

THE COOPERATIVE UNIVERSITY OF KENYA(CUK)

SCHOOL OF COMPUTING AND MATHEMATICS

FINAL PROJECT REPORT.

Machine Learning Based Loan Prediction.

BY:

Mbaya loyd koome

BSITC01/0063/2020.

# SUPERVISOR: Nicholus Nyapete

Project proposal submitted in partial fulfilment of the requirements for the award of the Bachelors of Science in Statistics and Information technology.

©DECEMBER, 2023

# DECLARATION.

**I hereby declare that this report on my project is my original work and has not been presented to any other organization or institution.**

**Name Signature Date**

**Mbaya Loyd Koome ...................... ..........................**

**This report has been submitted for Bachelor in Statistics and Information Technology with the approval Nicholus Nyapete of Cooperative University, main campus.**

**Name Signature Date**

**Nicholus Nyapete ...................... ..........................**

# Abstract:

In our current banking system, banks have many products to sell but main source of income of banks is its credit line. They earn from interest of those loans which they credit. Bank's profit or loss depends to a large extent on loans example whether the customers are paying back the loan or defaulting. By predicting loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research has shown that there are so many methods to study the problem of controlling loan defaults. As per the research the right predictions are very important for maximization of profits, it is essential to study the nature of different methods and the comparison. The very important approach in predictive analytics is used in studying the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle platform for studying and prediction. Logistic Regression models will be performed and different measures of performances will be computed. The models will compare on the basis of the performance measures such as sensitivity and specificity. The final results will show that the model produce different results. Model is imaginably better because it includes variables such as personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc. rather than checking account information showing wealth of a customer that should be taken into account to calculate the probability of default on loan correctly. However, by using a logistic regression approach to target right customers for granting, loan can be easily detected by evaluating their likelihood of default on a loan. This model concludes that banks shouldn’t only target rich customers for granting loan but also should assess other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.

**Table of Contents**

[SUPERVISOR: Nicholus Nyapete 1](#_Toc164034767)

[DECLARATION. 1](#_Toc164034768)

[Abstract: 2](#_Toc164034769)

[LIST OF ABBREVIATIONS. 6](#_Toc164034770)

[CHAPTER 1: INTRODUCTION 7](#_Toc164034771)

[1.1Background of study 7](#_Toc164034772)

[1.2 Problem Statement 8](#_Toc164034773)

[1.3 OBJECTIVES 9](#_Toc164034774)

[1.3.1 General Objectives. 9](#_Toc164034775)

[1.3.2 Specific objectives. 9](#_Toc164034776)

[1.4 SIGNIFICANCE OF THE STUDY 10](#_Toc164034777)

[1.5 SCOPE OF THE STUDY 10](#_Toc164034778)

[1.6 ASSUMPTIONS. 10](#_Toc164034779)

[1.7 LIMITATIONS AND DELIMITATIONS. 10](#_Toc164034780)

[1.8 SOLUTIONS TO THE LIMITATIONS 11](#_Toc164034781)

[1.9 Definition of terms 11](#_Toc164034782)

[CHAPTER TWO: LITERATURE REVIEW. 12](#_Toc164034783)

[2.1 Introduction 12](#_Toc164034784)

[2.2 TRENDS IN LOANING 12](#_Toc164034785)

[2.3 Traditional Loan Default Assessment 13](#_Toc164034786)

[2.3.1 Linear regression 13](#_Toc164034787)

[2.3.2 Discriminant Analysis 13](#_Toc164034788)

[2.3.3 Probit Analysis and Logistic Regression 13](#_Toc164034789)

[2.3.4 Judgment-Based Models 13](#_Toc164034790)

[2.3.5 Machine Learning Approaches in Credit Scoring 14](#_Toc164034791)

[2.3.6 Decision Trees supervised learning 14](#_Toc164034792)

[2.3.7 Random Forest 14](#_Toc164034793)

[2.3.8 Logistic Regression 14](#_Toc164034794)

[2.3.9 Neural Network 15](#_Toc164034795)

[2.4 Naïve Bayes 15](#_Toc164034796)

[2.5 Related systems. 16](#_Toc164034797)

[2.6 Limitations. 16](#_Toc164034798)

[3: METHODOLOGY 17](#_Toc164034799)

[3.1 Introduction 17](#_Toc164034800)

[3.2 SYSTEM DESIGN 17](#_Toc164034801)

[3.2.1 DEVELOPMENT METHODOLOGY OVERVIEW. 17](#_Toc164034802)

[3.2.2 Description. 17](#_Toc164034803)

[3.2.3 Reasons for choosing this methodology. 17](#_Toc164034804)

[3.3 DESIGN PROCEDURES. 18](#_Toc164034805)

[3.4 Design Requirements on hardware and Software Requirements for model 20](#_Toc164034806)

[3.5 DATA COLLECTION AND ANALYSIS 21](#_Toc164034807)

[3.5.1 Primary data collection. 21](#_Toc164034808)

[3.5.2 Secondary data collection. 21](#_Toc164034809)

[3.6 Data analysis 22](#_Toc164034810)

[3.6.1 Data Preprocessing 22](#_Toc164034811)

[3.6.2 Data Cleaning 22](#_Toc164034812)

[3.6.3 Data Reduction 22](#_Toc164034813)

[3.6.4 Feature Engineering 22](#_Toc164034814)

[3.6.5 Exploratory Data Analysis 23](#_Toc164034815)

[3.6.6 Univariate Analysis 23](#_Toc164034816)

[3.6.7 Bivariate Analysis 23](#_Toc164034817)

[3.6.8 Converting Categorical Variables 23](#_Toc164034818)

[3.6.9 Standard Scaler 23](#_Toc164034819)

[3.7 Handling Outliers 23](#_Toc164034820)

[CHAPTER FOUR: SYSTEM ANALYSIS AND REQUIREMENT MODELLING. 24](#_Toc164034821)

[4.1 INTRODUCTION. 24](#_Toc164034822)

[4.1.1 Detailed analysis of current system 24](#_Toc164034823)

[4.1.2 Data Flow Diagram of the existing system. 26](#_Toc164034824)

[4.1.3 System Flow Chart. 27](#_Toc164034825)

[4.1.4 UML diagrams. 27](#_Toc164034826)

[4.1.5 Structural properties 27](#_Toc164034827)

[4.2 SYSTEM REQUIREMENTS. 31](#_Toc164034828)

[4.2.1 Functional requirements. 31](#_Toc164034829)

[4.2.2 Non-functional requirements 31](#_Toc164034830)

[CHAPTER FIVE: SYSTEM DESIGN. 32](#_Toc164034831)

[5.1 Introduction. 32](#_Toc164034832)

[5.2 Architectural design. 32](#_Toc164034833)

[5.2.1 Overview of the system. 33](#_Toc164034834)

[5.2.3 Behavioral properties (UML) 33](#_Toc164034835)

[Activity diagram 33](#_Toc164034836)

[5.2.4 Data collection and pre-processing. 37](#_Toc164034837)

[5.2.5 Model training and evaluation. 37](#_Toc164034838)

[5.2.6 Prediction and deployment. 37](#_Toc164034839)

[5.2.7 DATABASE DESIGN 37](#_Toc164034840)

[5.2.8 User interface design. 38](#_Toc164034841)

[CHAPTER SIX: IMPLEMENTATION AND TESTING. 40](#_Toc164034842)

[6.1 Development Environment. 40](#_Toc164034843)

[6.2 System Components. 40](#_Toc164034844)

[6.2.1 Data 40](#_Toc164034845)

[6.2.2 Algorithms. 40](#_Toc164034846)

[6.2.3 Models. 41](#_Toc164034847)

[6.2.4 Packages 41](#_Toc164034848)

[6.2.5 System Development components 41](#_Toc164034849)

[6.3 Test Plan 42](#_Toc164034850)

[6.4 System Testing 42](#_Toc164034851)

[CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS. 44](#_Toc164034852)

[7.1 Achievements and lessons learnt. (Technical lessons and achievements). 44](#_Toc164034853)

[7.1.2 Achievements 44](#_Toc164034854)

[7.1.3 Lessons Learnt. 44](#_Toc164034855)

[7.2 Recommendations. 44](#_Toc164034856)

[7.3 Conclusions 45](#_Toc164034857)

[REFERENCES. 45](#_Toc164034858)

[Appendices 46](#_Toc164034859)

# LIST OF ABBREVIATIONS.

1. CBK – Central Bank of Kenya
2. CPU - Central Processing Unit
3. CRB - Credit Reference Bureau
4. CRISP-DM - CRoss Industry Standard Process for Data Mining
5. IRB - Internal Ratings-based Approach
6. KBA - Kenya Bankers Association's
7. RAM – Random Access Memory
8. SBI - State of the Banking Industry

# CHAPTER 1: INTRODUCTION

## 1.1Background of study

Kenya mobile banking has rapidly grown in terms of telecommunications and infrastructure development powered by the global decline in smartphone and cellphone in general prices. It has been harnessed by companies to provide value-added services such as digital credit in minutes through some online platforms and offline too such as tala apps and many more. Mobile cellular subscriptions (per 100 people) in World lately reported at 108 per 100 people in 2022 which is some months ago (World Development Indicators, 2022). According to the World Bank collection of development indicators compiled from officially recognized sources says that credit providers have traditionally interacted between agents and clients which could not achieve the highest profits margins targeted and the investors full potential utilization of available resources due to bulk of work on workers lending services.

So, risk assessment based on previous financial history and loans delivered into a bank account is a instant loan decisions model and are automated based on a set of rules applied on available data, also managed remotely (CGAP, 2022). It has evolved to incorporate a number of different business models. The first model is a bank and mobile network operator partnership such as Mshwari which is jointly connected to M-pesa by NCBA Bank and Safaricom. The second model is a non-bank lender such as Kopa Cash by Jumo, hustler fund and Airtel Kenya which has been slowly taking part in the new industry bringing higher competitions. The third model is bank utilizing mobile network operator channels such as Mco-op Cash by Co-operative bank that uses USSD and KCB mobile banking. Therefore, 4th model is non-bank internet mobile applications which is involved with non-bank lenders disbursing loans through smartphone mobile application like Branch and Tala.

So, in today’s banking structure financial organizations are using mobile lending platforms predictive analytics like transaction history, call logs, text messages, contact lists, age, education level, income, personal properties and other properties to come up to a credit worthiness score limit including status on other banks liquidity of personal assurance (E Feyen · 2021). When analyzing first-time borrowers, alternative digital data is especially significant but repayment based credit history becomes more important for subsequent loan applicants to determine their credibility. This research project aims to evaluate the application of machine learning technique to improve the predicted loan defaults in some readymade models to improve on some few challenges.

## 1.2 Problem Statement

Banks, Housing Finance Companies and some NBFC deal in various types of loans like housing loan, personal loan, business loan etc in all over the part of country. These companies have existence in Rural, Semi-Urban and Urban areas. After applying loan by customer these companies validate the eligibility of customers to get the loan or not. Financial mobile and over counter lending institutions use credit scoring models to evaluate potential loan default risks (L Gambacorta · 2019). These models are fixed and do not easily evolve with changing customer behavior to predict loan defaults more accurately and will generate a score that translates the likelihood of defaulting on a loan lending decision easier indicating the maximum qualification for each individual. However, there exist several traditional models with several gaps. These default prediction models have been long used by lenders and most institutions to access creditworthiness of borrowers followed by predicting likelihood of loan defaults (GT Kisutsa · 2021). The models typically rely on historical data with statistical techniques to determine probability of borrower defaulting a loan. The main limitation is reliance on historical data where the models are built using historical loan performance data and may not sometimes represent current borrower behavior (MF Vidal · 2019). Another gap in these models is their inability to capture complex relationships and nonlinearities in borrower characteristics for example a borrower with high income may default if they have excessive debt obligation leading to inaccurate predications (A Sadhwani · 2021). These models also tend to assume that borrowers’ characteristics are static and don’t change over time (JY Campbell). However, borrower’s financial situation can change accordingly example a person who had a low possibility of defaulting can experience life misfortunes like divorce, job loss property loss and so on which could increase their of risk default which is not the case in existing models. This project provides a solution to automate this process by employing advanced machine learning algorithms and alternative data source to improve accuracy and capture complex relationships. By considering broader range of variable as well as using sophisticated algorithms, advanced default prediction models can provide more accurate and dynamic predictions.

## 1.3 OBJECTIVES

### 1.3.1 General Objectives.

The goal of this project is modeling a prediction application system that will be used for predicting a loan approval of individuals according to the recorded details provided and boost default prediction clarity.

.

### 1.3.2 Specific objectives.

i). Examining literature of already existing models in loan default’s predictions.

1. Designing, developing and validating ml model that forecasts loans defaults accurately.
2. Conduct training and testing the machine learning models to predict loan defaults. iv. Deploy and evaluate the ability of machine learning model predicting loan defaulters.

1.4 SIGNIFICANCE OF THE STUDY**.**

Credit risk assessment is always crucial to the success of lending institutions because customer credit risk affects profitability directly. When a customer is lend some amount and fails to pay back, the organization incur losses. Traditional methods, which most African banks use take a lot of time and are inefficient. This system will make work easier for the managers and workers at large as they will save on time and resources because the loan default process would be automated. The reason as to why I will be developing this machine learning model is purposely to reduce chances of a credit risk in banks and ensure every customer is fairly looked into according to their qualities, a more dynamic and adaptability to changing client data. These techniques will definitely provide higher accuracy in predicting loan default.

## 1.5 SCOPE OF THE STUDY

This study is more focused on loan application approval and how to determine it in a fast and efficient way to improve performance reducing any chances of loss. It will reduce nonperforming loans and loan defaulters. Perceptions which are negative towards borrowers in certain locations may reduce because more details will be used checking the loan default of individuals. The whole process of prediction will always be done privately so no stakeholders can interfere with the loan processing.

## 1.6 ASSUMPTIONS.

* The prediction model will be very accurate and highly effective.
* The interfaces to be used will to be user friendly.
* Users of the system are expected to have basic knowledge regarding loan prediction approval

## 1.7 LIMITATIONS AND DELIMITATIONS.

* Choosing a machine learning model to use is challenging. Someone has to select a model with a highest accuracy.
* If someone one lacks technical knowledge on python programming and machine learning it’s challenging to start learning it from scratch.
* It’s difficult to choose the right dataset to building a perfect model. This study is dependent on datasets which could be limiting.
* Some details were missing in the dataset hence prediction was done based only on the available details.
* Model emphasizes distinct weights for every aspect but in reality, loans can occasionally be authorized just based on one powerful factor so this is Impossible using our approach. (S Ali · 2023)

## 1.8 SOLUTIONS TO THE LIMITATIONS

* I researched for machine learning models which will be highly recommended, approved and with highest accuracy level.
* It’s will be a major key to have programming skills before choosing to pursue such kind of a project.
* I will select a dataset offering close to required attributes to be used in the model building.
* The model will be based on powerful factors in which prediction will rely on the powerful factor when need be.

## 1.9 Definition of terms

**Loans and Loan Default.**

Loans can be defined as the amount of money one is given at some interests from a certain financial lending institution. They are offered for a certain time period in between the issuing and repayment duration. Loans are described as a part of the major components of the total assets of every bank. (Mabvure et.al, 2012). However, it’s impossible for lenders to ask for payment before the expiry of the period unless there have been some delayed payments.

Loan default can be defined as where loan repayment has not been met to by the borrower as terms were agreed on. It happens if a borrower will not or cannot repay their loan so the lenders does not expect the payment to be received (Maina and Kalui, 2014). Technical default applies mainly to banks and is applied where a negative covenant is not followed as agreed upon.

**Non-Performing Loans.**

A non-performing loan is a scenario that is in default or close to being in default. Many loans become non-performing after being in default for 90 days in most of lending loan issuers organizations but this can depend on the contract terms. However, IMF defines NPLs as the sum of borrowed money upon which the debtor has not made his scheduled payments for at least 90 days. When a loan is nonperforming the odds to be repaid in full are considered as substantially lower. (AC Mwangi · 2014)

# CHAPTER TWO: LITERATURE REVIEW.

## 2.1 Introduction

In this chapter we review of the trends in lending loan, loan aproval, existing systems for loan aproval and some related systems. It explains in details the machine learning algorithms that are increasingly being used to evaluate credit risk and the relevance of the literature to the problem statement. Traditional Credit Risk Assessment of data used in traditional credit scoring is historical data which includes bank transactional data such as past credit, records of late payment, credit bureau checks, commercial data such as financial statements and length of credit history. (world bank 2022)

The number of people looking for loan increases every month and year. In a case of covid-19 pandemic a lot of businesses, investor’s, organizations and even individuals due to financial instabilities and the crisis needed to look for a sympathy loan as the simplest way to keep going. Before loan is approved several attributes of a clients are first determined to check their credibility hence this model will help to solve loan defaults prediction in real time.

## 2.2 TRENDS IN LOANING

In this banking sectors there has been a lot of challenges affecting lending since the beginning of covid 19 pandemic. To start with there has been delayed rates of loan repayment and delayed lending due to the slowdown of the economy and country debts. There has been reduction in earnings due to low revenue collected from lending institutions that has been caused by defaults as customers struggle with the effects of the pandemic to settle their balances. loan defaults number has highly increased at a very high rate and the banks have been seeking for a way to curb this tragedy to minimize such kind of loses.

In this case some measures may be put in place to curb this kind of problem and most likely determine if a person can default a loan or not as explained below:

* According to (Thomas, 2000) education enhances the ability of a borrower to repay and it’s said that educated borrowers are more likely to repay loans as they have high levels of income and occupy higher positions.
* Gender is a fair discriminatory base on the statistical default rates of men comparison to women. This means men can easily default a loan in comparison to women as women default less frequently on loans possibly because they are more risk adverse. (Coval and Shumway (2000)
* Borrowers married tend to be more responsible and mature. They are likely to repay loans as compared to single borrowers. (Dean and Kleimeir, 2007)

One of the major problems in lending institution is poor management which leads to inaccurate loan allocations and losses at last. Most of the managers don’t take their time to follow up loan underwriting, monitoring and controlling the loaning process as they should.

These factors can be used in finding causes of loan inaccurate loan awards and making some measures that can reduce loan defaulting due to large amounts allocation to the unqualified. With access to data with these factors it’s easier to build a system that can predict whether a person is likely to default for a loan or not and the amount an individual qualifies for.

## 2.3 Traditional Loan Default Assessment

Historically data such as past credit’s, late payment record’s, credit bureau check, credit history and bank statements are used when checking credit score (World Bank, 2019).

### 2.3.1 Linear regression

Linear regression is used to determine the linear relationship between explanatory target variables. However, assumption in a general linear regression issue is that there is a dependent or response variable (Yi) that is impacted by independent variables. (Xi1), (Xi2) ...., (Xi)n.

Regression model can express this relation: Yi= (Xi1) + 2(Xi2) +.... + n(Xin) where 1, 2...n

These are fixed regression parameters and a random error also refereed to as noise parameter. Before attempting construction of a linear model, it's critical to assess whether there is a link between variables of interest.

### 2.3.2 Discriminant Analysis

In this analysis a credit scoring approach developed by (Sir Ronald Fisher in 1936) distinguishing between two groups. The most basic form is two-category label such as default versus nondefault. Linear discriminant analysis was first statistical tool to be used systematically explaining which firms went bankrupt on accounting track ratios and other financial indicators of default prediction.

### 2.3.3 Probit Analysis and Logistic Regression

Inverse standard normal distribution of the probability is modelled as a linear combination of characteristics in the probability unit (probit) model. The log of odds is used by the logistic unit (logit) function. Log of these label's the odd ratio described as linear combination attributes in logit model. This formula below is used to predict log odds ratios: logit(p) = β0 + β1 × X1 + β2 × X2 + . . . + βn × Xn (5)

### 2.3.4 Judgment-Based Models

Multiple strategies are used to create judgment-based models. The analytic hierarchy process which is a structured approach for organizing and analyzing complex decisions. The decision makers break down their choice problem into a hierarchy of easier-to understand subproblem each of which may be studied individually. Human judgments are not just the underlying data used to complete the evaluations in the AHP according to Bana e Costa, Barroso, and Soares (2000).

### 2.3.5 Machine Learning Approaches in Credit Scoring

Machine learning is based on the development of algorithms that can take in data and apply statistical analysis to anticipate a certain required output as well as update outcomes when new data becomes available. There are three different forms of machine learning. Purpose of supervised learning is presenting computer with well labeled data and to estimate mapping function to the point which you can predict the output variables of (y) for the data when you have new input data variable (x). Clustering and association problems are examples of unsupervised learning in which a computer is supplied with unlabeled, uncategorized data and the system's algorithms act on the data without prior training. Reinforcement learning is a type of learning algorithm that learns by interacting with its surroundings and the data allocated to it. When the agent performs successfully then he is rewarded and when he performs wrong, he is penalized. Therefore, by maximizing its reward and reducing its penalty agent learns without the need for human intervention.

### 2.3.6 Decision Trees supervised learning

This approach will be used to tackle classification and regression problems is decision trees. To tackle the prediction problem by decision trees we use tree representation in which the external node and leaf node of a tree represent attribute and class labels respectively. Categorical variable decision tree model contains some categorical target variables which are separated in categories with a yes or no.

A continuous variable decision tree has a continuous target variable like an individual's income which can be forecast-ed using information such as occupation, age and other continuous factors to be considered. Random Forest model is the machine learning algorithm that is supervised. Bagging approach is used to train a forest which is a collection of decision trees.

### 2.3.7 Random Forest

It generates a large number of decision trees and combines them to get a harmonious and accurate classification. Random Forest algorithm has advantage of being used for both classification and regression analysis.

### 2.3.8 Logistic Regression

A classification procedure called logistic regression used to describe data and explain the relationship between one or further independent variables and dependent double variable. double logistic regression has major hypotheticals

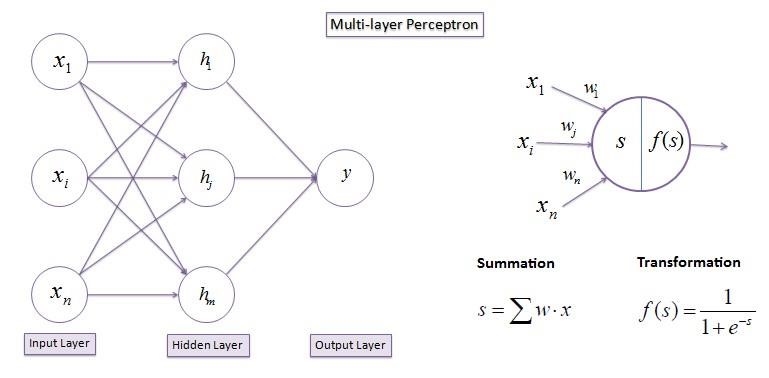
1. Dependent variables must either be double or dichotomous (e.g., yes or no).

ii). Data should have no outliers which can be determined by converting nonstop predictors to standardized scores.

iii). The predictors shouldn't have a lot of strong correlations(multicollinearity). The independent variables must be unconnected to one another. A correlation matrix among the predictors can be used to examine this.

### 2.3.9 Neural Network

Artificial Intelligence includes neural networks which is a type of learning model affected by the activity of organic neurons. The neural network is made up of nodes that reuse the data handed to them and shoot the results to other bumps. Each knot’s affair is known as the activation or knot value. Weights are assigned to the nodes which can change to help the network learn. The magnitude of an input's influence on an affair is represented by these weights. A direct, ramp, move sigmoid, hyperbolic or Gaussian activation function is used to perform the net direct computation because it can recognize nonlinear regions and the Multilayer Perceptron Model is utilized to detect fraud.



*Figure 1 NEURAL NETWORK*

It Propagates input by adding all the weighted inputs and then computing outputs using sigmoid threshold.

### 2.4 Naïve Bayes

This theorem is used to build a collection of classification algorithms known as Naive Bayes classifiers. They are implemented in a variety of fields for effects like prediction and anomaly discovery. Each knot represents a variable. The bends reflect the relationship between them and they're represented by a graph with bumps and directed linkages between them. Although their perhaps no information the Nave Bayes is accessible in fraud discovery to set of variables that cause the frauds and can be reckoned using the same theorem.

## 2.5 Related systems.

**CRB (Credit Reference Bureau).**

This is a company that collects credibility information on borrowers from various organizations and provides it to lenders for various purposes. The information collected by the CRB is important for banks, microfinance institutions and SACCOs (savings and credit cooperatives) as it’s used to assess a borrower's ability to repay loans. When a loan applicant contacts a lender, the lender sends the customer's name and identification number to her CRB. CRB then determines if the applicant is listed in the database for any reason of loan defaults or not. The credit report from the loan applicant is then forwarded to the bank. If the applicant is negatively listed on CRB the loan application will be declined. Information may also be made available to customers if they desire to see.

## 2.6 Limitations.

Banks rely on the information provided by the CRB to determine loan eligibility of a person. On the other hand, CRB relies on information provided by financial institutions about a borrower. Studies have concluded that CRB is having a negative impact on loans performance of Microfinance Institutions in Kenya. This means that an increase in credit ratings will reduce loan performance and increase the proportion of non-performing loans in MFIs. By adding the cost of borrowing to borrowers, credit bureaus can create roadblocks that can cause more people to default because they may not be able to afford the money to pay the credit bureaus.

# 3: METHODOLOGY

## 3.1 Introduction

This chapter gives overview of the methods of development of the project, method to use to obtain requirements, techniques of data collection and system design. It also enlists the hardware and software requirements needed for this project.

## 3.2 SYSTEM DESIGN

### 3.2.1 DEVELOPMENT METHODOLOGY OVERVIEW.

For this project I used CRISP- DM waterfall methodology. The Cross Industry Standard Process for Data Mining (CRISP-DM) is the underlying process model for data science processes.

### 3.2.2 Description.

CRISP-DM provides approach that is structured to planning data mining project. Its sequence of events and idealized with tasks that will be performed in different order’s. The order of carrying out this task depends on process model you will select for your project. This process model to be used can be either waterfall or agile depending on how we will carry on with your processes. I used the waterfall method which is a step-by-step methodology that follows a chronological process and work based on fixed data, requirements and outcomes. In this approach every execution is generally self-contained unless its specific integration was required. I followed the steps of this methodology from the first to the last systematically.

### 3.2.3 Reasons for choosing this methodology.

CRISP-DM is a methodology used for data mining projects. Using waterfall implementation for this methodology enabled me to use a step-by-step procedure that was followed throughout the project. Waterfall focuses most on a well-defined sequence of steps in which one is able to plan their time for the project duration as they already know what has to be done and when. With waterfall method, I was able to commit to the end product without getting side-tracked. The reason I used this methodology was to follow through all these steps sequentially so as to be able to keep time while working on every step.

## 3.3 DESIGN PROCEDURES.

**Business understanding** - This helped understand the background of the research, describe the problem and how the project that was developed arrived at its goals. With this I was also be able to understand the relevance of the data in this particular business model.

**Steps.**

1. Setting business goals - First, I needed to have a thorough understanding of what the customer really wants to achieve from a business perspective. (CRISP-DM Guidance), Defining Business Success Criteria.
2. Assess the situation. - Determine resource availability and project requirements, assess risks and contingencies, and perform cost-benefit analysis.
3. Set data mining goals. -In addition to defining your business goals, defining what success looks like from a technical data mining perspective was needed.
4. Create a project plan - Select technologies and tools and define a detailed plan for each project phase.

**Data understanding -** It involved collection of data, understanding its attributes and the requirements it needs. This would be able to describe how accurate the output would be. By this we were able to know what to expect and what to achieve with the data.

**Steps.**

1. Collect initial data: Collect the required data and load it into the analysis tool (if needed). Data Description: Examined the data and document surface properties such as data format, number of records.
2. Explore data: Dug deeper into data. Queried and visualized the data to identify relationships between data.
3. Data Quality Check: Determined how clean or dirty the data was. Checked for document quality issues.
4. Data preparation – Here I checked for missing values, outliers, incorrect data. Then I performed data pre-processing on the raw data and made it suitable for a machine learning model through methods such as data cleaning.
5. Data Selection: Decide which datasets to use and document reasons for inclusion/exclusion. Data cleaning: This is often the most tedious task. Without them, you can fall victim to garbage in, garbage-out. A common approach for this task is to fix, assign, or remove erroneous values.
6. Data building: Derived new attributes that are useful. For example, derive a person's body mass index from height and weight fields. In this case I obtained the total income of the borrower by adding up their income with that of their co-applicant.
7. Data Consolidation: Combined data from multiple sources to create new datasets. Format data: Reformatted the data as needed.

**Modelling** - Determined the model to use and testing it.

**Steps.**

1. Choice of modelling method -Decide which algorithm you want to try (e.g., regression, neural networks). I decided to use Random Forest Classifier in this case.
2. Generate a test design -Depending on your modelling approach, you may need to split your data into training, test, and validation sets. I had my data already separated into training and test sets.
3. Build model - it is usually maybe just a few lines of code like "reg = Linear Regression (). fit (X, y)". I used stratified k Folds for my Random Forest Classifier.
4. Pricing model -Multiple models commonly compete with each other, requiring data scientists to interpret model results based on domain knowledge, pre-defined success criteria, and test designs.

**Evaluation** - Results from the model with the best accuracy, this was determined by going through the results.

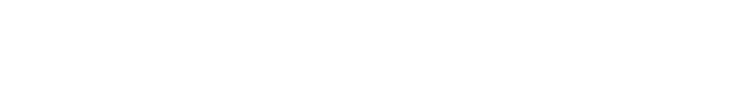
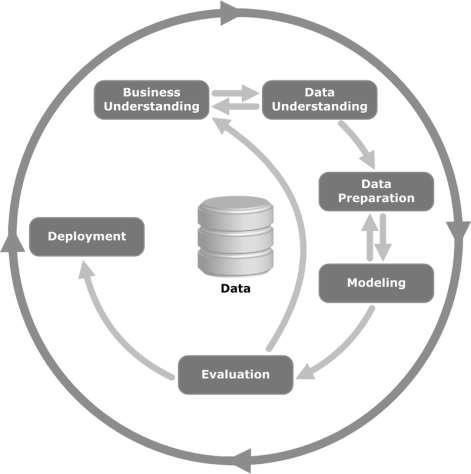
**Steps.**

1. Evaluate the results – Check whether the model meet a criterion for the success of the business and if it qualifies to be approved for the organization.
2. Verification process -A review of completed work was done to check if I missed something and whether I correctly performed all the steps. A summary of the results was done and some corrections made where necessary.
3. Decide your next step –I chose to continue with deployment. This will depend on the previous tasks that have been carried out.

**Deployment** - Implementation of the model.

**Steps**

1. Plan your deployment – This was to develop and document a plan for deploying the model.
2. Plan monitoring and maintenance -Develop a rigorous monitoring and maintenance plan to avoid problems during the operational phase of the model.
3. Preparation of final results – The project was documented and included the final presentation of the data mining results.



-



Figure 1

## 3.4 Design Requirements on hardware and Software Requirements for model

This hardware requirements for the project include a laptop with at least 4GB ram running windows or Linux operating system. This project uses Anaconda which is a code editor redefined and optimized for building and debugging modern web and cloud applications with different sub-software’s in it like R and others.

**Python programming**

**Python Libraries**

The machine learning models are implemented using python version 3.7 on a Jupyter notebook with the listed libraries: NumPy, pandas, matplotlib, seaborn and sklearn. To start with the libraries in details:

1. Jupyter notebooks are a web-based interface which one can write, visualize, analyze and execute python code in cells. It is best for exploratory analysis that enable to run individual code cells.
2. Numpy: This is a Python library that may be used to work with mathematical multidimensional arrays such as linear algebra, the Fourier transform and matrices.
3. Pandas is a data manipulation and analysis package written in Python.
4. Matplotlib is a Python package that allows you to create static animated and interactive visualizations with best graphics.
5. Seaborn is a matplotlib-based python data visualization package which has high-level interface for creating visually appealing and instructive statistics visuals.
6. Sklearn is a toolkit that allows you to create machine learning predictions and statistical models including clustering, classification and regressions.

3.5 DATA COLLECTION AND ANALYSIS**.**

### 3.5.1 Primary data collection.

It is data collected first hand directly from the source. This data is regarded to be the best kind

in research. The primary data collection methods can be further divided into qualitative data

collection methods that deal with non-numerical data and quantitative data collection methods that deal with numeric data.

Observation – Working at a financial institution during my attachment gave me an idea for this project. I could see guarantors who had to pay back the loans if the borrowers defaulted. Also,

in a case where guarantor was not in a position to repay the loan, the organization experienced

a loss. To avoid having guarantors settle the defaulter’s loans or to reduce the rate of financial institutions incurring losses I developed a model that could predict the likelihood of the borrower loan award. Hence, the idea for building a loan default prediction model.

### 3.5.2 Secondary data collection.

This method of data collection relies on data that has already been collected by someone else.

It is less expensive and easier to collect as compared to primary data.

Internet – Most of the information for this project, I had gotten from the internet. I used that

information to weigh factors that are needed and come up with a solution that relies on these factors to build my project.

Kaggle is an online platform for data scientists to showcase their work and learn. The site contains datasets that are usually used in data analysis and also in building machine learning

models. For this project I used a dataset from Kaggle. The dataset contained attributes of the borrowers. These attributes were analyzed and used in making predictions.

## 3.6 Data analysis

### 3.6.1 Data Preprocessing

Data preprocessing is a way of converting raw unprocessed data into a comprehensible format that a machine learning model can read and understand. The data is loaded on a Jupyter notebook in Anaconda files page from the file manager and the python libraries numpy, pandas, matplotlib, seaborn and sklearn are imported respectively. The dataset is loaded to the jupyter notebook for preprocessing. The data preprocessing involves data cleaning which involves handling missing values, data transformation that involves normalizing the data and data reduction by using only relevant features and discarding duplicate values of less relevant attributes.

### 3.6.2 Data Cleaning

First step of preprocessing is data cleaning through checking and eliminating where necessary any missing values because they affect the accuracy of the model. This can be achieved by either filling the missing values with a mean, mode function or dropping all missing values.

### 3.6.3 Data Reduction

The next step of the data preprocessing is data reduction used to remove duplicate features example ‘LoanId’ while there is ‘LoanNumber’, ‘DateofBirth’ when the feature ‘Age’ is present.

Features relating to dates excluding ‘Default Date’ are deleted. The multiple values of income are also deleted since they are already aggregated in ‘Income Total’.

### 3.6.4 Feature Engineering

When using this machine learning to create a predictive model this feature engineering is the act of choosing and modifying variables in dataset. The ‘Status’ and ‘Default Date’ variables will be used to create the target variable that is ‘Default’. The ‘Status’ variable cannot be used since it has three unique values current, late and repaid. Late may also not be treated as default since in some records the loan status is late however the default date is null which implies the loan was not defaulted but was only late. The ‘Default Date’ informs the period when a borrower defaulted. Combining both the ‘Status’ feature and ‘Default Date’ feature will enable create the target variable ‘Default’. This can be achieved by filtering the loan status to current and checking the default dates to create a new target variable called ‘Default’ that will have the values 0 if default and 1 if loan isn’t not default. The ‘Status’ and ‘Default Date’ features are removed once the target variable is created.

### 3.6.5 Exploratory Data Analysis

There are two types of independent variables in the data set. Categorical features which include Gender, Education, Marital Status and numerical features which include Income Total and Amount.

### 3.6.6 Univariate Analysis

This refers to a data that consists solely of observations on a single attribute. The basic goal of univariate analysis is to characterize the data and discover patterns within it. Then data is visually shown using graphing. The primary goal of graphs is to convey data, summarize data, enhance verbal descriptions, describe and explore data, facilitate comparisons and stimulate thought about the data. The bar graph is the specific graph that will be used. On the y (vertical) and x (horizontal) axes respectively, the graph is labeled (horizontal axis). The categorical and ordinal features explored include Gender, Education, Marital Status, Employment Status, Employment Duration Current Employer, New Credit Customer

### 3.6.7 Bivariate Analysis

This is analysis of two variables with the goal of identifying the empirical link known as bivariate analysis. The following categorical variables: Gender, Education, Employment Status, Marital Status, while new credit customer will be compared to dependent variable ‘Default’.

### 3.6.8 Converting Categorical Variables

Sklearn requires most inputs to be numeric and the categorical are converted to numerical variables using label encoder. The values ‘New Credit Customer’, ‘Restructured’, ‘Employment Duration Current Employer’ will be converted to numerical values.

### 3.6.9 Standard Scaler

This is used to turn data into a distribution with a mean of 0 and a standard deviation of 1, here we use a standard scaler.

## 3.7 Handling Outliers

Outliers are data points that are far apart from other similar points which could be due to measurement variability or experimental errors. The range and distribution of attribute values are particularly important to machine learning algorithms. Outliers in the data might cause the training process to be misled, resulting in longer training times, fewer accurate models and inferior outcomes. For analyzing the data and detect any outliers, data visualization is employed. There are four methods for dealing with outliers in a dataset. Remove the outlier records entirely to get rid of them. By setting a value range, you can limit the data of outliers. If the data is out of scope for the intended variable, we assign a new value in this case. Using techniques such as log transformation, data can be transformed. The Income Total and Amount variables include some outliers and are skewed. The log transformation is used to normalize the data. The log transformation is used to skewed data in order to approximate the normality because the dataset has a log-normal distribution, the log-transformed data will have a normal or near-normal distribution also reducing skewness.

# CHAPTER FOUR: SYSTEM ANALYSIS AND REQUIREMENT MODELLING.

## 4.1 INTRODUCTION.

This chapter shows the different diagrams showing how system works and the relationships of the attributes to be used. It also explains functional and nonfunctional requirements of the system.

### 4.1.1 Detailed analysis of current system

**Key features and functionalities**

* BorrowerInformation: These includes o Personal details: Age, gender, marital status, education level, employment status.
  + Financial information: Income, assets, debt-to-income ratio, credit history.
* LoanCharacteristics:
  + Loan amount: It’s amount of money requested by the borrower. o Loan purpose: Reason for loan request if it’s for a home purchase, education, vehicle or specific purposes. o Loan term: This duration of the loan (e.g. 30-year mortgage, 5-year personal loan).
  + Interest rate: Annual percentage rate (APR) charged on the loan. o Loan-to-value (LTV) ratio: The ratio of loan amount to appraised value of collateral (if applicable).
* RepaymentHistory:
  + Past loan performance: History of real time payments, delinquencies, defaults, bankruptcies.
  + Credit utilization: Percentage of available credit’s currently being utilized by the borrowers.
* Economic Indicators:
  + Macroeconomic factors which include unemployment rate, inflation rate, GDP growth, interest rate trends, etc. o Then local economic conditions: Housing market trends, employment opportunities, industry stability etc.
* RiskAssessment Models:
  + Credit scoring models: FICO scores, Vantage Score are used to quantify the borrower's creditworthiness.
  + Risk segmentation models: Segmenting borrowers to risk categories based on their likelihood of default.
* MachineLearning Algorithms:
  + Logistic regression: Predicts probability of default based on the input features. o Decision trees: Identifies decision rules classifying borrowers as high or low risk. o Random forest: Ensemble of decision trees for improved prediction accuracy.
  + Gradient boosting: Sequentially builds weak learners to create strong predictive model.
  + Neural networks: Deep learning models capable of learning complex patterns in data.
* ModelInterpretability:
  + Feature importance analysis for identifying the most influential features in predicting loan default. o Explainable AI techniques: This provides insights into how model arrives at its predictions (ex. SHAP values, LIME).
* Scalabilityand Performance:
  + Ability to handle large volumes of data and process loan applications in real-time analyses. o High prediction accuracy and low false positive/negative rates minimizing financial losses and maximize lending opportunities.

**Strengths:**

Data-Driven Decision Making**:** Loan default prediction models uses historical data to make predictions on the likelihood of a borrower defaulting on a loan

Automation and Efficiency: These models automate the loan approval process to large extent, reducing need for manual underwriting and making the process more efficient.

Scalability: Once developed models can be scaled up to handle large volumes of loan applications without additional costs.

Risk Management**:** By identifying potential defaulters, these models help financial institutions manage the risk exposure and make informed decisions about lending practices.

Continuous Learning: loan default prediction models can improve over time, adapting to any changing market conditions and borrower’s behavior.

**Weaknesses**:

Data Quality and Availability: Accuracy of loan default prediction models heavily relies on the quality and availability of data therefore inaccurate data can lead to unreliable predictions.

Assumption of Stationarity: Some models assume the underlying relationships between variables remain stable over time.

Bias and Fairness Concerns: Models trained on historical data can inherit biases present in those data, leading to unfair treatment of certain groups, example the minorities or low-income earners.

Black Box Nature: Some models based on complex algorithms like neural networks can be difficult to interpret in some cases hence this lack of transparency may raise a lot of concerns among regulators, stakeholders and borrowers about how decisions are made.

Overfitting: There's a risk that models may overfit the training data and capturing noise rather than meaningful patterns causing poor generalization of new data and reduced predictive accuracy.

### 4.1.2 Data Flow Diagram of the existing system.

This shows how system works to help in predicting loan default where the train model performs a prediction. The predictions give’s feedback on the loan status prediction provided by the model.



**:**

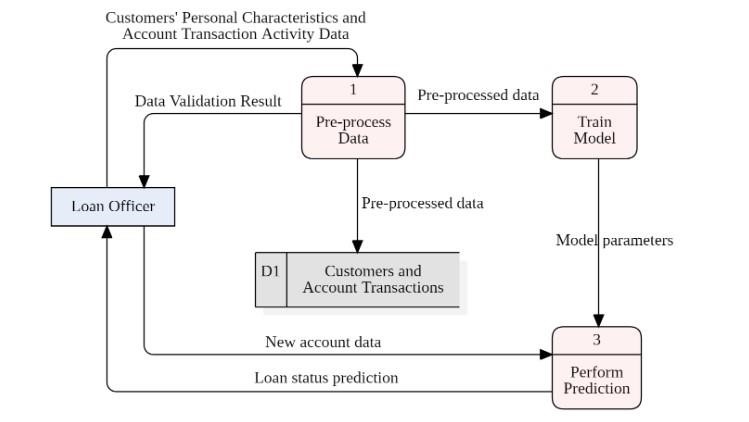


Figure 2 : data flow

### 4.1.3 System Flow Chart.

This shows that the system works to help in predicting loan default. The predictions determine whether customer applying for a loan will default or not. The loan details including the attributes will be entered into the system. The trained models will then predict if the customer is a likely defaulter or not.

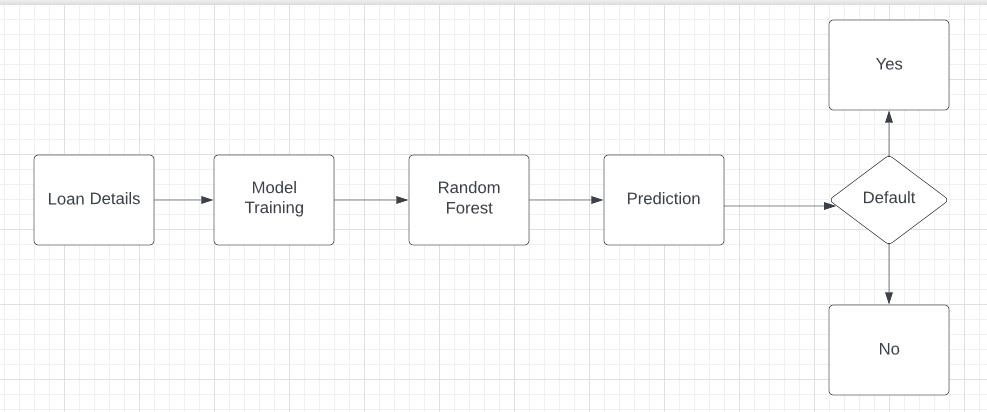


Figure 3

**: System flow chart**

### 4.1.4 UML diagrams.

UML stands for Unified Modelling Language and it’s used to visualize the design of the system.

4.1.5 Structural properties **A Class diagram.**

Class diagram shows classes in the system, their operations, attributes and relationship between each of them. Class diagrams are static display diagrams that interacts but not what happens when they do interact

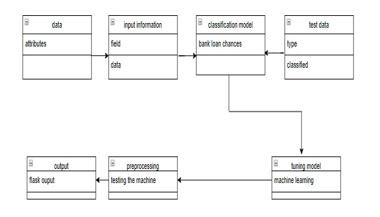


Figure 4

**A package diagram**

A case study of modeling software for library management is presented as an illustration of how to apply the proposed frameworks. Modeling tool features include model transformations, code generation, model validation, dependency matrix, model metrics, model comparison and also model refactoring are presented as enablers for efficient model-driven development.

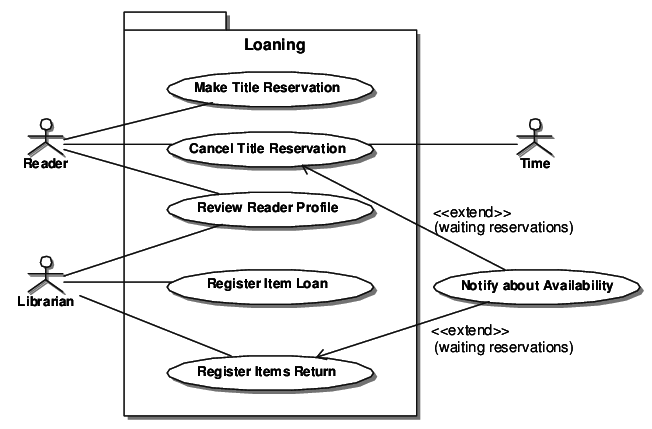


Figure 5

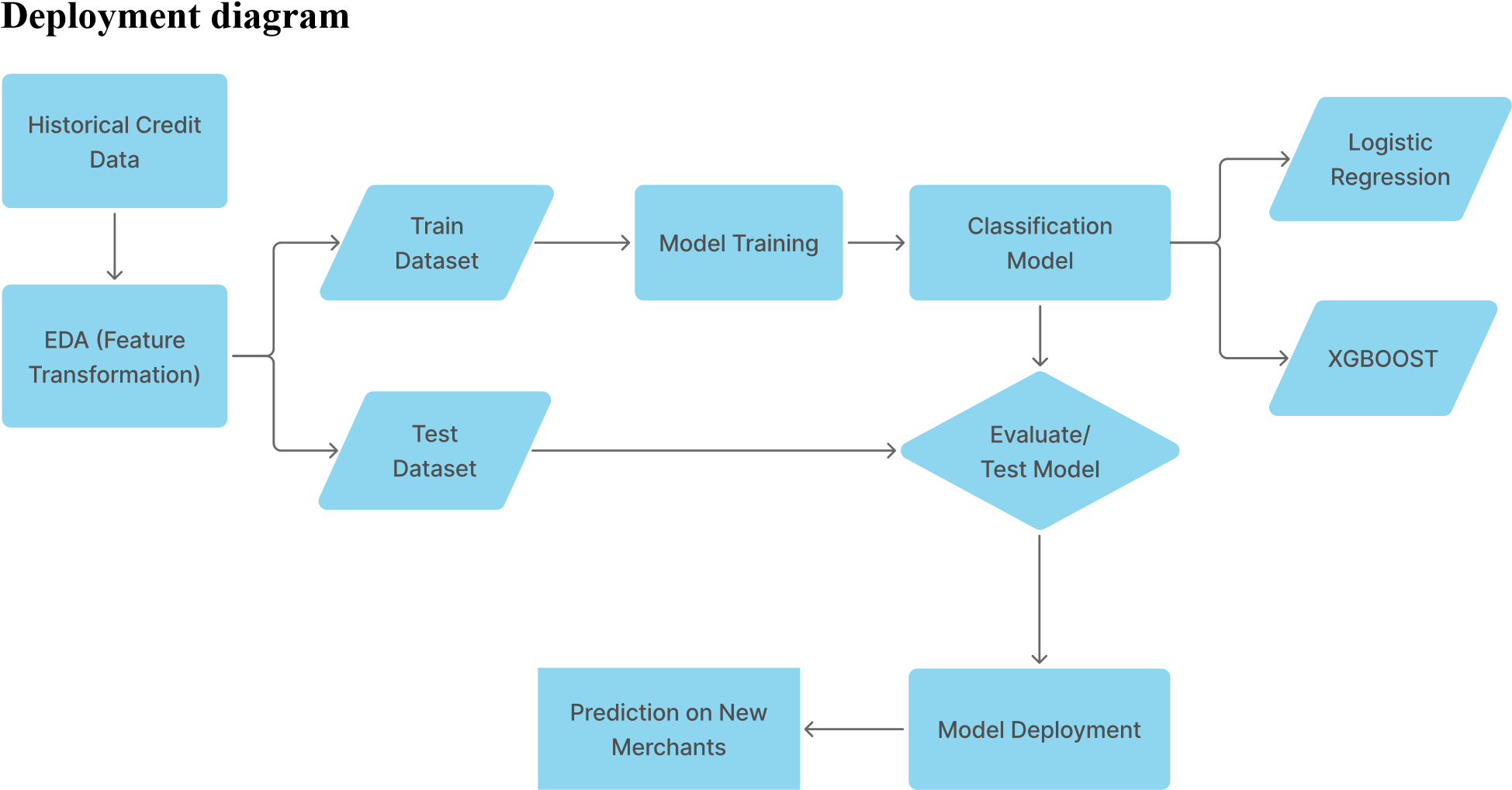


Figure 6

* + - From historical data, I employed advanced analytics to identify patterns in customer behavior allowing us to gain insights and more know abouts into their repayment tendencies.
    - Based on the identified patterns, I categorized these loanees into different groups according to their likelihood of defaulting on loans awarded. This enables us to effectively manage risk.
    - Merchants with low probability of default determined by our model are considered eligible for renewal cash advances and increased limits. This approach ensures continuity in support of reliable businesses.
    - In this system we use Logistic Regression, a powerful machine-learning technique to extract valuable patterns from the historical data analysis. This method provides better understanding of deeper factors influencing default probability.
    - This system operates in near real-time, enabling us to make quick informed decisions on loan approvals, ensuring there is efficiency and risk-conscious lending practices.
    - Finally, we've harnessed some other algorithms like XG-Boost and Random Forest improving ability to assess and predict applicants default probabilities with greater precision and robustness.

## 4.2 SYSTEM REQUIREMENTS.

### 4.2.1 Functional requirements.

The system must determine personal attributes of the cardholder and transaction activities that determine the default rate of credit card loans.

The system must allow loan officers to upload credit card specifics (attributes and their transaction activity).

The system should allow loan officers to predict loan status of customer based on their personal characteristics and trading.

### 4.2.2 Non-functional requirements

The system processes important customer details which could cause concern, concerns are handled improperly. The system will be used by loan officers who most likely has prior experience in handling computer systems. The information in the system should be handled with care and so the system offers:

High data security – the system should ensure high levels of confidentiality, availability and reliability.

High Performance - The system must have a short response time when issuing outputs.

High level of data protection – It ensures only access to your data, the intended users and does not allow sharing of data to unauthorized users.

High reliability – for systems to achieve its purpose, high availability and mostly accuracy of the results must be assured.

# CHAPTER FIVE: SYSTEM DESIGN.

## 5.1 Introduction.

This chapter is written considering the steps that would be followed while building system and operating it. The design was to ensure that the system met the functional requirements listed in the previous chapters.

System design is process of defining interfaces, architecture and data of a system to meet the specified requirements. It helps create systems that are easy to use and more efficient in meeting the needs of organizations.

## 5.2 Architectural design.

The architectural design of in this system describes the system structure, it’s behavior and its analysis.

It emphasizes the design of the system architecture.

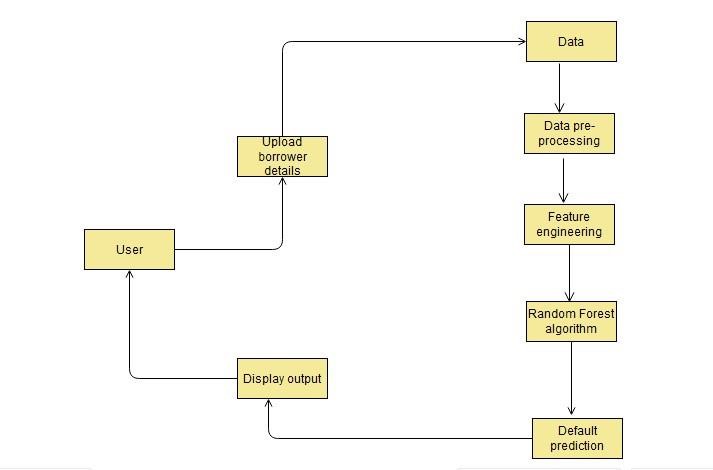


Figure 7

We can see in the diagram above; the system architecture is quite simple. It relies on data trained using Random Forest Algorithm to make the predictions. After predictions are made, the output is displayed to the user, who is the loan officer in the instance. The output states whether or not a borrower it’s likely to default for a loan.

### 5.2.1 Overview of the system.

The system can’t be described as a Loan default prediction system. It was built using python and has a web app made using stream lit that is accessible to loan officers. It checks the attributes of borrowers and evaluates them efficiently, thereby predicting whether they are likely to default for a loan or not.

### 5.2.3 Behavioral properties (UML)

### Activity diagram

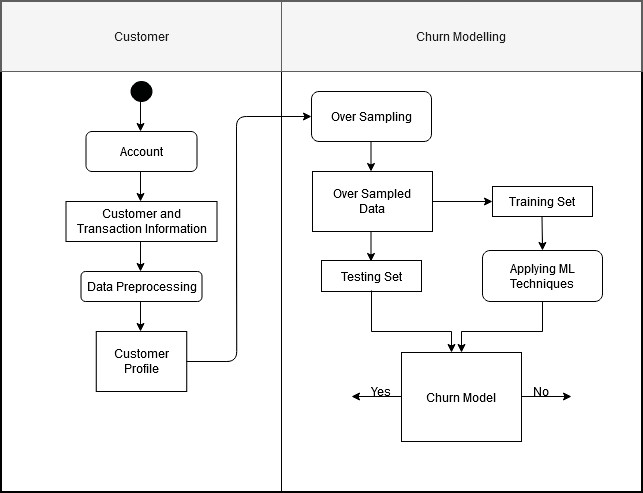
Illustrate a loan process or workflow between users(customer) and the system (churn modeling) where the prediction is done sampling**.** 

Figure 8

**Sequence diagram:** This diagram for loan default prediction explains several components to represent the sequence of actions and interactions between different entities/ objects

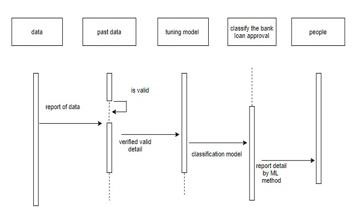


Figure 9

**Communication diagram**

This explains the interactions between the object’s roles associated with lifelines and the messages that pass between lifelines from borrower person to loan officer and back to the co-signer.

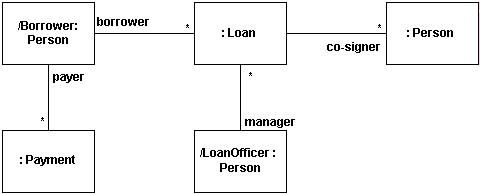


Figure 10

**Use case diagram:**

It gives graphic overview of the personnel directly involved with system and shows how the functions that are required by them are integrated.

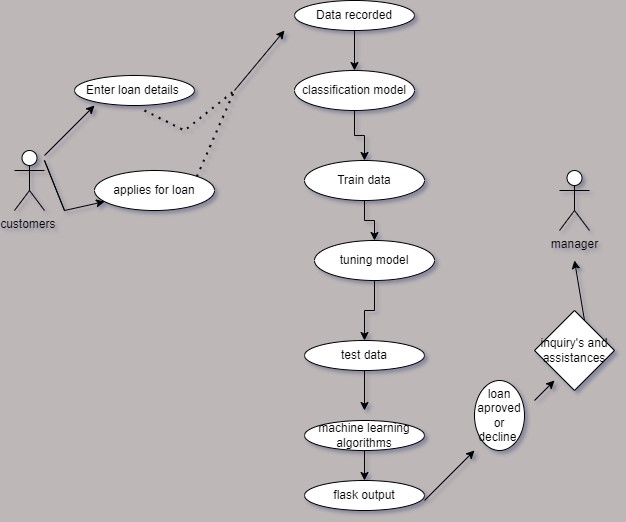


Figure 11

* Customer – Uses the organization’s system interface to apply for a loan.
* System checks on customer’s historic transactions through the process and views customer default likelihood.
* Then system predicts whether the customer is likely to default by either approves

with a yes or rejects with a no on the application.

* Incase customer is unsatisfied with results may get inquiries from loan officer manager.
* Loan manager – Uses the prediction system to check for default likelihood once more to unsatisfied customer or the manual processes.

**Timing diagram:**

This illustrates timing constraints and interactions between various components in a system over a period of time.

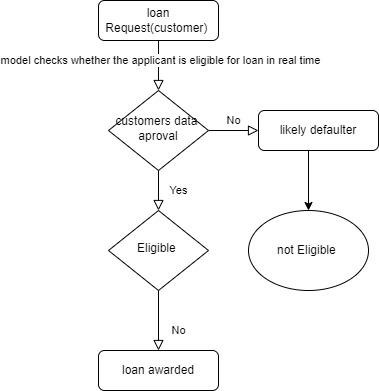
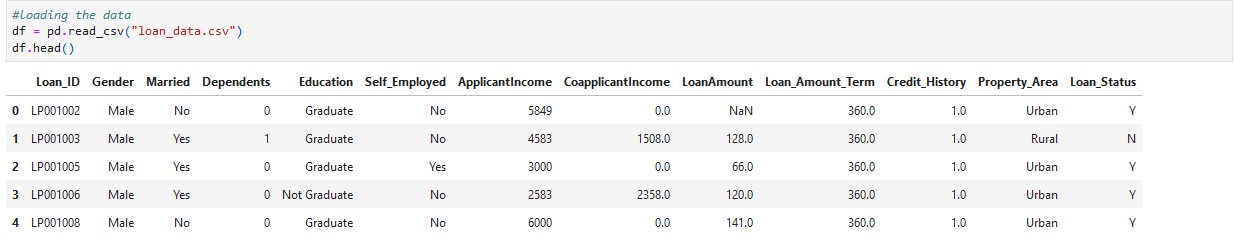


Figure 12

### 5.2.4 Data collection and pre-processing.

This data was obtained from an online platform named Kaggle that provides access to a wide variety of datasets**.** Data pre-processing is the process ofCleaning and transforming data to prepare it for analysis. The main purpose of data pre-processing is to improve data quality and make it more suitable for data mining tasks at hand. Data pre-processing is carried out to take care of missing values, null values and duplicates. This data does not have missing values hence said to be clean to perform the analysis. Below is the display of my data set



### 5.2.5 Model training and evaluation.

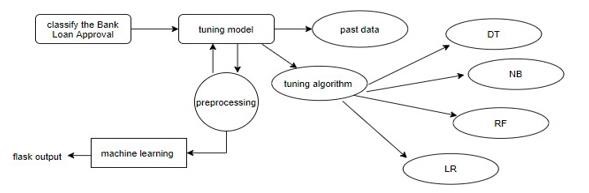
This machine learning model was built using random forest algorithms. Data was trained and tested to check the accuracy of the machine learning model. Evaluations were done by checking the areas under the ROC curve that summarizes performance of model by evaluating trade-offs between true positive rate and false positive rate.

### 5.2.6 Prediction and deployment.

Prediction is made from the trained model and saved in a pickle file. Finally, model was deployed using stream lit on web app where it could be used to make predictions.

5.2.7 DATABASE DESIGN  **Entity Relationship Diagram (ERD):**

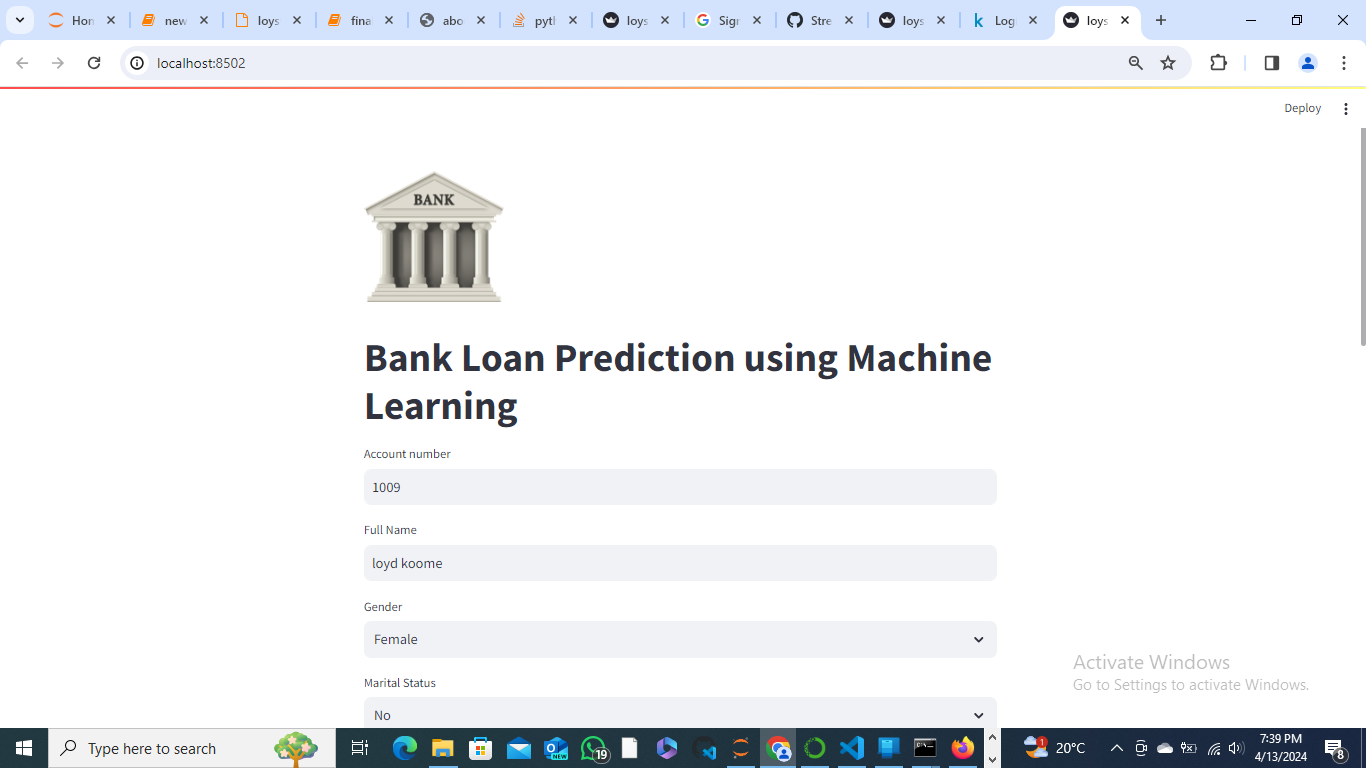
It’s a graphical representation of information system that shows relationships among people, objects, places, concepts and events within depicted system. Entity relationship model is data modeling technique that defines business processes also used foundation for a relational database.

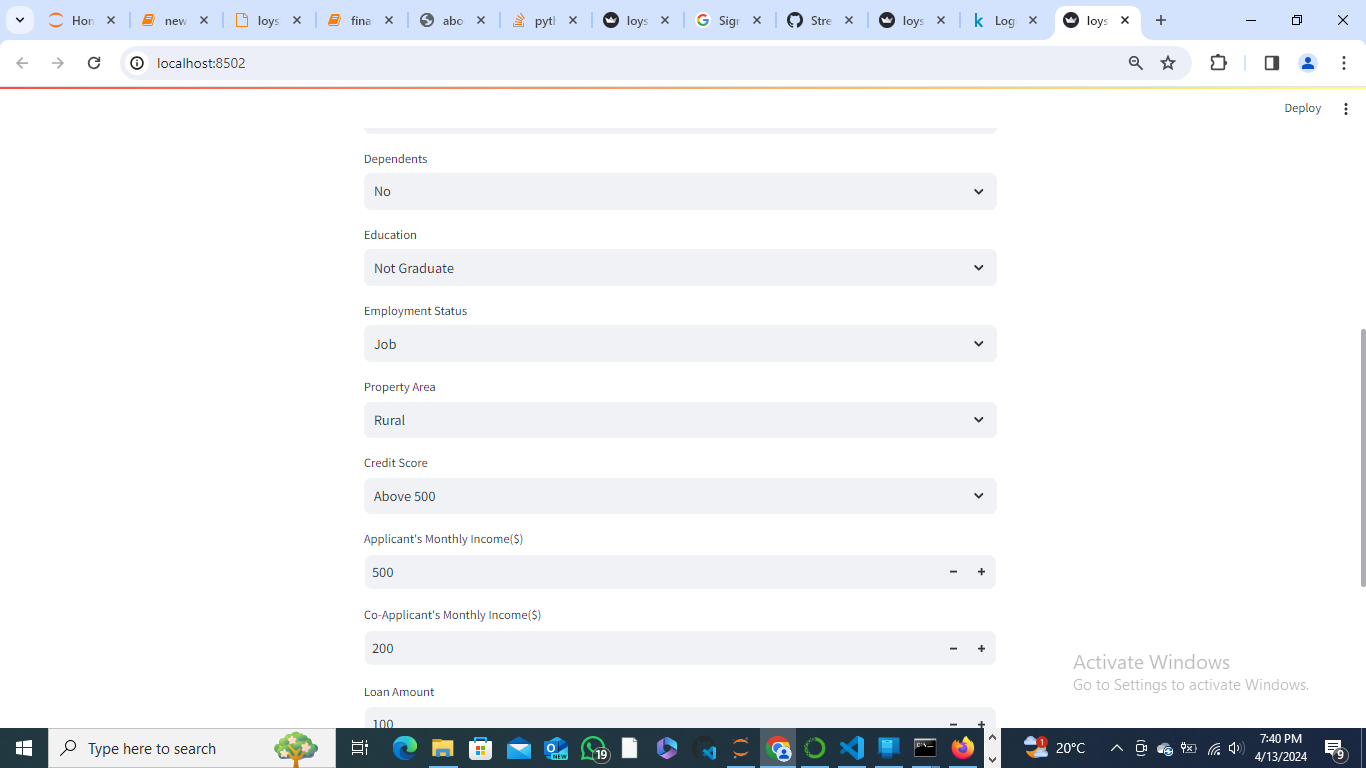


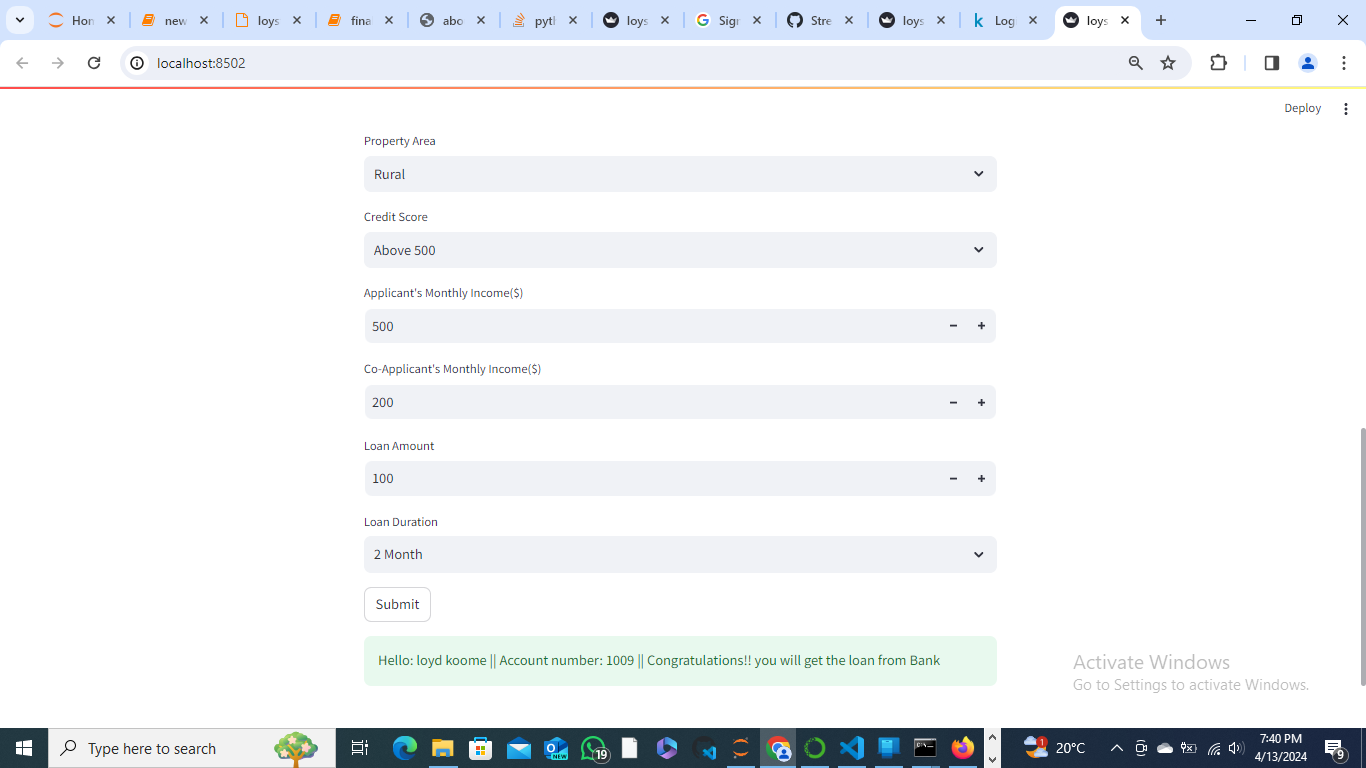
### 5.2.8 User interface design.

User interface design (UI) is process of designing user interfaces for machines and software. It focuses mainly on maximizing machine usability and making sure user experience is optimized. It also aims at achieving goals of its users by making the system simple and efficient to use.

The system was designed to allow entry of several details. The details required are as per the total loan amount borrowed, the credit score and dependents of the borrower. Using these details system will make predictions based on the models created. These predictions will state whether someone is likely to get approved







# CHAPTER SIX: IMPLEMENTATION AND TESTING.

## 6.1 Development Environment.

The system was developed using streamlit for front end and for the backend I developed the system using python and machine learning. The data used in the system is in a dataset was obtained from an online platform named Kaggle. The data was trained and tested using machine learning and the model created was deployed. The web application relies on the machine learning model’s predictions to output results.

For the development of the system, I used:

* Anaconda navigator – It’s a graphical interface for launching python programs without need of installing packages and managing environment using command lines.
* Jupyter Notebooks – This application is provided in Anaconda Navigator or google Collab. By default, I used it for writing my programs loading the data, pre-processing, data visualization, building the model and saving the model.
* PyCharm – Frontend development was done using PyCharm also inbuilt in Anaconda.

## 6.2 System Components.

System components are the processes, programs or utilities of given system that help in managing the different areas of the system. They usually act as means of breaking down the complexity of the system into manageable parts. The following are components of the system:

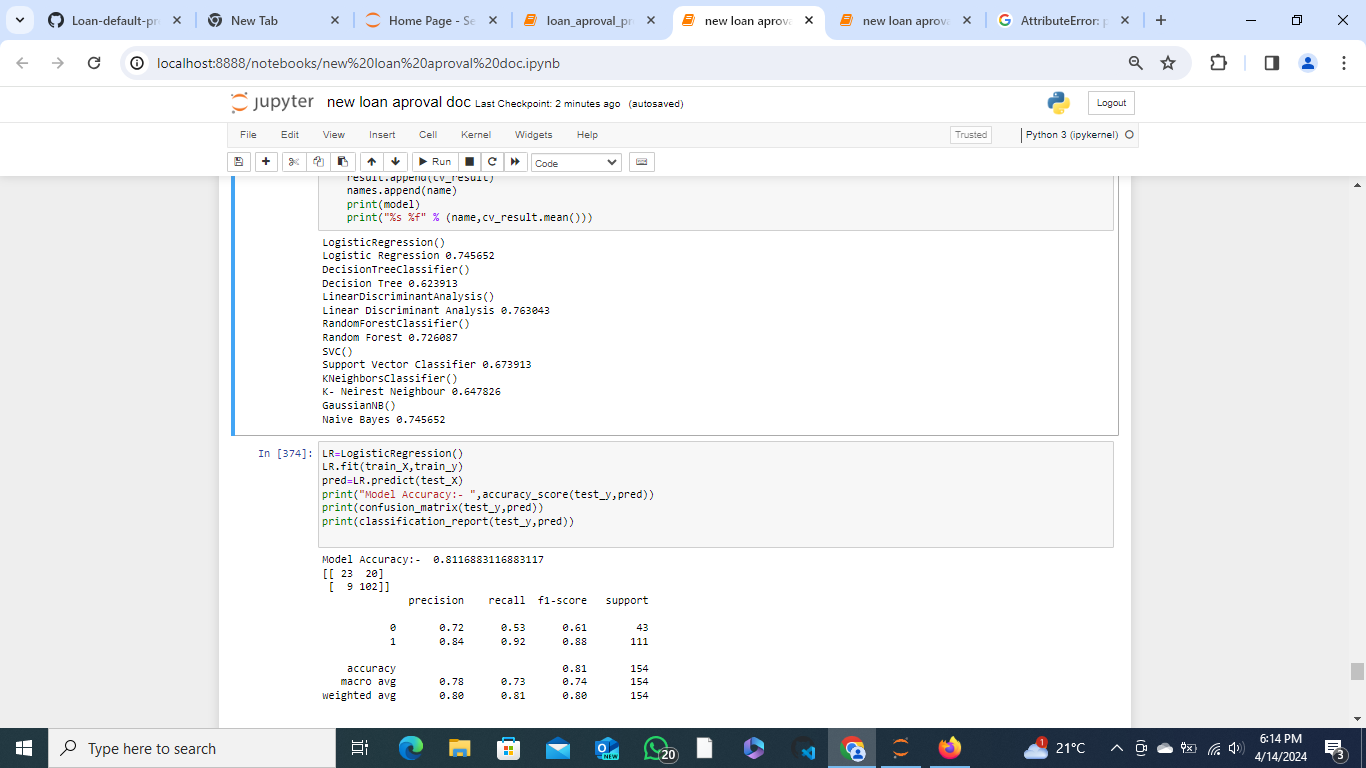
### 6.2.1 Data

The system was using transactional data that contained loan status information of borrowers.

The data is contained in a train dataset of and with 614 rows and 13 columns and a test dataset with 367 rows and 12 columns.

### 6.2.2 Algorithms.

A machine learning algorithm is a method which is generally used for generating output from a given input data. Machine learning algorithms have two main processes; classification and regression. They are also supervised, unsupervised or semi supervised. For this project I used a classification, supervised machine learning algorithm known as Random Forest

Classifier. 

### 6.2.3 Models.

These is testing algorithms that are used to improve the accuracy of a model and also its efficiency. Processes such as outlier treatment, handling null values and missing values and normalizing data have been used in this system

### 6.2.4 Packages

The system uses some data science packages such as NumPy, pandas, matplotlib, sklearn.

### 6.2.5 System Development components

The focus was established in two major aspects during system development as illustrated below:

**•** Data-driven platform-based process

The system was designed to give users a clear chance to make strategic decisions based on data analysis and interpretation. Loan officers made choices on which decision to based from information gained in the predictions that are made.

**•** Value-driven cocreation process

The system was designed to directly interact with users. This enables them to make decisions on whether borrowers are qualifying for loan or not. The model was designed by the use of both frontend programming languages and backend languages which improved its structure and functionality’s.

## 6.3 Test Plan

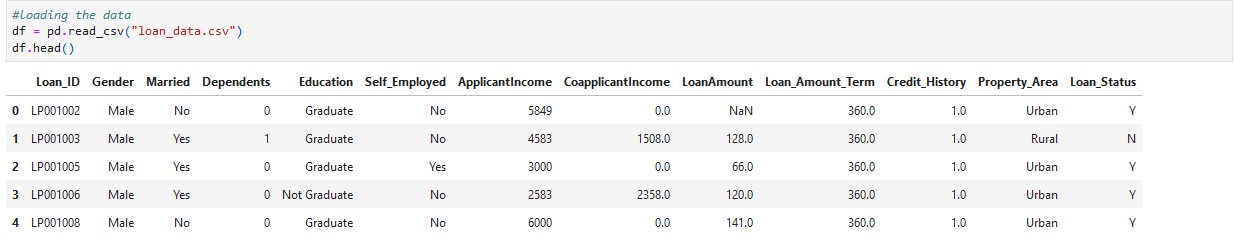
This involves testing the system before delivering it to the user. It aims at satisfying all the system requirements, therefore satisfying the users.

**Validation**

The system was to be tested and successfully implemented to ensure that the requirements listed in software requirements specifications are met.

Test cases – Accuracy of the machine learning model

Test data – Dataset



**Figure 11: Data**

Test results – The model is more than 75% accurate which is quite suitable for a model.

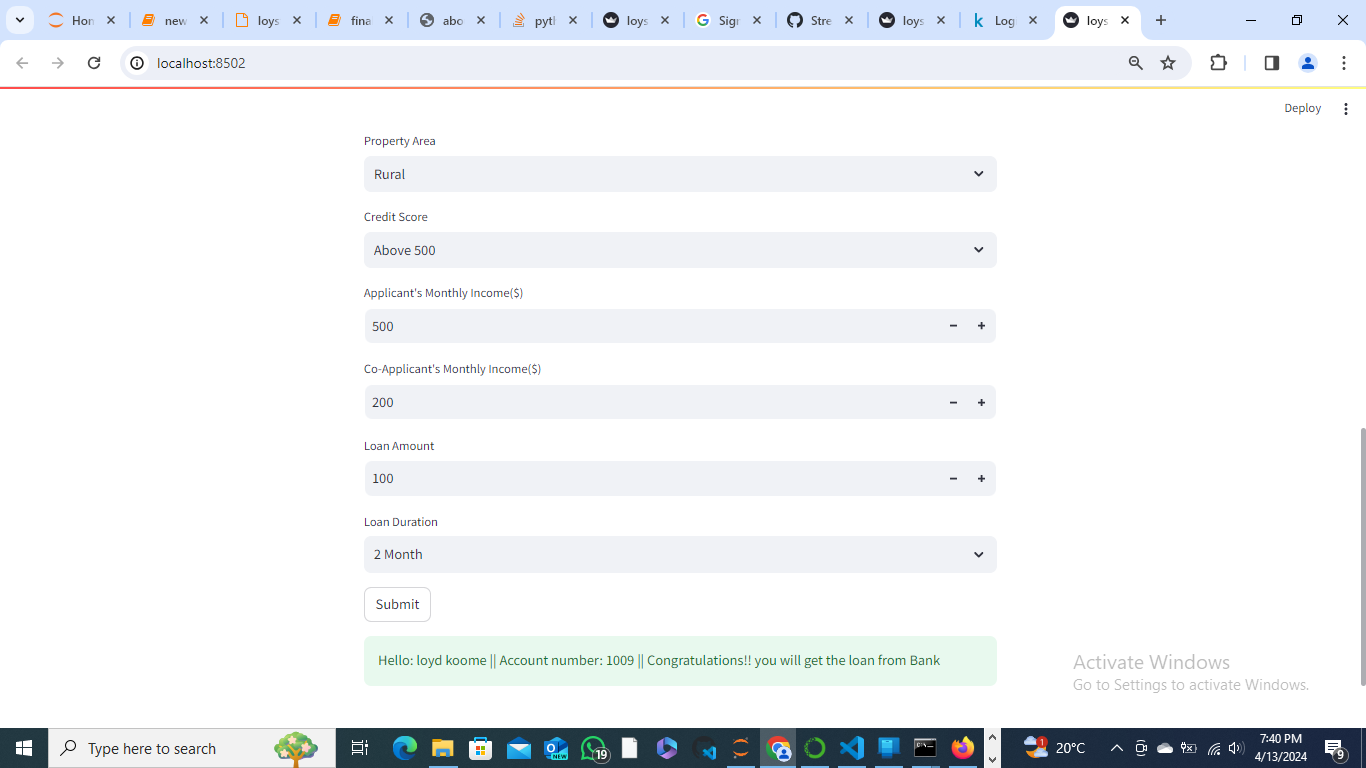
## 6.4 System Testing

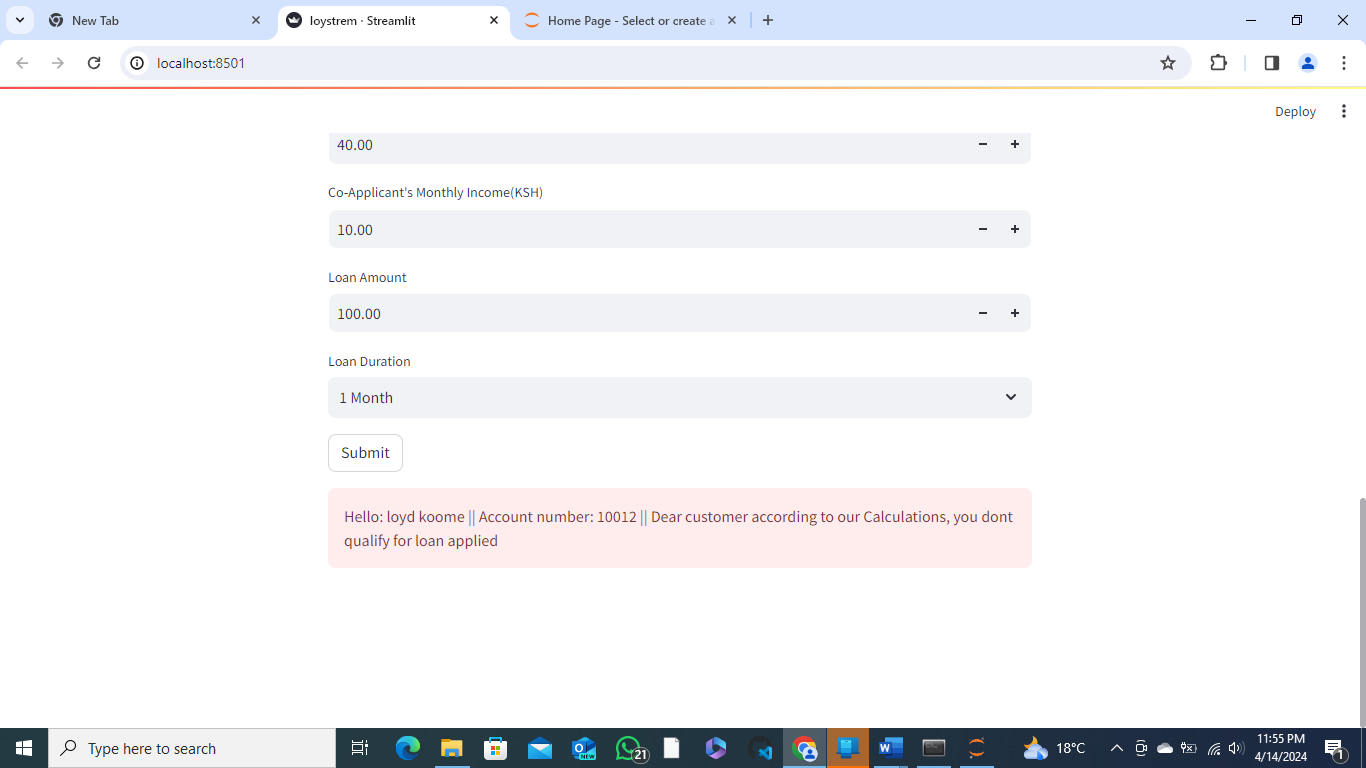
The following are the tests carried out for my project:

Unit testing – It’s the testing of individual software component. For this test I was to check the accuracy of the model components individually.

Regression testing – This was done to check whether tests from previously encountered bugs have been solved.

Integration testing – The components developed function properly on my machine learning pipeline, which from this case the integration of this web application with my machine learning model.





# CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS.

## 7.1 Achievements and lessons learnt. (Technical lessons and achievements).

### 7.1.2 Achievements

* Learnt machine learning – By learning machine learning I was able to understand machine learning algorithms work and how they can be used in the banking sector.
* Developed model for predicting loan approval and integrated it with my web app.
* Deployed my machine learning model and created a user-friendly interface using streamlit.

### 7.1.3 Lessons Learnt.

* The system is useful for checking the credit worthiness of a borrower in a way that is fast and efficient.
* Different python libraries - I have worked with different python libraries and I have been able to understand purpose of each of them and the instances in which should be used.
* Learnt new technologies- technologies can be used to gain insights from data used to make predictions. I also learnt that the use of these new technologies can be used in the system development using lesser development time.

## 7.2 Recommendations.

* Additional changes can be made on the web application to make it more user-friendly and more robust when more test cases are applied to it.
* Multiple machine learning algorithms can be used, compared and the best performing algorithm could be deployed. We can also try combining different models.
* More visualizations can be done on the independent variables to discover more patterns.

## 7.3 Conclusions

The developed model automates the procedure to determine creditworthiness of an applicant. It focuses on information containing the main characteristics of the borrower.

Random Forest Classifier is used in this system and it is a classification model that is used in machine learning. The classification analysis of Random forest algorithm relies on supervised learning algorithms where each tree is constructed independently from input data with and all the trees in the forest have the same distribution. Therefore, it is suitable for predicting the correct earnings and contribution to it in the current world scenario. Banks will be able to lend money in the right hands and people can get loans much faster. The accuracy of the system is quite high which is the advantage.

# REFERENCES.

1. Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, *4*(1), 111-138.
2. Nejad, M. G. (2022). Research on financial innovations: an interdisciplinary review.

*International Journal of Bank Marketing*.

1. Obiero, S. O. (2020). Influence of Women Enterprise Fund Project on

Socioeconomic Development: A Case of Lurambi Constituency, Kenya.

1. Prorokowski, L. (2019). Validation of the backtesting process under the targeted review of internal models: practical recommendations for probability of default models. *Journal of Risk Model Validation*
2. Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & EmmertStreib, F. (2021). Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in artificial intelligence*, *4*, 22.

# Appendices

PROJECT SCHEDULE

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Task | 20/09/2  023 | 16/01/2  024 | 04/02/2  024 | 27/02/2  024 | 02/03/2  024 | 21/03/2  024 | 11/04/2  024 | 28/04/2  024 |
|  | 4/12/20  23 | 02/02/2  024 | 25/02/2  024 | 01/03/2  024 | 20/03/2  024 | 10/04/2  024 | 26/04/2  024 | 24/05/2  024 |
| Project planning |  |  |  |  |  |  |  |  |
| System Analysis |  |  |  |  |  |  |  |  |
| PROJECT Design |  |  |  |  |  |  |  |  |
| Programming andDevelopment |  |  |  |  |  |  |  |  |
| Testing and debugging |  |  |  |  |  |  |  |  |
| Documentation  and Training |  |  |  |  |  |  |  |  |
| Implementation and Deployment |  |  |  |  |  |  |  |  |
| Monitoring and  Evaluation |  |  |  |  |  |  |  |  |