear Regression (Kanade, 2022)

Logistic regression is a common machine learning model that falls in the category of supervised models can be used when it comes to binary predictions (classification problems). This classification model makes use of log function. In logistic regression the set of input features are used to model a probability that it falls under a certain classification. This method can be looked at as an enhanced

multiple linear regression, save for the binomialism of its variables. Rather than fitting a line to a data logistic regression fits an “S” shaped logistic function. The line goes from 0 to 1. This gives the probability of how a certain set of inputs features can be classified. Its key merit can be considered the fact that it aims to reduce compounding effects by factoring in how the variable are related (Understanding logistic regression analysis, 2014)

When applied to our research scenario, logistic regression can be applied to predict if a borrower will end up defaulting on a loan. This can be done by creating a simple model which uses the most important (highly correlated to defaulting) variable or a more complicated model that factors in more variables to come up with a prediction of whether the client will default or not. When it comes to predicting the possibility of defaulting loans many inputs shall be employed to confirm. Logistic regression can be used as it can work with continuous data as well as discrete data. The researcher tested if a variable's impact on the forecast differs noticeably from zero. If not, it was concluded that the variable did not play a role in determining the prediction. This helped in fine tune the variable to help scrap out irrelevant input variables. The methods ability to provide probabilities as well as classify new samples using both continuous and non-continuous samples made it suitable machine learning method that was applied in this study.

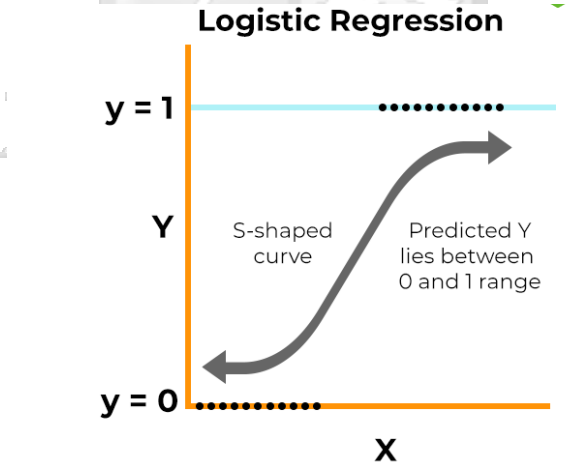


Figure 2.2: Logistic Regression (Kanade, 2022)

2.2.2.2 Decision Tree

This also falls on the category of supervised machine learning models that are employed to create a categorization algorithm. Decision trees can be used to curry out two variations: classification problems (having a binary result) and a regression tree that is used to provide a non-categorical prediction (

Xoriant, n.d.).A decision tree at its core is a binary tree that recursively splits the dataset until you arrive at the leaf nodes (data having only one type of class). There two types of nodes namely, decision node and a leaf node. The former contains a condition to further split the data while the latter provides a classification group. As you progressively move from the root node (the first node) the number of elements that fall within the nodes reduce.

The choice of the split is based on information theory. The model chooses the split that boosts and expands the information gain. In order to compute the information gain, we have to gain knowledge on the information that is contained in a particular state. In order to quantify this information gain entropy is used. Entropy quantifies impurities in a particular set of data (Ogola, 2021). If entropy is high then we are very unsure about the classification of a randomly picked point and thus more bits in order to describe the state. The aim is to reduce the entropy to 0, and calculate the split that provide the highest information gain using the formula: Gain = Entropy(parent) – Entropy (children). This informs the choice of split to make. As you go down the decision tree, the impurities in each classification reduce.

The model traverses every feature and feature value and comes up with the best feature and the corresponding threshold. Decision trees are considered to fall under the class of greedy algorithms. It takes the current optimal split that optimizes information gain and does not back truck to change a previous classification. It however does not guarantee the best possible split but makes the training fast and works really well in spite of its simplicity. Decision trees can be used to provide a classification model that is simple as well as effective in predicting the possibility of defaulting loan payments.

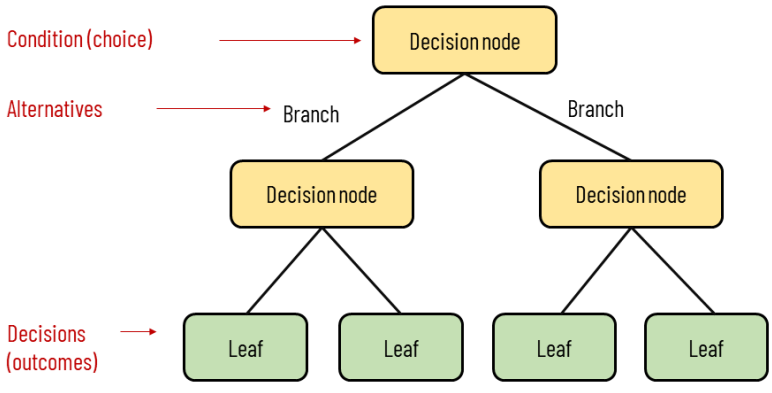


Figure 2.3: Decision Tree (Kosarenko, 2021)

2.2.2.3 Random Forest

This is an improvement on decision trees as it comprises of a couple decision trees that work as a unit (Yiu, 2019). Random forest aim at reducing bias and over fitting that is characterized with decision trees. Overfitting happens when a model begins memorizing the data instead of try to generalize the data in order to perform predictions. In such situation, the model will grasp both the noise and other changes contained in the train set and build up on them as concepts (Brownlee, 2019). By making use of an ensemble of decision trees to make a prediction. Random forest can make use of a "voting" scheme where the prediction that gets the most votes get selected as the best prediction. The performance of the random forest will improve as more decision trees are utilized with various criteria since it boosts prediction accuracy.

There is consideration that have to be satisfied in order for a random forest to be considered to have been done as expected. One is the presence of signal in the input variables that dictate correlation with the output variable thus eliminate the element of no difference with guess work. Two, the errors and predictions resulting from the individual trees need not be correlated.

To set up a random forest a number of parameters have to be set. They include the node size, the number of individual trees and the number of features. Random forests are popular and this is due to its ability to perform different types of classification problems. This model can be used to predict the probability of defaulting as needed in our research.

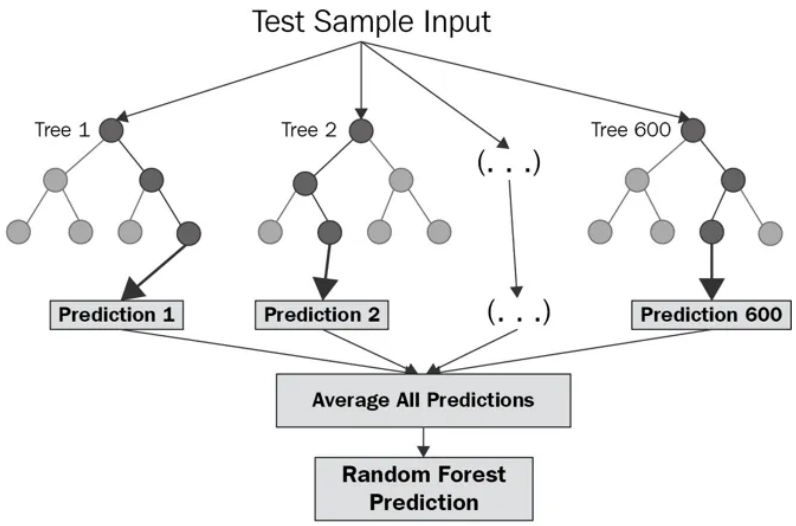


Figure 2.4: Random Forest (Corporate Finance Institute, 2021)

Table 2.1 shows the different accuracy levels for the different algorithms and the type of environment in which the different they accuracy levels were tested. The accuracy levels determine accuracy levels of the algorithms being able to correctly predict the likelihood of loan defaulting.

Table 2.1: Accuracy Levels of Different Algorithms (Aasim, 2019)

ALGORITHM	ACCURACY LEVELS (%)	SD ACCURACY (%)	DATA SIZE	PURPOSE
Logistic Regression	82.57	11.37		Classification
K-Nearest Neighbor	90.50	7.73		Classification
Naïve Bayes	85.25	10.34		Classification
Decision Trees	84.50	8.50		Classification
Random Forest	88.75	8.46		Classification

2.3. Models and Frameworks

2.3.1. TensorFlow

TensorFlow Probability (TFP), a Google-cloud ready python-based library is built on TensorFlow framework to help in combining different probabilistic models and deep learning algorithms when trying to determine the likelihood of an occurrence (Shwe, Dillon, & Seybold, 2018). Deep networks, gradient-based inference with automatic differentiation, and scalability to big datasets are all combined in TPF.

The main component in the TFP is the probabilistic inference that is the third layer of the TensorFlow framework. This layer contains all the algorithms and functions that are need to get the probability of an occurrence. According to (TensorFlow, n.d.),the main algorithms features found in the probabilistic inference component are:

i. Markov chain Monte Carlo – which contains the Hamiltonian Monte Carlo algorithm for estimating integrals through use of sampling.
ii. Variational inference – contains integral optimization algorithms
iii. Optimizers – contains Stochastic optimization methods
iv. Monte Carlo – contains the tools and techniques for calculating the expectations

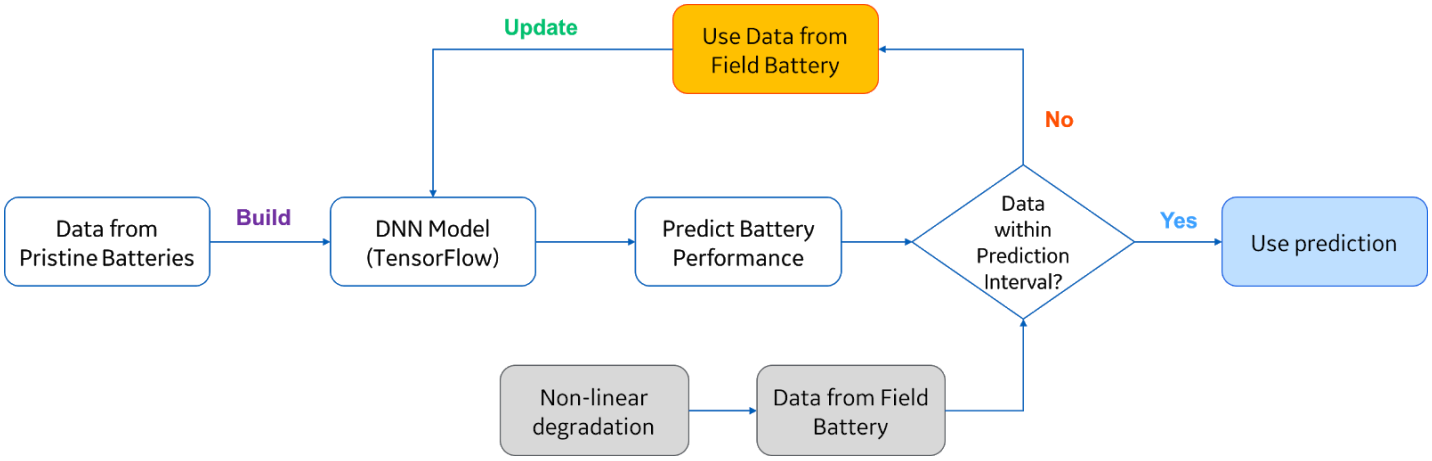


Figure 2.5: Using TensorFlow to known unknown unknows (TensorFlow, 2019)

Using the different components, TensorFlow combines deep learning models and machine learning, and displays the output using large data sources to train the model to think and create accurate and sensible data by itself (Project Pro, 2022).

Although TensorFlow guarantees a seamless performance, quick and regular updates and is very efficient as it is backed by Google, it isn't suitable for use in Kenya as many companies use Windows as their primary operating system (Statcounter Global Stats, n.d.) and TensorFlow is not suitable for Windows (Project Pro, 2022).

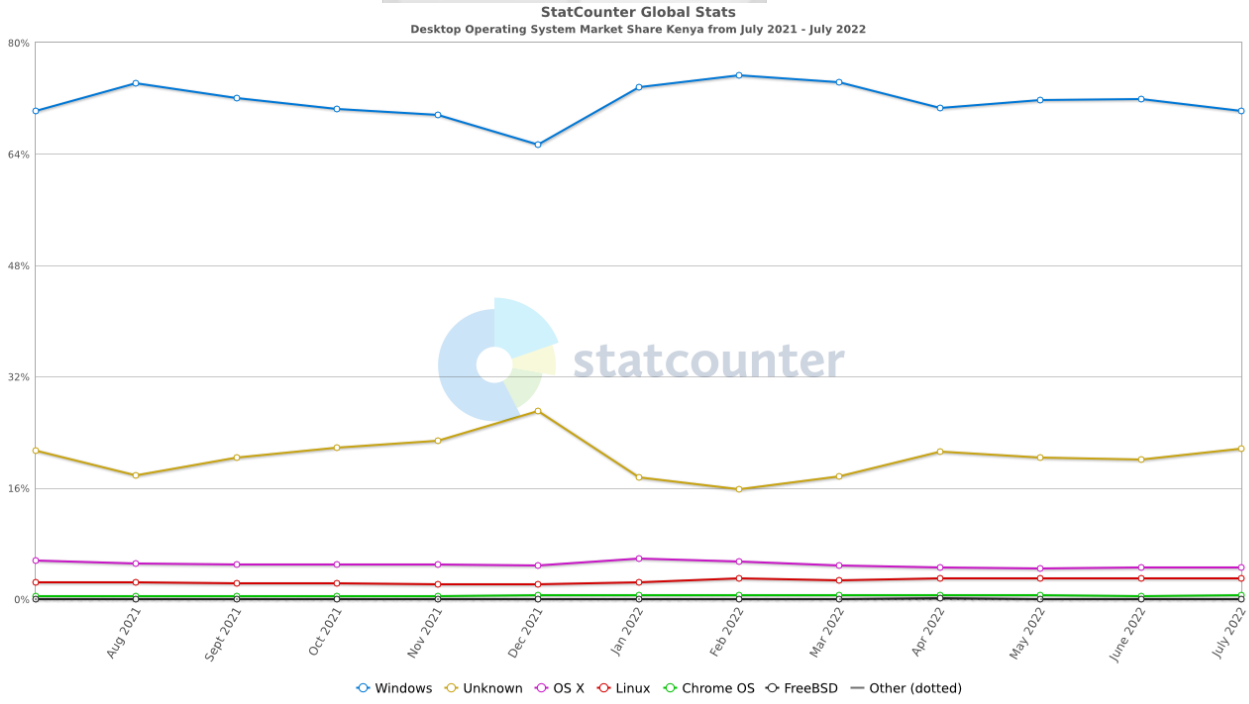


Figure 2.6: Statistics of OS usages in Kenya from July 2021 to July 2022 (Stat Counter Global Stats, n.d.)

2.3.2. Amazon Machine Learning

The output of the Amazon machine learning (AML) is a prediction score. The AML is a cloud-based framework and service that relies on data sources, machine learning models, evaluations, batch predictions and real time predictions by using identification, classification thresholding and comparison (Amazon Web Services, n.d.). First a cut off (classification threshold) is picked, then the observation is compared against the threshold. Any value that is above the threshold is a positive value, and any below the threshold is a negative value.

Amazon ML uses the following learning algorithms:

i. Amazon ML employs logistic regression (logistic loss function + SGD) for binary classification.

ii. Amazon ML employs multinomial logistic regression (multinomial logistic loss plus SGD) for multiclass classification.

iii. Amazon ML employs linear regression (squared loss function plus SGD) for regression.

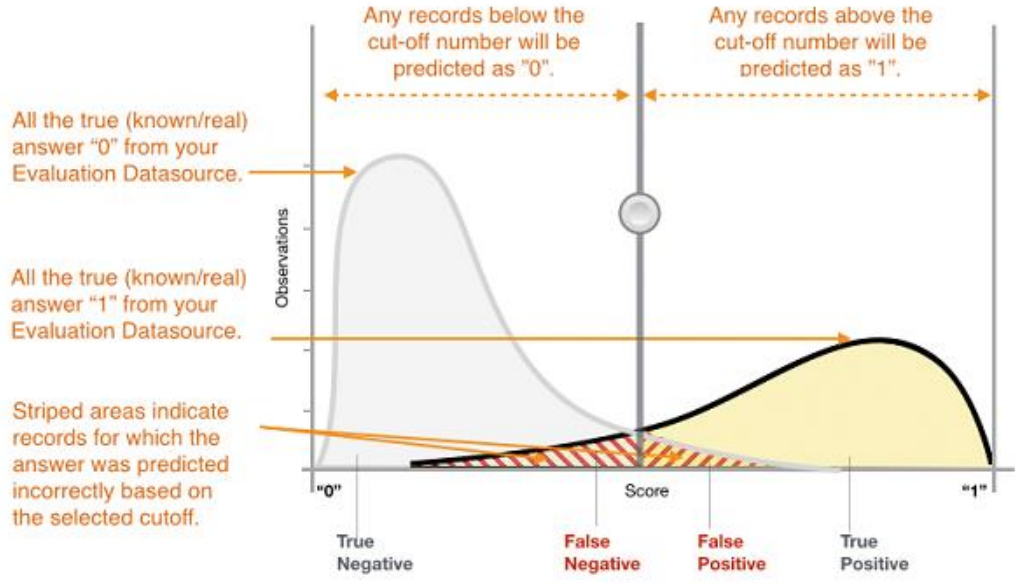


Figure 2.7: Score distribution for AML binary classification (AWS, n.d.)

Despite this ML framework being among the best, it cannot be used as AWS is no longer updating the AML services or accepting new users for the services offered (Amazon Web Services, n.d.). Another con of using the AML is that it is not integrated with multiple language support.

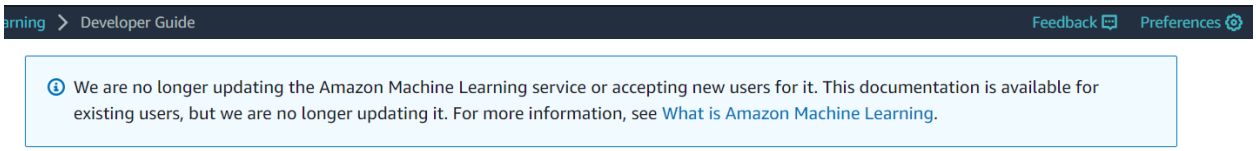


Figure 2.8: AML no longer supported (AWS, n.d.)

2.4. Architectural Design

Figure 2.10 shows the design that will be used in coming up with the architecture of the model. The feature selection involves determining the data and the variables that will be used, the subset features are variables that have multiple types e.g., employment type (full time, contractual, casual).

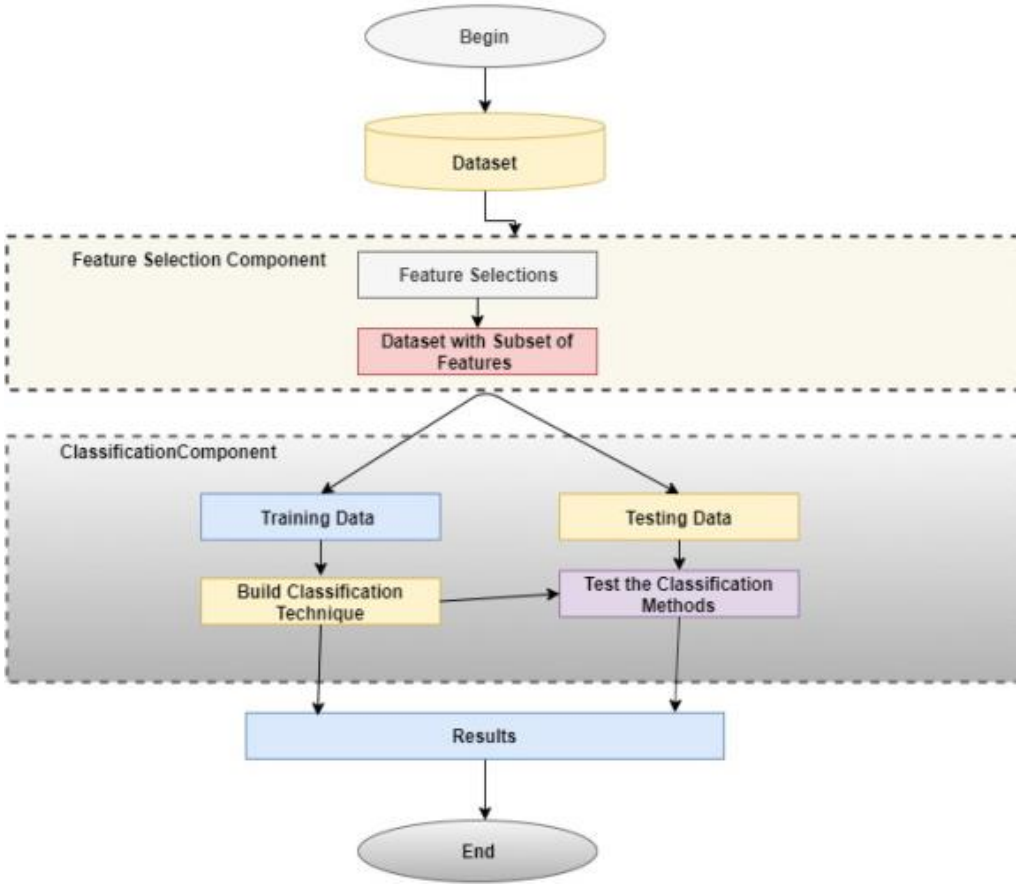


Figure 2.9: Architectural design (Arora, Sushant, Survesh , & Vinay, 2022)

2.5. Algorithms

2.5.1 Bayes Algorithm

The Bayes Theorem is the foundation of the probabilistic machine learning algorithm called the Naïve Bayes which is used in a wide variety of classification problems (Chauhan, 2022). It relies heavily on conditional probability where the likelihood than an event will occur given another event is used to determine the probability.

Some of the advantages of Naïve Bayes Classifier are:

i. Naïve Bayes is a quick and simple machine learning technique that can predict a class of datasets.

ii. It can be used for both Binary and multi-class Classifications.

iii. In comparison to other algorithms, it performs well in multi-class predictions.

iv. It is the approach for text categorization problems that is most frequently utilized.

The main disadvantage of Naïve Bayes Classifier is that Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features (DataRobot, 2021).

Applications of Naïve Bayes Classifier:

i. Credit Scoring is done using it.

ii. It is employed in the classification of medical data.

iii. Because Naïve Bayes Classifier is an eager learner, it can be used to make real-time predictions.

iv. It is utilized in text classification processes such as Sentiment analysis and spam screening

2.5.2 K-Nearest Neighbor (KNN)

The k-nearest neighbors' algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which classifies or predicts how a particular data point will be grouped using proximity. Although it can be applied to classification or regression problems, it is typically used as a classification algorithm because it operates under the premise that similar points can be found close to one another. (IBM, n.d.).

KNN has been used in the finance industry in credit risk to help banks assess risk of a loan to an organization or individual. It is used to determine the credit-worthiness of a loan applicant. (Christopher, 2021). The test data will fall into one of the "K" training data classes determined by the KNN algorithm, and the class with the highest probability is selected. The average of the "K" selected training points serves as the value in the case of regression.

KNN can also be used in recommendation suggestion. For borrower's who have a likelihood to default, this algorithm to help recommend a lower amount that they can afford to pay or that they are less likely to default. According to (Kumar, 2022), KNN can be used for clustering, with this we can be able to put a loan repayment default into a cluster based on the variables we use, and we can see the amount that people in that cluster are comfortable paying. This will also help in increasing the lender's source of income since the interest earned from the reduced recommended amount will be better than no disbursed loan at all if we only predict default likelihood without recommending a solution.

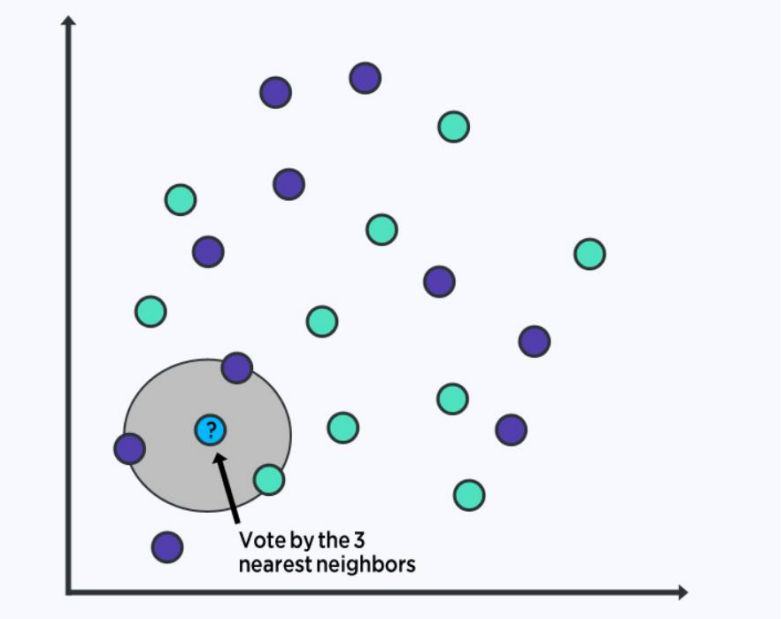


Figure 2.10: K-Nearest Neighbor

2.6. Conceptual Framework

Figure 2.11 shows the conceptual design flow of how the model works. The main algorithms selected for the prediction were based on the accuracy levels of the different algorithms as illustrated on Table 2.1. The loan data (borrower details and loan detail will be uploaded to the prediction model and once a prediction is made on the borrower's likelihood to default, the prediction and the loan data are

then uploaded to the recommendation model where for the borrower's likely to default, a lower amount is recommended as the new loan amount.

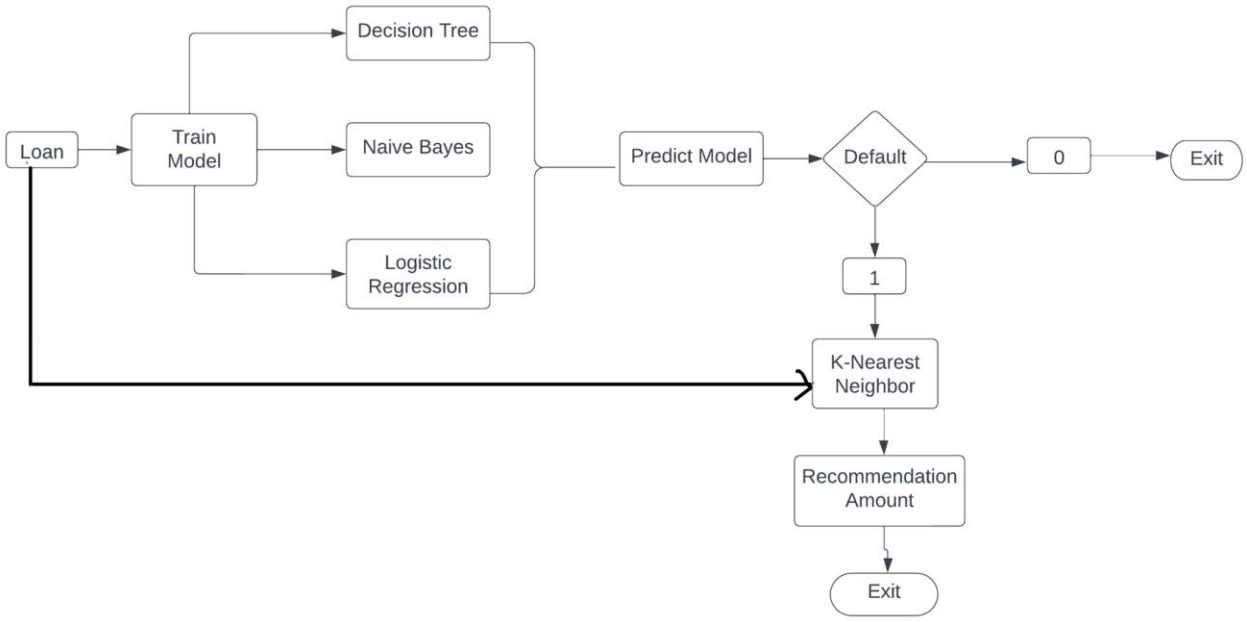


Figure 2.11: Conceptual design flow

2.7. Existing Works Used to Model Default Likelihood and Recommendations

There are a quite a few existing tools that are currently being used in loan prediction and recommendation. Figure 2.13 shows a summary of some of the existing models, their models and algorithms used, as well as their contributions and limitations. Some of the methods used currently are credit forecasting, score cards, excess value modelling based on modeling the maximum of a sample called the upper order statistics over a period and modeling excess values of a sample over a threshold within a period (Wanjohi et al. 2016).

Table 2.2: Existing tools and the gaps present

Theories and Models used	**Source**	**Industry and Country**	**Year**	**Gaps**
Logistic model	(Wanjohi et al. 2016)	Banking, Kenya	2012	Only relied on demographic information like gender and age but did not include economic factors like income levels and employment types

Data mining algorithms, artificial neural network decision tree and naïve Bayesian classifiers	(O. & A., 2012)	Banking, Kenya	2012	As they used parametric modelling, the model was less intuitive and limited flexibility
Generalized Extreme Value Regression Model	(Wanjohi et al. 2016)	Banking, Kenya	2016	Seemed to be more suitable for rare events.

As seen in Table 2.2, the existing model in the country are only suitable for the banking industry yet there are many distinguishing factors that distinguish banking form micro finance meaning that a lot the existing systems should be modified to cater for the structure of microfinances. Other existing tools that are currently being used include:

i. (Anatoly, 2014) – who used binary choice models. He suggested that if we used automated clustering, the accuracy of the models' predictive powers could be improved. (Wanjohi et al. 2016)
ii. Peter Croshie who applied his model in trying to determine risk that surrounds the uncertainty of a firm's in servicing its obligations and debts (Andrea, 2010). His model cannot be used in SACCOs and microfinance as the characteristics are very different.

Most existing models that exist currently only use one algorithm to train and validate their data or those that used more than one used one machine learning and the other a pure statistical model. Previous work has been done in Russia using binary choice models, in Kenya by using only logistic regression and in other countries using random forests (Wanjohi et al. 2016). However, in all these none of them used a combination of different algorithms for their models or offered both prediction and recommendation. Table 2.3 shows the summarized review of the different algorithms and their limitations.

Table 2.3: Summary of Algorithms and their limitations

Theory	**Focus**	**Limitations**
Regression theory **(Focus on logistic regression)**	- Categorical dependent variables - Probabilistic outcomes	It assumes that there is complete linearity between the dependent and independent variables
Score Card Theory	- Weighted variables	We cannot for sure know the parameters that were used especially for 3rd party solutions

Linear Regression Theory	- Continuous variables and continuous output	Assumes normal distribution of and a complete linear relationship between dependent and independent variables
Bayes Theory	- Conditional probability of an event based on occurrence of another event	There is no universal representation or method for dealing with uncertainty within the context of prior probability and background knowledge.

Chapter 3: Design and Methodology

3.1. Introduction

Following a specific process is important in ensuring a credible, organized and structured research. According to (Thmoson, n.d.), most research is usually broad and sometimes cyclical and repetitive hence the need to follow a guided step by step approach that will make it clear by providing a path to follow from the start of the research to the end of the same. She further argues that the approach followed should be based on what is most suitable for you based on your research. (Jansen, 2010) recommends considering the following factors when selecting the best appropriate methodology for your research:

i. Feasibility, suitability and constraints
ii. Research nature – objectives and research questions, scope, etc.

3.2. Design and Philosophy

The main design method that will be used in the study is the positivism philosophy. In this philosophy, only knowledge obtained by observation (the senses), including measurement, is reliable and trustworthy. In positivism, the role of the researcher is limited to data collection and interpretation in an objective way. In other words, the researcher is an objective analyst and they distance themselves from personal values in conducting the study. In these types of studies research findings are usually observable and quantifiable.

According to (Dudovskiy, 2022), the positivist philosophy is based on five main concepts that can be summed up as follows:

i. The logic of inquiry is the same throughout all sciences.
ii. The goal of the research should be to predict and explain.
iii. Research should be empirically visible by means of human senses. To create assertions (hypotheses) that will be put to the test during the research process, inductive reasoning should be applied.
iv. Science differs from common sense. The research findings shouldn't be biased by common sense.
v. Science must be value-free and evaluated solely on the basis of logic and rationality.

3.3. Population and Sampling

According (Matara, 2022), there are 175 licensed to carry out deposit-taking SACCO business in Kenya as of 2022. Despite the number or licensed SACCOs being this high, the researcher focused on using Okoa Management Ltd., as the sampled population size since the research proposal is aimed at fixing their needs.

Factors considered when selecting the sample size (here the sample size will refer to the number of loan records to be used from all the loans they have disbursed from January 2018 to January 2023 – which has about 850,000 rows of data on the disbursed loans after the data was formatted and cleaned):

i. Confidence interval: This is the margin of error which is the measure of the uncertainty or certainty degree. It can be said to be what tells you how confident one can be in the study results (Kibuacha, 2021).
ii. Confidence level: this is the percentage of probability, or certainty that the confidence interval would contain the true population parameter when you draw a random sample many times (Kibuacha, 2021).
iii. Standard deviation: used to be able to help in approximating how much the predictions will vary from each and from the average

3.4. Data Collection and Data Analysis

3.4.1. Data Collection

As the research focus will be based on mostly SACCOs and micro credit institution SME on focus, the main data collection method is going to be coming from SACCOs themselves. The idea is to obtain data directly from the source (through Excel and CSV documents as according to my correspondent from the Okoa, this is how they currently store the data), in this case Okoa Ltd., to ensure that the model is trained and validated against features that are relevant to the institution used for purposes of this research. With the schools help with an official document, the researcher managed to reach out to Okoa and ask for their loan defaulters and non-defaulters' history. According to (Maione, 2022) the main advantage of collecting data directly from the source is that it's more reliable and accurate meaning that ultimately there will be better results and solutions. Another pro for collecting primary data this way is that it is generally at a lower cost since there is no need for professional researchers or any investment in additional equipment (Maione, 2022). Despite this being a good method, different SACCOs offer

services to different people, meaning that for the researcher to collect data from a wide pool of different people (different income levels and employment types, different social backgrounds) a larger data set should be collected. Because of this, Okoa became a great case study as they offer services to the low-income range, middle income range and upper-middle to lower-upper class thus ensuring there was an inclusive and comprehensive data set for training and validating the model. Another reason for opting to lean towards this is a collection model was to get accurate historic data on loan defaulting and the characteristics of the people that defaulted, so that the collected data could be used for training the model. This data was collected using excel sheets as Okoa rely on it for reporting and data capturing. However, it was not feasible in the time frame for this project to collect enough data from the main SACCO under study, so additional data from online repositories that offer training data sets for machine learning models for loan prediction was used to obtain the required training, testing and validation data.

The following research tools were used to collect the above data:

i. Interviews: Interview loan processors to understand the factors that they considered when processing a loan – these were used to determine the factors, and/or the characteristics that were used for training the developed tool.
ii. Document Reviews: reviewing existing literatures on current applications of ML in the lending industry and the gaps that could be fixed with the development of the suggested tool.
iii. Statistics: to convert the collected data from quantitative such as low income, mid-income, high-income into numerical representation such as 1,2,3, female or male into 1,2 etc. This is was to ensure that the model can be able to generate quantitative results and reports.

3.4.2. Data Analysis

Most of the data collected for the purpose of this research was structured and quantitative data and thus quantitative data analysis was. Quantitative data analysis simply involves analyzing number-based data or data that can be easily converted to into numbers without losing any meaning (Jansen & Warren, Quantitative Data Analysis 101: The lingo, methods and techniques, explained simply, 2020). Quantitative analysis generally is used for (Jansen & Warren, Quantitative Data Analysis 101: The lingo, methods and techniques, explained simply, 2020):

I. Firstly, it measures the variations between groups. For example, the popularity of the various loan types over others.

II. Secondly, it's used to evaluate the relationships and connections between different variables. For example, the relationship between salaries and loan repayment.

III. And third, it's used to test hypotheses in a scientifically rigorous way. For example, a hypothesis about the impact of a certain rates on the different loans.

According to (Bhatia, 2018), there are 2 main steps involved in quantitative data analysis:

i. Data preparation

Data preparation involves converting the raw data collected into meaningful and readable information. It prepared in the following for steps:

a. Data validation: done to confirm that the data collection was done according to set standards and without any bias. Figure 3.3 shows the different types of validation that can be done to the collected data.

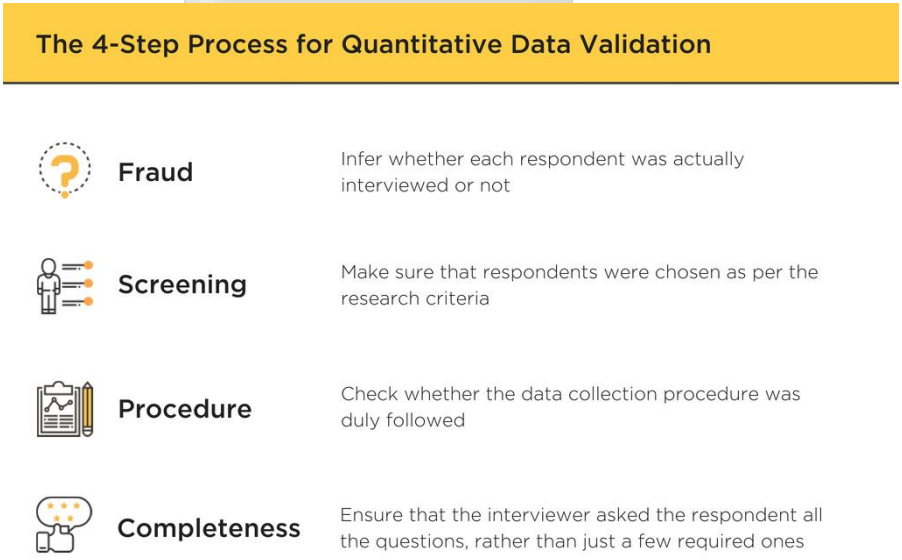


Figure 3.1: Processes for Quantitative data Validation (Bhatia, 2018)

b. Data editing: the researcher aims at ensuring there are no errors in the data. It could involve them conducting basic data checks, checking for outliers, and maybe removing data points that could affect the accuracy of the results.

c. Data coding: which involves grouping out data and assigning values to these groups e.g., salary ranges between 30K – 60K could be represented as 1, range from 61K- 150K could be 2 and so on and so on for your variables.

ii. Data analysis

After the data is validated, then the actual analysis can begin. In quantitative data analysis, the researcher focused on descriptive analysis and inferential analysis. In descriptive analysis the main aim is to summarize the data and find patterns such as lowest and highest loan amounts borrowed (range), average loan amount (mean), the number of times people default(frequency) etc. In inferential analysis, the main idea is to analyze the complex relationships between the different variables under research to generalize ideas and be able to make predictions.

Figure 3.4 highlights the main differences between inferential and statistical analysis:

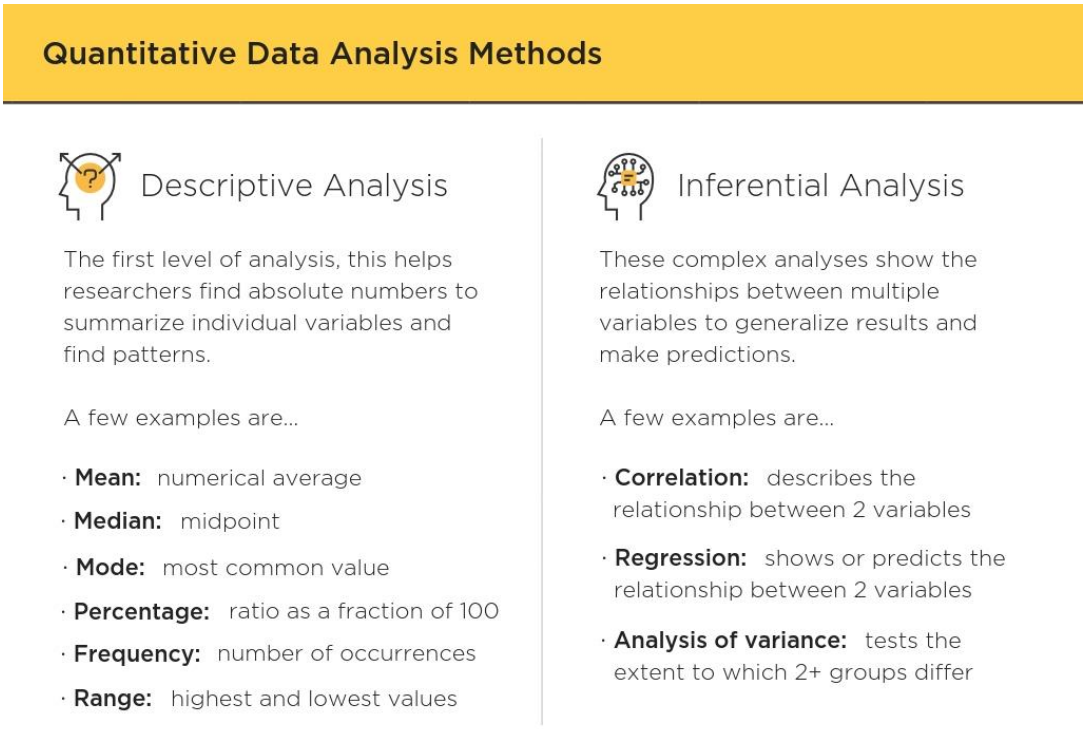


Figure: 3.2: Differences between statistical and inferential analysis (Bhatia, 2018)

3.5. System Development Methodology

The system was developed using the Agile approach. According to (Atlassian, n.d.), Agile is an iterative approach to software development and project management that helps teams deliver faster and continuously. Agile is a good methodology as it creates a common ground amongst scrum, extreme programming, crystal clear and other framework methodologies. The project will follow the agile software development manifesto that is based on four core values (Drumond, 2022):

i. People and the interactions come before processes, procedures and tools

ii. Working and functional software over thorough documentation

iii. Customer collaboration over contract negotiation

iv. Adapting to change over sticking to a plan

The main reason the researcher used Agile methodology is because it has been said to produce software more swiftly and responsively than traditional waterfall techniques, and has also been credited to help software projects more successfully meet user objectives, customer needs, and business goals. (Pratt & Torode, 2020). Proponents of Agile methodologies say the four values outlined in the Agile Manifesto promote a software development process that focuses on quality by creating products that meet consumers' needs and expectations.

The agile philosophy is based on the following 12 core principles:

i. Delivering quality job on time and continuously to satisfy clients.
ii. Dividing a large task into smaller manageable tasks that can be completed quickly.
iii. Understanding that self-organized teams provide the best work.
iv. Trusting motivated employees to do the task at hand while providing them with the atmosphere and assistance they require.
v. Creating processes that promote sustainable efforts.
vi. Maintaining a constant pace for completed work.
vii. Accepting modified specifications, even late in a project.
viii. Assembling the project team and business owners on a daily basis throughout the project.
ix. Having the team often reflect on how to be more productive, then altering and refining behavior as necessary
x. Measuring progress based on the volume of work accomplished.
xi. Constantly pursuing excellence.
xii. Harnessing change for a competitive advantage.

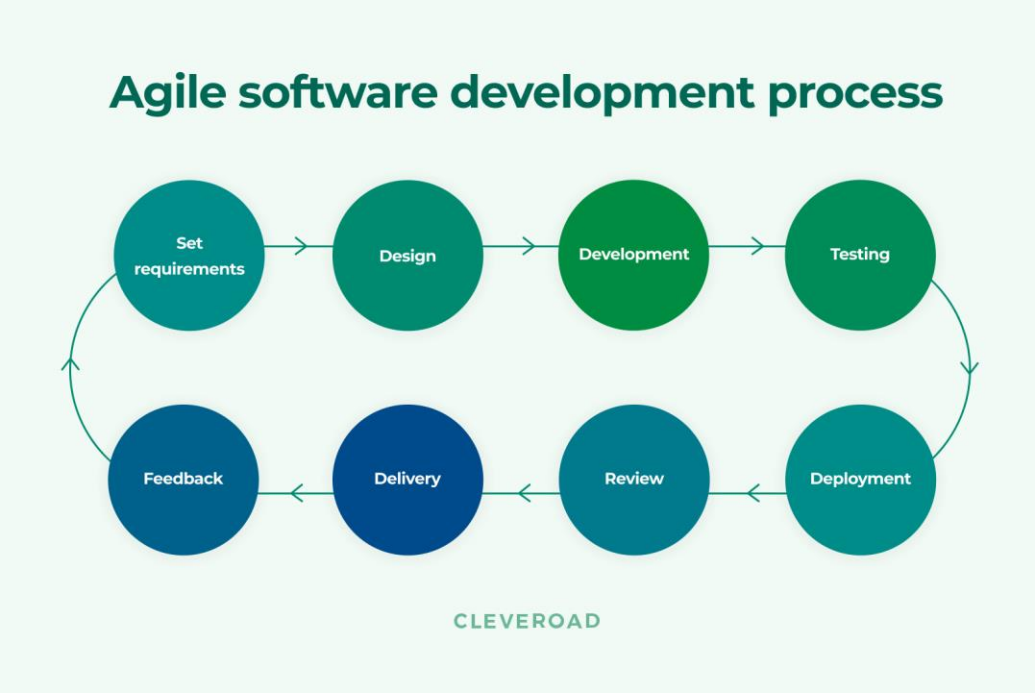


Figure 3.3: Agile Methodology (Rovyana, 2022)

3.6. Dissemination and utilization of results

This was done through thesis publication. It is through publication that the research, including its scientific and practical contributions, was disseminated to others in a particular field. (Stellenbosch Business School, 2018). This makes scientific researchers and practitioners with similar interests aware of new knowledge in their field and it helps in advancing knowledge and its application

Reports and proof of concept model was also be pitched to SACCOs. Sharing reports with the SACCOs sampled on findings that could help in planning and determining the risk profile of their members. Such reports include: average amount borrowed by members, frequency of loan defaults, disparities between the members who repay and those that don't, among other reports.

The provided results can be used by the SACCO in minimizing their risk and reducing the number of defaulters. Adoption of the tool by the organization will also help them in making efficient and faster decisions on lending as the automation makes it easier to process the risk assessment and provide recommendations. The generated results can then be used for planning to target new markets to advertise to as it can provide them with data on the characteristics of the loan borrowers such as low risk borrowers, etc. Then organization can then create better loan options and types for the high-risk clients to make them more appealing to a broader borrowing clientele.

3.7. Ethical considerations and issues

The main ethical consideration for the researcher was data privacy issues. (Karn, 2022) reiterates that maintaining the privacy of personal information of the users has become vital for any company which collects and stores personally identifiable data. He further identifies some of the major challenges that are faced by financial institutions such as sharing data with third parties, lack of technology to handle data breaches as being among the most reasons why data privacy can be easily violated in the industry.

Another major concern is fairness and transparency issues as a result of automated decision making. According to (Tilimbe, 2019), the clients and users rarely ever get to understand why the decision was made in that specific way. Not being able to see or understand how the model works could reduce the trust the users have in the predictions made. Arithmetic bias as a result of data being included or excluded when cleaning up the data to train the model could be a basis for the transparency issues that the users may have with regard to the model results.

To ensure that the privacy, fairness and transparency issues raised were addressed, the researcher took the following countermeasures:

i. Observation of all regulations regarding data including is storage, processing and disclosure e.g., NACOSTI (National Commission for Science, Technology, and Innovation) regulations
ii. Got the informed consent of all the parties involved.
iii. Offered training to the users to ensure they were aware on how to use the system (what to enter, etc.).
iv. Put sanity checks and validations in the model to validate inputs (to prevent data corruption and promote integrity of data used to train the model)

Chapter 4: System Design

4.1. Functional and Non-Functional Requirements

A Functional Requirement (FR) is an outline of the service that the software must provide. It describes either a software system or one of its components. A function is nothing more than the inputs, behavior, and outputs of the software system. The functions a system is likely to perform can be determined by calculations, data manipulation, business processes, user interactions, or any other specialized functionality (Martin, 2022). Non-functional requirements or NFRs are a collection of guidelines that describe the system's limitations and operational capabilities in an effort to enhance its functionality. These are basically the requirements that outline how well the system will operate.

4.1.1. Functional Requirements

i. The system should authenticate users

ii. The system should accurately predict default likelihood and recommend lower amounts to borrowers predicted to default.

iii. The system should allow for report generation and viewing

iv. A learning model that allows the model to be continually updated with new data sets

4.1.2. Non-Functional Requirements

i. The predictions made by the model should be as accurate as possible.

ii. System downtime should be minimal and availability very high.

iii. The model should be easy to use and user friendly

iv. The model should be secure.

v. The model should return prediction and recommendation results as fast as possible (speed).

4.2. Use case

4.2.1. Use Case Diagram

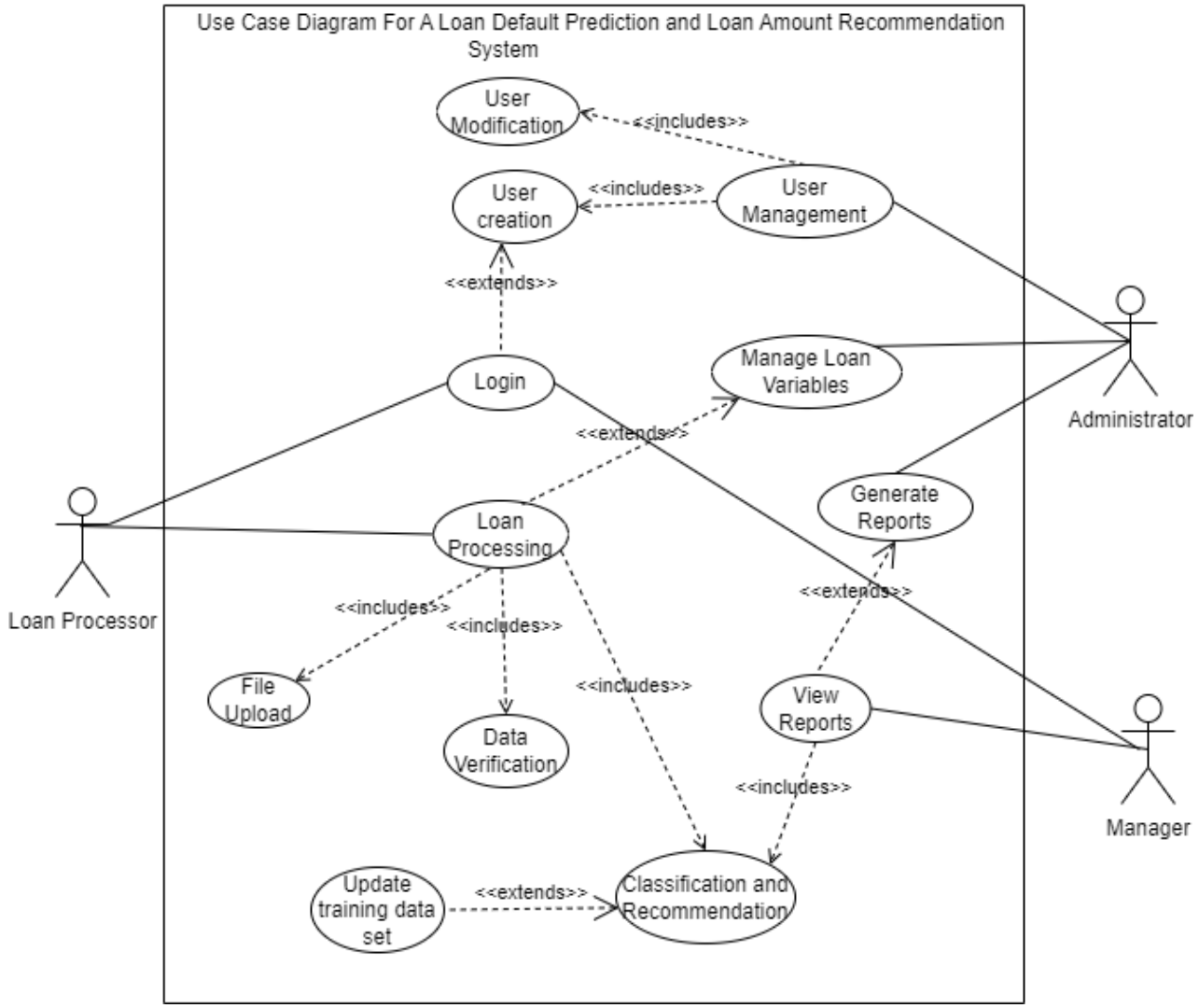


Figure 4.1: Use case diagram

4.2.2. Use Case Scenarios

Table 4.1 shows the different user case scenarios, their pre-conditions, success scenarios and post conditions. It shows the detailed descriptions of the various uses cases as referenced in the Figure 4.1

Table 4.1: Use case scenarios

Use Case	Pre conditions	Main Success Scenario	Post Conditions
Member Registration	None	i. Member ID is generated ii. Member details are saved	None
Login	User is already registered	- User can access the system with their credentials	None
File Upload		i. Loan details are extracted from the CSV ii. The model can read the data from the uploaded file	None
Downloads	i. Prediction has been made ii. Recommendation has been made	i. A loan prediction file id successfully downloaded ii. A loan recommendation CSV file is successfully downloaded	i. A prediction column is added to the downloaded CVS file. ii. A loan recommendation amount is added to the downloaded CSV file

4.3. Sequence diagram

Figure 4.3 shows the sequence of events and how the model works from the time the CSV is uploaded to when the model predicts the likelihood of default and recommends a lower amount for the loans predicted to amount applications predicted to default. It shows how the file is uploaded and data extracted and fed into the model for prediction.

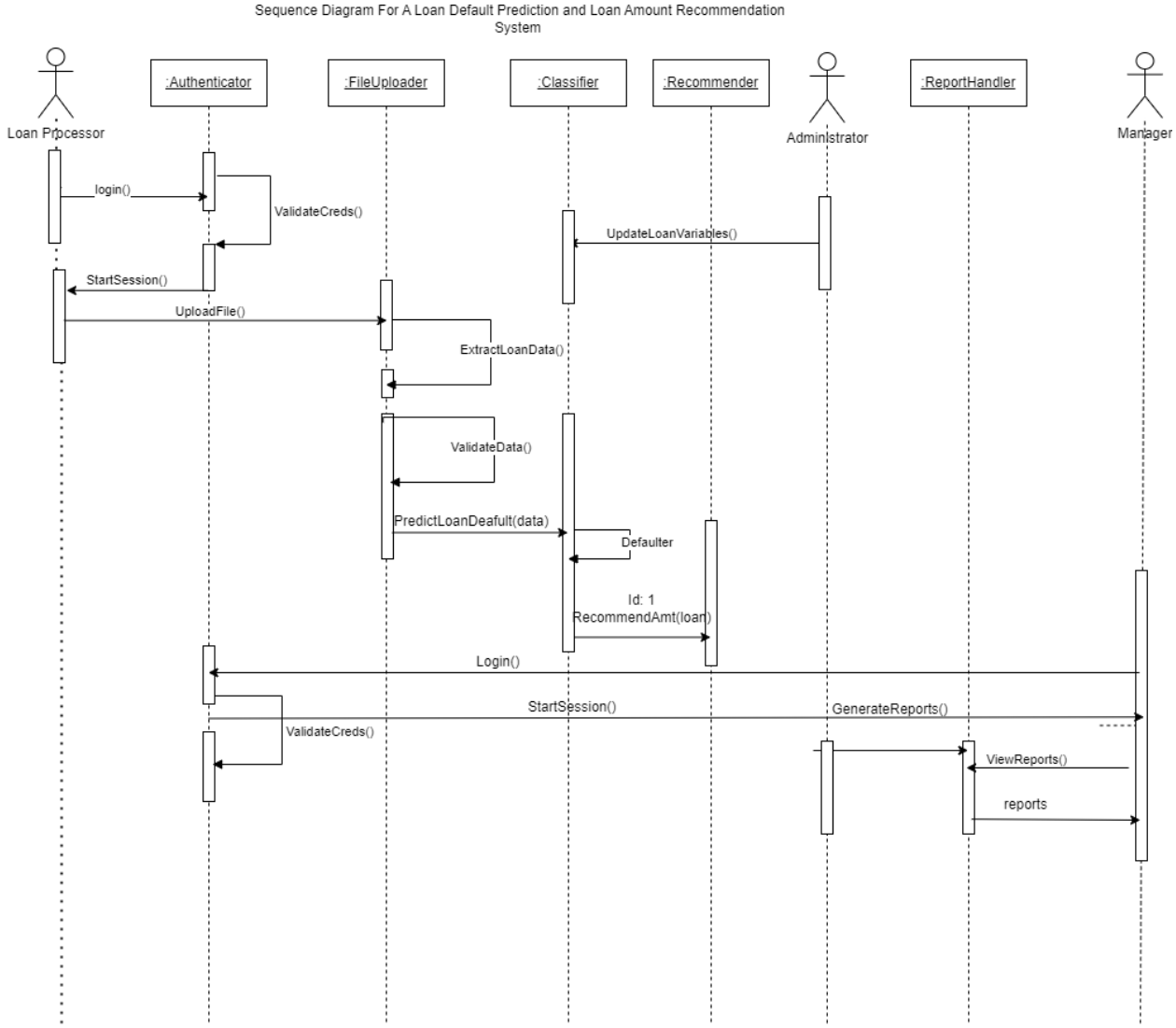


Figure 4.2: Sequence Diagram

4.4. ERD Diagram

The Figure 4.4 highlights how the different entities in the model interact with each other. It shows the different attributes, entities and relationships between the different tables in the database. It shows the how the different elements of the database relate with each other and shows some of the attributes (columns) of the existing entities.

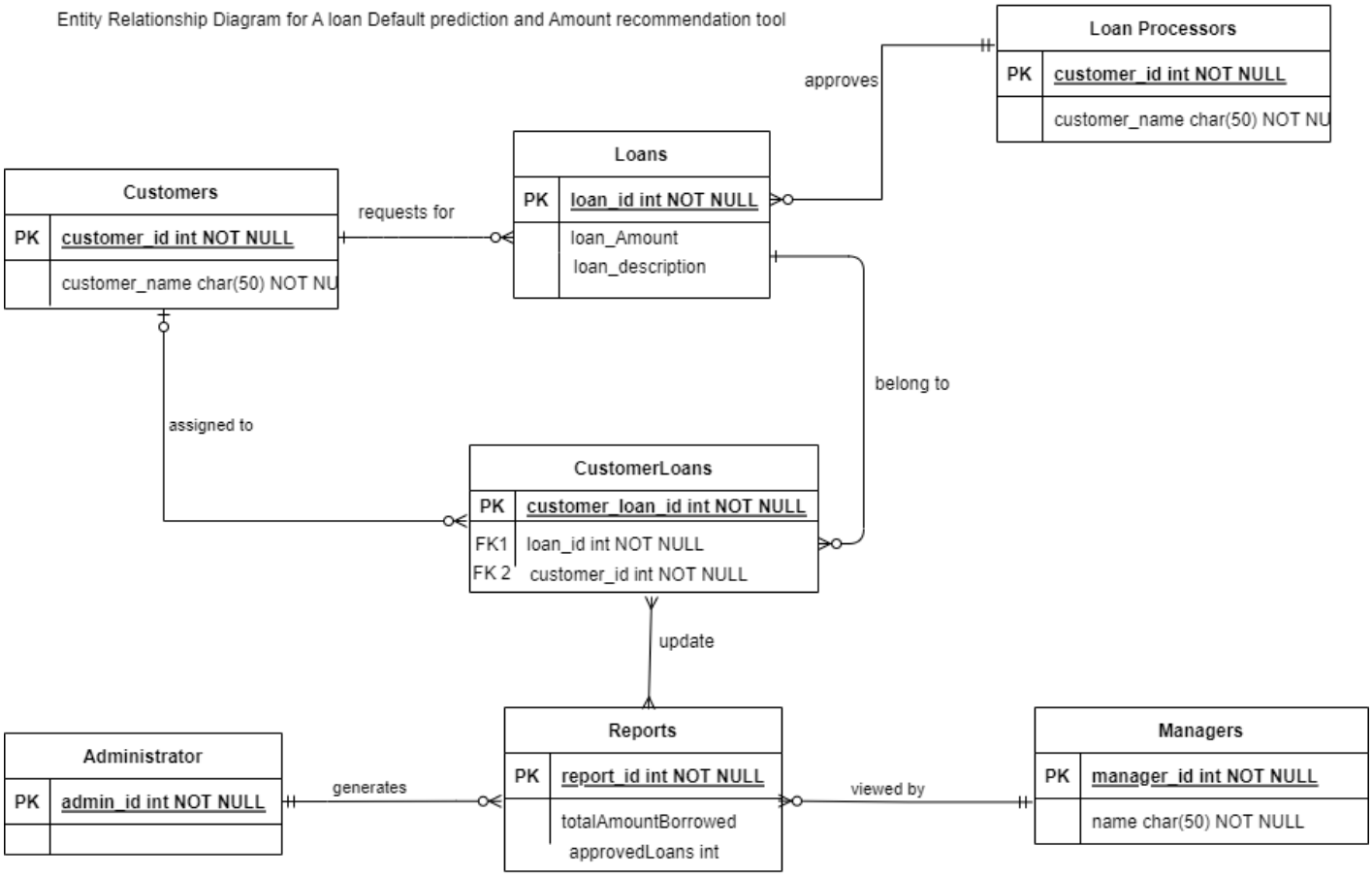


Figure: 4.3: Entity Relationship Diagram

Chapter 5: Implementation and Testing

5.1. Introduction

Generally, software bugs will almost always exist in any software module, but it is not because of the carelessness or irresponsibility of the programmer but because of the complexity of the system. This chapter focuses on discussing the resting of the solution as well as the implementation methodologies used. According to (University of Kentucky, 2018), systems implementation can be defined as the process of:

i. Defining the physical system design – how the system should be built

ii. Ensuring that the system is used and operational

iii. Quality assurance

5.2. Hardware and Software Requirements

The model development was carried out in a localized platform environment then deployed to the Google cloud platform. The main reason for opting to deploy to a cloud environment was the high demand required for resources such as CPU and RAM which made the need to auto-scale a real issue. Python was the main underlying language used for development and training of the model with machine library extensions such as Numpy, Keras, Scikit-learn, Pandas being among the imported libraries used in the model. Table 5.1 shows the summarized hardware and software components used.

Table 5.1: Hardware and software requirement

Software	**Specific Library**	**Version**
Python (3.7 or higher)	TensorFlow	2.1.0
	Keras	2.3.0
	Numpy	1.16.3
	Pandas	1.5.3
	Matplotlib	3.7.1
IDE	Jupyter Notebook	6.5.3 or higher
	VS Code	1.67.2 or higher
Hardware	**Description**	**Version**
	CPU	
	RAM	8GB minimum – for cloud 16GB minimum – for local

5.3. System Implementation

There were several processes involved in the implementation of this model. Some of the key processes that were a critical for the developed model were: data collection, cleaning and preprocessing, selection of the algorithms to use for the model, model training and validation, among others.

5.3.1. Loading the data set

Figure 5.2 shows how the data for training and testing the model was loaded into the model while Figure 5.3 shows the structure of the data (first 5 rows, the number of rows and columns (set1_train. shape)). Some of the columns are hidden for Figure 5.3 so as to scale and fit the images on the pages without reducing the clarity of the pictures.

There were different packages that required to be included in the model so that the data could be loaded and read by the model. One of them main packages that was needed for the extraction of the data form the CSV was the Python language panda package. Since the package is well suited for working with labelled data, it was necessary for the model for both data manipulation and analysis. The other key package for data loading is the Numpy package for arrays using Keras provided for by TensorFlow.

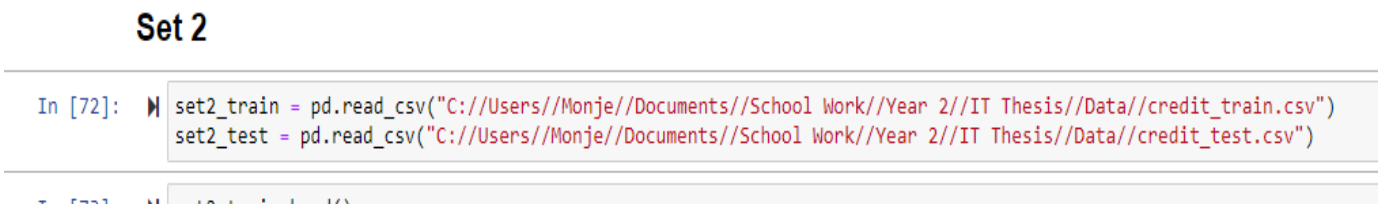


Figure 5.1: Loading data sets for training and testing

```
In [74]: set2_test.head()
```

Out[74]:

	Loan ID	Customer ID	Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	Home Ownership	Purpose	Monthly Debt	Years of Credit History
0	f738779f-c726-40dc-92cf-689d73af533d	ded0b3c3-6bf4-4091-8726-47039f2c1b90	611314.0	Short Term	747.0	2074116.0	10+ years	Home Mortgage	Debt Consolidation	42000.83	21.8
1	6dcc0947-164d-476c-a1de-3ae7283dde0a	1630e6e3-34e3-461a-8fda-09297d3140c8	266662.0	Short Term	734.0	1919190.0	10+ years	Home Mortgage	Debt Consolidation	36624.40	19.4
2	f7744d01-894b-49c3-8777-fc6431a2cff1	2c60938b-ad2b-4702-804d-eeca43949c52	153494.0	Short Term	709.0	871112.0	2 years	Rent	Debt Consolidation	8391.73	12.5
3	83721ffb-b99a-4a0f-aea5-ef472a138b41	12116614-2f3c-4d16-ad34-d92883718806	176242.0	Short Term	727.0	780083.0	10+ years	Rent	Debt Consolidation	16771.87	16.5
4	08f3789f-5714-4b10-929d-e1527ab5e5a3	39888105-fd5f-4023-860a-30a3e6f5ccb7	321992.0	Short Term	744.0	1761148.0	10+ years	Home Mortgage	Debt Consolidation	39478.77	26.0

```
In [75]: set2_train.shape
Out[75]: (100514, 19)

In [76]: set2_test.shape
Out[76]: (10353, 18)
```

Figure 5.2: Data set structure and shape

5.3.2. Data Preprocessing and Clean up

The Figure 5.4 shows the information about the data, if there are any missing values requiring cleaning up, and how the null values were removed as they could not be used to train the model. The main reason for removing the null entries is to remove and/or reduce bias in the model. The same figure also shows that after clean up, the number of rows has reduced from 100514 (as seen in Figure 5.3) to 100000 (Figure 5.4).

After null loan id values were removed, the missing values (work years, annual income, etc.) were then replaced with either the mode or the mean as shown in Figure 5.5.

```
In [58]: set2_train[set2_train['Loan ID'].isnull()]
```

Out[58]:

Current Loan Amount	Term	Credit Score	Annual Income	Years in current job	Home Ownership	Purpose	Monthly Debt	Years of Credit History	Months since last delinquent	Number of Open Accounts	Number of Credit Problems	Current Credit Balance	Maximum Open Credit	Bankruptcies	Tax Liens
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
In [59]: # Drop null rows in Loan ID because most are null in the other columns as well
         set2_train.dropna(subset=['Loan ID'], inplace= True)

In [51]: set2_train.shape

Out[51]: (100000, 19)
```

Figure 5.3: Cleaning data to remove null values

```
In [52]: # Fill in missing values
         for col in ['Credit Score', 'Annual Income', 'Maximum Open Credit', 'Bankruptcies', 'Tax Liens']:
             mean = set2_train[col].mean()
             set2_train[col].fillna(mean, inplace=True)

In [53]: # Fill Months since last delinquent null values with 0
         set2_train['Months since last delinquent'].fillna(0, inplace= True)

In [54]: # Fill in Years at current Job with mode
         set2_train['Years in current job'].fillna(set2_train['Years in current job'].mode()[0], inplace= True)

In [55]: set2_train.info()
```

Figure 5.4: Updating null values with the mean and mode as calculated by the model

Figure 5.5 shows the description of the data in the model as loaded from the training CSV data.

```
Name: Loan Status, dtype: float64

In [60]: set2_train.describe()
```

Out[60]:

	Current Loan Amount	Credit Score	Annual Income	Monthly Debt	Years of Credit History	Months since last delinquent	Number of Open Accounts	Number of Credit Problems	Current Credit Balance	Maximum Open Credit
count	1.000000e+05	80846.000000	8.084600e+04	100000.000000	100000.000000	46859.000000	100000.00000	100000.000000	1.000000e+05	9.999800e+04
mean	1.176045e+07	1076.456089	1.378277e+06	18472.412336	18.199141	34.901321	11.12853	0.168310	2.946374e+05	7.607984e+05
std	3.178394e+07	1475.403791	1.081360e+06	12174.992609	7.015324	21.997829	5.00987	0.482705	3.761709e+05	8.384503e+06
min	1.080200e+04	585.000000	7.662700e+04	0.000000	3.600000	0.000000	0.00000	0.000000	0.000000e+00	0.000000e+00
25%	1.796520e+05	705.000000	8.488440e+05	10214.162500	13.500000	16.000000	8.00000	0.000000	1.126700e+05	2.734380e+05
50%	3.122460e+05	724.000000	1.174162e+06	16220.300000	16.900000	32.000000	10.00000	0.000000	2.098170e+05	4.678740e+05
75%	5.249420e+05	741.000000	1.650663e+06	24012.057500	21.700000	51.000000	14.00000	0.000000	3.679588e+05	7.829580e+05
max	1.000000e+08	7510.000000	1.655574e+08	435843.280000	70.500000	176.000000	76.00000	15.000000	3.287897e+07	1.539738e+09

Figure 5.5: Model description to show the numerical details used to update the null values

5.3.3. Model Training

Figure 5.7 shows the correlation between the different variables and Figure 5.8 shows the same variables' correlation in a heat map.

Name: Loan Status, dtype: int64

In [63]:
```
# Check correlation of variables
numeral = set2_train.corr()
numeral
```

```
C:\Users\Monje\AppData\Local\Temp\ipykernel_3928\868641932.py:2: FutureWarning: The default value of numeric_only in DataFra
me.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of nume
ric_only to silence this warning.
  numeral = set2_train.corr()
```

Out[63]:

	Current Loan Amount	Credit Score	Annual Income	Monthly Debt	Years of Credit History	Months since last delinquent	Number of Open Accounts	Number of Credit Problems	Current Credit Balance	Maximum Open Credit	Bankruptcies	Tax Liens
Current Loan Amount	1.000000	-0.096652	0.013112	-0.006643	0.019282	0.011248	0.001478	-0.002795	0.003880	-0.001271	-0.000608	-0.002048
Credit Score	-0.096652	1.000000	-0.017078	-0.001674	-0.009720	-0.003739	0.006435	-0.003022	-0.000104	-0.002827	-0.006935	0.005146
Annual Income	0.013112	-0.017078	1.000000	0.485230	0.161669	-0.077577	0.146175	-0.017006	0.312340	0.053064	-0.047672	0.040167
Monthly Debt	-0.006643	-0.001674	0.485230	1.000000	0.199289	-0.056818	0.411353	-0.055383	0.481348	0.039268	-0.078979	0.020119
Years of Credit History	0.019282	-0.009720	0.161669	0.199289	1.000000	-0.044292	0.132349	0.061588	0.208470	0.031124	0.066247	0.017245
Months since last delinquent	0.011248	-0.003739	-0.077577	-0.056818	-0.044292	1.000000	-0.032569	0.104642	-0.028662	-0.008785	0.123951	0.012624
Number of Open Accounts	0.001478	0.006435	0.146175	0.411353	0.132349	-0.032569	1.000000	-0.013995	0.228136	0.031341	-0.024575	0.006545
Number of Credit Problems	-0.002795	-0.003022	-0.017006	-0.055383	0.061588	0.104642	-0.013995	1.000000	-0.112516	-0.012072	0.752942	0.581290
Current Credit Balance	0.003880	-0.000104	0.312340	0.481348	0.208470	-0.028662	0.228136	-0.112516	1.000000	0.139204	-0.122603	-0.015645
Maximum Open Credit	-0.001271	-0.002827	0.053064	0.039268	0.031124	-0.008785	0.031341	-0.012072	0.139204	1.000000	-0.014574	-0.001029
Bankruptcies	-0.000608	-0.006935	-0.047672	-0.078979	0.066247	0.123951	-0.024575	0.752942	-0.122603	-0.014574	1.000000	0.046110
Tax Liens	-0.002048	0.005146	0.040167	0.020119	0.017245	0.012624	0.006545	0.581290	-0.015645	-0.001029	0.046110	1.000000

Figure 5.6: Variable correlation

The model was trained using three different algorithms; Logistic regression, decision tree and TensorFlow. According to Figure 2.13, when the logistic model was used for prediction in the banking industry, it only relied on demographic information, however from the Figure 5.6 and 5.7, we can see that the model developed by researcher also include economic information such as employment duration.

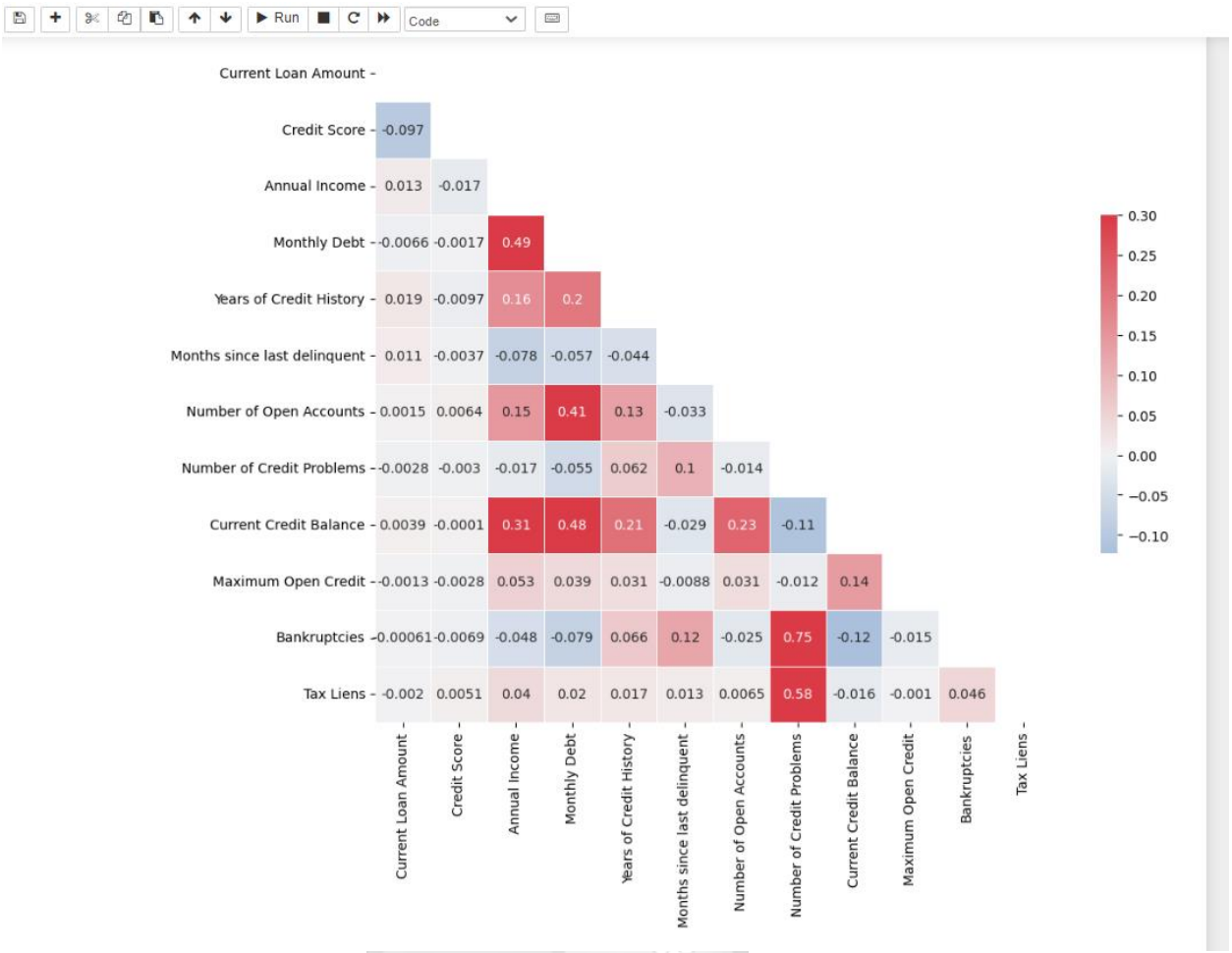


Figure 5.7: Variable correlation heatmap

5.4. Model Validation, Accuracy and Testing

5.4.1. Model Validation

For model validation, the model proved to be accurate in prediction of loan default likelihood of new, unseen and unknown loan customer loan borrowing profiles. Once the model finished running, the user can download a CSV file with the added column Predicted Loan Default at the end of the existing data. The column shows whether or not the user will default if given the full loan amount. Figure 5.8 shows how to add the prediction column to the CSV file.

```
In [84]:
encoded_data2_test = StandardScaler().fit_transform(encoded_data2_test)

encoded_data2_test = pd.DataFrame(encoded_data2_test)

In [85]:
# Test on test data
prediction2 = model2.predict(encoded_data2_test)

# Converting predicted probabilities to binary predictions
threshold2 = 0.5
prediction2_binary = (prediction2 >= threshold).astype(int)

In [ ]:
prediction2_df = pd.DataFrame(prediction2_binary, columns=['Predicted Loan Default', 'dummy'])
prediction2_df.drop(columns=['dummy'], inplace=True)

set2_result = pd.concat([set2_test, prediction2_df], axis=1)

set2_result.head()

In [87]:
set2_result.to_csv("Set 2 Prediction Results.csv")
```

Figure 5.8: Prediction Results and CSV download

5.4.2. Model Accuracy

The model was trained using 3 different algorithms: Logistic regression, decision trees and tensor flow. The data was used such that 80% of the records were used for training and 20% for testing of the model for the prediction of results. The same sizes were also used for the recommendation algorithm. Figure 5.9 shows the accuracy, precision and f1-scores for the different algorithms that were run to predict the default likelihood, Figure 5.10 the TensorFlow scores and 5.11 shows the decision tree scores.

```
In [68]:
# Train using Logistic regression
logistic2 = LogisticRegression()
logistic2.fit(X2_train, y2_train)

y_pred2 = logistic2.predict(X2_test)

# Confusion Matrix
print(confusion_matrix(y2_test, y_pred2))

# Classification Report
print(classification_report(y2_test, y_pred2))
```

```
[[  924  3555]
 [   13 15508]]
              precision    recall  f1-score   support

           0       0.99      0.21      0.34      4479
           1       0.81      1.00      0.90     15521

    accuracy                           0.82     20000
   macro avg       0.90      0.60      0.62     20000
weighted avg       0.85      0.82      0.77     20000
```

Figure 5.9: Logistic regression scores

In [ ]:
```
# Train using tensorflow
model2 = tf.keras.Sequential([
    tf.keras.layers.Dense(64, activation='relu', input_shape=(45,)),
    tf.keras.layers.Dense(32, activation='relu'),
    tf.keras.layers.Dense(2, activation='softmax')
])

# Compile your model
model2.compile(optimizer='adam',
               loss='sparse_categorical_crossentropy',
               metrics=['accuracy'])

model2.fit(X2_train, y2_train, epochs=100, batch_size=32, validation_data=(X2_test, y2_test))
```

In [74]:
```
y_test_onehot2 = to_categorical(y2_test)

y_pred_onehot2 = model2.predict(X2_test)
y_pred_labels2 = np.argmax(y_pred_onehot2, axis=1)
y_true_labels2 = np.argmax(y_test_onehot2, axis=1)

print(classification_report(y_true_labels2, y_pred_labels2))
```

```
              precision    recall  f1-score   support

           0       0.69      0.28      0.40      4479
           1       0.82      0.96      0.89     15521

    accuracy                           0.81     20000
   macro avg       0.76      0.62      0.64     20000
weighted avg       0.79      0.81      0.78     20000
```

Figure 5.10: TensorFlow scores

n [69]:
```
# Train using Decision Tree
classifier2 = DecisionTreeClassifier()

classifier2.fit(X2_train, y2_train)

y_predc = classifier2.predict(X2_test)

print(confusion_matrix(y2_test, y_predc))

print(classification_report(y2_test, y_predc))
```

```
[[ 1999  2480]
 [ 2417 13104]]
              precision    recall  f1-score   support

           0       0.45      0.45      0.45      4479
           1       0.84      0.84      0.84     15521

    accuracy                           0.76     20000
   macro avg       0.65      0.65      0.65     20000
weighted avg       0.75      0.76      0.75     20000
```

Figure 5.11: Decision Tree scores

5.4.3. Model Testing

Table 5.12 shows the different tests done on the model and the results achieved on each test case while including the importance level of each of them. The test cases were based mostly on the objectives of the study.

Table 5.2: Test cases and their results

Test Case	**Importance Level**	**Results**
Can data be uploaded to the model in form of a CSV?	High	Successful upload of data
Does the model predict the likelihood of default?	High	Prediction done with accuracy levels of an average of 83%
Does the model recommend an amount for predicted defaults?	High	Recommended loan amount suggested for loan borrowers predicted to default.

Chapter 6: Discussion of Results

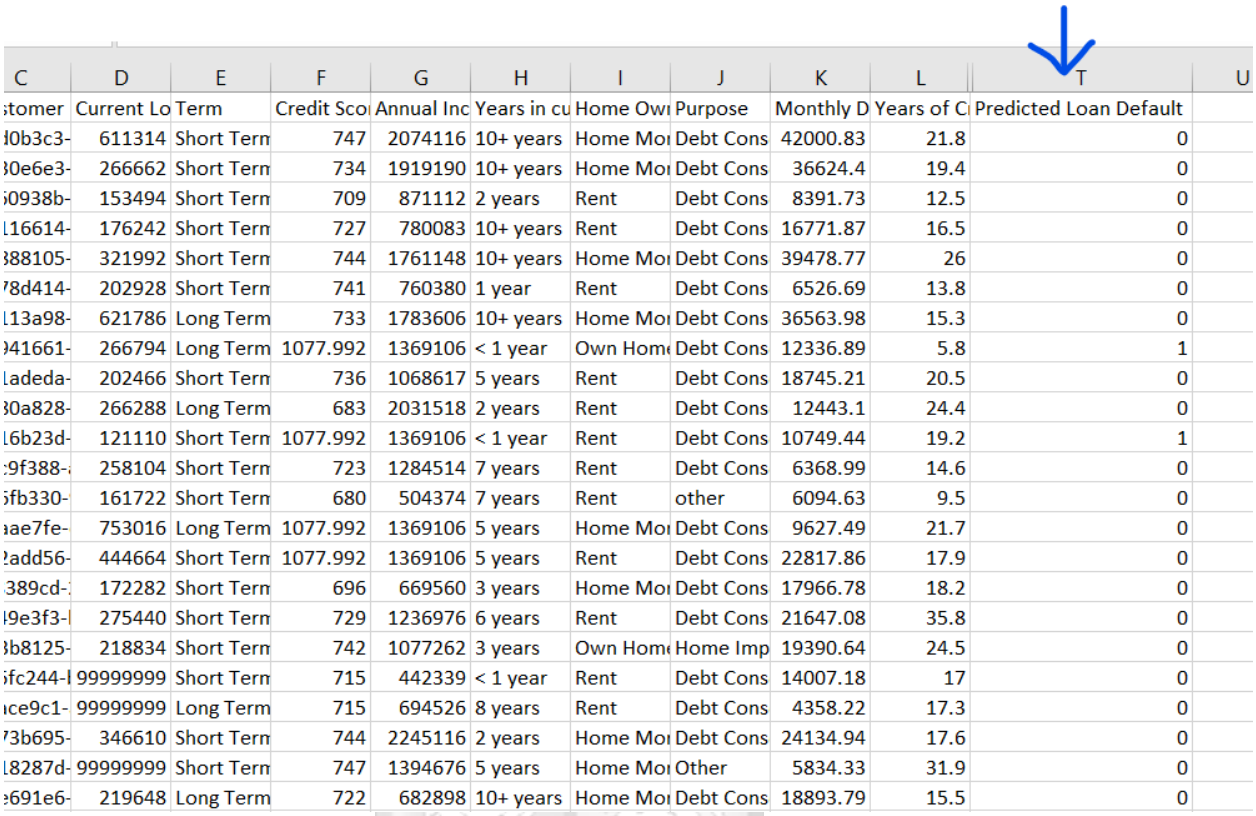
6.1. Introduction

The study's main purpose was the development of a loan default prediction and loan amount recommendation tool to help Okoa Management Ltd in the processing of their loan applications. The development of the tool will not only help them in faster processing on loans, but it will also help in pattern behavior and feature determination which cause loan repayment inability. Using a combination of the historical data of the client and their credit scores, has made the model improve its accuracy and performance as it incorporates a lot more variables.

6.2. Study Results

The tool was created using 3 different algorithms: Logistic regression, Decision trees and TensorFlow. The loan predicted was passed through all the 3 algorithms and the algorithm with the highest accuracy (Logistic Regression) was used to make the prediction of whether or not the borrower was likely to default in the repayment of the borrowed loan. After the prediction was made, a column was added to the CSV uploaded with the prediction could be downloadable. The fact that the model predicted the results, the human element biasness present in manually approving loans was removed resulting in fair, unbiased predictions.

Figure 6.1 shows a sample of how the downloaded CSV with the prediction column looks like after the model makes the prediction on the loan default likelihood. When the model was developed, the values were encoded as 1 – (non-defaulter; approve full amount borrowed) and 0 – (defaulter – partially approve the amount or not depending on the borrower's characteristics as well as the loan characteristics)



C	D	E	F	G	H	I	J	K	L	T	U
stomer	Current Lo	Term	Credit Sco	Annual Inc	Years in cu	Home Ow	Purpose	Monthly D	Years of C	Predicted Loan Default	
40b3c3-	611314	Short Term	747	2074116	10+ years	Home Mo	Debt Cons	42000.83	21.8	0	
30e6e3-	266662	Short Term	734	1919190	10+ years	Home Mo	Debt Cons	36624.4	19.4	0	
50938b-	153494	Short Term	709	871112	2 years	Rent	Debt Cons	8391.73	12.5	0	
116614-	176242	Short Term	727	780083	10+ years	Rent	Debt Cons	16771.87	16.5	0	
388105-	321992	Short Term	744	1761148	10+ years	Home Mo	Debt Cons	39478.77	26	0	
78d414-	202928	Short Term	741	760380	1 year	Rent	Debt Cons	6526.69	13.8	0	
113a98-	621786	Long Term	733	1783606	10+ years	Home Mo	Debt Cons	36563.98	15.3	0	
941661-	266794	Long Term	1077.992	1369106	< 1 year	Own Hom	Debt Cons	12336.89	5.8	1	
1adeda-	202466	Short Term	736	1068617	5 years	Rent	Debt Cons	18745.21	20.5	0	
30a828-	266288	Long Term	683	2031518	2 years	Rent	Debt Cons	12443.1	24.4	0	
16b23d-	121110	Short Term	1077.992	1369106	< 1 year	Rent	Debt Cons	10749.44	19.2	1	
29f388-	258104	Short Term	723	1284514	7 years	Rent	Debt Cons	6368.99	14.6	0	
5fb330-	161722	Short Term	680	504374	7 years	Rent	other	6094.63	9.5	0	
3ae7fe-	753016	Long Term	1077.992	1369106	5 years	Home Mo	Debt Cons	9627.49	21.7	0	
2add56-	444664	Short Term	1077.992	1369106	5 years	Rent	Debt Cons	22817.86	17.9	0	
389cd-	172282	Short Term	696	669560	3 years	Home Mo	Debt Cons	17966.78	18.2	0	
19e3f3-	275440	Short Term	729	1236976	6 years	Rent	Debt Cons	21647.08	35.8	0	
3b8125-	218834	Short Term	742	1077262	3 years	Own Hom	Home Imp	19390.64	24.5	0	
5fc244-	99999999	Short Term	715	442339	< 1 year	Rent	Debt Cons	14007.18	17	0	
1ce9c1-	99999999	Long Term	715	694526	8 years	Rent	Debt Cons	4358.22	17.3	0	
73b695-	346610	Short Term	744	2245116	2 years	Home Mo	Debt Cons	24134.94	17.6	0	
18287d-	99999999	Short Term	747	1394676	5 years	Home Mo	Other	5834.33	31.9	0	
2691e6-	219648	Long Term	722	682898	10+ years	Home Mo	Debt Cons	18893.79	15.5	0	

Figure 6.1: Loan Prediction Results

Once the predictions are made, the generated CSV above is then uploaded to the recommendation model when loans that were predicted to be defaulted in repayments are recommended for a lower amount. The recommendation algorithms focused on two main things: the credit score of the customer/borrower, and the similarity of the loan details (amount, borrower) to other borrowers whose loan amounts were approved, and uses these to check how for what amount to recommend the borrower with. Figure 6.2 shows how the data will look like after the recommendation is done and the CSV is downloaded by the user. For both Figures 6.1 and 6.2, some columns have been hidden to scale the image to fit on the page.

	F	G	H	I	J	K	L	U	V
ent Lo	Term	Credit Sco	Annual Inc	Years in cu	Home Ow	Purpose	Monthly D	Predicted Loan Default	Recommendation
1314	Short Term	747	2074116	10+ years	Home Mo	Debt Cons	42000.83	0	Loan upto 500,000
6662	Short Term	734	1919190	10+ years	Home Mo	Debt Cons	36624.4	0	Loan upto 250,000
3494	Short Term	709	871112	2 years	Rent	Debt Cons	8391.73	0	Loan Rejected, likely to default
6242	Short Term	727	780083	10+ years	Rent	Debt Cons	16771.87	0	Loan Rejected, likely to default
1992	Short Term	744	1761148	10+ years	Home Mo	Debt Cons	39478.77	0	Loan Rejected, likely to default
2928	Short Term	741	760380	1 year	Rent	Debt Cons	6526.69	0	Loan Rejected, likely to default
1786	Long Term	733	1783606	10+ years	Home Mo	Debt Cons	36563.98	0	Loan upto 250,000
6794	Long Term	1077.992	1369106	< 1 year	Own Hom	Debt Cons	12336.89	1	Loan Approved
2466	Short Term	736	1068617	5 years	Rent	Debt Cons	18745.21	0	Loan Rejected, likely to default
6288	Long Term	683	2031518	2 years	Rent	Debt Cons	12443.1	0	Loan upto 250,000
1110	Short Term	1077.992	1369106	< 1 year	Rent	Debt Cons	10749.44	1	Loan Approved
8104	Short Term	723	1284514	7 years	Rent	Debt Cons	6368.99	0	Loan upto 250,000
1722	Short Term	680	504374	7 years	Rent	other	6094.63	0	Loan Rejected, likely to default
3016	Long Term	1077.992	1369106	5 years	Home Mo	Debt Cons	9627.49	0	Loan Rejected, likely to default
4664	Short Term	1077.992	1369106	5 years	Rent	Debt Cons	22817.86	0	Loan Rejected, likely to default
2282	Short Term	696	669560	3 years	Home Mo	Debt Cons	17966.78	0	Loan Rejected, likely to default
5440	Short Term	729	1236976	6 years	Rent	Debt Cons	21647.08	0	Loan upto 250,000
8834	Short Term	742	1077262	3 years	Own Hom	Home Imp	19390.64	0	Loan Rejected, likely to default
9999	Short Term	715	442339	< 1 year	Rent	Debt Cons	14007.18	0	Loan upto 250,000
9999	Long Term	715	694526	8 years	Rent	Debt Cons	4358.22	0	Loan upto 250,000

Figure 6.2: Loan Recommendation Results

From the results obtained in the study – from the correlation matrix, the following observations were made:

i. The following variables have a strong relationship with the Loan Status - Home Ownership and Number of credit problems
ii. The following have a slight relationship with the Loan Status – Bankruptcies, Months since last Delinquent and Annual Income
iii. The decision tree algorithm had the highest accuracy level at 90% which is a good level of prediction
iv. The borrowers most likely to default are those who do not own houses or have rental cars.
v. Borrowers with low credit scores and high loan amount request with no collateral (have rental houses) and low incomes were more likely to default.

There were also more observations made from the study results. It was shown that the accuracy levels of the different algorithm levels ranged from between 78%- 90% shown in Figures 5.9, 5.10 and 5.11. Though some existing models have higher accuracy levels, my results were affected by the amount of data was that was used to train my model. Since my data source was from one organization, the amount was collected was not enough in comparison with data used in existing models. While existing models were trained with millions of data, the researcher's model only had about 300000 data entries.

Cleaning of the data to remove null value entries also reduced my data set. The reduced data for training and validation, could be among the factors that affected my model's accuracy.

6.3. Objectives Accomplishment

The main objectives of the study were to develop a tool to help Okoa Management in prediction of loan default and recommending of lower amount for borrowers likely to default in the full amount they borrowed. The objectives as defined in chapter 1 were accomplished as follows:

i. A thorough review was done and a lot of communication with a respondent to determine which factors to use as attributes of the model
ii. A model was developed to help in predicting the loan default likelihood – shown in chapter 5
iii. For loans predicted to likely default, the predictions were run on another algorithm where a recommendation of a lower amount was suggested to the borrowers – based on the borrower's credit score and the similarity of the loan details to loan details that were approved.
iv. The model was validated with an accuracy level of about 90% when using the decision tree algorithm for prediction

The study managed to achieve everything that is set out to achieve based on the objectives described and the scope defined in Chapter 1. With the model being able to be run both locally and, in the cloud, there is little to no cost of set up making it an ideal solution to Okoa as there are minimal cost implications for them.

6.4. Research Limitations

Despite all the objectives set being accomplished by the study, the tool developed in the research still has some shortcomings. Some of the limitations of this research were:

i. As the research was focused on Okoa Management (narrow and specific focused research), it might not be suitable to other SACCOs that use different factors in evaluating a borrower's likelihood to default making it a customized solution instead of a generic one of one size fits all (which should ideally be the case as most SACCOs in Nairobi use almost similar factors when evaluating a borrower's risk when processing their loans).

ii. The use has to manually upload results from the prediction to the recommendation model instead of them being passed dynamically – this may increase the time taken (though not by much as the approach is faster than manual evaluation).

iii. For the recommendation, the main attribute that was considered was the borrower's credit score, yet there are other attributes that had high correlations with the default likelihood.

Chapter 7: Conclusion and Recommendations

7.1. Conclusions

The loan amount prediction and recommendation tool has proven to be effective and efficient with increased accuracy levels and reduced amount of work in processing loan applications. With the model being able to predict default likelihood and recommend amounts, it is suitable for small SACCOS and micro-credit institutions with similar criteria as SACCO used as the case study for this research (Okoa Management). The number of benefits that SACCOs could exploit when they adopted this tool are immense. Not only with the SACCOs benefit but their members as well as the turnaround time for processing the applications will be reduced through the use of the model.

With the use of the developed tool, we ensure that there is accurate prediction of loan default likelihood as well as reduced/ minimized biasness in the loan application approval process by eliminating the human element present in the manual application and approval process.

With the continuous advancements in AI and machine learning, better and more accurate models will continue being developed to help in the banking, micro-finance and credit lending institutions. Continuous research and development in these industries is important as well to because if the ever-changing dynamics of the people, meaning the models need to be trained and re-trained as frequently as possible to learn the changing behaviors of the people.

7.2. Recommendations

Based on the research results and findings, the following recommendations can be made to promote adaptation of not just this tool but other automated tools:

i. The SACCO governing body (The SACCO Societies Regulatory Authority (SASRA)) should encourage SACCOs to use automated systems through the policies they make – either making it mandatory, offering subsidized membership fees or other ways to encourage the members to automate their systems.

ii. Improving participant involvement during the entire process to make the acceptance of the developed easier as part of the organization's culture. This will increase the support of its adaptation. In this research there was minimal participation as I only interacted with one member of the entire organization.

7.3. Suggestions for Future Works

Further research can be conducted in the following areas:

i. How the tool can be developed and improved to become a generic tool that can be used across multiple SACCOs with different evaluation criteria.

ii. Using different algorithms and parameters with different methods of parameter tuning and data clean up to check if higher accuracy levels can be achieved.

iii. Further research and exploration can be done to determine the expected returns earned from the loan should the borrowers default. Returns could be considered as the fines charged, value of the collateral to be confiscated and/or repossessed, etc.

iv. There is need to determine how the use of machine learning in loan default prediction and recommendation can be used by organizations to create loan options that reduce their risk to lose money through defaults.

References

Aasim, O. (2019, September 17). *Machine Learning Project 17 — Compare Classification Algorithms*. Retrieved from Towards Data Science: https://towardsdatascience.com/machine-learning-project-17-compare-classification-algorithms-87cb50e1cb60

Amazon Web Services. (n.d.). *Binary Classification*. Retrieved from Amazon Web Services: https://docs.aws.amazon.com/machine-learning/latest/dg/binary-classification.html

Anatoly. (2014). The probability of default models of Russian banks. *Journal of Institute of Economics in Transition, 21*(5), 203-278.

Andrea, R. (2010). Measuring the likelihood of small Business default. *Journal of Applied Sciences, 33*(7), 1289-1386.

Arora, S., Sushant, B., Survesh , S., & Vinay, K. N. (2022). Materials Today Proceedings. *Prediction of credit card defaults through data analysis and machine learning techniques,, 51*, 110-117. doi:https://doi.org/10.1016/j.matpr.2021.04.588.

Arya, N. (2022, April 4). *Logistic Regression for Classification*. Retrieved from KD Nuggets: https://www.kdnuggets.com/2022/04/logistic-regression-classification.html

Aryal, S. (2022, March 17). *Questionnaire method of data collection*. Retrieved from The Biology Notes: https://thebiologynotes.com/questionnaire-method-of-data-collection/

Atlassian. (n.d.). *What is Agile*. Retrieved from Atlassian: https://www.atlassian.com/agile#:~:text=Agile%20is%20an%20iterative%20approach,small%2C%20but%20consumable%2C%20increments

AWS. (n.d.). *Binary Classification*. Retrieved from Amazon Web Services: https://docs.aws.amazon.com/machine-learning/latest/dg/binary-classification.html

Bhatia, M. (2018, September 5). *Your Guide to Qualitative and Quantitative Data Analysis Methods*. Retrieved from Humans of Data: https://humansofdata.atlan.com/2018/09/qualitative-quantitative-data-analysis-methods/

Brownlee, J. (2019, August 12). *Overfitting and Underfitting With Machine Learning Algorithms*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Business Daily. (2020, September 7). *Sacco loan defaulters have nowhere to hide in era of credit bureaus*. Retrieved from Business Daily: https://www.businessdailyafrica.com/bd/lifestyle/personal-finance/sacco-loan-defaulters-have-nowhere-to-hide-in-era-of-credit-bureaus--2067316

Central Bank of Kenya. (2019, April). *2019 FinAccess HouseHold survey*. Retrieved from Central Bank of Kenya:

/https://www.centralbank.go.ke/uploads/financial_inclusion/2050404730_FinAccess%202019%20Household%20Survey-%20Jun.%2014%20Version.pdf

Central Bank of Kenya. (2019, June 14). *FinAccess 2019 Household Survey*. Retrieved from Central Bank: https://www.centralbank.go.ke/uploads/financial_inclusion/2050404730_FinAccess%202019%20Household%20Survey-%20Jun.%2014%20Version.pdf

Chauhan, N. S. (2022, April 8). *Naïve Bayes Algorithm: Everything You Need to Know*. Retrieved from KDNuggets: https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

Christopher, A. (2021, February 2). *K-Nearest Neighbor*. Retrieved from Medium: https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4

Corporate Finance Institute. (2021, September 1). *Random Forest*. Retrieved from Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/knowledge/other/random-forest/

Corporate Finance Institute. (2022, May 6). *Credit Risk Analysis Models*. Retrieved from Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/knowledge/credit/credit-risk-analysis-models/

DataRobot. (2021). *Predict Likelihood of Loan Default*. Retrieved from DataRobot: https://pathfinder.datarobot.com/use-case/predict-likelihood-credit-default?tab=overview

Drumond, C. (2022, April). *Is the Agile Manifesto still a thing?* Retrieved from Atlassian: https://www.atlassian.com/agile/manifesto

Dudovskiy, J. (2022, January). *Positivism Research Philosophy*. Retrieved from Business Research Methodology: https://research-methodology.net/research-philosophy/positivism/

Financial Access. (2013). *Profiling developments in financial access and usage in Kenya*. Retrieved from Financial Access: www.kenyacic.org/sites/default/.../ 131031_ FinAccess _2013

Geeks for Geeks. (2022, May 19). *ML|Linear Regression*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/ml-linear-regression/#:~:text=Linear%20Regression%20is%20a%20machine,relationship%20between%20variables%20and%20forecasting.

GeeksforGeeks. (2020, September 2). *Advantages and Disadvantages of Logistic Regression*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Gouda, D. I., A, A. K., Madivala, A. M., & R, D. K. (2021, January). LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING. *International Research Journal of Engineering and Technology, 8*(1), 2395-0072.

Gundaniya, N. (n.d.). *What is SACCOs and why should it be digitized?* Retrieved from Digipay Guru: https://www.digipay.guru/blog/digital-evolution-of-saccos-after-covid-19/

Hayes, A. (2021, February 09). *FICO Score*. Retrieved from Investopedia: https://www.investopedia.com/terms/f/ficoscore.asp

IBM. (n.d.). *What is the k-nearest neighbors algorithm?* Retrieved from IBM: https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20algorithm%2C%20also%20known%20as%20KNN%20or,of%20an%20individual%20data%20point.

Jansen, D. (2021, June). *How To Choose Your Research Methodology*. Retrieved from Grad Couch: https://gradcoach.com/choose-research-methodology/

Jansen, D., & Warren, K. (2020, December 01). *Quantitative Data Analysis 101: The lingo, methods and techniques, explained simply*. Retrieved from Grad Coach: https://gradcoach.com/quantitative-data-analysis-methods/

Jemoek, N. (2013, October). *The Relationship Between Loan Default and the Financial Performance of SACCOs in Kenya* . Retrieved from UON Reposiroty: http://erepository.uonbi.ac.ke/bitstream/handle/11295/58986/NANCY%20FINAL%20PROJECT.pdf?sequence=3&isAllowed=y

Kanade, V. (2022, April 8). *What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022*. Retrieved from Spice Works: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/

Karn, T. (2022, January 2). *Data Privacy & security issue in Financial Industry-THE KARN*. Retrieved from LinkedIn: https://www.linkedin.com/pulse/data-privacy-security-issue-financial-industry-the-karn-the-karn

Kibuacha, F. (2021, April 06). *How to Determine Sample Size for a Research Study*. Retrieved from GeoPoll: https://www.geopoll.com/blog/sample-size-research/

Kimisitu Sacco. (2021, March 17). *Frequently Asked FOSA Questions*. Retrieved from Kimisitu Sacco Ltd.: http://www.kimisitusacco.or.ke/documents/FAQs.pdf

Kosarenko, Y. (2021, November 13). *How to Create Decision Trees for Business Rules Analysis*. Retrieved from Why Change Consulting: https://why-change.com/2021/11/13/how-to-create-decision-trees-for-business-rules-analysis/

Kumar, A. (2022, July 8). *Loan Eligibility Prediction using Machine Learning*. Retrieved from Data Analytics: https://vitalflux.com/loan-eligibility-prediction-using-machine-learning/#:~:text=Machine%20learning%20can%20be%20used%20to%20automate%20the%20decision%2Dmaking,are%20predictive%20of%20loan%20default.

Kumar, A. (2022, July 8). *Loan Eligibility Prediction using Machine Learning*. Retrieved from Data Analytics: https://vitalflux.com/loan-eligibility-prediction-using-machine-learning/

Least square method. (2018, March). Retrieved from ByJU's: https://byjus.com/maths/least-square-method/

Maina, M. (2021, November). *Kenya: The cycle of mobile app loans is costing the youth its health and finances*. Retrieved from The African Report: https://www.theafricareport.com/149766/kenya-the-cycle-of-mobile-app-loans-is-costing-the-youth-its-health-and-finances/

Maione, I. (2022, August 23). *What is Primary Data Collection? Types, Advantages, and Disadvantages*. Retrieved from Click Worker: https://www.clickworker.com/customer-blog/primary-data-collection/

Marwa, M., & Aziakpono, M. (2015). Financial sustainability of Tanzanian saving and credit cooperatives. *Journal of Social Economics*.

Matara, V. (2022, March 10). *List of All Licensed SACCOs In Kenya 2022*. Retrieved from Mictor Matara: https://victormatara.com/list-of-all-licensed-saccos-in-kenya-2020/#:~:text=There%20are%20175%20licensed%20to,in%20Kenya%20as%20of%202022.

Mwaniki, C. (2022, August 25). *Bank loan defaults cross half a trillion for the first time*. Retrieved from Business Daily: https://www.businessdailyafrica.com/bd/markets/capital-markets/bank-loan-defaults-cross-half-a-trillion-for-the-first-time-3925518

O., A., & A., W. (2012). Parametric modeling of the probability of bank loan default in Kenya. *Journal of Applied Statistics, 14*(1), 61-74.

Ogola, W. (2021, July 3). *Entropy and Information Gain to Build Decision Trees in Machine Learning*. Retrieved from Section: https://www.section.io/engineering-education/entropy-information-gain-machine-learning/#entropy

Olando, C. O., Mbewa, O. M., & Jagongo, A. (2013). Financial Practice as a Determinant of Growth of Savings and. *A Pointer to Overcoming Poverty Challenges in Kenya and the*.

Olando, O. C., & Mbewa, M. O. (2012). The Contribution of SACCO Financial Stewardship to Growth of SACCOS in Kenya. *International Journal of Humanities and Social Science, 3*(17), 112-132.

Opiyo, I. (2014, May 04). *What sacco members must do to increase dividend payouts*. Retrieved from Business Daily: https://www.businessdailyafrica.com/bd/lifestyle/personal-finance/what-sacco-members-must-do-to-increase-dividend-payouts-2057442

Pratt, M. K., & Torode, C. (2020, April). *Agile Manifesto*. Retrieved from TechTarget: https://www.techtarget.com/searchcio/definition/Agile-Manifesto

Project Pro. (2022, July 26). *15 Popular Machine Learning Frameworks for Model Training*. Retrieved from Project Pro: https://www.projectpro.io/article/machine-learning-frameworks/509

Question Pro. (n.d.). *Questionnaires: The ultimate guide, advantages & examples*. Retrieved from Question Pro: https://www.questionpro.com/blog/what-is-a-questionnaire/

Rovyana, A. (2022, June 8). *Agile Software Development Methodology: Definition, Types, Workflows*. Retrieved from Cleveroad: https://www.cleveroad.com/blog/agile-software-development/

Salaton, K. E., Gudda, P., & Rukaria, G. (2020). Effect of Loan Default Rate on Financial Performance of Savings and Credit Cooperative Societies Innarok, County Kenya. *International Journal of Academic Research in Accounting, Finance and Management Sciences, 10*(2), 65-75. doi:10.6007/IJARAFMS/v10-i2/7345

Shwe, M., Dillon, J., & Seybold, B. (2018, December 10). *An introduction to probabilistic programming, now available in TensorFlow Probabilit*. Retrieved from TensorFlow Blog: https://blog.tensorflow.org/2018/12/an-introduction-to-probabilistic.html

Stat Counter Global Stats. (n.d.). *Desktop Operating System Market Share Kenya*. Retrieved from Stat Counter Global Stats: https://gs.statcounter.com/os-market-share/desktop/kenya/#monthly-202107-202207

Statcounter Global Stats. (n.d.). *Desktop Operating System Market Share Kenya*. Retrieved from Statcounter Global Stats: https://gs.statcounter.com/os-market-share/desktop/kenya/#monthly-202107-202207

Stellenbosch Business School. (2018). *Why publication in an academic journal matters*. Retrieved from Stellenbosch Business School: https://www.usb.ac.za/usb_food_for_thought/why-publication-in-an-academic-journal-matters/

TensorFlow. (2019, June 13). *Modeling "Unknown Unknowns" with TensorFlow Probability — Industrial AI, Part 3*. Retrieved from TensorFlow: https://blog.tensorflow.org/2019/06/modeling-unknown-unknowns-with-tensorflow-probability.html

TensorFlow. (n.d.). *TensorFlow Probability*. Retrieved from TensorFlow: https://www.tensorflow.org/probability/overview

The Investopedia Team. (2021, April 08). *Credit Scoring*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/credit_scoring.asp

Thmoson, E. (n.d.). *Why methodologies are important?* Retrieved from Future Learn: https://www.futurelearn.com/info/courses/research-project/0/steps/4055

Thomas, L. (2020, September 18). *Stratified Sampling | Definition, Guide & Examples*. Retrieved from Scribbr: https://www.scribbr.com/methodology/stratified-sampling/

Tilimbe, J. (2019). Ethical Implications of Predictive Risk Intelligence. *ORBIT Journal, 2*(2). doi:10.29297

Understanding logistic regression analysis. (2014, February 15). Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/#:~:text=Logistic%20regression%20is%20used%20to,the%20observed%20event%20of%20interest.

Wanjohi, S. M., Waititu, A. G., & Kibira, A. W. (2016). Modeling Loan Defaults in Kenya Banks as a Rare Event Using the Generalized Extreme Value Regression Model. *Science Journal of Applied Mathematics and Statistics., 4*(6), 289-297. doi:10.11648/j.sjams.20160406.17

Wanjohi, S. M., Waititu, A. G., & Wanjoya, A. K. (2016, December). Modeling Loan Defaults in Kenya Banks as a Rare Event Using the Generalized Extreme Value Regression Model. *Science Journal of Applied Mathematics and Statistics, 4*(6), 289-297.

World Bank. (n.d.). *Credit scoring approaches guideline*. Retrieved from World Bank: http://pubdocs.worldbank.org/en/935891585869698451/CREDIT-SCORING-APPROACHES-GUIDELINES-FINAL-WEB.pdf

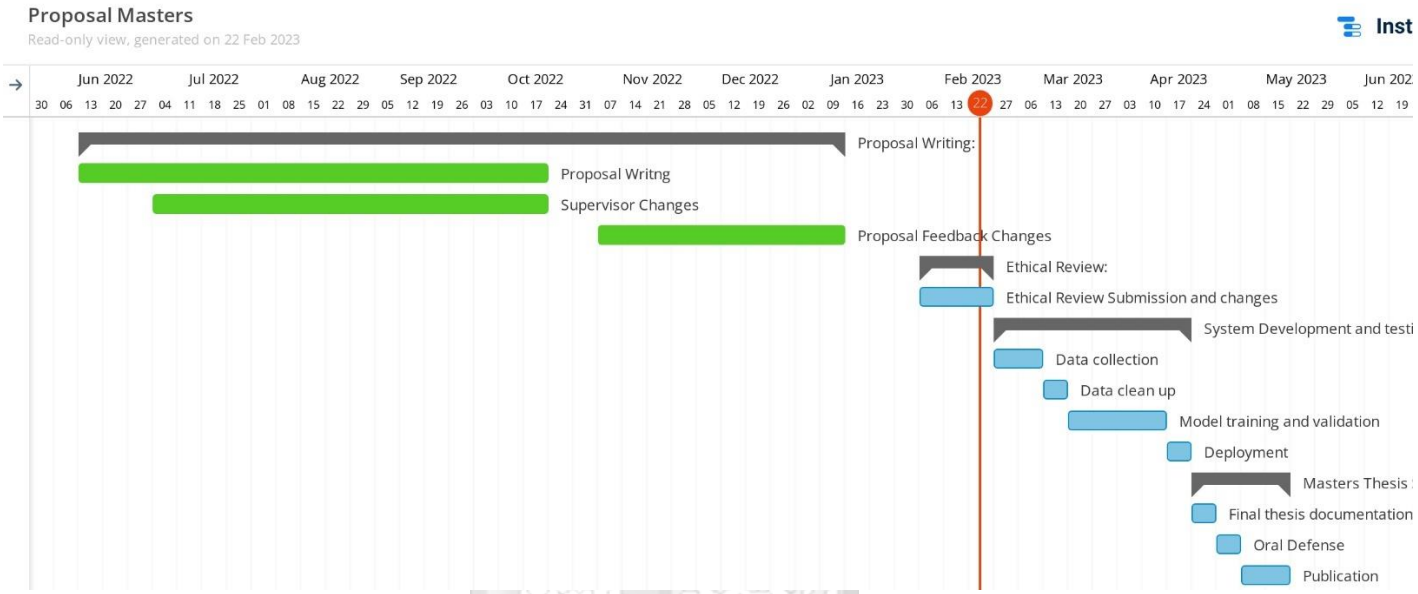
Xoriant. (n.d.). *Decision Trees for Classification: A Machine Learning Algorithm*. Retrieved from Xoriant: https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm#:~:text=Introduction%20Decision%20Trees%20are%20a,namely%20decision%20nodes%20and%20leaves

Yiu, T. (2019, June 12). *Understanding Random Forest*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.

Zhao, S. (2020, August 23). *Predicting Loan Defaults Using Logistic Regression*. Retrieved from Medium: https://selenaezhao.medium.com/predicting-loan-defaults-using-logistic-regression-71b7482a8cf7#:~:text=We%20were%20able%20to%20conclude,that%20measured%20accuracy%20and%20error.

Appendices

Appendix A: Gantt Chart



Appendix 1: Project Gantt Chart

Appendix B: Ethical Approval

Strathmore UNIVERSITY

17th March 2023

Ms Mwalozi Purity,
purity.mwalozi@strathmore.edu

Dear Ms Mwalozi,

RE: A Loan Default Prediction and Loan Amount Recommendation Tool for Saccos in Nairobi, A Case of Okoa Management

This is to inform you that SU-ISERC has reviewed and **approved** your above SU- **master's** research proposal. Your application reference number is **SU-ISERC1624/23.** The approval period is from **17th March 2023 to 16th March 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-ISERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

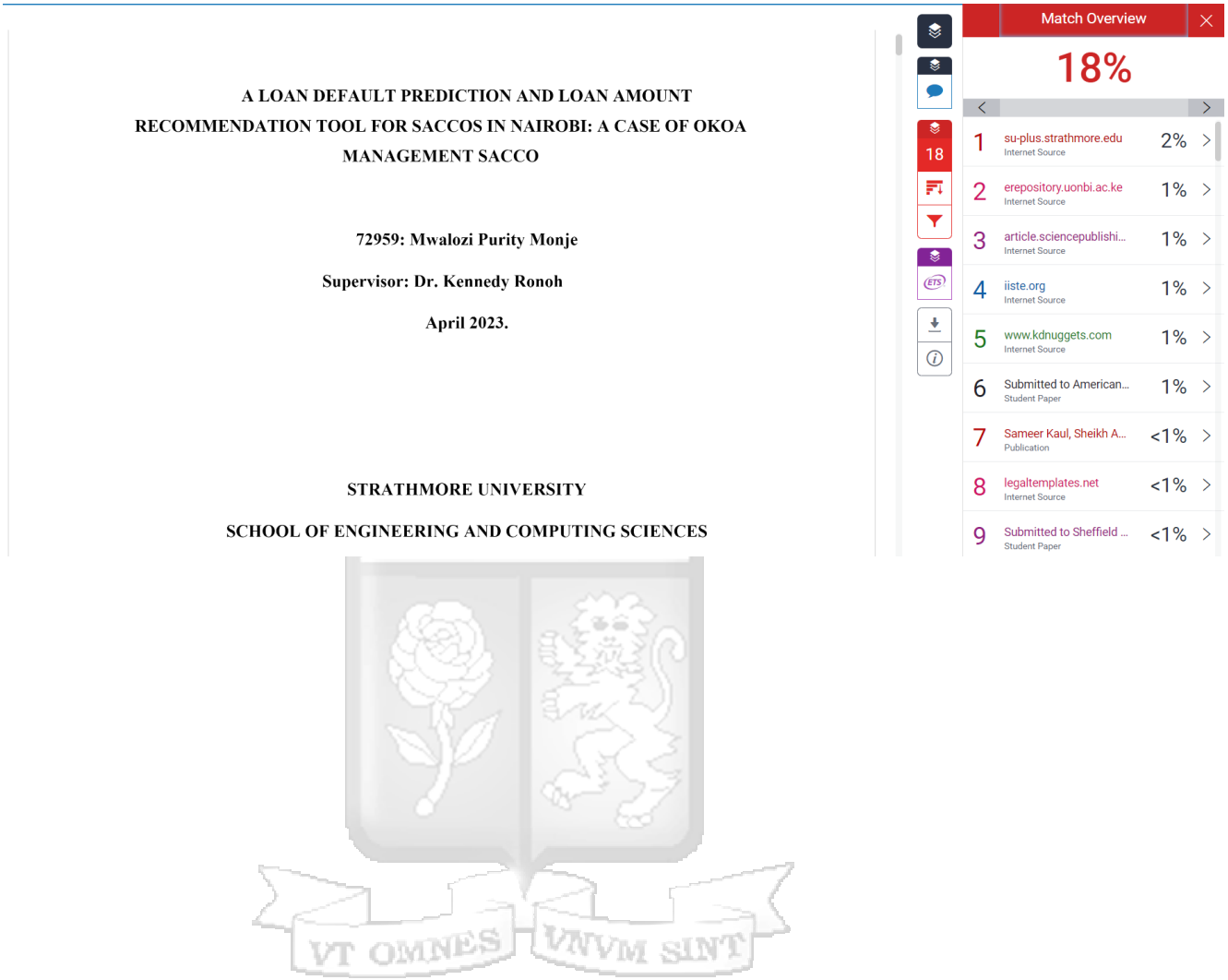
Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-ISERC

Cc: Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C: Turn-it in Plagiarism Level



A LOAN DEFAULT PREDICTION AND LOAN AMOUNT RECOMMENDATION TOOL FOR SACCOS IN NAIROBI: A CASE OF OKOA MANAGEMENT SACCO

72959: Mwalozi Purity Monje

Supervisor: Dr. Kennedy Ronoh

April 2023.

STRATHMORE UNIVERSITY

SCHOOL OF ENGINEERING AND COMPUTING SCIENCES

Appendix D: Participant Information and Consent Form

A Loan Default Prediction and Loan Amount Recommendation Tool for Saccos in Nairobi, A Case of Okoa Management Sacco

SECTION 1: INFORMATION SHEET

Investigator: Mwalozi Purity Monje

Institutional affiliation: Strathmore School of Computing and Engineering Sciences

SECTION 2: INFORMATION SHEET–THE STUDY

2.1: Introduction and Purpose of the study.

The purpose of this research is the design, and development of a machine learning tool to predict loan default likelihood and recommend lower amounts to users who are likely to default. The tool will be designed to retrieve loan characteristics from an uploaded CSV file. The data will then be sent to the processing model for validation and a response will be sent to the user on whether or not the borrower should be given the loan in full.

2.2: Who is eligible to take part in this study?

- Okoa Management SACCO Ltd. Relying on the data to be provided by the SACCO on member loans

2.3: Who is not eligible to take part in this study?

- Other SACCOs. They were not included in the research as the tool developed is to be customized to the how Okoa operates and their rules.

2.4: What will it entail for me to participate in this study??

Taking part in this study as a participant will involve the following steps:

- Being approached by the investigator and being asked to take part in the study in the data collection and testing stage.

- ❖ Reading and fully understanding the goals of the study
- ❖ Signing the informed consent forms (both participant consent and privacy and confidentiality agreements).

2.5: Are there any risks or dangers in taking part in this study?

Taking part in this study has no risks. You are only required to provide member data on their borrowing histories. The members' identity will be kept confidential and no personal information will be disclosed or shared to unauthorized personnel.

2.6: Are there any benefits of taking part in this study?

Yes. Some of the benefits from this study include;

- ❖ Contributing to scientific research and knowledge advancement in the field of machine learning in the microfinance industry in loan disbursement.
- ❖ Providing valuable feedback on the tool developed, which may help improve its design, functionality and reliability.
- ❖ Having the opportunity to try out the latest technology and potentially improve your experience with loan recommendations, loan default likelihood and the entire loan disbursement process in general

2.7: Who will be able to access my data while this research is being conducted??

The information collected during the study will be kept confidential and secure. Appropriate measures, such as encryption and secure storage of data, will be taken to protect your privacy. Your information will only be used for the purposes of this study and will not be shared with third parties without your explicit consent, except as required by law.

2.8: Who can I speak with if I have any other inquiries?

You can contact me, Purity Monje Mwalozi, via e-mail(purity.mwalozi@strathmore.edu), or by phone (+254711589113). You can also contact mysupervisor, Dr. Kennedy Ronoh, at the Strathmore School of Computing and Engineering Sciences, Nairobi, or via e-mail (kronoh@strathmore.edu)

Please get in touch with an impartial person if you have any questions concerning this study: The Secretary–Strathmore University Institutional Ethics Review Board, P. O. BOX 59857,00200, Nairobi, email ethicsreview@strathmore.edu Tel number: +254 703 034 375

SECTION 3: CONSENT

I,_____, willingly consent to take part in this research investigation.

I understand that even though I have given my approval to participate at this time, I have the option to stop at any time without suffering any repercussions.

The study's objectives and methodology have been thoroughly explained to me in written form, and I have had the chance to clarify any uncertainties or questions I may have had.

I am aware that participation calls for the sharing of personal data of the SACCO members.

I am aware that the data I contribute to this study will be kept private and confidential.

Research participant's signature **Date:**
(Representing Okoa Management)

---------/----------/----------

DD/MM/YEAR

Research participant's Name: **Date:**

---------/----------/----------

DD/MM/YEAR

I,_____(Name of person taking consent) hereby attest that I have adhered to the SOP for this study, have given the study information to the participant named above, and that the participant has understood the nature and aim of the study and has given his/her consent to participation. S/he had the chance to ask questions, and those questions were satisfactorily answered.

Researcher's signature: ***Date:***

_____ _____/_____/_____

DD / MM / YEAR

Researcher's Name

_____

Appendix E: Privacy and Confidentiality Agreement

Background

This privacy and confidentiality agreement ("the Agreement") is created and effective as from March 2023 (the "Commencement Date").

The Agreement is between:

a. Purity Mwalozi (1st Party) - Receiver
b. Okoa Management Limited (2nd) - Discloser

Collectively referred to as "Parties";

The parties agree to the following terms with the intention to be legally bound:

1. Party 1 intends to receive confidential information from party 2.
2. Regarding the release, disclosure and use of Confidential Information, the Parties have agreed to abide by this Agreement.
3. The purpose of this Agreement relates to the loan prediction and recommendation model and the Parties intention to keep this information confidential (the "Transaction").
4. Each party, their respective affiliates and their respective directors, officers, agents, or advisors (collectively "Representatives") may provide or grant access to certain confidential and proprietary information.
5. In consideration for the disclosure of the Confidential Information the Receiver agrees to the following:

Confidential Information

6. "Confidential Information" relates to all Confidential Information relating to the Transaction which the Discloser or its representatives directly or indirectly discloses or makes available to the Recipient or its Representatives after the date of this Agreement. This includes:
    a. The Discloser's business, assets, affairs, clients, customers, suppliers, and plans (intentions or market prospects). and
    b. The Discloser's procedures, methods, knowledge, expertise, technical data, designs, trade secrets, or software;

c. any knowledge, conclusions, information, or research based on the confidential Information;
d. any other information that is identified as being confidential or proprietary in nature
e. 'Customer Information' includes the names of the organizations or people to whom the disclosing party renders its services, including their affiliates and representatives, as well as any associated information, such as leads, contact lists, sales plans and notes, shared and learned sales information like pricing sheets, projects or plans, agreements, or such other data.

Permitted Disclosure

7. The Recipient may make the private data available to its Representatives on the grounds that it:
    a. before disclosing the Confidential Information, notifies those Representatives of its sensitivity;
    b. takes all reasonable steps to procure that those Representatives comply with the confidentiality obligations in clause 8.

Obligation

8. The Receiving party and their representatives agree to:
    a. maintain the secrecy and confidentiality of the Confidential Information;
    b. not use or otherwise make use of the Confidential Information for any purpose other than the Transaction;
    c. not unless as expressly permitted by, and in accordance with, this Agreement, directly or indirectly divulge or make available any Confidential Information to anyone, in whole or in part.
9. To protect the Confidential Information from unauthorized access or use, the Receiver must install and maintain effective security measures, including any reasonable security measures the Discloser may occasionally propose.
10. The Receiver shall take reasonable measures to enforce the compliance with the requirements of this Agreement by its Representatives.
11. Any violation of this Agreement by one of the Receiving party's Representatives is the Receiving party's responsibility.

Destruction of Confidential Information

12. Upon the Discloser's reasonable written request, the Recipient shall:
    a. obliterate and destroy any records and materials based on, containing, or reflecting the Confidential Information; [and]
    b. [to the extent technically possible,] delete any electronic copies of Confidential Information from its computer systems, communications networks, and other devices [; and OR].
    c. confirm in writing to the Discloser that it has met with this clause's requirements. [Clause 12.]

Disclaimer

13. Either party, at its sole discretion, may:
    a. Disapprove of or reject any suggestions made regarding the Transaction by the opposite party or its Representatives;
    b. Cancel talks and negotiations with the opposite party or its representatives whenever, for any reason, or for no reason at all;
    c. At any time, without giving the other party prior notice, modify the processes pertaining to the consideration of the Transaction.

Third Parties

14. No one other than a party to this Agreement, its successors, and allowed assignees shall have any right to enforce any of its terms, unless otherwise specifically specified elsewhere in this Agreement.

Termination

15. This Agreement shall expire on the earlier of:
    a. Both parties' written consent to terminate this Agreement; or
    b. Completion of the Transaction;

Governing Law and Jurisdiction:

16. Kenyan law shall govern and be followed in the interpretation of this Agreement and any disagreement or claim arising out of or related to it, its subject matter, or its formation (including extracontractual disputes or claims). Each party thus irrevocably consents to the

exclusive jurisdiction of the Kenyan courts to resolve any controversy or claim arising out of or related to this Agreement or the subject matter or formation thereof. (Including non-contractual disputes or claims).

Disputes

17. Before starting any legal proceedings or starting any other alternative dispute resolution procedure in connection with a dispute that arises under or in connection with this agreement (a "Dispute"), a party must first send the other party written notice ("Dispute Notice") of the dispute, describing the dispute and requesting that it be resolved under the dispute resolution procedure. 17. Any disagreements resulting from this Agreement must be settled by:
    a. Starting legal action in Kenyan courts.
    b. Mediation. Arbitration will be used to settle disputes if the parties are unable to reach one through mediation.
    c. Arbitration in accordance to the workings of the Chartered Institute of Arbitrators

IN WITNESS WHEREOF, the parties have executed this Agreement as of the Commencement Date.

_____	_____
Discloser Name	**Discloser Signature**
_____	_____
Receiver Name	**Receiver Signature**