2024

A Credit scoring model for mobile lending.

Oindi, Brian
School of Computing and Engineering Sciences
Strathmore University

A Credit Scoring Model for Mobile Lending

BRIAN OINDI

145572

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computing and Information Systems at Strathmore University.

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya

June, 2024

Declaration

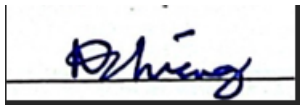
I declare that this work has not been previously submitted and approved for the award of a degree by this or any other university. To the best of my knowledge and belief, the dissertation contains no material previously published or by another person except where due reference is made in the dissertation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

Oindi, Brian Nyabicha

ORB April 8, 2024

Approval

The dissertation of Oindi Brian Nyabicha was reviewed and approved for examination by the following:

Sign:  April 8, 2024

Dr. Nelson Ochieng Odunga

Lecturer

School of Computing and Engineering Sciences

Strathmore University

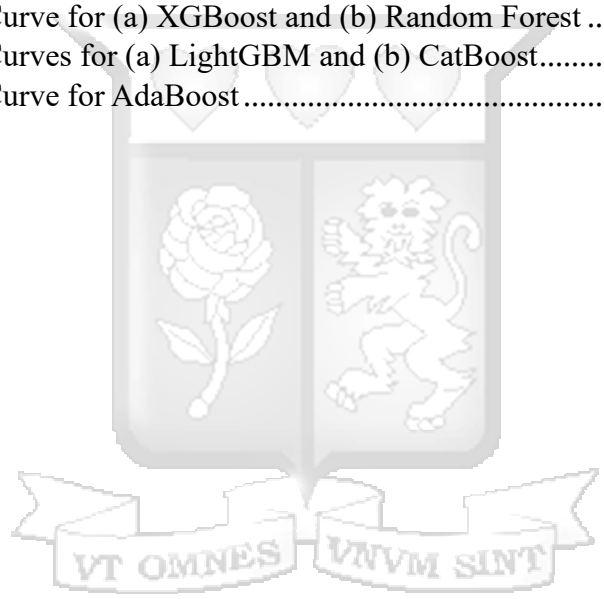
Abstract

An exponential increase in mobile usage has led to more accessible access to mobile loans for most Kenyans; this has created a lifeline for those excluded by traditional financial institutions; the easier way to borrow loans comes with its risks. The major one is borrower defaulting. This creates a need for credit scoring, which plays a crucial role in decision-making for lenders to determine borrowers' creditworthiness, therefore minimizing credit risk and managing information asymmetry. On mobile lending, borrowers' financial information is usually limited, making machine learning a favorable tool for credit assessment. Traditionally, the process required statistical algorithms and human assessment, which fall short when subjected to large datasets and are time-consuming. The traditional methods also need help adjusting to changes in borrowers' behavioral needs. Against this backdrop, this research developed a novel credit scoring model for mobile lending using Random Forest, XGBoost, LightGBM, Catboost, and AdaBoost algorithms. SMOTE was used to address the class imbalance problem. The model achieved the best accuracy of 86%. The research further analyzes the challenges in credit scoring and reviews related works by several authors. The research also looked at the feature importance of the models, which effectively analyzed the model's behavior. This model can analyze vast volumes of data, which would otherwise be resource-intensive if done manually. The machine learning model was then deployed into a Streamlit Web Application with a user interface where real-time predictions are made based on borrower data. The model can give lenders insights into determining borrowers' creditworthiness and enable them to make informed decisions before lending.

Keywords: Mobile loans. Credit Scoring. Probability of Default. Machine Learning. Statistical Algorithms. SMOTE

Table of Contents

List of Figures



List of Tables

List of Abbreviations/Acronyms

AHP	-	Analytic Hierarchical Process
AI	-	Artificial Intelligence
AUC	-	Area Under Curve
BWM	-	Balancing and Weighting Model
CAK	-	Communications Authority of Kenya
CART	-	Classification and Regression Trees
CRB	-	Credit Reference Bureau
CSP	-	Credit Service Providers
DT	-	Decision Trees
EBIT	-	Earnings Before Interest and Tax
EFB	-	Exclusive Feature Bundling
FICO	-	Fair Isaac Corporation
FP	-	False Positive
FN	-	False Negative
GBDT	-	Gradient Boost Decision Trees
GOSS	-	Gradient Based One Side Sampling
KNN	-	K-Nearest Neighbor
LIME	-	Local Interpretable Model-agnostic Explanations
LR	-	Logistic Regression
MCC	-	Mathew's Correlation Coefficient
MLA	-	Machine Learning Algorithm
MNN	-	Modular Neural Network
NACOSTI	-	National Commission for Science Technology and Innovation
NN	-	Neural Networks
PD	-	Probability of Default
RAM	-	Random Access Memory

RF	-	Random Forest
SMOTE	-	Synthetic Minority Oversampling Technique
SSD	-	Solid State Drive
SVM	-	Support Vector Machines
TB	-	Terabyte
TN	-	True Negative
TP	-	True Positive

Definition of Terms

Analytic Hierarchical Process	A structured process of analyzing complex decisions (Passage Technology, n.d.)
Credit Risk	Risks that include the borrower not repaying the loan on time (Aslam, 2019)
Credit Scoring	The predictive probability is that an existing borrower or an applicant will become delinquent (Kenton, 2019).
Data Imbalance	Refers to the difference in sample records. (Dewi, 2020).
Decision Trees	This supervised learning produces accurate and easily interpretable results-based probability estimation of distinct event occurrences (Shehadeh et al., 2021) and can handle discrete and continuous data.
Machine Learning	Machine learning is forming a series of actions and algorithms to solve problems with less human intervention and automatic optimization. Such techniques can find patterns in complex data (SAS, 2019)
Precision	The precision score measures the consistency of positive predictions (Minaee, 2019)
Recall	It is the fraction of true positives and the sum of true positives and false negatives (Alazab, 2015).
Support Vector	Supervised ML algorithms that analyze data and patterns to find optimal input segmentation for classification and regression (Gor & Lee, 2019).

Acknowledgments

I thank God because this far He has brought us. For good health, time and, strength to faithfully complete this research. His Grace is Sufficient.

I wish to express my gratitude to the following:

My Dad – Thank you for the Prayers, Encouragement and Moral Support even in difficult times you have always been there no matter what.

My Mum – Thank you for always wishing me the best in my studies, constant prayers and always motivating us to believe in ourselves. Thank you for the Sacrifice and Love.

My Siblings – Thank you for constantly checking on me, sharing progress and exchanging ideas.

Dr. Nelson Ochieng, my supervisor- Thank you for agreeing to be my supervisor and responding to my queries, I am grateful for your willingness to share knowledge through this research journey.

Dr. Bernard Shibwabo – Thank you for guiding us through the Dissertation classes, your reasonings and critiques made this research a success.

SCES - Thank you for maintaining a professional working relationship and sending us timely information when needed.

MSCIS and MSIT Class of 2024 – Thank you for your knowledge sharing, encouragement, and light moments.

I will not get the opportunity to thank everyone who helped and guided me, but I seal this moment to be everlasting.

Chapter 1: Introduction

1.1 Background

In a report by the (Communications Authority of Kenya, 2023), mobile money users stood at 38.4 million at the end of March 2023, which translates to a penetration rate of 76%. This shows there is an increase in mobile money services by Kenyans for internet usage and accessing digital credit. A mobile loan is preferably the way to go for most borrowers due to its instantaneous nature, taking less than 24 hours to get a loan, automation in the credit approval process, and, thirdly, remote access, which refers to easy access to the service anywhere (Chen & Mazer, 2016). These features make digital credit effective in after-shock situations, primarily when covering essential overheads, as researched by (Bharadwaj et al., 2019). However, mobile credit has its risks, too, as literature from (Izaguirre et al., 2018 and Burlando et al., 2021) suggests that Kenya, Tanzania, and Mexico have high default rates among digital credit users.

Much as mobile loan borrowing bears benefits for both the lender and borrower, it certainly has risks, which include the borrower not repaying the loan on time; this risk is known as 'Credit Risk' (Aslam, 2019). Therefore, mobile lenders must assess the client's creditworthiness before authorizing a loan. Indeed, according to the (Central Bank of Kenya, 2022) Survey report, which presents indicators on the quality of financial services and products used at the county level, Marsabit, Garissa, and Samburu counties record the highest levels of debt distress at 74%, 59%, and 58%, respectively among adult population which was attributed to climate-related shock facing the counties reducing the likelihood of borrowers to repay loans.

Parties in mobile lending operate in a virtual space (Yao et al., 2019), meaning more information must be provided. Lenders do not know if a borrower can repay the loan, let alone repay it on time. This leads to information asymmetry, jeopardizing truthful borrowers' adverse selection and raising moral issues. The predictive probability that an existing borrower or applicant will become delinquent is called 'Credit Scoring,' commonly used in consumer lending (Kenton, 2019). It typically indicates creditworthiness; a good rating represents a higher possibility of repayment, whereas a poor one with past debt obligations might also face challenges in the future based on the numeric score (Kagan, 2019). Credit Scoring is effective in managing information asymmetry and minimizing credit risk. Since most mobile lending is characterized by a short

interaction between lenders and borrowers, the principal means of determining credit score is through the CRB, which provides lenders with information on credit repayment history (Fosu et al., 2020). This is evident in most developing countries, including Kenya. A negative report affects future access to credit from mobile lending, not financial institutions. Much as this method has existed for years, its major downside is the inability to assess borrowers with no credit history but who need borrowing.

Traditional financial institutions mainly adopted the 5c's method for credit scoring with the following metrics: (i) Character, (ii) Capital, (iii) Capacity, (iv) Collateral, and (v) Conditions (Li, 2019). This assessment was hugely dependent on knowledge of customer dealing and personal experience. Another demerit of this method is the inability to assess borrowers in rural areas with little or no loan history or banking transactions. Generally, credit risk is calculated using various mathematical tools to estimate the default probability of the borrower (El-Qadi et al., 2022). This approach can provide valuable insights into client creditworthiness, but these traditional data analysis methods can be resource-intensive and time-consuming.

Mobile credit lending sector researchers have recently turned to machine learning and other automated techniques, including AI and algorithms, to accurately determine borrowers' credit scores (El-Qadi et al., 2023). These new technologies can analyze vast volumes of data and identify trends humans may fail to detect, thus enabling credit lenders to make informed lending decisions.

Therefore, to address the above challenges, this research paper aims to develop a model for credit scoring in mobile credit lending processes and work out the best approach to identify whom to lend to while identifying defaulters and reducing credit risk.

1.2 Problem Statement

Mobile lending is beneficial to many users since it offers loans in a short period, which helps in times of economic shock. However, the default cases are higher than traditional lending in financial institutions (Burlando et al., 2021). Classic credit risk scoring models have shortcomings when dealing with large datasets and some unintended biases when using human judgment-based models. Since credit scoring is an ephemeral task, adaptation to behavior change (Heaven, 2020) and data imbalance (Dewi, 2020) are also not considered in the traditional models, making them time-consuming and resource-intensive when developing credit scoring models.

Recent research papers have challenges, such as low accuracy (Kisutsa, 2021) and some algorithm bias by putting non-default in the default class Madaan et al., 2021. Therefore, this research focuses on implementing a credit scoring model to combat the challenges above and validate its use in prediction.

1.3 General Aim

This study aims to develop and validate a machine learning-based model for credit scoring in mobile lending. The model can help lenders determine consumer creditworthiness and make informed decisions before lending.

1.4 Research Objectives

1. To evaluate the challenges experienced in credit scoring.
2. To review the existing credit scoring methods.
3. To develop a machine learning model for mobile loan credit scoring
4. To validate the developed machine learning model.

1.5 Research Questions

1. What are the challenges experienced in credit scoring?
2. How can we review the existing methods used for credit scoring?
3. How can we develop a machine-learning model for mobile loan credit scoring?
4. How can we validate the developed machine learning model?

1.6 Scope and Limitations

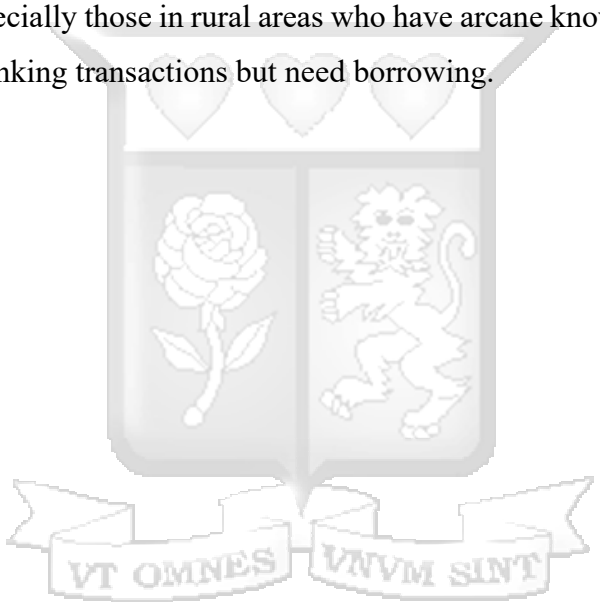
The core focus of this research was to propose a model for credit scoring using machine learning that will give mobile lenders insight into borrowers' ability to repay their loans while identifying the probable defaulters using machine learning algorithms.

The research was carried out using secondary data, which had missing values. Furthermore, limited resources and research on credit scoring for mobile lending using machine learning made the research consume much time.

1.7 Justification

An increase in mobile money users sets a trajectory for an increase in mobile lending; this creates a need for an effective credit scoring model since borrowing affects the lenders, borrowers, and the economy at large. Traditional scoring models may not align with mobile lending, which has unique data patterns. They also tend to be biased when dealing with users with no banking history, which creates alienation of lenders and a lack of equality in credit access.

Machine learning models can handle significant data needs and create interpretable results for lenders to determine the creditworthiness of borrowers and reduce some borrowers' debt cycles. The development of this model can also help in policy formulation to deal with financially excluded individuals, especially those in rural areas who have arcane knowledge of borrowing and little to no financial or banking transactions but need borrowing.



Chapter 2: Literature Review

2.1 Introduction

This chapter provides brief literature on credit scoring theory. Then, it covers the subtopic areas highlighted in the research objectives: challenges experienced in credit scoring, existing credit scoring techniques, machine learning models in credit scoring, a review of related works, and finally, the conceptual framework.

2.2 Credit Scoring Postulation

Credit scoring gained popularity in the 1990s. It predicts whether the borrower will default, and it is an essential metric in financial institutions used in credit lending. It helps determine borrower creditworthiness (Sengupta & Bharadwaj, 2015). Credit scoring minimizes default and provides a borrower's creditworthiness based on a numeric score; the higher the credit score, the lower the non-repayment risk. Credit scoring is applied to individuals who need borrowing. The data includes demographics, past credit behavior, financial statements, and alternative data. The methodology includes statistical and machine learning techniques; the producers include CSPs and credit managers, while the users are financial lending institutions, credit providers, and central banks. The scale used is any numerical range (World Bank, 2019). The main aim of credit scoring is to determine the probability that a borrower will default. If the likelihood is low, the borrower is classified as "good," otherwise “bad” (Simão, 2023).

Credit lending is one way to meet financial obligations and personal needs (Zhu et al., 2020). The development of a credit scoring model is necessary to aid in borrower evaluation. Amongst the different methods used to mitigate credit risk, historical data is the primary method many financial institutions adopt. Mobile lending is the way for people from all walks of life because of its ease of accessibility; financially excluded individuals can also leverage mobile loans to survive during economic shocks. Most lending platforms try to find the appropriate strategy to encourage credit borrowing, but some borrowers end up defaulting (Çığşar & Ünal, 2019). This makes a borrower's credit risk assessment crucial to managing the business entity and effective risk management. Financial institutions used to employ credit officers whose sole duty was to determine borrowers’ creditworthiness by manually using their credit history (Egwa

et al., 2022). However, technological advancements have shifted this trend and adopted machine learning for massive data analysis and visualization.

Credit scores are used in the credit life cycle, and below are some of the use cases:

i) early alerts to CSP of an event that affects the borrowers' credit risk, ii) applying scores based on borrowers' application information and determining the creditworthiness, iii) detecting fraud based on some behavior change and alerting CSP (World Bank, 2019).

2.2.1 Regulatory Growth

Regulators have used several tools to manage credit risk - Basel I, II, and III, which are discussed below.

The Basel I Accord was formed to enhance stability and establish a standard operating procedure for banks. The main aim was to 'strengthen the soundness and stability of the banking system. However, frailties such as risk weightings were observed, and the regulatory requirements became less meaningful. The Basel II objective was implemented to counter the inefficiencies, which contributed significantly to the CSP development of credit scores (Sidiqqi, 2017).

Basel II's approach is governed by three pillars: minimum capital requirements, market discipline, and supervisory review (Chen, 2023). The financial institutions decide their method for minimum capital requirement; this can be done using two approaches: one is the standard approach, where a fixed percentage of outstanding loans is set aside. The second one is the Internal Rating approach, where the financial institution selects the percentage of exposure in each asset to put aside. In each, the approach's losses are calculated; in this way, Probability of Default models are built (Marte, 2019).

Conversely, the 2008 financial crisis exposed the weaknesses of the financial systems internationally; as such, the Basel III accord was created in November 2010 and implemented in January 2022. Basel III identified the main reasons for the financial crisis and aimed to improve Basel I and II accords by introducing liquidity requirement ratios and leverage. In this regulation, banks were required to categorize different groups based on size and economic importance. These regulations are not constant but continuously reviewed and improved based on past experiences and market conditions.

2.2.2 Credit Scoring Data Types

According to (World Bank, 2019), some traditional data types in credit scoring include Bank transactional data, Credit Bureau checks, and Commercial data. Data for credit scoring is mainly classified into structured, semi-structured, and unstructured. The structured data is stored in databases that include daily operations data; unstructured data have no predefined order and include social media usage, text, images, and audio. A report by (CGFS & FSB, 2017) has shown that the use and quality of unstructured data for credit scoring have yet to be fully proven. For the semi-structured data, it has markers or tags. In modern credit scoring models, all these data types are used. Below is some of the detailed data used for credit scoring.

Table 2. 1: Common Data Types Used for Credit Scoring

Data	Description
Mobile Data	A rise in smartphone usage has led to the rise of unstructured and structured data, and these mobile applications collect data on geolocation, movement, and transactions. As such, the data may allow some phone apps to perform credit checks without the owners' consent (Grab, 2018).
Social Media Usage Data	Research papers (Blazquez & Domenech, 2018) have shown that the frequency of social media usage and posts can lead to a better understanding of borrowers' lifestyles and provide insights into their creditworthiness. However, this may give a false impression of lifestyle since many may fake a lifestyle based on social media usage.
Payment History Data	CSP makes payment history records on past and present credit available. A late payment negatively affects the borrower's score, whereas a timely payment improves the credit score rating (World Bank, 2019)
Transactional Data	This is an individual's account usage, such as credit card usage and e-commerce data. Such data provides an organized transactional view based on past payment history (Siddiqi, 2017). Transaction data may offer up-to-date information on borrower creditworthiness and give the CSP a helpful context in credit scoring (Barasch, 2017).

2.3 Challenges Experienced in Credit Scoring

Credit scoring is a linchpin for ascertaining borrowers' creditworthiness. However, challenges have always existed, leading to changes in calculating credit scores. These challenges range from changing consumer behavior to information asymmetry and data imbalance. These challenges offer a significant background and opportunity for policy formulation to mitigate bias. Therefore, understanding these challenges and ways to combat them can create fairness in credit scoring models and enhance proper decision-making by lenders.

2.3.1 Information Asymmetry

Digital lending participants are usually strangers; lenders cannot fully comprehend borrowers' information or economic situation, and they can easily forge information to borrow funds and later default in payment, leading to losses in mobile lending platforms. Only the borrowers know if they can repay the borrowed loan (Serrano & Gutiérrez-Nieto, 2016). Thus, information asymmetry is an adverse selection as some lenders may make inaccurate selections of high-risk borrowers, leading to losses. For seasoned and experienced borrowers, traditional financial institutions are still the first choice, with considerably low interest rates. For this reason, most mobile credit customers have difficulty getting funds in banks (Bhaird et al., 2019). Banks deal with asymmetry by building long-term relationships with borrowers, which makes it easy to collect 'soft' information.

2.3.2 Data Imbalance

This is the difference in sample records; since most traditional models treat all data samples equally, the generalization error is huge (Dewi, 2020). Credit scoring data is usually non-linear, with some classes having many samples and others having fewer. Users with high credit risk are relatively few compared to actual credit review tasks; this makes it extremely hard for traditional models to achieve the best performance. It also occurs in the credit risk classification due to significant differences between the number of 'good' and 'bad' borrowers.

Data imbalance also affects machine learning models due to imbalanced datasets, which may lead to biased accuracies. Credit scoring is ephemeral; variables change over time, and creating a reliable credit scoring model takes more than a year. For this reason, some credit scoring models have been unchanged for decades.

SMOTE uses KNN and bootstrapping to address this challenge. It creates new minority classes using feature-space linear interpolation data instead of data space. The SMOTE combination of over-sampling and under-sampling methods has created a balanced dataset (Niu et al., 2020).

2.3.3 Adaptation to Behavioral Change

In the case of an economic recession, most credit scoring models will have problems; this was evident during the COVID-19 pandemic. Job losses and unemployment were expected at the height of the pandemic; occurrences such as a drop in spending or an increase in deposits are all indicators of a shift in consumer dynamics. Behavior change can also be a case of strategic delinquency, where consumers will choose to go delinquent in hoarding cash. Since machine learning is a data-driven model, it is vital to monitor the models early to allow for retraining of the models (Heaven, 2020). A need arises for model triaging by applying human judgment in creating intuitive models, making machine learning models explainable in such behavioral shifts, leading to robustness (Christoph, 2019).

2.3.4 Risk of Unintended Bias

When dealing with different datasets, some machine learning datasets have zip codes as input; in former times, this did not exist; instead, the algorithms were the zip tool detection. Redlining, which is using zip codes in credit scoring, is prohibited. This is because it creates bias and zip codes should be protected. Since most machine-learning datasets have alternative inputs, the results can be biased (Feldman et al., 2015; Prince & Schwarcz, 2019). Using big data and sophisticated approaches can lead to unfair treatment and unobserved risks (O'neil, 2017). Amazon once used AI to find job candidates (Dastin, 2022). It was shut down since it favored female candidates affiliated with women's groups, which meant many females were rejected from the application.

When replaying the same tale in credit scoring, a Machine Learning algorithm can refer to a protected class using branch transactions and even social media data. Using social media data for credit scoring and assuming it produces the same results as the FICO or CRB score is considered a case of bias since the digital footprints could correlate to a protected class. Researchers (Lehr & Ohm, 2017) have argued that this is one of the most significant impediments of machine learning algorithms in credit scoring, whose solution is legal rather than statistical.

2.3.5 Large Data Records

Dealing with significant data needs can be challenging, especially for traditional credit scoring models, which leads to a lot of time and resources, this data can be high-velocity and high-variety which need proper collection and storage for better decision-making. This raises regulatory concerns to protect consumers against discriminatory credit scoring methods (Onay & Öztürk, 2018). Machine-learning incorporates linear and non-linear data as input variables, as seen in most datasets with extensive records. However, many comparative machine learning models aiming to identify the best-fit model must note that the answer lies in the data availability. When dealing with massive data sets, neural network algorithms are the best; with spurious data and some correlation, boosted trees may be the best fit. Even with machine learning, it is unlikely that we can have a single winner model-wise because datasets and outputs vary even in a specific context, like credit scoring.

2.4 Existing Credit Scoring Methods

Many researchers have made tremendous contributions in the field of credit scoring using various models in the past and created valuable models; some of these complex models consider financial and non-financial information to achieve the same end goal: assessing the risk of lending money to individuals. In this section, the researcher explores the existing techniques Credit Scoring Techniques.

2.4.1 CreditXpert

Mortgage lenders mainly use this Credit Scoring method to evaluate borrower's creditworthiness. In its methodology, CreditXpert considers loan-to-value and debt-to-income ratios. This method has also implemented a Credit Analyzer tool that enables users to see changes in their credit scoring if they take specific actions like paying their credit balance (FasterCapital, 2023).

2.4.2 Z-Score Altman Model

Dr. Edward Altman developed this model in 1968. It is one of the earliest contemporary quantitative credit scoring models (Altman, 2018), based on a multivariate analysis of the five accounting ratios. Although this model is more than 50 years old, it still holds importance to many

market participants. Some credit risk models, like structured models, have been created using age-old models. The major criticism of this model is its backward-looking nature and sporadic.

The original model is seen in (i).

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

Where:

X_1 = *total assets/ working capital*

X_2 = *total assets/ retained earnings*

X_3 = *EBIT*

X_4 = *book value of total liabilities*

X_5 = *sales*

(i)

2.4.3 Expert Judgement-Based Model

Traditional financial institutions employ highly skilled professionals to assess applicants based on their expert judgment and decide whether they default. One such technique was the AHP Analytic Hierarchical Process, a structured process of analyzing complex decisions (Passage Technology, n.d.). This model works on information represented in a hierarchy, making it easier for professionals to comprehend sub-issues and address them effectively. It is more like a fishbone diagram. The main element of this method is human judgment, which is used to make evaluations. This method for credit scoring was time-consuming, with colossal data records since each customer must be analyzed independently.

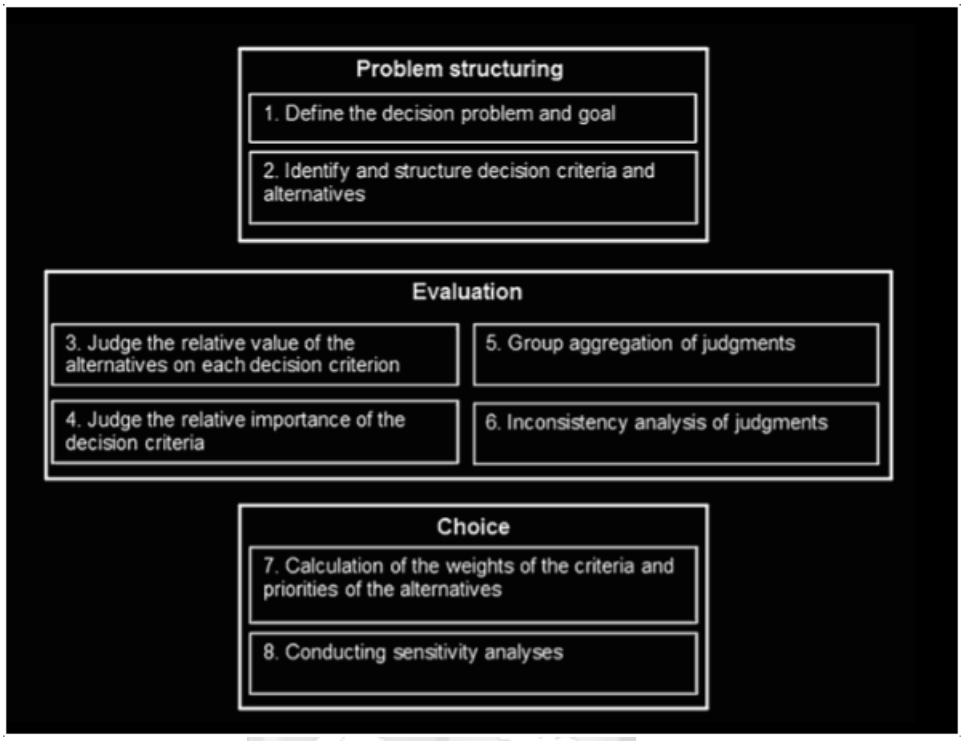


Figure 2. 1: Group Decision-Making with AHP. Adapted from. (Hummel et al., 2014).

2.4.4 Probability Unit and Logistic Regression

Abbreviated as 'probit model,' the probability unit is closely related to Logistic Regression; probit predicts a dependent variable. It is the inverse of the normal distribution of the probability and is modeled as feature combinations. This model was mainly used because the dependent variable had to be binary. However, a linear relationship between dependent and independent variables is not assumed. LR is a classical model for PD; it uses borrowers' basic information and loan information data for analysis (Hou, 2020). It is commonly used because of its ease of development, use, and result interpretation. The logit model picks parameters that increase the chances of observing sample values. In the last few years, it has been observed that LR was the standard among the traditional credit scoring models since it fulfilled the Basel II accord requirements (Gor & Lee, 2019). It is used to solve binary problems in this context, default and non-default, and maps inputs to (0,1). It is, however, noted that these traditional methods violate banking practices. For this reason, (Antunes, 2021) recommends Random Forest due to its better prediction levels than LR.

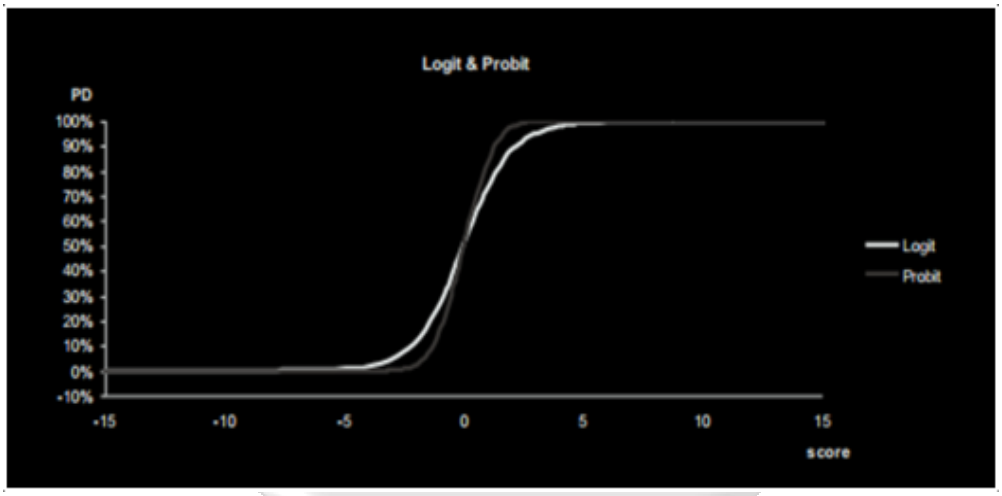


Figure 2. 2: Graphical Illustration of Logit and Probit models. Author Illustration

2.4.5 Linear Regression

As a statistical method, linear regression has been helpful in credit scoring. One reason is the ease of explanation and PD parameters in credit scoring (World Bank, 2019). Here, the target variable is projected on some covariates.

2.4.6 FICO Score

This credit scoring algorithm is widely used to determine borrowers' creditworthiness, especially in the US. FICO scores range from 300 to 850. The higher the borrower's credit score, the higher the creditworthiness (FasterCapital, 2023). FICO utilizes financial information such as payment history and length of credit history to determine whether to approve a loan and at what interest.

2.4.7 Vintage and Survival Models

Vintage models like Age Period Cohort operate on time series by original date so that credit risk by date and age of loan can be used for forecasting (Fu, 2018). Survival models are based on account performance and when an event occurred instead of if an event occurred. Both models are middle ground between time series and credit scoring models. These models provide cash flow modeling, pricing, and credit risk forecasts. Over time, ensemble survival and vintage models have been reported to be successful in credit risk scoring.

2.4.8 Discriminant Analysis

Context-wise, this is a form of supervised learning whose task is to separate individual groups based on the probability of default. This aims to combine scores into a single variable called the discriminant score. The significant assumptions of Discriminant Analysis are: i) observations are random and predictor variable distributed, ii) the dependent variable of the training data is classified correctly, and iii) at least two mutually exclusive groups where each case belongs to a single group.

Each group will have a standard score distribution in a successful discriminant analysis. If some group's discriminant score is more significant than some cut-off value, assign 'non-Default' if it is less than or equal to 'Default' (Gurný & Gurný, 2013). These analyses are susceptible to outliers, and unordered predictors cannot be used as inputs. They also require a strong assumption that predictors in each class have a normal distribution.

2.5 Machine Learning Models for Credit Scoring

Machine learning is forming a series of actions and algorithms to solve problems with less human intervention and automatic optimization. Such techniques can find patterns in complex data (SAS, 2019). Some of the machine learning models for credit scoring are discussed below.

2.5.1 Neural Networks

Inspired by how the human brain operates and its purpose to copy its learning process, NN comprises three layers: an input layer that receives data, the information flows to the hidden layers, and an output layer, the answer calculated by NN based on the given data. The initial objective of neural networks was to solve problems in a structured layer akin to the human brain. Neural Networks produce good results and are fault-tolerant (Mijwel, 2018).

However, they have become used to performing several tasks, such as credit scoring. They are also the first machine learning model employed for this task. One can increase the number of hidden layers by adding hidden layers. In the first layer, external information, such as independent variables, is received, and then neurons in the input layer send feedback to the hidden layer.

Credit scoring uses MNN architecture, where neural networks transmit signals to a hidden layer receiving information from input neurons. Numerous credit default studies have applied NNs and have reported higher accuracy. NNs are data hungry; where data is abundant and of good quality, NN always wins in prediction. However, the major criticism of NN is the tendency to overfit where analysis corresponds closely to training data, making future predictions inaccurate and slow in training.

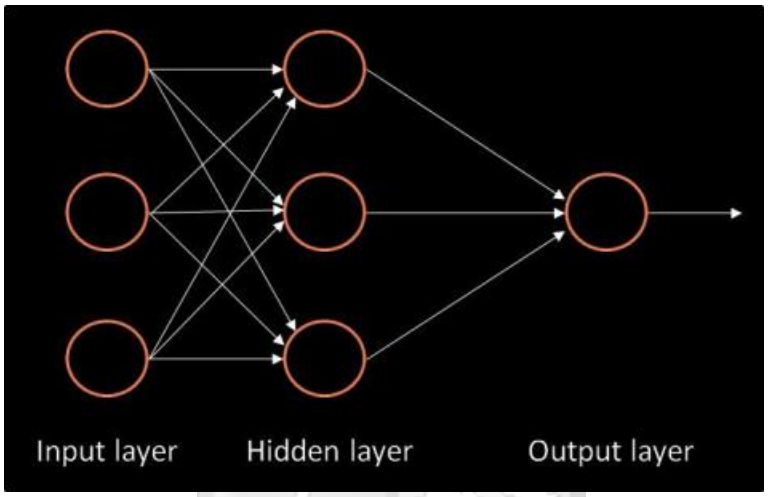


Figure 2. 3: Simplified NN with three layers. Author Illustration

2.5.2 Decision Trees

Also known as CART, this supervised learning produces accurate and easily interpretable results-based probability estimation of distinct event occurrences (Shehadeh et al., 2021). Capable of handling discrete and continuous data. The concept of DT is to create recursive partitioning of the input till it is enough to make a prediction. It is a graphical representation with the core goal of creating a goal that predicts the target variable based on input variables. They have a tree-like structure with nodes reflecting attributes, branches reflecting test outputs, and core nodes representing the categories. DT splits training data into smaller pieces, memorizes each, and predicts each.

They have also historically been used for credit risk scoring; modern DT uses multiple partitioning metrics, e.g., ANOVA and Gini index, which are used to calculate information gain

which is applied to each node to provide a measure of the quality of split (Jijo & Abdulazeez, 2021) merit of DT is mapping between rules and trees. DT has an advantage in handling data with outliers and low computing costs, which makes training faster. However, it needs help dealing with missing values and overfitting problems and can build complex trees.

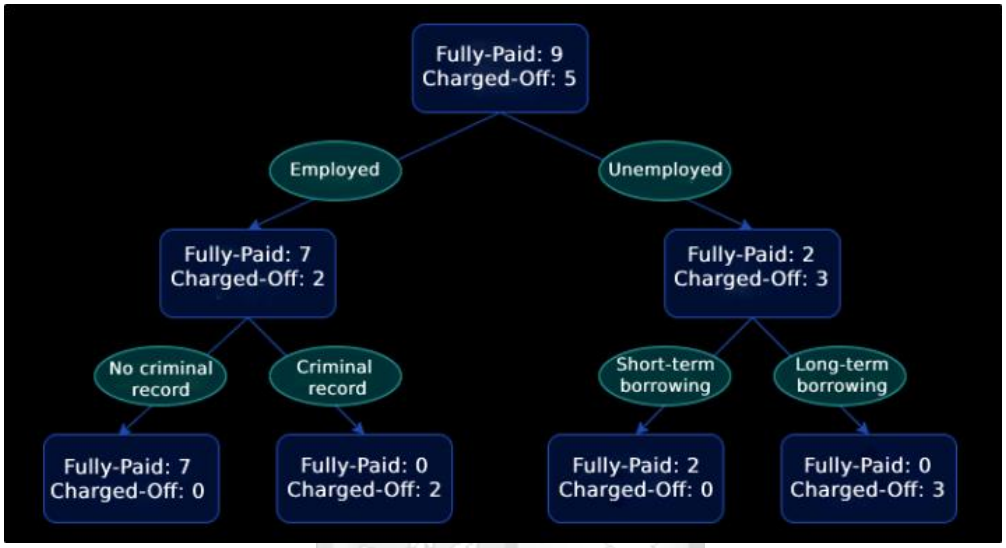


Figure 2. 4: Schematic Illustration of Decision Trees. Adapted from. (Ko et al., 2022)

2.5.3 Random Forest

Considered state-of-the-art, RF is used for both regression and classification problems. In this case, we are interested in the classification problem. RF is like bagging, with a significant difference being randomness. This ensures diversity in the trees, leading to low correlation; this is also the critical contrast between RF and DT. While RF selects subsets of features to perform splits on the nodes, DT examines existent variables. RF generates and averages unrelated trees; each tree has a random feature selection for splitting (World Bank, 2019). Builds DT of various samples and takes majority vote for binary classification.

Some advantages of RF are the easy-to-understand importance of each feature, efficiency when dealing with large datasets, overcoming the overfitting issue in DT, and the capability to handle many independent features without feature selection. However, it takes longer to train the model. Figure 2.5 illustrates a Random Forest tree.

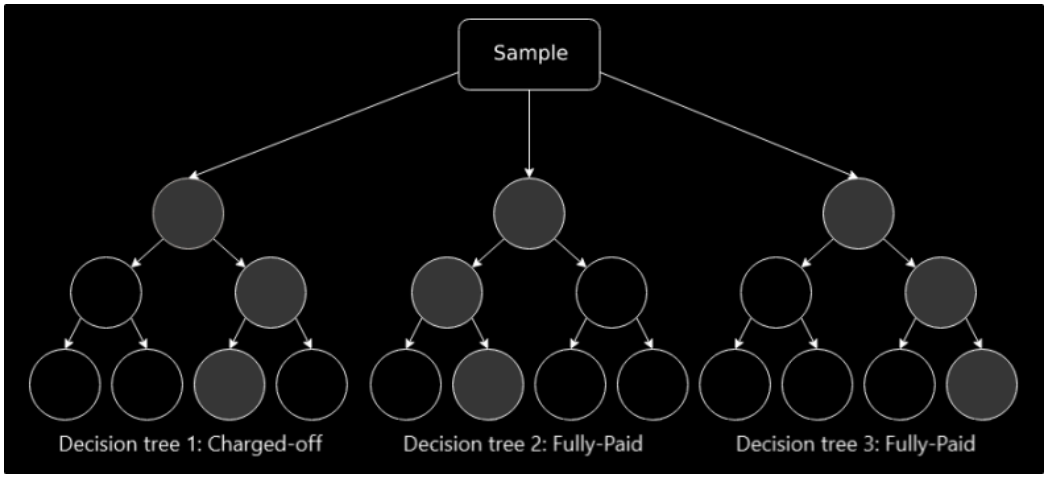


Figure 2. 5: Random Forest Illustration. Adapted from. (Ko et al., 2022)

2.5.4 Support Vector Machines

Support Vector (SVMs) are supervised ML algorithms that analyze data and patterns to find optimal input segmentation for classification and regression. They were first introduced in 1998 by Vapnik and have been effective in credit scoring (Goh & Lee, 2019). SVMs are a kernel-based approach, where training data represents the kernel. This model is beneficial in classification where similar data belong to the same class and where the interaction of input variables is complex to know beforehand.

The hyperplane segmentation of inputs leads to an effective classification method. The classification divides data into homogenous groups, separating training sets into distinct classes. In this way, the different classes are predicted based on the inputs. Data is complex to be linearly separable and less prone to outlier adjustments. Authors (Singh et al., 2021) reported the successful application of SVM for a modernized loan approval system with an accuracy of 80%. Figure 2.6 shows an overview of the Support Vector Machine model.

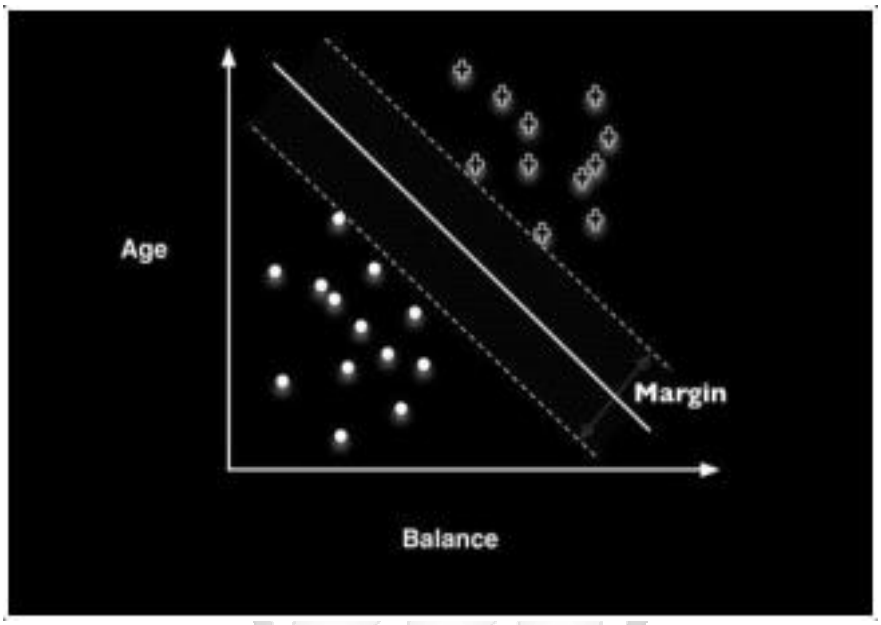


Figure 2. 6: Support Vector with 2 Features. Author Illustration

2.5.5 Xtreme Gradient Boost

Gradient Boosting is a concept where weak learners are trained, which results in high accuracy. Gradient Boosting was introduced to enhance quality in classification tasks due to advancements in computing power. It is an extension of Gradient Boost Decision Trees; in this new method, each new tree is born from the previous one, keeping the original model unchanged but adding a new function to alleviate the shortcomings of the previous one. An ensemble of weak prediction models is used in the boosting function. XGBoost uses an approximation histogram algorithm, reducing overfitting and increasing efficiency (Chen & Guesgrin, 2016).

XGBoost can use categorical and numerical features to handle classification and regression problems. It also includes L1 and L2 regularization to prevent overfitting faster than traditional boosting methods. XGBoost handles missing values through a built-in function.

2.5.6 Categorical Boosting

CatBoost was developed by Yandex; it can handle large datasets and quickly make predictions. It can also handle categorical data without transformation through a learning process that performs encoding on the categorical data, leading to accurate predictions while reducing overfitting; this increases feature dimension. CatBoost improves traditional GBDT by converting categorical features to numerical ones through the Ordered Target Statistic Method, thereby increasing support for categorical features (Prokhorenkova et al., 2018). It uses a symmetric tree to implement the same splitting for layers, increasing prediction speed and accuracy.

CatBoost allows novice learners to choose the best model parameters with built-in cross-validation. It also supports L1 and L2 regularization to minimize overfitting and automatic feature scaling.

2.5.7 Light Gradient Boosting Machine

This gradient boosting algorithm can also handle large datasets in machine learning classification and regression tasks. It is widely known for its fast speed and low memory usage for prediction. LightGBM works for both multi-class and binary classification tasks. It is a decision tree-based framework that uses Exclusive Feature Bundling and Gradient-Based One Side Sampling to solve the limitations of Gradient Boost Decision Trees and achieve cutting-edge results. LightGBM has better accuracy and supports distributed and parallel learning. It requires less training time than other machine learning models due to GOSS and EFB enhancements (Ke et al., 2017).

2.6 Related Works Review

Table 2. 2: Related Works

Author	Content Focus	Findings	Challenge
Kisutsa, (2021)	Predict loan defaults on online mobile-based lending: ***Bondora Dataset. 2009-2021***	Decision Trees 64%.> Logistic Regression 63%. Naïve Bayes 61%.	The main challenge with this paper is the low accuracy of 64%.
Madaan et al., (2021)	Loan Default Prediction: ***Lending Club Dataset 2007 & 2015***	Random Forest 80% > Decision Trees 73%	The dataset used was highly imbalanced and no SMOTE was done to address the issue.
Runchi et al., (2023)	Credit Scoring using Logistic- Balancing and Weighting Model in Default Recognition **<u>6 Datasets</u>** ***Australian Dataset*** ***German Dataset*** ***Chinese personal loan*** ***Default of credit card*** ***Give Me Some Credit*** ***Credit card Fraud Detection***	Logistic- BWM > 10 MLA using metrics AUC, Specificity, Sensitivity, G-Mean, F-Score, MCC.	While recognition of the Default sample is improved, there is a Sacrifice of non-default samples. An increase in sample size weakens the model.

2.7 Conceptual Framework

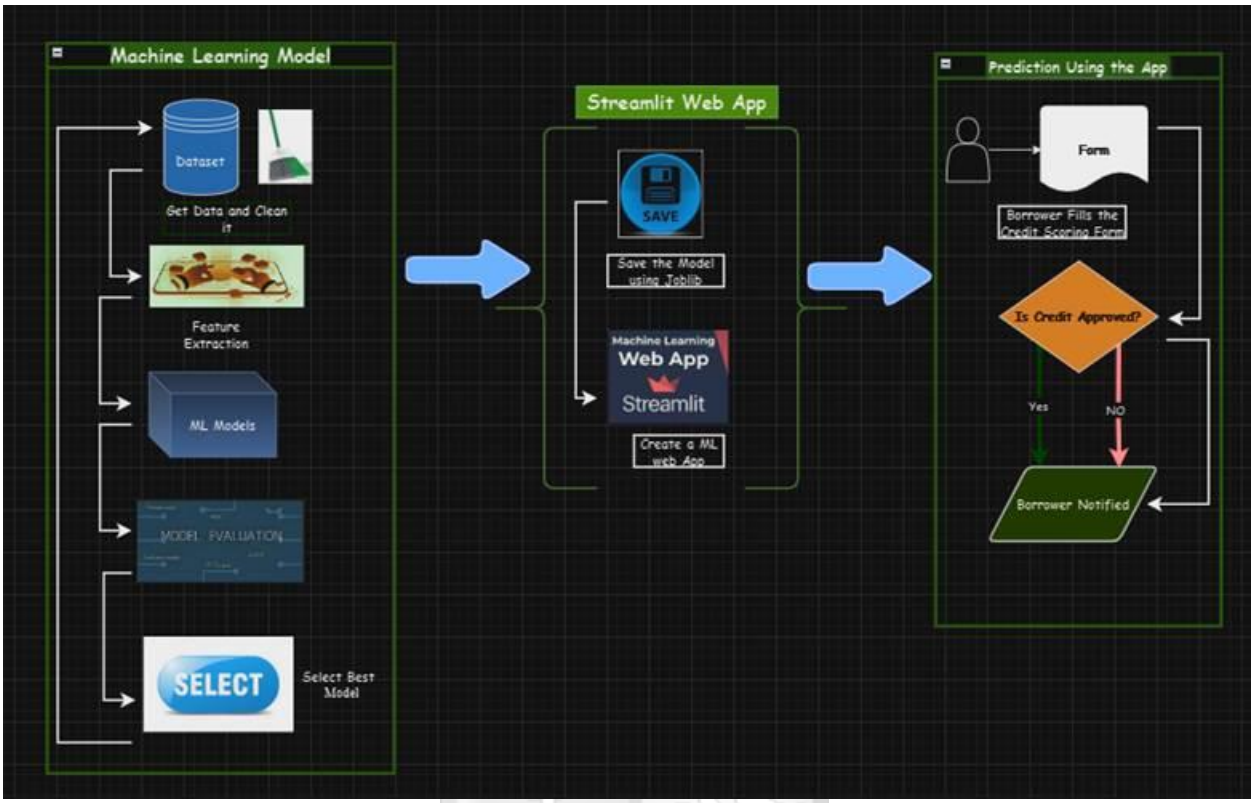


Figure 2. 7: Conceptual Framework

Steps Involved in the Conceptual Framework

I. The data downloaded from a verified online source was cleaned and pre-processed.
II. Feature extraction was done to identify the training features.
III. The training features implemented Random Forest, XGBoost, LightGBM, CatBoost, and AdaBoost Machine Learning algorithms.
IV. Model Evaluation used Accuracy, Recall, Precision, F1-Score, Confusion Matrix, and AUC-ROC.
V. The best-performing model was selected and saved using the joblib library.
VI. The model was used to develop a Streamlit machine learning web application.
VII. The web application is then used for real-time prediction, where a borrower is prompted to input prediction data. The borrower's credit is either approved or rejected. Finally, the borrower is notified of the status of their application.

Chapter 3: Research Methodology

3.1 Introduction

This chapter discusses the Research Methodology for the credit scoring model for mobile lending is presented. It highlights the steps followed to come up with the model. It highlights the research design (data source, data preparation, data modeling, data analysis, and evaluation metrics), software methodology (requirements assessment, user design, construction, and implementation), design requirements (hardware and software requirements), and finally, ethical considerations and research quality.

3.2 Research Design

Below are the steps taken to develop the model:

i) Data Source
ii) Data Preparation
iii) Data Modeling
iv) Data Analysis
v) Evaluation Metrics

3.2.1 Data Source and Description

This research adopted secondary data publicly downloaded from Bondora https://www.bondora.com/en/public-reports an online lending platform. Due to privacy issues, credit lending institutes are typically unwilling to share their internal data. It is not accessible to people outside the financial institutions. The adoption of the dataset was also triggered by the similarities in features with the local financial institutions, as displayed in Table 3.1 below. The lengthy process of obtaining a local dataset also contributed to the use of secondary data. The economic nature of the secondary dataset is another factor, as is the complex nature of collecting primary data and the required commitment.

The data contains loans between March 2009 and February 2024, containing borrowers' financial information and demographics.

```
[2]: df  = pd.read_csv('LoanData.csv',nrows=50000)

[3]: df.head(10)
```

[3]:

	ReportAsOfEOD	LoanId	LoanNumber	ListedOnUTC	BiddingStartedOn	BidsPortfolioManager	BidsApi	BidsManual	
0	2024-02-05	D8EBF360-104C-420F-BEC9-000924E6EFC7	3015853	2022-09-09 12:27:01	2022-09-09 15:27:01	11	0	88.0	{1AD71AF0-82 AD
1	2024-02-05	C1A98DDA-5E20-429C-BBFF-0009A05354E0	3743447	2023-05-19 11:04:33	2023-05-19 14:04:33	0	0	0.0	{3A9CF70B-0BD A
2	2024-02-05	9B0B252E-45B9-4172-8E2D-0014A8F18117	4335414	2024-02-03 14:43:32	2024-02-03 16:43:32	0	0	0.0	{FBC12A4E-7A5 A
3	2024-02-05	66AE108B-532B-4BB3-BAB7-0019A46412C1	483449	2016-03-23 16:07:19	2016-03-23 16:07:19	970	1150	5.0	{EBF05573-554 A
4	2024-02-05	C7EA512A-465D-4043-A9F2-001B14C3C14E	4043783	2023-09-26 12:14:12	2023-09-26 15:14:12	0	0	0.0	{7ACD344 ACAB-A
5	2024-02-05	A6635EA6-2F39-4DEA-AA3B-001C9521BE7C	2819530	2022-05-31 15:11:34	2022-05-31 18:11:34	16	0	0.0	{F6507F91-F9 A
6	2024-02-05	636993AA-338B-45FD-A60C-001E6282489C	3398045	2023-01-17 14:46:36	2023-01-17 16:46:36	12	0	3.0	{04336B48-A7 A
7	2024-02-05	B17292D8-7999-4372-86EB-0038041ABDB6	3618149	2023-03-27 21:08:20	2023-03-28 00:08:20	0	0	0.0	{42A0E51 BA6D-A
8	2024-02-05	D152382E-A50D-46ED-8FF2-0053E0C86A70	378148	2015-06-25 11:02:28	2015-06-25 11:02:28	1295	0	1705.0	{46C6CBA4-0FB A
9	2024-02-05	6A18E350-FA5E-4CCD-8383-007798DE275F	4020115	2023-09-14 14:58:04	2023-09-14 17:58:04	0	0	0.0	{5EC8528F-1B Al

```
[4]: df.shape

[4]: (50000, 112)
```

Figure 3. 1: Importing Dataset and Sample Features

Some of the Select Features in the dataset are shown in Table 3.1; the rest of the data dictionary is available on the dataset link provided in the Data Source and Description link above.

Table 3. 1: Dataset Features

Feature type		Description
Demographic Features	Age of the borrower	*Age of borrower when applying for a loan*
	Applicant's Gender	*0.0=male,1.0=female, 2.0=unknown*
	Education Level	*1.0= primary, 2.0= basic, 3.0= vocational, 4.0= secondary, 5.0 higher education*
	Marital Status	*1.0=married, 2.0=cohabitant, 3.0=single, 4.0=divorced, 5.0=widow, -1=not specified*
	Number of Dependents	*Number of children or other dependents*
	Employment Status	*-1.0=employed, 2.0= partially employed, 3.0= fully employed, 4.0= self-employed, 5.0=entrepreneur, 6.0= retiree*
	Home Ownership Type	*1.0=owner, 2.0= living with parents, 3.0= tenant pre furnished, 4.0=tenant unfurnished property, 5.0= council house, 6.0=joint-tenant, 7.0=joint-ownership, 8.0=mortgage, 9.0=owner, 10.0=other*
Financial Features	Amount Received	*Amount the borrower received*
	Income Total	*The borrowers' total income*
	Use of Loan	*1=real estate, 2=home improvement, 3=business, 4=education, 5= travel, 6=vehicle, 7=other, 8=health* *-1=not set*
	Status	*Current application status*

3.2.2 Data Preparation

The dataset obtained from the online repository must be explored using data analysis to understand the depth and variables. Here, the inaccurate data was corrected to ensure the reliability of the model during the training and testing phase of the model. Some steps include importing libraries for use in the modeling process, data normalization, and construction of a sample set.

```
# IMPORTING ALL NECESSARY LIBRARIES

# packages for Exploratory Data Analysis
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# To display all the columns of dataframe
pd.set_option('display.max_columns', 500)
import warnings
import os

# Data Preprocessing
from sklearn import preprocessing,metrics
from sklearn.preprocessing import LabelEncoder
from IPython.core.display import HTML
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.feature_selection import RFE
from sklearn.datasets import make_regression
from imblearn.over_sampling import SMOTE

# Model Evaluation
%matplotlib inline
from sklearn.linear_model import LinearRegression
from sklearn.metrics import average_precision_score
from sklearn.metrics import precision_recall_curve
from sklearn.datasets import make_classification
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix, roc_curve, roc_auc_score
from xgboost import XGBClassifier
from catboost import CatBoostClassifier
from lightgbm import LGBMClassifier
from sklearn.metrics import precision_score, recall_score
import xgboost as xgb
```

Figure 3. 2: Importing Necessary Libraries

```
missing_columns = df.columns[100*(df.isnull().sum()/len(df.index)) > 60]
print(missing_columns)
```

```
Index(['ContractEndDate', 'DateOfBirth', 'County', 'City', 'NrOfDependants',
       'EmploymentPosition', 'WorkExperience', 'PlannedPrincipalTillDate',
       'CurrentDebtDaysPrimary', 'DebtOccuredOn', 'CurrentDebtDaysSecondary',
       'DebtOccuredOnForSecondary', 'DefaultDate',
       'PlannedPrincipalPostDefault', 'PlannedInterestPostDefault', 'EAD1',
       'EAD2', 'PrincipalRecovery', 'InterestRecovery', 'EL_V0', 'Rating_V0',
       'EL_V1', 'Rating_V1', 'Rating_V2', 'ActiveLateCategory',
       'CreditScoreEsEquifaxRisk', 'PrincipalWriteOffs',
       'InterestAndPenaltyWriteOffs', 'PreviousEarlyRepaymentsBefoleLoan',
       'GracePeriodStart', 'GracePeriodEnd', 'NextPaymentDate',
       'ReScheduledOn', 'PrincipalDebtServicingCost',
       'InterestAndPenaltyDebtServicingCost', 'ActiveLateLastPaymentCategory'],
      dtype='object')
```

```
miss_col=['ContractEndDate', 'NrOfDependants', 'EmploymentPosition',
       'WorkExperience', 'PlannedPrincipalTillDate', 'CurrentDebtDaysPrimary',
       'DebtOccuredOn', 'CurrentDebtDaysSecondary',
       'DebtOccuredOnForSecondary',
       'PlannedPrincipalPostDefault', 'PlannedInterestPostDefault', 'EAD1',
       'EAD2', 'PrincipalRecovery', 'InterestRecovery', 'RecoveryStage',
       'EL_V0', 'Rating_V0', 'EL_V1', 'Rating_V1', 'Rating_V2',
       'ActiveLateCategory', 'CreditScoreEsEquifaxRisk',
       'CreditScoreFiAsiakasTietoRiskGrade', 'CreditScoreEeMini',
       'PrincipalWriteOffs', 'InterestAndPenaltyWriteOffs',
       'PreviousEarlyRepaymentsBefoleLoan', 'GracePeriodStart',
       'GracePeriodEnd', 'NextPaymentDate', 'ReScheduledOn',
       'PrincipalDebtServicingCost', 'InterestAndPenaltyDebtServicingCost',
       'ActiveLateLastPaymentCategory']
```

Figure 3. 3: Deleting Columns with Higher Percentage of Missing Values

```
cols_del = ['ReportAsOfEOD', 'LoanId', 'LoanNumber', 'ListedOnUTC', 'DateOfBirth',
       'BiddingStartedOn','UserName','NextPaymentNr',
       'NrOfScheduledPayments','IncomeFromPrincipalEmployer', 'IncomeFromPension',
       'IncomeFromFamilyAllowance', 'IncomeFromSocialWelfare',
       'IncomeFromLeavePay', 'IncomeFromChildSupport', 'IncomeOther','LoanApplicationStartedDate','ApplicationSignedHour',
       'ApplicationSignedWeekday','ActiveScheduleFirstPaymentReached', 'PlannedInterestTillDate',
       'ExpectedLoss', 'LossGivenDefault', 'ExpectedReturn',
       'ProbabilityOfDefault', 'PrincipalOverdueBySchedule',
       'StageActiveSince', 'ModelVersion','WorseLateCategory']
```

Figure 3. 4: Deleting More Columns with No Impact on Final Prediction

Defining the Target Variable

Here, status is the variable that helped in creating the target variable. Not making status the target variable is because it has three unique values: Current, Late, and Repaid. There is no default feature, but a default date feature tells us when the borrower defaulted, which means on which date the borrower defaulted. So, combining Status and Default date features to create a target variable was crucial. Late was not treated as default because it also had some records in which the actual status is Late, but the borrower has never defaulted, i.e., the default date is null. So, Current Status records were filtered since they have yet to mature. They are current loans. Finally, the target variable was created with 0 being 'not default' and 1 being 'default' as seen in Figure 3.6.

```
df['Status'].value_counts()

Status
Repaid     33688
Late       13202
Current     3110
Name: count, dtype: int64
```

Figure 3. 5: Status Variable before Preprocessing

Target Variable Definition

```
# filtering out Current Status records
df = df[df['Status'] != 'Current']
```

```
df["Default"] = df['Status'].apply(lambda x: 0 if x=='Repaid' else 1)
```

```
df['Default'].value_counts()

Default
0    33688
1    13202
Name: count, dtype: int64
```

Figure 3. 6: Defining the Target Variable

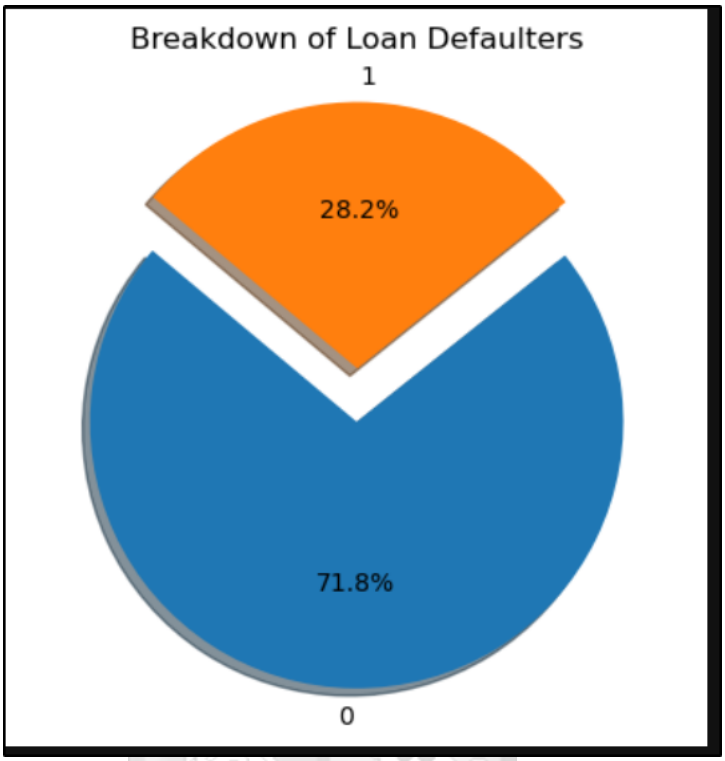


Figure 3. 7: Percentage of Default and Non-Default

Feature Elimination

Relevant features likely to contain meaningful information were selected for the predictive task. The dataset's many features aided this, and a dimensionality reduction using Recursive Feature Elimination was performed.

The Recursive Feature Elimination process helped prevent overfitting by eliminating the least important features, improving the model's interpretability. Therefore, the feature size was reduced to 46890 columns and 25 rows as seen in Figure 3.7.

Figure 3. 8: Selection of Training Features

3.2.3 Data Modeling

In this process, data cleaning was performed, filling missing values with the mean and other inconsistencies, leading to standardization of data analysis. The data was split into 80% training and 20% testing data, which determined the model's effectiveness. Modeling also transformed categorical data attributes into numerical data using Label Encoder.

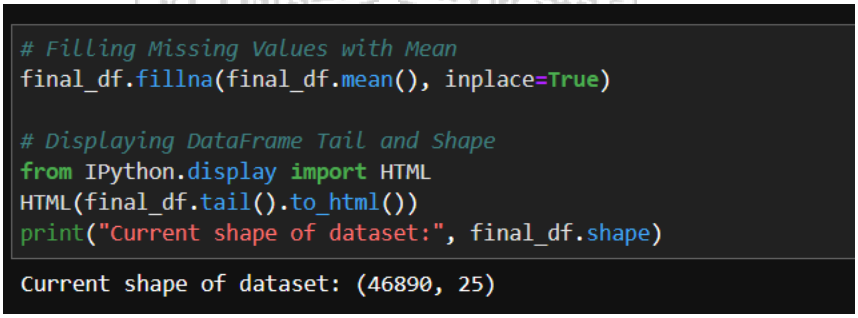


Figure 3. 9: Filling Missing Values with Mean

```python
label_encoder = LabelEncoder()

# Convert categorical columns to integers using label encoding
final_df['EmploymentDurationCurrentEmployer'] = label_encoder.fit_transform(final_df['EmploymentDurationCurrentEmployer'])
final_df['Rating'] = label_encoder.fit_transform(final_df['Rating'])
final_df['NewCreditCustomer'] = label_encoder.fit_transform(final_df['NewCreditCustomer'])
final_df['Restructured'] = label_encoder.fit_transform(final_df['Restructured'])
final_df['CreditScoreEsMicroL'] = label_encoder.fit_transform(final_df['CreditScoreEsMicroL'])
```

Figure 3. 10: Converting Categorical values to Numerical values.

```python
print("Data types of each column:")
print(final_df.dtypes)
```

```
Data types of each column:
Age                                 int64
LoanDuration                        int64
NewCreditCustomer                   int64
VerificationType                  float64
Gender                            float64
Interest                          float64
UseOfLoan                           int64
Amount                            float64
AppliedAmount                     float64
LanguageCode                        int64
Education                         float64
EmploymentDurationCurrentEmployer     int32
Rating                                int32
MaritalStatus                     float64
EmploymentStatus                  float64
OccupationArea                    float64
HomeOwnershipType                 float64
CreditScoreEsMicroL                 int32
ExistingLiabilities                 int64
DebtToIncome                      float64
IncomeTotal                       float64
Restructured                        int64
NoOfPreviousLoansBeforeLoan       float64
ModelVersion                      float64
Default                             int64
dtype: object
```

Figure 3. 11: Selected Features Datatypes

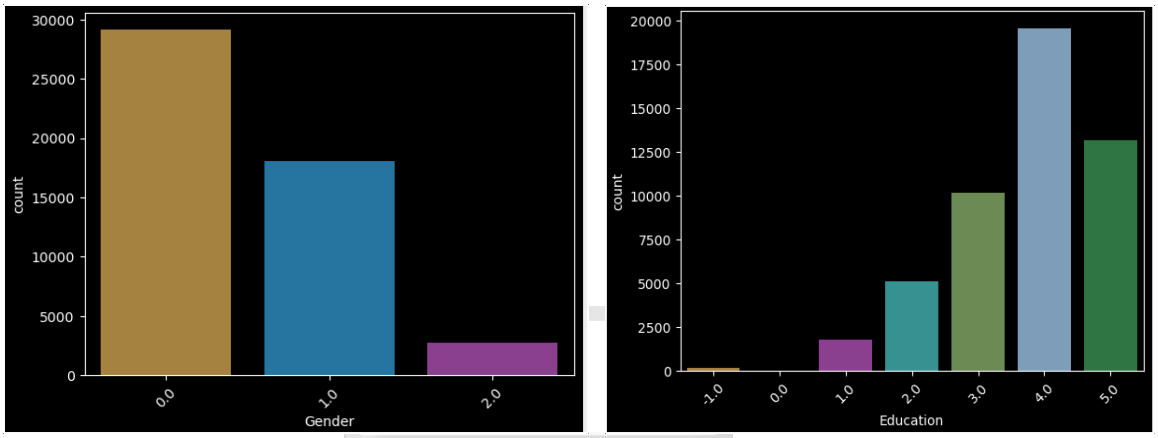
```python
# Training , Test Split

X_train, X_test, y_train, y_test = train_test_split(final_df.iloc[:, :-1], final_df.iloc[:, -1], test_size=0.2, random_state=42)
```

Figure 3. 12: Splitting the Dataset into 80% Training and 20% Testing

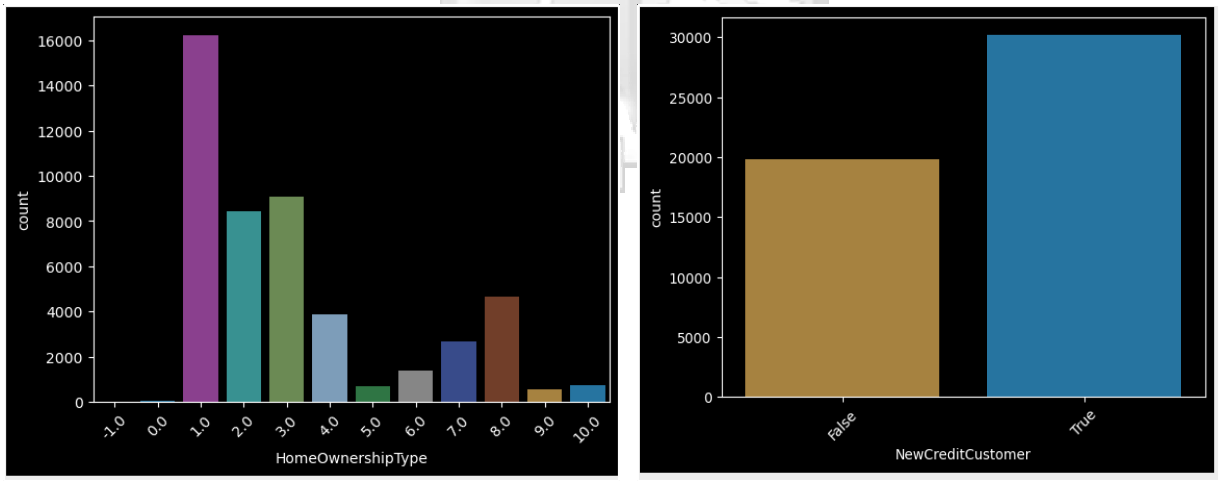
3.2.4 Univariate Data Analysis

This is the analysis between individual variables. Here, Gender, Education, Home Ownership, and New Credit Customer features were analyzed. The analysis was done using bar unstacked bar graphs for proper visualization.



(a) (b)

Figure 3. 13: Univariate Analysis of (a) Gender and (b) Education



(a) (b)

Figure 3. 14: Univariate Analysis of (a) Home Ownership and (b) New Credit Customer

General Observations

13,302 of the loans are defaults, representing 28.2% of the total value counts, while non-defaults are 33,688, representing 71.8%.

Male applicants account for 58.3%, females for 36.2%, and others for 5.5%.

39% of borrowers have attained secondary education, while 26.4% have higher education. 20.4% are vocational, 10% have basic education, and 3.5% have primary education. Overall, most of the borrowers are educated.

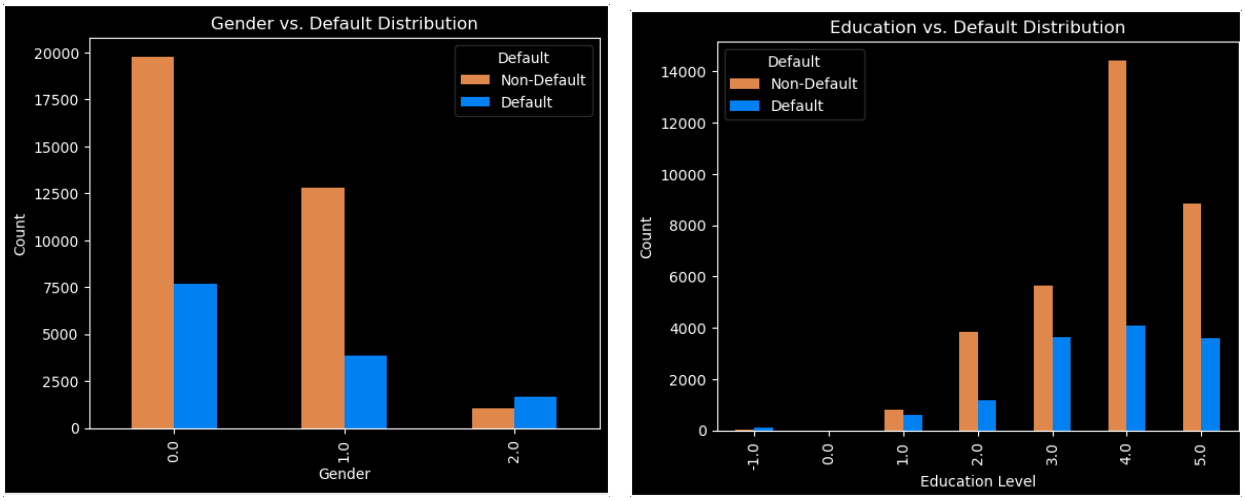
33.6% of the borrowers are homeowners, 18.8% are tenants in pre-furnished properties, 17.4% live with parents, 9.7% are on the mortgage, and 8% are tenants in unfurnished properties—the rest account for negligible percentages.

New borrowers account for 60%, meaning it is their first time borrowing.

More than 59% of borrowers are fully employed, and 28% are employed.

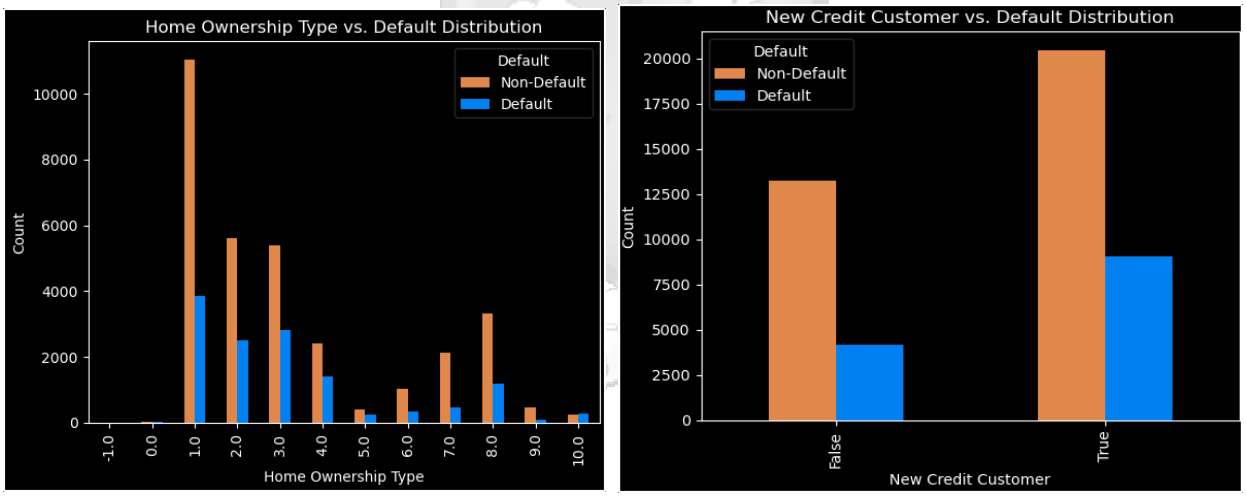
3.2.5 Bivariate Data Analysis

This is a data analysis between two relationships: predictor and target variables. Gender, Education, Home Ownership, and New Credit Customer features were analyzed against the Default variable.



(a) (b)

Figure 3. 15: Bivariate Analysis For (a) Gender and (b) Education vs Default



(a) (b)

Figure 3. 16: Bivariate Analysis of (a) Home Ownership and (b) New Credit Customer vs Default.

General Observation

Male borrowers default more than females.

Most defaulters have secondary education, followed by vocational, higher, basic, and primary education, respectively.

Homeowners account for most defaults, followed by tenants with pre-furnished property and those living with parents.

New credit customers also default more than seasoned borrowers.

Fully employed borrowers account for more defaults than employed borrowers.

3.2.6 Data Balancing using SMOTE Technique

As seen in Figure 3.6, the target variable 'Default' is imbalanced, with the number of non-defaults higher than defaults. This leads to a model bias towards the majority class. SMOTE was used to address the class imbalance problem.

The algorithm selects random minority examples close to feature space and selects random neighbors using K-Nearest Neighbor to balance the classes (Fernández et al., 2018).

```
sm = SMOTE(random_state=42, k_neighbors=5)
X_resampled, y_resampled = sm.fit_resample(X_train, y_train)

# Select features and target variable
X = final_df[features_to_keep]
y = final_df[target_variable]

# Apply SMOTE
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)

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

Figure 3. 17: SMOTE for Data Imbalance

3.2.7 Evaluation Metrics

The evaluation metrics to use for this research will include the following.

I. Accuracy

This research study is modeled as a binary classification. Accuracy is used as one of the test evaluation metrics; it identifies compelling pattern

correlation among data samples using training data. It will be derived using the formula:

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population}$$

(ii)

II. Precision

The precision score measures the consistency of positive predictions (Minaee, 2019). It shows that the classifier is good enough that it can be computed by the formula seen in (iii):

$$Precision = \frac{\sum True\ Positive}{\sum Predicted\ Condition\ Positive}$$

(iii)

III. Recall (Sensitivity)

It gauges whether the positive percentage was correctly identified (Alazab, 2015). It is the fraction of true positives and the sum of true positives and false negatives, as seen in (iv).

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

(iv)

IV. Specificity

This is the percentage of true negatives divided by the summation of true and false positives. As seen in (v), this shows how a model avoids classification.

$$Specificity = \frac{True\ Negative}{\text{True Negative} + \text{False Positive}} \quad \text{(v)}$$

V. Confusion Matrix

This is a common way to evaluate a model with binary outcomes. The default cases observed positives, while the non-observed ones were negatives (Finance, 2017). The True Positive represents the defaulted customer predicted to have defaulted, and the True Negative if non-default has been predicted as non-default. The False Positive is if non-default has been predicted to default, and the False Negative is if defaulted, which is predicted as non-default. Table 3.2 shows a confusion matrix.

Table 3. 2: Confusion Matrix

Actual Class	Prediction Results	
	Positive Class (Default)	Negative Class (Normal)
Positive Class (Default)	TP	FN
Negative Class (Normal)	FP	TN

3.3 Software Development Methodology

The software methodology adopted a CRISP-DM model widely used for machine learning and data mining projects.

The stages are iterative, meaning each stage may need revisiting as new information becomes available or problems arise. The model provides a structured approach to planning and executing data mining projects, but it needs to be prescriptive and can be adapted to suit different contexts and needs.

Figure 3. 18: CRISP-DM Methodology. Adapted From. (Hotz, 2023).

Steps Description

Business Understanding: Understand the business problem, objectives, requirements, and constraints. Define the project's scope and formulate a plan to achieve the goals. This is explicit in the background and literature.

Data Understanding: Collect and explore the data that will be used for modeling. Verify data quality, completeness, and relevance. Identify data issues and prepare the data for modeling. This step also involves finding meaningful patterns in the data and visualizing it using bar graphs to familiarize oneself with it.

Data Preparation: Select, clean, transform, and integrate the data to create a dataset suitable for modeling. Create new features or variables that may improve model performance. This process can take time.

Modeling: Select appropriate modeling techniques and build predictive models using the prepared dataset. Evaluate model performance using appropriate metrics as seen in Figure 3.17.

Evaluation: Assess the quality of the models and determine whether they meet the business objectives. Select the best model (s) for deployment.

Deployment: In this final step, Integrate the selected model (s) into an application. In this step, the deployed model will be used for academic purposes and not in a real-life case. Through analysis and discussion of results, whether they meet the business needs in step 1, and finally, a conclusion on whether the research objectives were achieved.

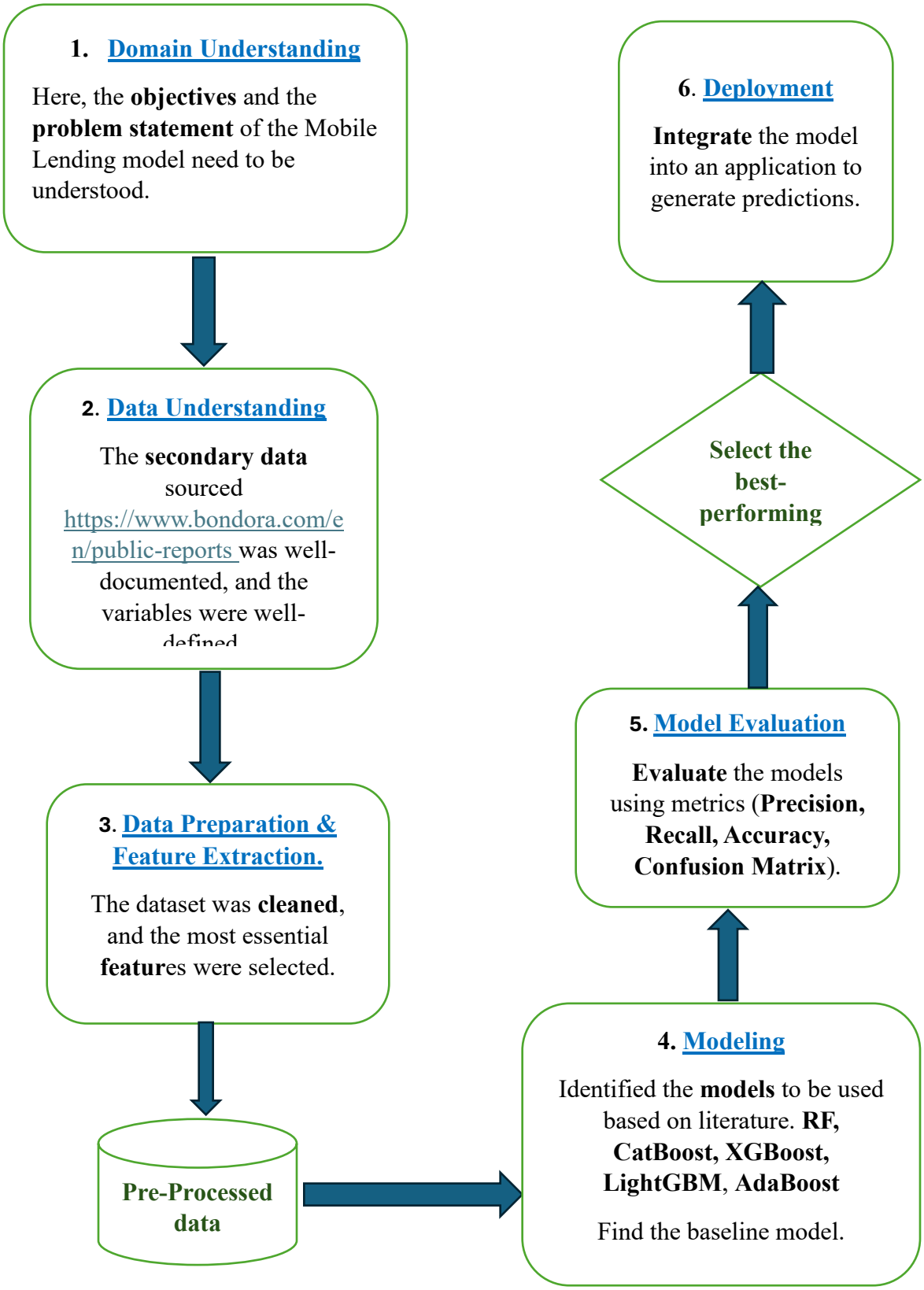


Figure 3. 19: Software Development Methodology Steps. Author Preparation

3.4 Design Requirements

For any software development to succeed, the design requirements must be well defined, and proper hardware and software requirement specifications must be listed since they act as a blueprint for creating system needs. These are also essential in planning to meet the intended purpose of this research from a computing perspective.

3.4.1 Hardware Requirements

The hardware requirements used for this research include a laptop with 8GB RAM with storage of 1TB SSD for faster data processing, a CPU multi-core processor intel core i7, and an internet connection to download the datasets and libraries.

3.4.2 Software Requirements

The software requirements for the machine learning part include the Operating System Windows 10 as a personal Preference, the Development Environment, and Visual Studio code as a preferred code editor. The machine learning models were implemented on the latest version of Python 3.12 on Jupyter Lab with Anaconda package manager to manage Python packages, deployed on the Brave browser with the necessary machine learning libraries- pandas, numpy, sklearn, Keras, and seaborn.

The prototype software was deployed using Pycharm Community Edition and a version control tool, Github, to upload the research code and track the changes.

3.5 Research Quality and Validity

The data source was secondary; it can be verified since it is open and collected from a reputable online repository for research purposes. The data cleaning process can also be verified and was well documented. A study budget was prepared to ensure the resources were well utilized, including the milestones, and creating realistic timelines for the study. The research objectives have governed the aspect of research design.

For research validity, the methodology was consistent throughout the study, and the data source can be cross-checked to validate the accuracy of the results. This was achieved through the evaluation metrics mentioned in Chapter 3, additional model validation methods were also discussed in Chapter 6 of this document. External examiners can also verify the credibility of the results. The source has been cited to ensure transparency in the secondary data. The source code for the user interface is attached in Appendix D. At the same time, the rest of the machine learning model code was also posted on GitHub, improving transparency, and verifying findings.

3.6 Ethical Considerations

The researcher understands the potential bias in the credit scoring models where qualified borrowers can be overlooked. These biases exist in most secondary data. A sensitivity analysis was conducted to understand how different assumptions on bias affect the model. This can also increase the robustness of the findings. Ethical considerations in data analysis were also considered in the machine learning pre-processing and model training. The researcher ensured that all references to other researchers and authors were cited and well-documented; a similarity report was also attached in Appendix A.

The research was submitted to the supervisor for continuous improvement and proper guidance. After making the necessary corrections after the proposal defense, the researcher applied for ethical clearance as part of the university's standard operating procedure. The ethical clearance letter is attached to the Appendix B section of this document. Later a research license was obtained from NACOSTI as a statutory mandate.

Chapter 4: System Design and Architecture

4.1 Introduction

This chapter outlines the design and architecture of the web-based Mobile credit scoring system based on the conceptual framework presented in Figure 2.7. The model can be used by stakeholders in the Mobile credit lending industry. The diagrams show interactions between various users and the system.

4.2 Requirement Analysis

This involves users' expectations to ensure the system considers stakeholders' needs based on the research objectives. Divided into functional and non-functional requirements.

4.2.1 Functional Requirements

These are statements of the services a system must provide to its users.

I. The system should allow the borrower to fill in the application form.
II. The system should implement a credit scoring algorithm based on the borrower's input to determine the borrower's creditworthiness.
III. The system should provide a clear credit decision output of either approval or denial and give a feedback mechanism to the borrower.
IV. The system should allow the borrower to view the output of their application.

4.2.2 Non-Functional Requirements

This is a statement of the operational system constraints.

I. Usability - The user interface for interaction with the Mobile Lending credit system should be user-friendly for the stakeholders.
II. Data Privacy- Data privacy should be prioritized to ensure that user information is secure.
III. Scalability- The system should be scalable to handle the growing number of applications without performance sacrifices.

IV. Performance- The system should be able to make timely decisions and respond to questions during the application process.

V. Maintainability- The system should be easy to maintain and install updates based on changing business needs.

VI. Logging and Monitoring- The system should include surveillance of system activities, including system trails, model performance, anomalies, and analysis.

VII. Quality Data Training- The system should ensure the reliability and accuracy of the training data since it directly impacts the model's effectiveness.

4.3 System Architecture

Figure 4.1 illustrates the mobile lending system architecture. The raw data is pre-processed and then trained using machine learning models Random Forest, XGBoost, LightGBM, and CatBoost. Finally, the trained model is used to develop a web application where various users can interact with the system.

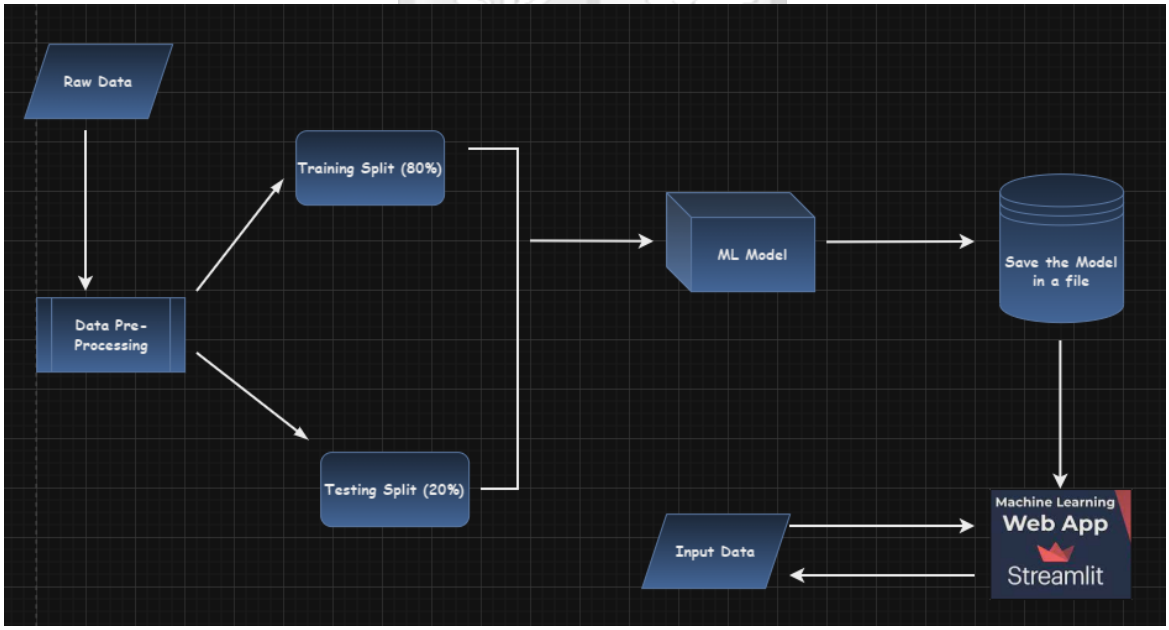


Figure 4. 1: System Architecture Diagram. Author Illustration

4.4 Use-Case Diagram

A use case diagram represents the steps in a specific system where the actor initiates a use case by requesting the system to perform a process. The objective is to identify the actors and the process they initiate. In this case, the system administrator and the borrower are the actors in the mobile lending system. Each use case has a use case description in a tabular format with some description.

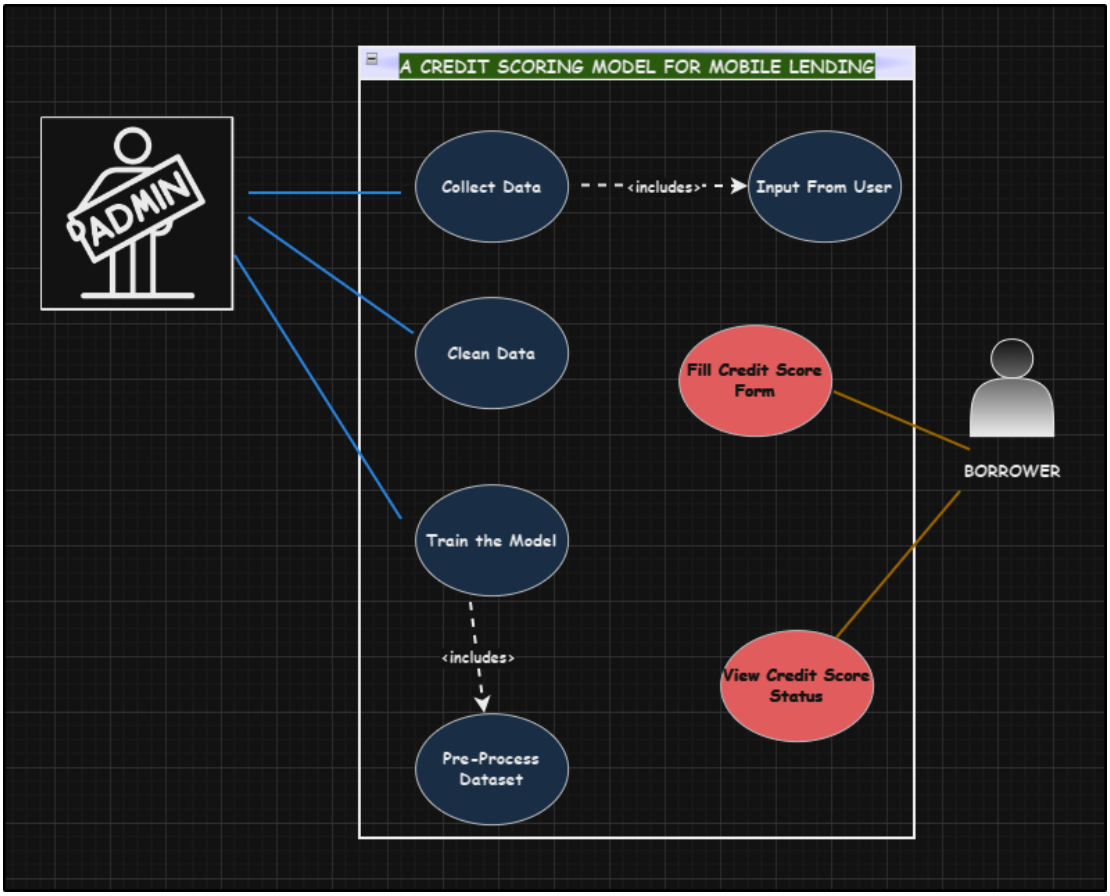


Figure 4. 2: Use Case Diagram

Table 4. 1: Data Collection Description

Use Case:	**Collect Data**
Primary Actors:	System Administrator
Description:	Describes how the System Administrator will collect data.
Pre-Condition:	Data is relevant to Mobile Credit Scoring
Post-Condition:	System Administrator Identifies and Retrieves the data.

Main Scenarios

Actor Responsibility	**System Responsibility**
1. The System Administrator identifies the data.	
2. The System Administrator retrieves the identified data.	
	3. The system saves the identified data.

Table 4. 2: Data Cleaning Description

Use Case:	**Clean Data**
Primary Actors:	System Administrator
Description:	Describes how the System Admin will clean raw data
Pre-Condition:	Sufficient data to clean
Post-Condition:	Pre-Processed data

Main Scenarios

Actor	**System**
1. The System Admin imports the Python libraries.	
2. The System Administrator imports the dataset.	
3. System Admin checks the dataset for missing values and fills them with either mean, median, mode, or interpolation.	
4. Splitting the data into training and testing sets.	
	5. The system saves the features extracted from the dataset.

Table 4. 3: Model Training Description

Use Case:	**Train the Model**
Primary Actors:	System Administrator
Description:	Describes how the System Admin will train the model
Pre-Condition:	Training System available, Pre-Processed data
Post-Condition:	Trained model can predict the borrower's creditworthiness
Main Scenarios	
Actor	**System**
1. The System Admin selects the pre-processed data.	
2. The System Admin selects training and testing percentages.	
3. Admin selects output format.	
4. Admin runs the training command.	
	5. The system splits the data into training and testing sets per the admin command.
	6. The system trains the model as per admin command.
	7. The system uses the test data to validate the trained model.
	8. The system outputs the trained model and saves it.

Table 4. 4: Borrower Creditworthiness Description

Use Case:	**Fill out the Credit Application Form**
Primary Actors:	The Borrower
Description:	Describes how the borrower will fill out the Credit Scoring form.
Pre-Condition:	System Available
Post-Condition:	Borrower can view their Credit Application Status

Main Scenarios

Actor	**System**
1. Fill out the form for credit scoring prediction.	
2. Submit the filled form.	
	3. The system Predicts the borrower's creditworthiness.
	4. The system displays the status of the credit application to the borrower.
5. Borrower views the status of their application.	

4.5 System Sequence Diagram

This dynamic model of a use case shows interactions among classes over a certain period (Al-Fedaghi, 2021). It graphically documents the use case by showing classes, messages, and the timing of the messages.

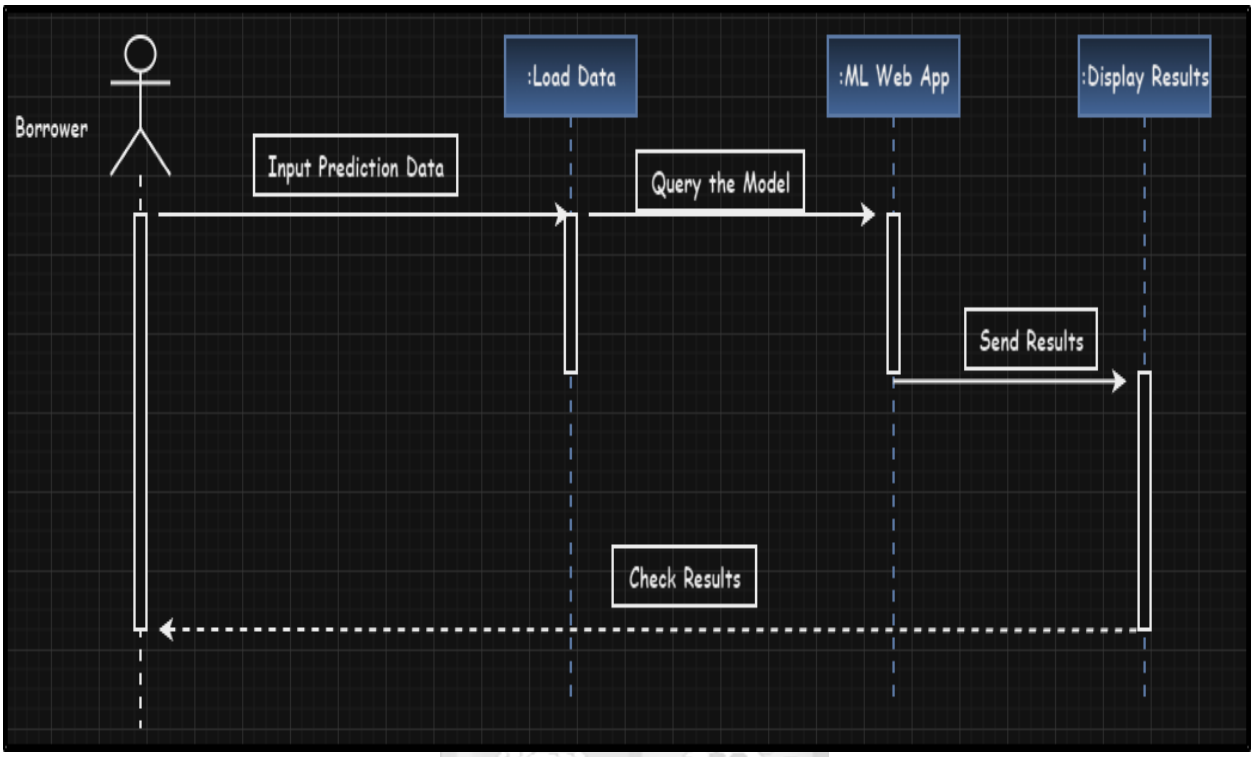


Figure 4. 3: User Illustration of Sequence Diagram

4.6 Context Diagram

This is a top-level view of an information system that shows its boundaries and scope. It shows the entities of the proposed mobile lending system and how the user interacts with it. A context diagram at level 1 in Figure 4.4 breaks down the single processes into sub-processes, showing the user's interaction with the system.

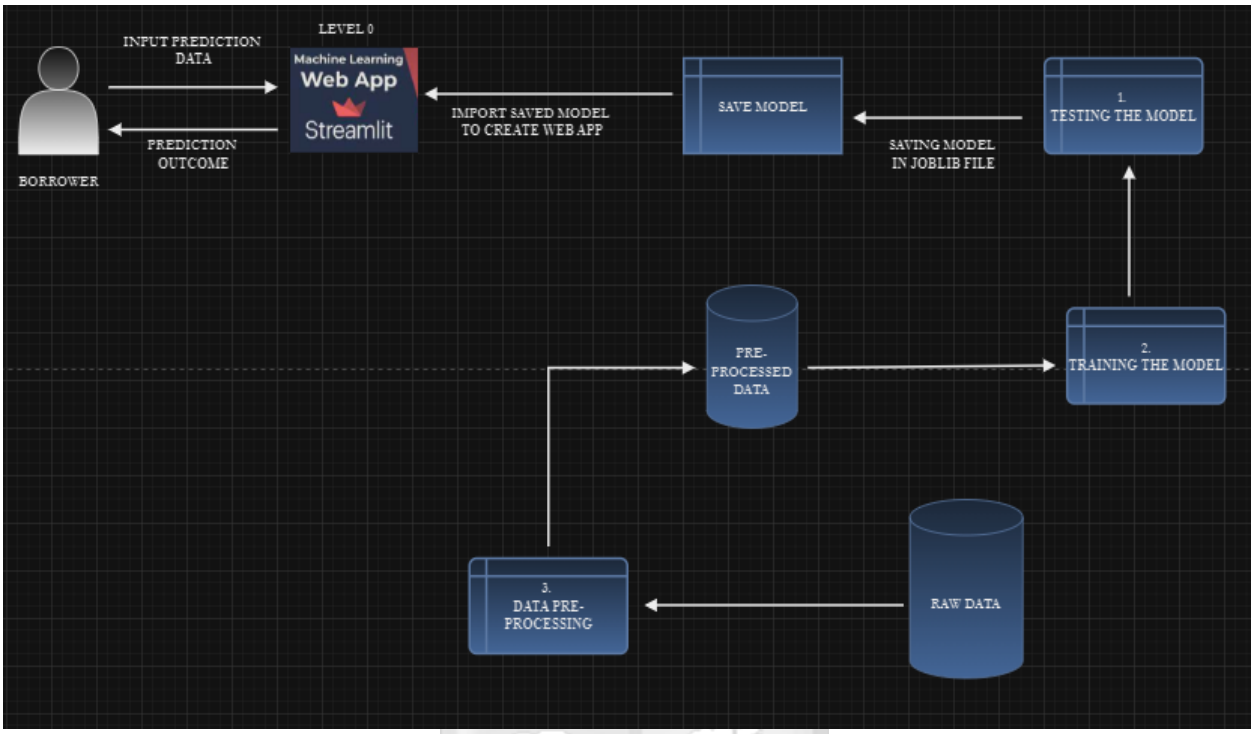


Figure 4. 4: Level 1 Context Diagram

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter gives a detailed look at the training results of the credit scoring model developed. The model was then deployed in the Streamlit Web Application to make real-time predictions. The web application was developed using the CatBoost algorithm. The System implementation was done and tested to ensure the research objectives were accomplished.

5.2 Model Training and Results

After pre-processing, data was trained using five machine learning algorithms, the outcomes of which are described below. The best model was used to implement the web-based machine learning application.

5.2.1 Random Forest Results

Figure 5.1 shows the training results achieved on the Random Forest algorithm. The accuracy rate, which measures the number of correctly classified instances, was 86.

```python
#Random Forest

# Initialize Random Forest Classifier
model = RandomForestClassifier(n_estimators=100, random_state=42)

# Train the model
model.fit(X_train, y_train)

# Predict on the test set
y_pred = model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Print classification report
print("Classification Report:")
print(classification_report(y_test, y_pred))

# Generate Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(conf_matrix)
```

```
Accuracy: 0.8627931136835857
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.85      0.86      6792
           1       0.85      0.87      0.86      6684

    accuracy                           0.86     13476
   macro avg       0.86      0.86      0.86     13476
weighted avg       0.86      0.86      0.86     13476
```

Figure 5. 1: Random Forest Results

The Confusion Matrix, which evaluates a model with binary outcomes for Random Forest, is shown in Table 5.1, and the interpretation is as follows.

The model correctly classified 86% of the data, whereas 14% was misclassified.

Type 1 Error

The model had a 7.6% False Positive error, meaning 994 defaults were classified as non-defaults, but they are defaults. If these borrowers are issued a loan, this results in revenue loss for the mobile credit lenders.

Type 2 Error

The model had a 6.4% False Negative error, meaning 855 borrowers were classified as defaults, but they did not default. This does not lead to revenue loss for mobile credit lenders; instead, it restricts these borrowers from being issued a loan despite their ability to fulfill their loan obligations.

Table 5. 1: Random Forest Confusion Matrix

Confusion Matrix N= 13476	**Non-Default**	**Default**
Non-Default	True Positive **5798 (43%)**	False Positive (Type 1 error) **994 (7.6%)**
Default	False Negative (Type 2 error) **855 (6.4%)**	True Negative **5829 (43%)**

The feature importance was done to identify features with the most influence on prediction.

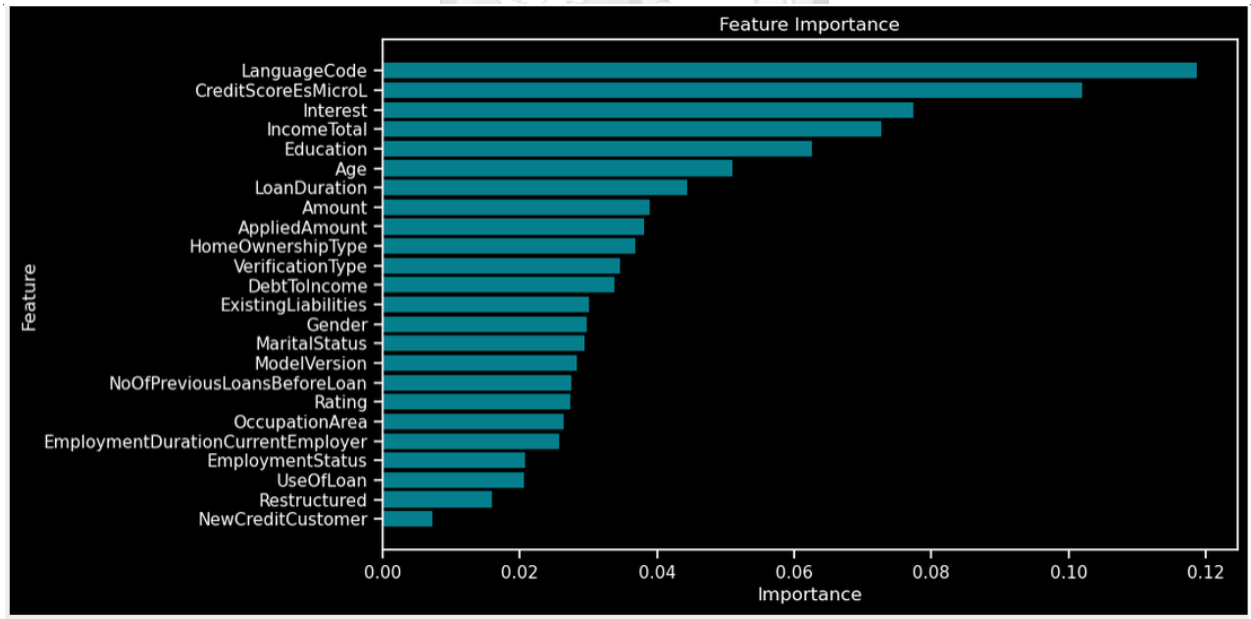


Figure 5. 2: Feature Importance for Random Forest

5.2.2 XGBoost Results

Figure 5.3 shows the training results achieved on the XGBoost algorithm. The accuracy rate, which measures the number of correctly classified instances, was 85%.

```
# XGBoost

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42)

# Train a model
xgb_classifier = xgb.XGBClassifier(random_state=42)
xgb_classifier.fit(X_train, y_train)

# Make predictions
y_pred = xgb_classifier.predict(X_test)

# Evaluate the model
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.86      0.85      6792
           1       0.86      0.85      0.85      6684

    accuracy                           0.85     13476
   macro avg       0.85      0.85      0.85     13476
weighted avg       0.85      0.85      0.85     13476
```

Figure 5. 3: XGBoost Results

The Confusion Matrix, which evaluates a model with binary outcomes for XGBoost, is shown in Table 5.2, and its interpretation is as follows.

The model correctly classified 85% of the data, while 15% was misclassified.

Type 1 Error

The model had a 7.3% False Positive error, meaning 959 defaults were classified as non-defaults, but they are defaults. If these borrowers are issued a loan, this results in revenue loss for the mobile credit lenders.

Type 2 Error

The model had a 7.7% False Negative error, meaning 1024 borrowers were classified as defaults, but they did not default. This does not lead to revenue loss for mobile credit lenders; instead, it restricts these borrowers from being issued a loan despite their ability to fulfill their loan obligations.

Table 5. 2: XGBoost Confusion Matrix

Confusion Matrix N= 13476	**Non-Default**	**Default**
Non-Default	True Positive **5833 (43%)**	False Positive (Type 1 error) **959 (7.3%)**
Default	False Negative (Type 2 error) **1024 (7.7%)**	True Negative **5660 (42%)**

The feature importance was done to identify features with the most influence on prediction.

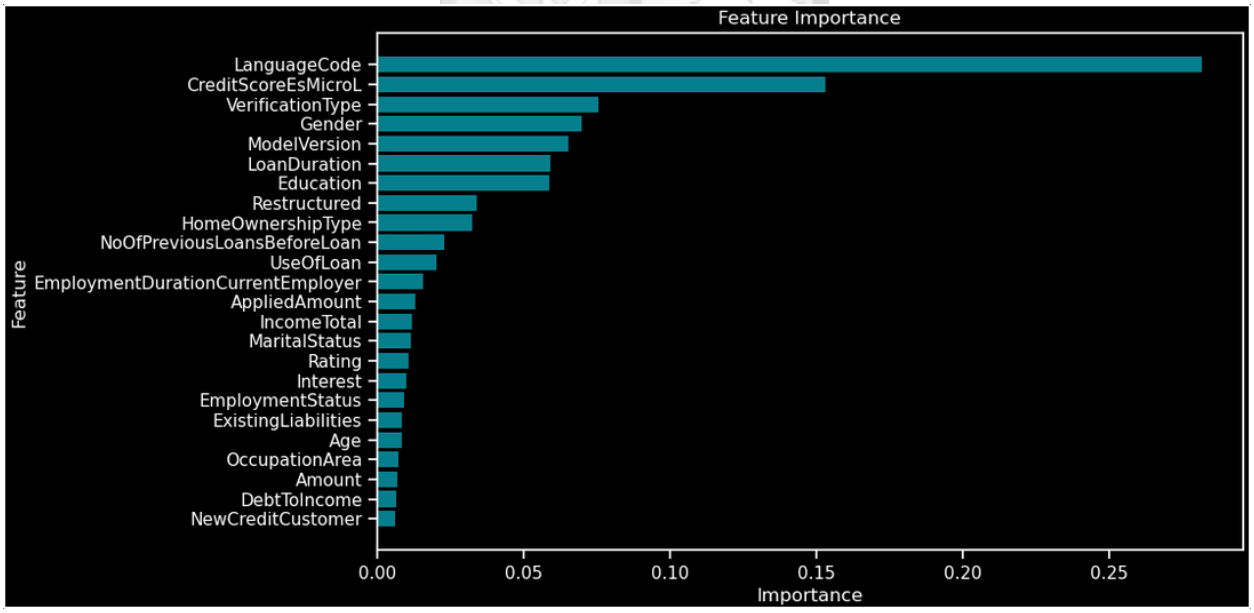


Figure 5. 4: Feature Importance for XGBoost

5.2.3 LightGBM Results

Figure 5.5 shows the training results achieved on the LightGBM algorithm. The accuracy rate, which measures the number of correctly classified instances, was 85%.

```
# lightgbm

# Select features and target variable
X = final_df[features_to_keep]
y = final_df[target_variable]

# Train a model
lgbm_classifier = LGBMClassifier(random_state=42)
lgbm_classifier.fit(X_train, y_train)

# Make predictions
y_pred = lgbm_classifier.predict(X_test)

# Evaluate the model
print(classification_report(y_test, y_pred))
```

```
[LightGBM] [Info] Number of positive: 27004, number of negative: 26896
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.075166 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 3807
[LightGBM] [Info] Number of data points in the train set: 53900, number of used features: 24
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.501002 -> initscore=0.004007
[LightGBM] [Info] Start training from score 0.004007
              precision    recall  f1-score   support

           0       0.85      0.85      0.85      6792
           1       0.85      0.85      0.85      6684

    accuracy                           0.85     13476
   macro avg       0.85      0.85      0.85     13476
weighted avg       0.85      0.85      0.85     13476
```

Figure 5. 5: LightGBM Results

Table 5.3 shows the Confusion Matrix, which evaluates a model with binary outcomes for LightGBM. The interpretation is as follows.

The model correctly classified 85% of the data, whereas 15% was misclassified.

Type 1 Error

The model had a 7.6% False Positive error, meaning 1026 defaults were classified as non-defaults, but they are defaults. If these borrowers are issued a loan, this results in revenue loss for the mobile credit lenders.

Type 2 Error

The model had a 7.3% False Negative error, meaning 978 borrowers were classified as defaults, but they did not default. This does not lead to revenue loss for mobile credit lenders; instead, it restricts these borrowers from being issued a loan despite their ability to fulfill their loan obligations.

Table 5. 3: LightGBM Confusion Matrix

Confusion Matrix N= 13476	**Non-Default**	**Default**
Non-Default	True Positive **5766 (43%)**	False Positive (Type 1 error) **1026 (7.6%)**
Default	False Negative (Type 2 error) **978 (7.3%)**	True Negative **5706 (42%)**

The feature importance was done to identify features with the most influence on prediction.

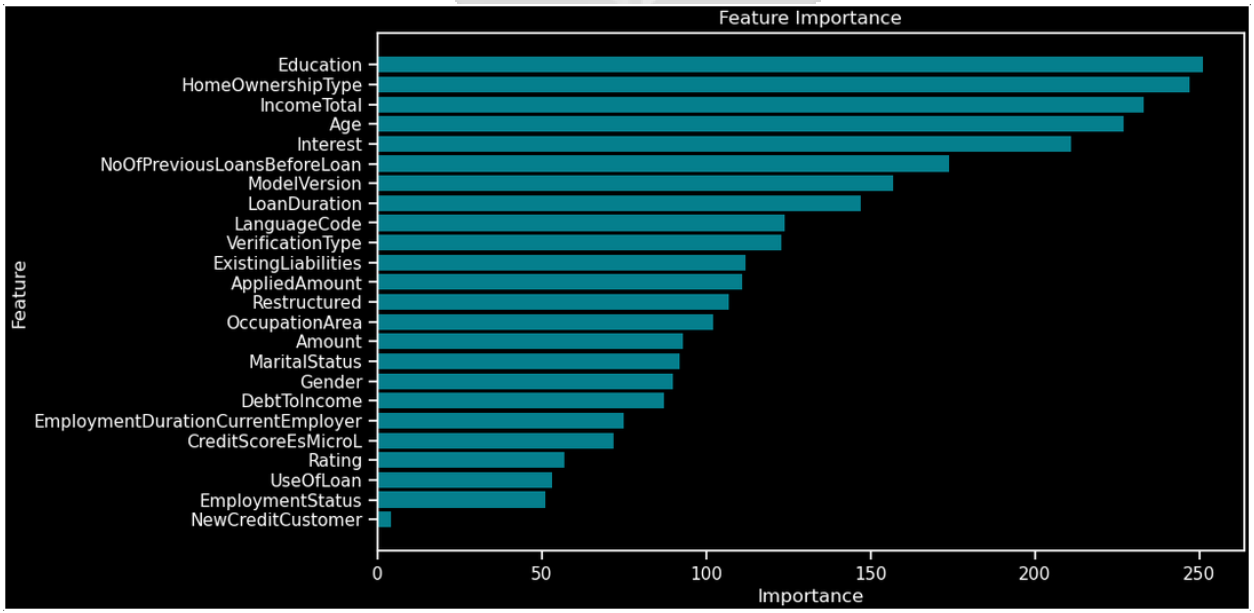


Figure 5. 6: Feature Importance for LightGBM

5.2.4 CatBoost Results

Figure 5.7 shows the training results achieved on the CatBoost algorithm. The accuracy rate, which measures the number of correctly classified instances, was 86%.

```
# Select features and target variable
X = final_df[features_to_keep]
y = final_df[target_variable]

# Train a model
catboost_classifier = CatBoostClassifier(random_state=42)
catboost_classifier.fit(X_train, y_train)

# Make predictions
y_pred = catboost_classifier.predict(X_test)

# Evaluate the model
print(classification_report(y_test, y_pred))
```

```
991:    learn: 0.2369022        total: 21.8s    remaining: 176ms
992:    learn: 0.2368588        total: 21.9s    remaining: 154ms
993:    learn: 0.2368071        total: 21.9s    remaining: 132ms
994:    learn: 0.2367807        total: 21.9s    remaining: 110ms
995:    learn: 0.2367308        total: 21.9s    remaining: 88ms
996:    learn: 0.2366922        total: 21.9s    remaining: 66ms
997:    learn: 0.2366638        total: 22s      remaining: 44ms
998:    learn: 0.2366186        total: 22s      remaining: 22ms
999:    learn: 0.2365843        total: 22s      remaining: 0us
              precision    recall  f1-score   support

           0       0.85      0.86      0.86      6792
           1       0.86      0.85      0.86      6684

    accuracy                           0.86     13476
   macro avg       0.86      0.86      0.86     13476
weighted avg       0.86      0.86      0.86     13476
```

Figure 5. 7: CatBoost Results

Table 5.4 shows the Confusion Matrix, which evaluates a model with binary outcomes for CatBoost. Its interpretation is as follows.

The model correctly classified 86% of the data, while 14% was misclassified.

Type 1 Error

The model had a 7% False Positive error, meaning 934 defaults were classified as non-defaults, but they are defaults. If these borrowers are issued a loan, this results in revenue loss for the mobile credit lenders.

Type 2 Error

The model had a 7.4% False Negative error, meaning 994 borrowers were classified as defaults, but they did not default. This does not lead to revenue loss for mobile credit lenders; instead, it restricts these borrowers from being issued a loan despite their ability to fulfill their loan obligations.

Table 5. 4: CatBoost Confusion Matrix

Confusion Matrix N= 13476	**Non-Default**	**Default**
Non-Default	True Positive **5858 (43.5%)**	False Positive (Type 1 error) **934 (7%)**
Default	False Negative (Type 2 error) **994 (7.4%)**	True Negative **5690(42%)**

The feature importance was done to identify features with the most influence on prediction.

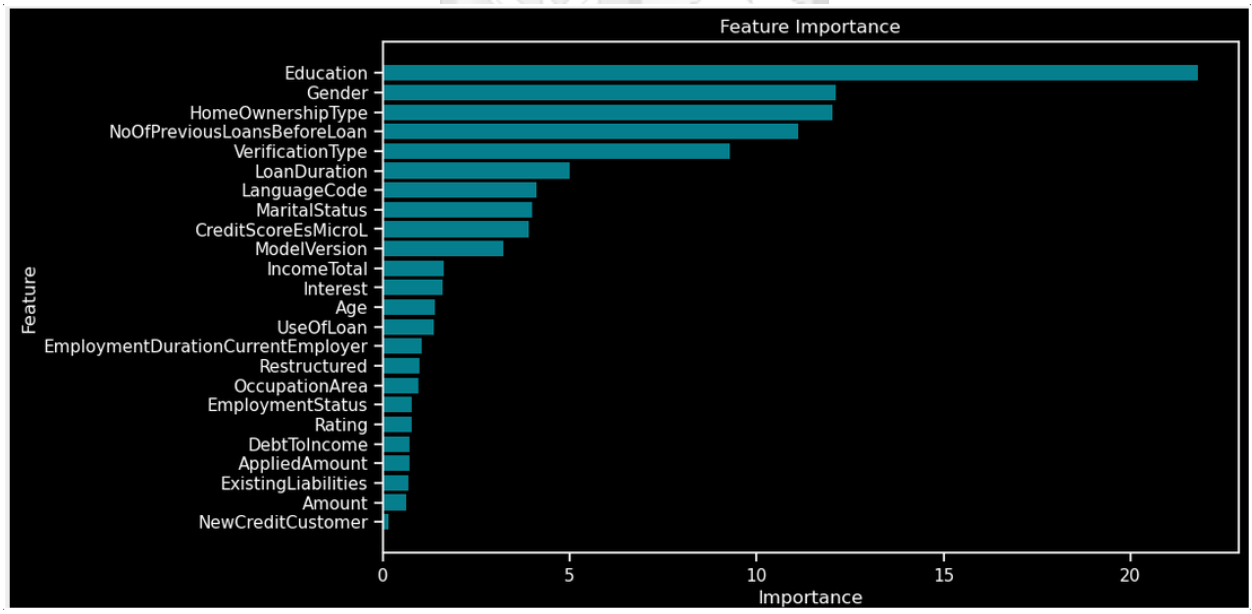


Figure 5. 8: Feature Importance for CatBoost

5.2.5 AdaBoost Results

Figure 5.9 shows the training results achieved on the AdaBoost algorithm. The accuracy rate, which measures the number of correctly classified instances, was 83%.

```
# Initialize AdaBoost Classifier
model = AdaBoostClassifier(n_estimators=100, random_state=42)

# Train the model
model.fit(X_train, y_train)

# Predict on the test set
y_pred = model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Print classification report
print("Classification Report:")
print(classification_report(y_test, y_pred))

# Generate Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(conf_matrix)
```

```
Accuracy: 0.8313297714455328
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.82      0.83      6792
           1       0.82      0.85      0.83      6684

    accuracy                           0.83     13476
   macro avg       0.83      0.83      0.83     13476
weighted avg       0.83      0.83      0.83     13476
```

Figure 5. 9: AdaBoost Results

Table 5.5 shows the Confusion Matrix, which evaluates a model with binary outcomes for AdaBoost. Its interpretation is as follows.

The model correctly classified 83% of the data, while 17% was misclassified.

Table 5. 5: AdaBoost Confusion Matrix

Confusion Matrix N= 13476	**Non-Default**	**Default**
Non-Default	True Positive **5551 (41%)**	False Positive (Type 1 error) **1241 (9.2%)**
Default	False Negative (Type 2 error) **1032 (7.7%)**	True Negative **5652 (42%)**

Type 1 Error

The model had a 9.2% False Positive error, meaning 1241 defaults were classified as non-defaults, but they are defaults. If these borrowers are issued a loan, this results in revenue loss for the mobile credit lenders.

Type 2 Error

The model had a 7.7% False Negative error, meaning 1032 borrowers were classified as defaults, but they did not default. This does not lead to revenue loss for mobile credit lenders; instead, it restricts these borrowers from being issued a loan despite their ability to fulfill their loan obligations.

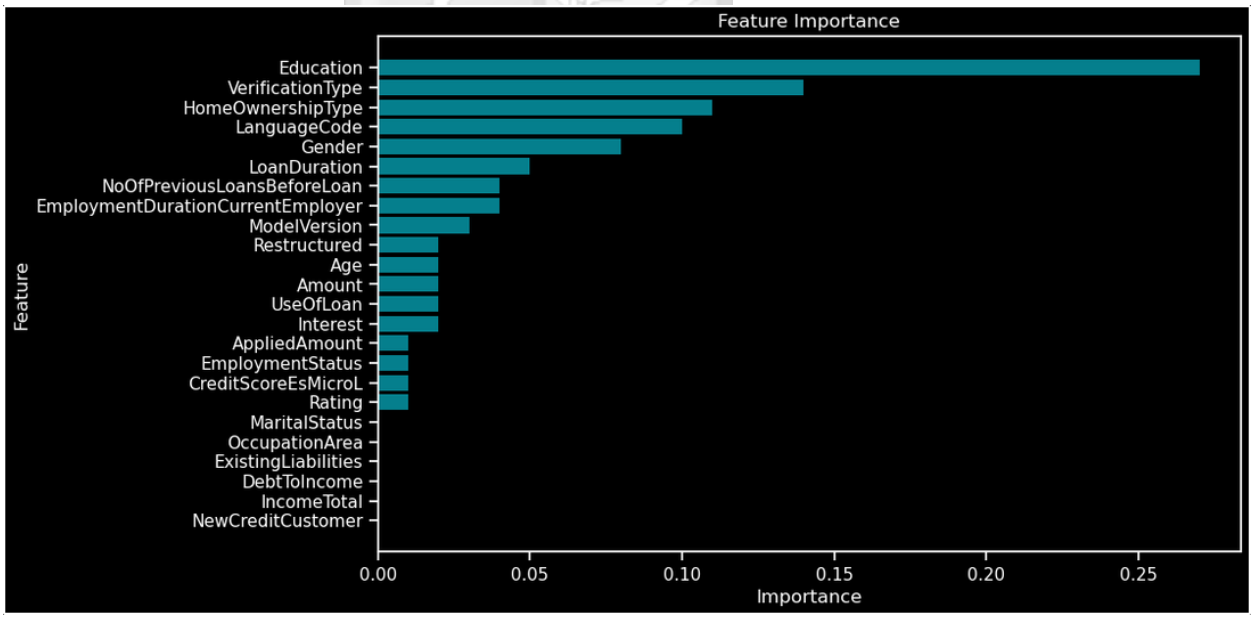


Figure 5. 10: Feature Importance for AdaBoost

5.3 Classification Results

Table 5.5 displays the classification results for the trained models.

Table 5. 6: Classification Results for Default Class

Classification Method	Correct Classification		Incorrect Classification		Score (%)			
	TP	TN	FP	FN	Accuracy	Precision	Recall	F1-Score
Random Forest	5798	5829	994	855	86	85	87	86
XGBoost	5833	5660	959	1024	85	86	85	85
CatBoost	5858	5690	934	994	86	86	85	86
LightGBM	5766	5706	1026	978	85	85	85	85
AdaBoost	5551	5652	1241	1032	83	82	85	83

Table 5. 7: Classification Results for Non-Default Class

Classification Method	Correct Classification		Incorrect Classification		Score (%)			
	TP	TN	FP	FN	Accuracy	Precision	Recall	F1-Score
Random Forest	5798	5829	994	855	86	87	85	86
XGBoost	5833	5660	959	1024	85	85	86	85
CatBoost	5858	5690	934	994	86	85	86	86
LightGBM	5766	5706	1026	978	85	85	85	85
AdaBoost	5551	5652	1241	1032	83	84	82	83

5.4 Model Testing

The model was split into training and testing sets. 80% was used for training, while 20% was used for testing, which aided in model validation. The best accuracy for the model testing was 86%; this showed how well the trained model would predict new credit scores when presented with new data. The CatBoost model was then saved in a joblib file and imported into the code editor to create a machine-learning web app to make real-time predictions.

```
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

# Print classification report
print("Classification Report:")
print(classification_report(y_test, y_pred))

# Generate Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(conf_matrix)
```

```
Accuracy: 0.8627931136835857
```

Figure 5. 11: Best Test Results

```
joblib.dump(catboost_classifier,"catboost_model.pkl")
```

```
['catboost_model.pkl']
```

Figure 5. 12: Saving the Trained Model

5.5 Streamlit Web Application for the Credit Scoring Model

The credit scoring model was deployed using the open-source framework to build a machine learning application that makes real-time predictions based on borrower input.

Figure 5.14 shows the user interface for the mobile credit scoring system when run on a local server. Streamlit allows users to deploy their machine learning applications for free using their GitHub account. As such, the Credit Scoring model can also be accessed via the link below and seen in Figure 5.13.

https://statuspred.streamlit.app/

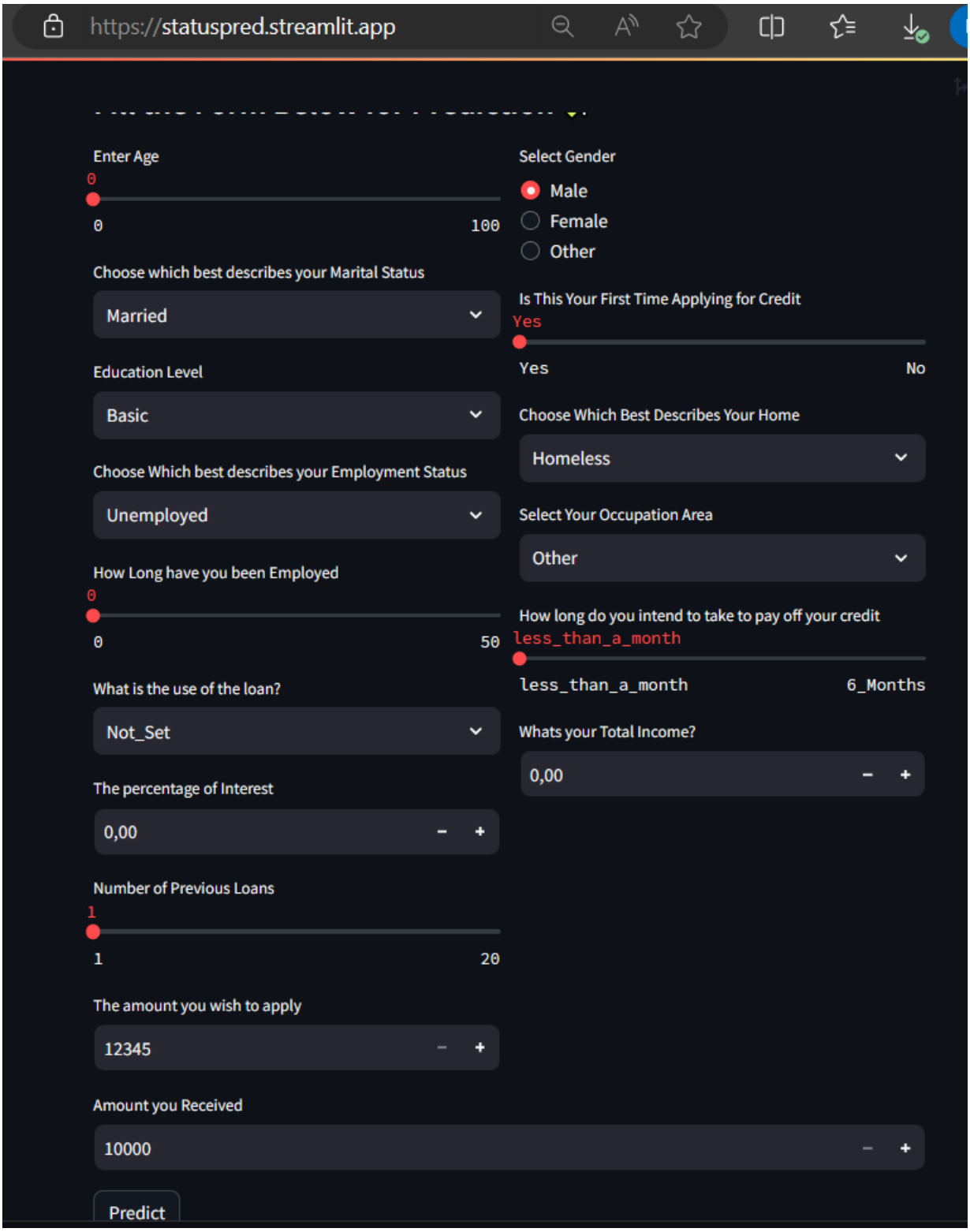


Figure 5. 13: Deployed Credit Scoring Application

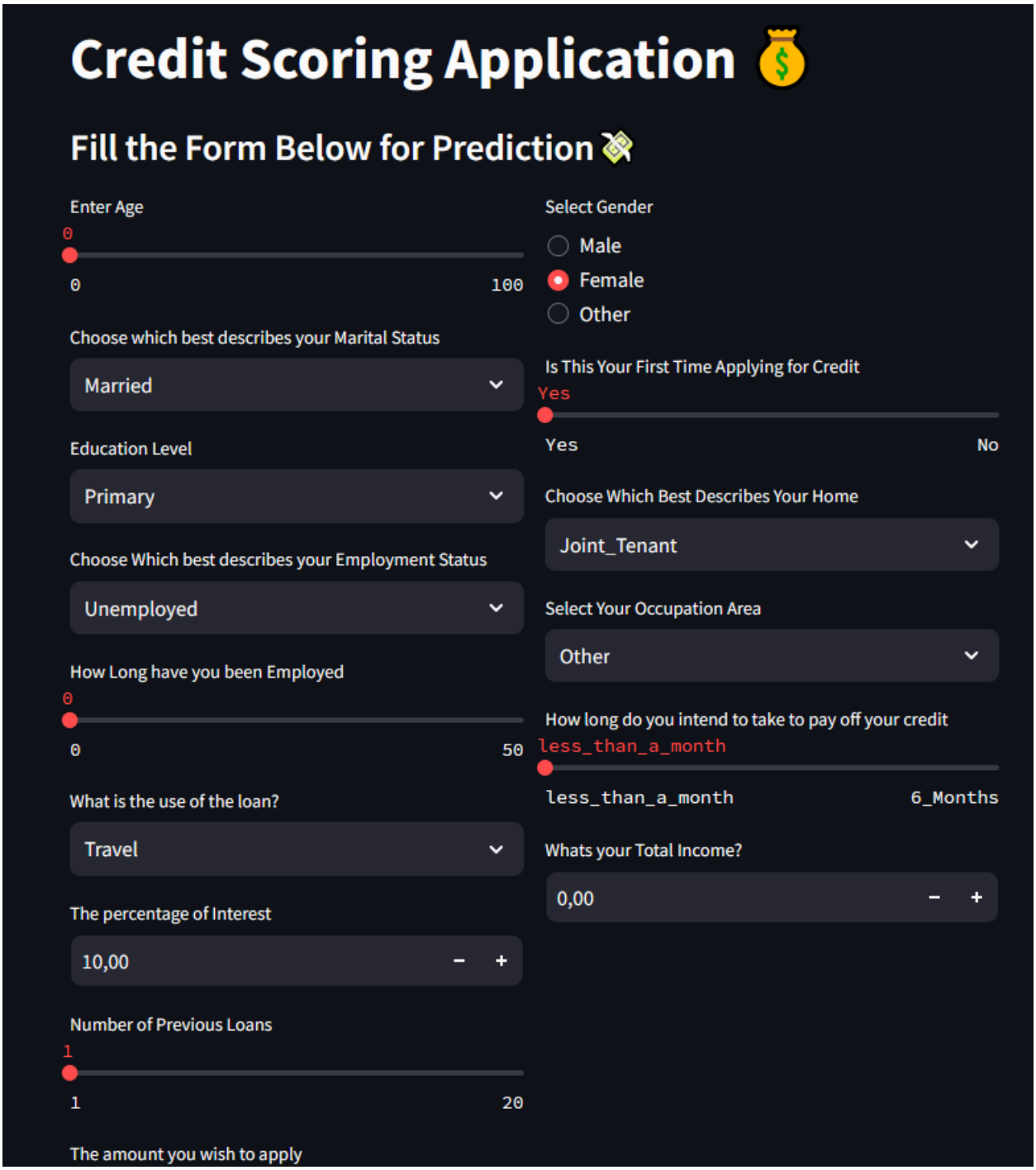


Figure 5. 14: User Interface for the Credit Scoring Application

5.6 Model Usage for Prediction

The model can then be used to make real-time predictions using some data provided by the borrower, as seen in Figure 5.15.

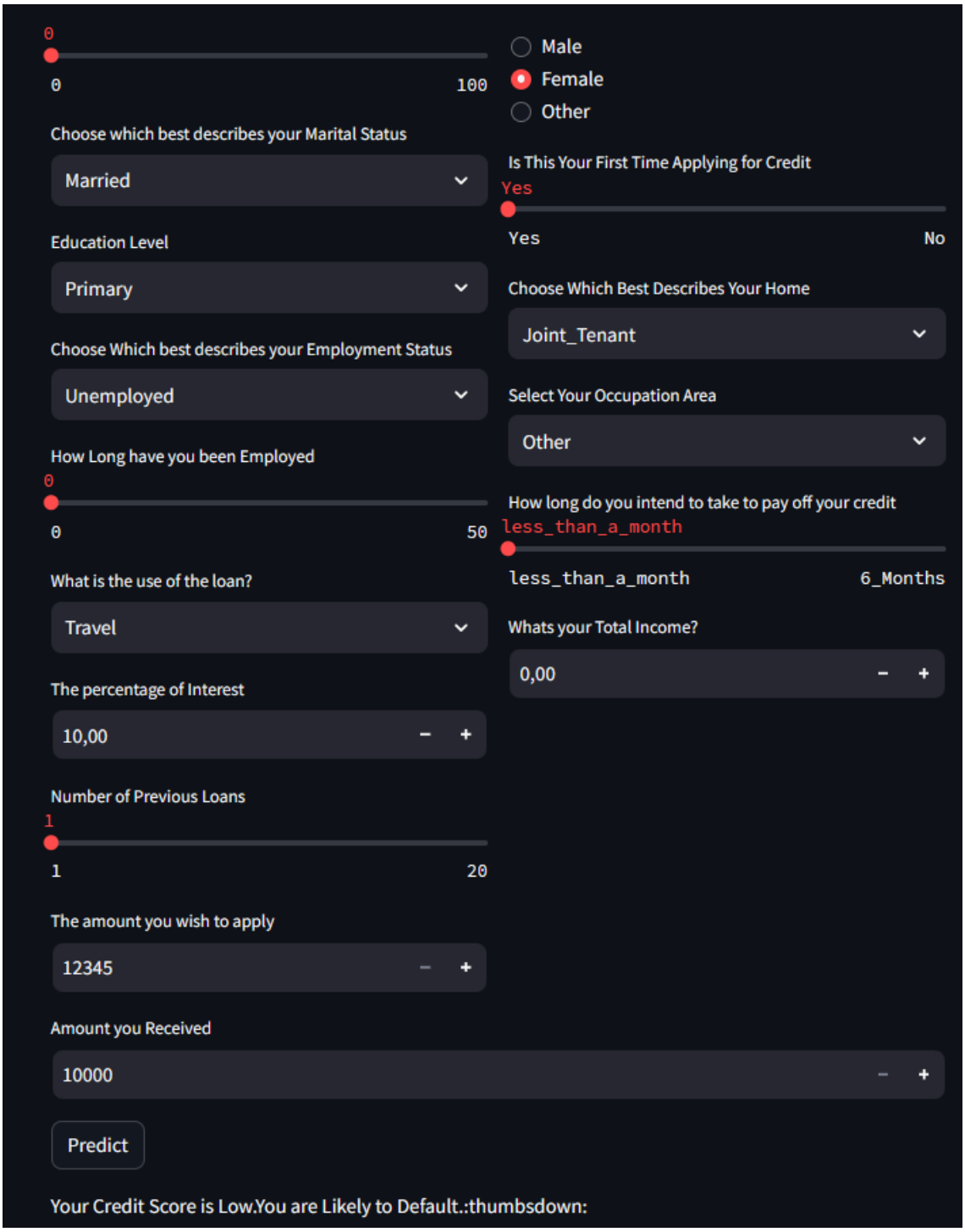


Figure 5. 15: Model Usage for Prediction

Chapter 6: Discussions

6.1 Introduction

This chapter briefly discusses the research findings and reviews the solution compared with the research objectives highlighted in Chapter 1. The main aim of this research was to develop a credit scoring model for mobile lenders. The research reviewed the challenges experienced in credit scoring and the existing methods. Finally, a machine learning web application was developed where borrowers can input prediction data and receive real-time output.

6.2 Model Validation

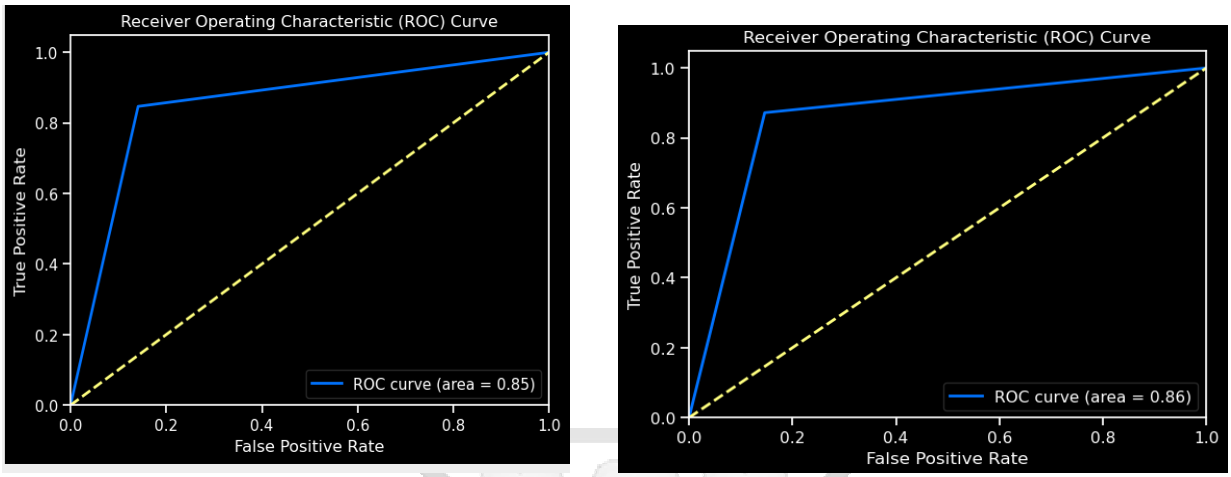
The data sourced from a verified online source was used for the study. The data was split into 80% training and 20% testing, used for model validation, with Random Forest and CatBoost achieving the highest results of 86%. The data cleaning and feature engineering process was well documented. The performance metrics used to evaluate the credit scoring model included Accuracy, Recall, Precision, F1-Score, AUC-ROC curve, Confusion Matrix, MCC, and Informedness. The highest accuracy obtained was 86% by both Random Forest and CatBoost algorithms, while AdaBoost had the lowest accuracy of 83%, representing how the model performs when fed with prediction data. Due to its high accuracy and low False Positive rate, the CatBoost model was used to create the Streamlit machine-learning web app. The performance metrics for all the models are presented in Table 5.6.

The importance of features for all the models was carried out. This measures the contribution of each feature towards the target variable, which helps interpret the model's behavior. The machine-learning code was uploaded to GitHub for verification and availability; the link is available in Appendix D.

The AUC-ROC curves were also deployed to measure each model's performance in binary classifications. An AUC-ROC range between 0.5 and 1 suggests that the model can distinguish between positive and negative classes.

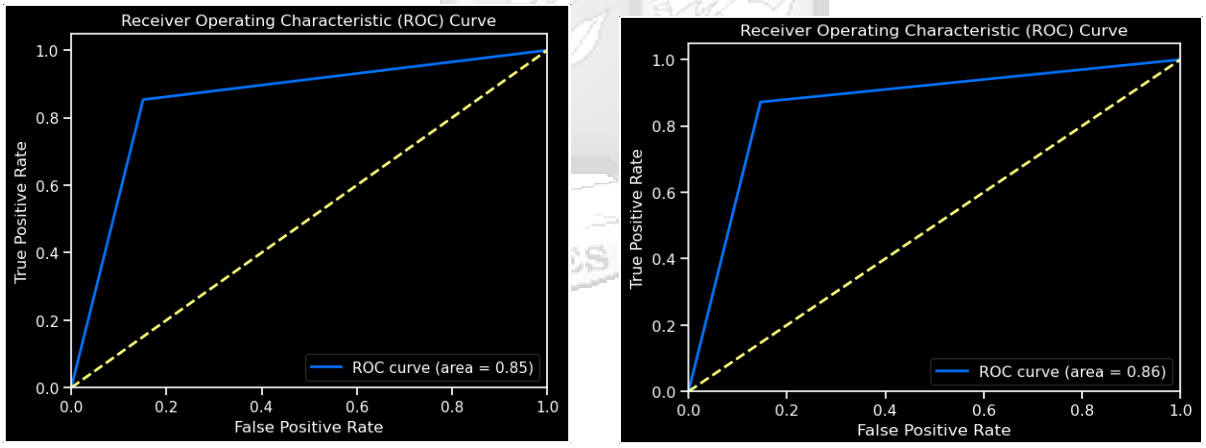
The results and model evaluation metrics signify the importance of using different models for credit scoring. Selecting a model based on one metric would lead to borrower bias and potential loss to mobile lenders. Striking a balance between the evaluation metrics is crucial for credit

scoring. A model with a low false positive rate and high accuracy can be effective for credit scoring to ensure that mobile lenders do not incur losses when lending.



(a) (b)

Figure 6. 1: AUC-ROC Curve for (a) XGBoost and (b) Random Forest



(a) (b)

Figure 6. 2: AUC-ROC Curves for (a) LightGBM and (b) CatBoost

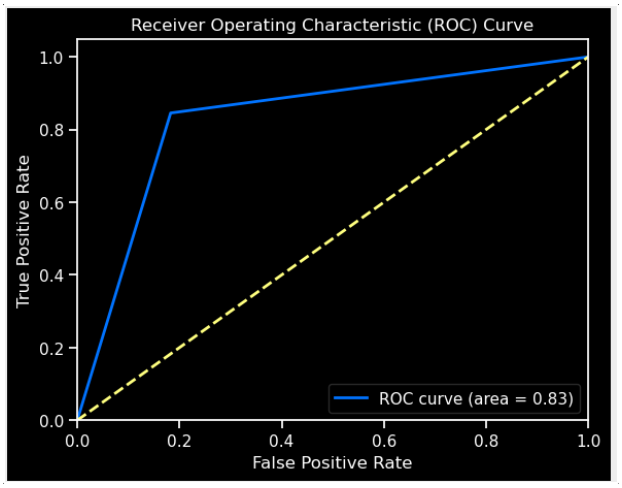


Figure 6. 3: AUC-ROC Curve for AdaBoost

Mathews Correlation Coefficient for each model was also calculated; this aided in measuring the quality of the binary class. A value of +1 represents a perfect classification; 0 represents a prediction made by chance, and a -1 value represents an opposite prediction where all the positive samples were predicted as negative and vice versa, meaning the model performs worse than a random guess. MCC maximizes all four confusion matrix essentials: specificity, sensitivity, negative predicted value, and precision (Chicco et al., 2021). Table 6.1 shows the MCC for the five machine-learning algorithms.

Table 6. 1: Mathews Correlation Coefficient for each Model

Rank	Classification Algorithm	MCC
1	Random Forest	72
2	CatBoost	71
3	LightGBM	70
4	XGBoost	70
5	AdaBoost	66

The Informedness level, also known as the Youden Index, measures how well the model discriminates between positive and negative classes by considering sensitivity and specificity rates was also calculated.

Table 6. 2: Informedness Level

Rank	Classification Algorithm	Informedness Level
1	Random Forest	72
1	CatBoost	72
3	LightGBM	70
4	XGBoost	70
5	AdaBoost	66

6.3 Merits of the Developed System to Existing Ones

I. The developed system can give instant results on whether the borrower qualifies for mobile credit. This leads to faster lender decision-making on borrower creditworthiness.

II. The model also generates valuable insights into borrower behavior.

III. The developed system takes various user inputs, such as age, gender, income, employment status, and loan duration, and predicts whether a borrower's application will be approved.

6.4 Research Flaws

The research initially planned to use primary data for credit scoring prediction purposes; however, due to the cumbersome process and privacy nature of local data, the researcher opted for secondary data.

A decrease in training features consequently makes the model perform poorly. Even with the Recursive Feature Elimination process, the best-ranked features do not increase the model accuracy as intended.

Parameter tuning to a higher number of values to increase the model accuracy was not done since it was computationally expensive.

Chapter 7: Conclusions, Recommendations, and Future Works

7.1 Conclusion

An increase in mobile usage leads to an upward shift in mobile credit lending. A mobile loan is instantaneously available for most phone users. However, defaults are lenders' primary concern. Lenders must analyze borrowers' creditworthiness before lending.

The study aimed to develop a credit scoring model for mobile lending. Reviewing the existing methods was crucial to articulating the scope and ensuring the research was guided by the objectives. The challenges in credit scoring were also well-reviewed, and a proposed solution was implemented. The data for the research was collected from a verified online source; this data was pre-processed and split into training and testing sets. The SMOTE technique was used to balance the dataset, boosting the default prediction accuracy. Model training used 80% of the data, while the remaining 20% was used for validation. The best-performing model with the lowest Type 1 error, Catboost, was selected to implement a scoring model on Streamlit, which the borrower can use to make real-time predictions. The Education, Gender, and Home Ownership features were the most relevant for credit scoring using the CatBoost model. Identifying key features is helpful for lenders in determining key predictors for potential loss. These features also aid in identifying different borrower groups and implementing proper policies, such as allocating resources to attract new customers. While the research implemented a valuable scoring model, more is needed for a real-world scenario due to the absence of primary data. Based on the discussions and findings, the following conclusions can be made:

I. Random Forest and CatBoost achieved the highest accuracy of 86%, while LightGBM and XGBoost presented accuracies of 85%.
II. CatBoost had the lowest Type 1 error while LightGBM had the highest. This justified the model selection for system development.
III. XGBoost had the highest Type 2 error while Random Forest had the lowest.
IV. AdaBoost model had the lowest performance on all evaluation metrics.
V. Model performances on all evaluation metrics are different.
VI. Feature importance across the five models also differs.

VII. A typical lending scenario consists of imbalanced data, making balancing crucial to avoid class bias.

7.2 Recommendation

The research demonstrates that machine learning can be used to develop credit-scoring methods. This can aid cumbersome manual methods and capture borrowers in rural areas with little financial history. Based on the study results, the researcher recommends.

I. A credit range, where a higher range indicates a creditworthy borrower.
II. Suggestions on improving one's credit score if they fall in the default category.
III. Model integration to a real-world lending case using a primary dataset for the prediction task.

7.3 Future Works

Extending the research using other models would increase the model's accuracy. This can be done through optimization methods using primary dataset features and applying privacy-preserving techniques such as Differential Privacy to ensure the privacy of primary data while maintaining valuable data. This can be achieved through a privacy budget to manage the privacy provided. This research used five machine-learning algorithms; other supervised and unsupervised algorithms can be implemented to improve the model's accuracy.

Due to the change in social environment and increased online information, social media and phone usage data can be integrated into the primary data features to get predictions.

Testing the model in a real-world setting would be helpful to ensure feasibility and utility. For instance, it would examine how much data the model can process, whether it can process new datasets accurately or poorly, and measure the utility of the results. The model can be monitored to establish performance indicators.

Using One-Shot learning and Stochastic Gradient Descent, where a model is trained to make predictions using only one class, can increase the training time of the model and the use of Parameter Tuning to improve model accuracy.

Credit risk assessment uses survival analysis, a statistical tool, to predict default time. Since the time of default is crucial, this can provide invaluable insights to lenders.

Finally, model explainability is achieved by using LIME to interpret and explain the decisions of the developed model. The feature importance provided in Chapter 5 provides limited insight into model explainability; an in-depth exploration would be necessary in the future. This will make the credit scoring model more transparent and lead to equal opportunities.

References

Alazab, M. (2015). Profiling and classifying the behavior of malicious codes. *Journal of Systems and Software*, *100*, 91-102.

Al-Fedaghi, S. (2021). UML sequence diagram: an alternative model. *arXiv preprint arXiv:2105.15152*.

Altman, E. I. (2018). A fifty-year retrospective on credit risk models, the Altman Z-score family of models, and their applications to financial markets and managerial strategies. *Journal of Credit Risk*, *14*(4).

Antunes, J. A. P. (2021). To supervise or to self-supervise: a machine learning based comparison on credit supervision. *Journal of Financial Innovation, 7(1), 1-21.*

Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3483-3488.

Barasch, Ron. 2017. *"Leveraging Alternative Data to Energize Your Lending Portfolio."* Yodlee.com. https://www.yodlee.com/retail-banking/leveraging-alternative-data-energize-lending-portfolio

Bharadwaj, P., Jack, W., & Suri, T. (2019). Fintech and household resilience to shocks: Evidence from digital loans in Kenya (No. w25604). National Bureau of Economic Research.

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99-113.

Burlando, Alfredo, Kuhn, Michael A, Prina, Silvia (2021): Too Fast, Too Furious? Digital Credit Delivery Speed and Repayment Rates. CEGA Working Paper Series No. WPS-151. Center for Effective Global Action. University of California, Berkeley. Text. https://doi.org/10.26085/C32P49

Central Bank of Kenya (2022, November) *FinAccess Household Survey County Perspective.* https://www.centralbank.go.ke/2022/11/11/finaccess-household-survey-report-county-perspective-november-2022/

CGFS (Committee on the Global Financial System) and FSB (Financial Stability Board). 2017. *"FinTech Credit: Market Structure, Business Models and Financial Stability Implications."* Working Group Report. http://www.fsb.org/wp-content/uploads/CGFS-FSB-Report-on-FinTech-Credit.pdf

Chen, G., & Mazer, R. (2016, February 8). Instant, automated, remote: The key attributes of digital credit. *CGAP Blog.*

Chen, J., (2023, August 2). *Basel II: Definition, Purpose, Regulatory Reforms*. Investopedia. https://www.investopedia.com/terms/b/baselii.asp

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, *14*, 1-22.Christoph, M. (2019). Interpretable machine learning: A guide for making black box models explainable. *Lulu. com*.

Çığşar, B., & Ünal, D. (2019). Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*, *2019*. https://doi.org/10.1155/2019/8706505

Communication Authority of Kenya (2023, June 28). *Mobile Subscriptions Hit 66m as at March 2023* https://www.ca.go.ke/mobile-subscriptions-hit-66m-march-2023

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296-299). Auerbach Publications.

Dewi, P. M. (2020). Credit insurance as an effort to overcome bad credit risk in modern banking economy in the industrial revolution 4.0 in Indonesia. *UNIFIKASI: Jurnal Ilmu Hukum*, *7*(1), 88-95.

Egwa, A. A., Kakudi, H. A., Ahmad, A. A., Bichi, A. M., & Madu, M. A. (2022). Prediction Model for Loan Default Using Machine Learning. *The International Journal of Science & Technol*ogy, *10*(2).

El Qadi, A., Trocan, M., Diaz-Rodriguez, N., & Frossard, T. (2023). Feature contribution alignment with expert knowledge for artificial intelligence credit scoring. *Signal, Image and Video Processing*, *17*(2), 427-434.

El-Qadi, A., Trocan, M., Frossard, T., & Díaz-Rodríguez, N. (2022, December). Credit Risk Scoring Forecasting Using a Time Series Approach. In *Physical Sciences Forum* (Vol. 5, No. 1, p. 16). MDPI.

FasterCapital (12 December, 2023), *Credit Scoring Algorithms: Cracking the Code for a Better Score*. https://fastercapital.com/content/Credit-Scoring-Algorithms--Cracking-the-Code-for-a-Better-Score.html

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259-268).

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905.

Finance, J. (2017). Machine Learning in Credit Risk Modeling: Efficiency should not come at the expense of Explainability.

Fosu, S., Danso, A., Agyei-Boapeah, H., Ntim, C. G., & Adegbite, E. (2020). Credit information sharing and loan default in developing countries: the moderating effect of banking market concentration and national governance quality. Review of Quantitative Finance and Accounting, 55(1), 55-103. https://doi.org/10.1007/s11156-019-00836-1

Fu, W. (2018). *A practical guide to age-period-cohort analysis: the identification problem and beyond*. CRC Press.

Goh, R. Y., & Lee, L. S. (2019). Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, *2019*.

Grab. (2018. March) *"Grab and Credit Saison Form Financial Services Joint Venture to Expand Access to Credit for Southeast Asia's Unbanked."* https://www.grab.com/sg/press/others/grab-and-credit-saison-form-financial-services-joint-venture-to-expand-access-to-credit-for-southeast-asias-unbanked/

Gurný, P., & Gurný, M. (2013). Comparison of credit scoring models on probability of default estimation for us banks. *Prague economic papers*, *22*(2), 163-181.

Heaven, W. D. (2020). *Our weird behavior during the pandemic is messing with AI models.* MIT Technology Review. https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/

Hou, X. (2020). P2P borrower default identification and prediction based on RFE-multiple classification models. *Open Journal of Business and Management*, *8*(2), 866-880.

Hummel, J. M., Bridges, J. F., & IJzerman, M. J. (2014). Group decision making with the analytic hierarchy process in benefit-risk assessment: a tutorial. *The Patient-Patient-Centered Outcomes Research*, *7*, 129-140.

Izaguirre JC, Kaffenberger M, Mazer R. (2018, September 25). It's Time to Slow Digital Credit's Growth in East Africa. *CGAP*. It's Time to Slow Digital Credit's Growth in East Africa | Blog | CGAP

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. 02 (01), 20–28.

Kagan, Julia. 2019. *"Credit Rating*." Investopedia. https://www.investopedia.com/terms/c/creditrating.asp

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kenton, Will. 2019. "*Credit Scoring.*" Investopedia. https://www.investopedia.com/terms/c/credit_scoring.asp

Kisutsa, G. T. (2021). *Loan Default Prediction Using Machine Learning: A Case of Mobile Based Lending* (Doctoral dissertation, University of Nairobi).

Ko, P. C., Lin, P. C., Do, H. T., & Huang, Y. F. (2022). P2P lending default prediction based on AI and statistical models. *Entropy*, *24*(6), 801.

Lehr, D., & Ohm, P. (2017). Playing with the data: what legal scholars should learn about machine learning. *UCDL Rev.*, *51*, 653.

Li Y 2019 Credit risk prediction based on machine learning methods The 14th Int. Conf. on Computer Science & Education (ICCSE) pp 1011–3

Mac an Bhaird, C., Owen, R., Dodd, S. D., Wilson, J., & Bisignano, A. (2019). Small beer? peer-to-peer lending in the craft beer sector. *Strategic Change*, *28*(1), 59-68.

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012042). IOP Publishing.

Marte, A. (2019). *Machine learning in default Prediction: the incremental power of machine learning techniques in mortgage default prediction* (Master's thesis).

Mijwel MM (2018). Artificial Neural Networks Advantages and Disadvantages."[LinkedIn Page]". Retrieved September, 10, 2022, from https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel/

Nick Hotz. 2023. *What is CRISP-DM?* https://www.datascience-pm.com/crisp-dm-2/

Niu, A., Cai, B., & Cai, S. (2020). Big data analytics for complex credit risk assessment of network lending based on SMOTE algorithm. Complexity, 2020, 1-9.

Onay, C., & Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*, *26*(3), 382-405.

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Passage Technology (n.d.). *What Is The Analytic Hierarchy Process (AHP)?*. https://www.passagetechnology.com/what-is-the-analytic-hierarchy-process

Prince, A. E., & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, *105*, 1257.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Runchi, Z., Liguo, X., & Qin, W. (2023). An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects. Expert Systems with Applications, 212, 118732.

S. Minaee, 2019. *20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics.* https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

SAS. 2019. *"Artificial Intelligence: What It Is and Why It Matters*." https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html.

Sengupta, R., & Bhardwaj, G. (2015). Credit Scoring and Loan Default. International Review of Finance, 15(2), 139–167.

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, *89*, 113-122.

Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation in Construction*, *129*, 103827.

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

Simão, S. B. S. (2023). Machine Learning applied to credit risk assessment: Prediction of loan defaults (Doctoral dissertation).

Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021, June). Prediction of modernized loan approval system based on machine learning approach. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-4). IEEE. https://doi.org/10.1007/s11156-019-00836-1

World Bank (2019). *CREDIT SCORING APPROACHES GUIDELINES* https://pubdocs.worldbank.org/en/935891585869698451/CREDIT-SCORING-APPROACHES-GUIDELINES-FINAL-WEB.pdf

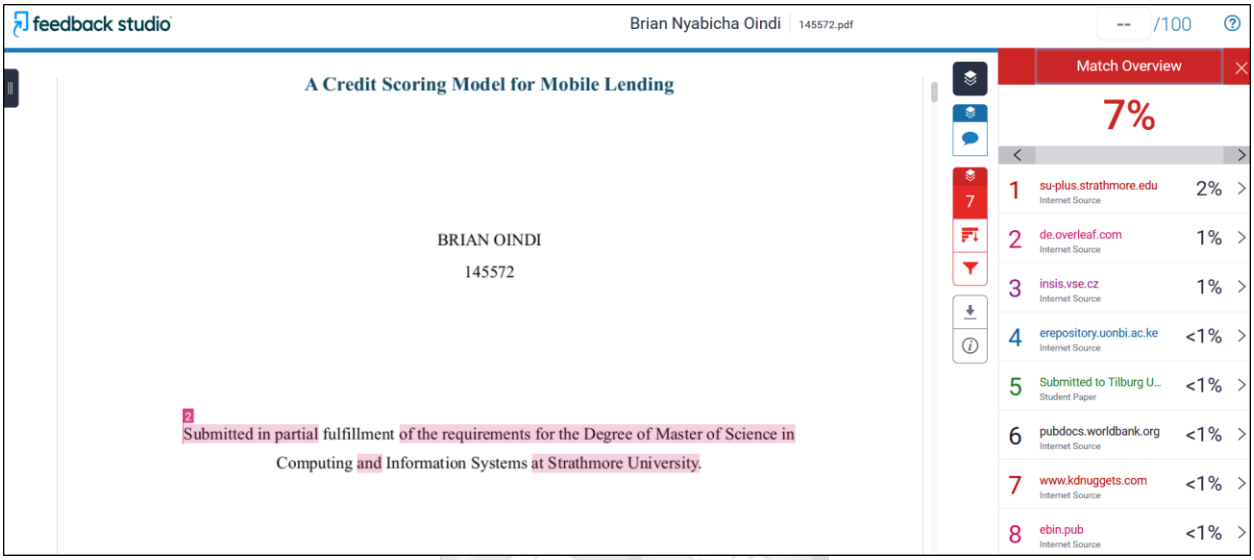
Yao, J., Chen, J., Wei, J., Chen, Y., & Yang, S. (2019). The relationship between soft information in loan titles and online peer-to-peer lending: evidence from RenRenDai platform. *Electronic Commerce Research*, *19*, 111-129.

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2020). A Study on Predicting Loan Default Based on the Random Forest Algorithm. International Conference on Information Technology and Quantitative Management. 162, 503–513.

Appendices

Appendix A: Similarity Report

Turnitin Similarity Report



145572.pdf

ORIGINALITY REPORT

7%	7%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	su-plus.strathmore.edu Internet Source	2%
2	de.overleaf.com Internet Source	1%
3	insis.vse.cz Internet Source	1%
4	erepository.uonbi.ac.ke Internet Source	<1%
5	Submitted to Tilburg University Student Paper	<1%

6	pubdocs.worldbank.org Internet Source	<1 %
7	www.kdnuggets.com Internet Source	<1 %
8	ebin.pub Internet Source	<1 %
9	academic.oup.com Internet Source	<1 %

10	erepository.uonbi.ac.ke:8080 Internet Source	<1 %
11	Submitted to University of Reading Student Paper	<1 %
12	etda.libraries.psu.edu Internet Source	<1 %
13	myfik.unisza.edu.my Internet Source	<1 %
14	www.coursehero.com Internet Source	<1 %

Exclude quotes	Off	Exclude matches	< 25 words
Exclude bibliography	On		

Appendix B: Ethical Clearance

Strathmore UNIVERSITY

16th November 2023

Mr Oindi Brian,
brian.oindi@strathmore.edu

Dear Mr Oindi,

RE: A Credit Scoring Model for Mobile Lending

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1916/23.** The approval period is from **16th November 2023 to 15th November 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 72 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 72 hours.
v. Clearance for the 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 the expiry of the approval period. Attach a comprehensive progress report to support the renewal.

vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before 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,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
16-Nov-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

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: NACOSTI Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **705924**

Date of Issue: **02/December/2023**

RESEARCH LICENSE

This is to Certify that Mr.. Brian Oindi of Strathmore University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: A CREDIT SCORING MODEL FOR MOBILE LENDING for the period ending : 02/December/2024.

License No: **NACOSTI/P/23/31741**

705924

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013 (Rev. 2014)
Legal Notice No. 108: The Science, Technology and Innovation (Research Licensing) Regulations, 2014

The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was the established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of International treaties of which Kenya is a signatory to
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
    i. Endanger national security
    ii. Adversely affect the lives of Kenyans
    iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
    iv. Result in exploitation of intellectual property rights of communities in Kenya
    v. Adversely affect the environment
    vi. Adversely affect the rights of communities
    vii. Endanger public safety and national cohesion
    viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. The license any rights thereunder are non-transferable
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.
13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and
Innovation(NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 0713788787, 0735404245
E-mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke

Appendix D: Code Used in the Research

This is the GitHub Link for the machine-learning classification task.

https://github.com/ORB-7/Credit-Scoring-Task

The screenshot contains the code for User Interface.

```
import streamlit as st  # import the module for the websites
import joblib  # module to load model data
import pandas as pd

# ------- SETTINGS-----------
page_title = "Credit Scoring Application"
page_write = " Fill the Form Below for Prediction"
page_icon = ":moneybag:"
layout = "centered"
# -----------------


st.set_page_config(page_title=page_title, page_icon=page_icon, layout=layout)

st.title(page_title + " " + page_icon)

st.write(""" ### Fill the Form Below for Prediction:money_with_wings:""")

st.sidebar.selectbox('Explore or Predict', ("Predict", "Explore"))

education_list = ["Basic", "Primary", "Vocational", "Secondary", "Higher"]
marital_list = ["Married", "Cohabitant", "Single", "Divorced", "Widow"]
employmentStatus = ["Unemployed", "Partially", "Fully", "Self",
"Entrepreneur", "Retiree"]
homeTypeOwnership = ["Homeless", "Owner", "Living_With_Parents", "Tenant",
"Prefurnished_Property",
                     "Unfurnished_Property", "Joint_Tenant",
"Joint_Ownership", "Mortgage", "Owner_with_Encumbrance",
                     "Other"]
occupationArea = ["Other", "Mining", "Processing", "Energies", "Utilities",
"Construction",
                  "Retail_and_Wholesale", "Transport_and_Warehousing",
"Hospitality_and_Catering",
                  "Finance_and_Insurance", "Real_Estate", "Research",
"Administrative", "Civil_Service_and_Military",
                  "Education", "Health_Care_and_Social_Help",
"Arts_and_Entertainment",
                  "Agriculture_Forestry_and_Fishing"]
useOfLoan = ["Not_Set", "Loan_Consolidation", "Real_Estate",
"Home_Improvement", "Business", "Education", "Travel",
             "Vehicle", "Other", "Health", "Finance_and_Insurance",
"Research", "Administrative",
             "Civil_Service_and_Military", "Education_2",
"Health_Care_and_Social_Help", "Arts_and_Entertainment",
             "Agriculture_Forestry_and_Fishing"]
```

```
loanDuration = ["less_than_a_month", "1 Month", "2_Months", "3_Months",
"4_Months", "5_Months", "6_Months"]

col1, col2, = st.columns(2)

Age = col1.slider('Enter Age', 0, 100)
Gender = col2.radio("Select Gender", ["Male", "Female", "Other"])
MaritalStatus = col1.selectbox("Choose which best describes your Marital
Status", marital_list)
Education = col1.selectbox("Education Level", education_list)
NewCreditCustomer = col2.select_slider('Is This Your First Time Applying for
Credit', ['Yes', 'No'])
EmploymentStatus = col1.selectbox('Choose Which best describes your
Employment Status', employmentStatus)
HomeOwnershipType = col2.selectbox('Choose Which Best Describes Your Home',
homeTypeOwnership)
EmploymentDurationCurrentEmployer = col1.slider('How Long have you been
Employed', 0, 50)
OccupationArea = col2.selectbox('Select Your Occupation Area',
occupationArea)
UseOfLoan = col1.selectbox("What is the use of the loan?", useOfLoan)
LoanDuration = col2.select_slider("How long do you intend to take to pay off
your credit", loanDuration)
Interest = col1.number_input('The percentage of Interest')
IncomeTotal = col2.number_input('Whats your Total Income?')
NoOfPreviousLoansBeforeLoan = col1.slider('Number of Previous Loans', 1, 20)
AppliedAmount = col1.number_input("The amount you wish to apply", 1234)
Amount = st.number_input('Amount you Received', 1000)

Rating = 10
ExistingLiabilities = 0
DebtToIncome = 0.12
Restructured = 1
CreditScoreEsMicroL = 0.9
ModelVersion = 2
VerificationType = 1
LanguageCode = 9600
df_pred = pd.DataFrame([[Age, LoanDuration, NewCreditCustomer,
VerificationType, Gender,
                         AppliedAmount, Interest, UseOfLoan, Amount,
Education,
                         EmploymentDurationCurrentEmployer, Rating,
MaritalStatus,
                         EmploymentStatus, OccupationArea, HomeOwnershipType,
ExistingLiabilities,
                         DebtToIncome, IncomeTotal,
                         Restructured, NoOfPreviousLoansBeforeLoan,
CreditScoreEsMicroL, ModelVersion]],

                        columns=['Age', 'LoanDuration', 'NewCreditCustomer',
'VerificationType', 'Gender',
                                 'AppliedAmount', 'Interest', 'UseOfLoan',
'Amount', 'Education',
                                 'EmploymentDurationCurrentEmployer',
'Rating', 'MaritalStatus',
                                 'EmploymentStatus', 'OccupationArea',
'HomeOwnershipType', 'ExistingLiabilities',
```

```
                                'DebtToIncome', 'IncomeTotal',
                                'Restructured',
'NoOfPreviousLoansBeforeLoan', 'CreditScoreEsMicroL', 'ModelVersion'])


def transform(data):
    if data in marital_list:
        return marital_list.index(data)
    elif data in education_list:
        return education_list.index(data)
    elif data in employmentStatus:
        return employmentStatus.index(data)
    elif data in occupationArea:
        return occupationArea.index(data)
    elif data in useOfLoan:
        return useOfLoan.index(data)
    elif data in loanDuration:
        return loanDuration.index(data)
    else:
        return 0


df_pred['Gender'] = df_pred['Gender'].apply(lambda x: 1 if x == 'Male' else
0)
df_pred['NewCreditCustomer'] = df_pred['NewCreditCustomer'].apply(lambda x: 1
if x == 'Yes' else 0)
df_pred['Education'] = df_pred['Education'].apply(transform)
df_pred['EmploymentStatus'] = df_pred['EmploymentStatus'].apply(transform)
df_pred['MaritalStatus'] = df_pred['MaritalStatus'].apply(transform)
df_pred['HomeOwnershipType'] = df_pred['HomeOwnershipType'].apply(transform)
df_pred['UseOfLoan'] = df_pred['UseOfLoan'].apply(transform)
df_pred['LoanDuration'] = df_pred['LoanDuration'].apply(transform)
df_pred['OccupationArea'] = df_pred['OccupationArea'].apply(transform)


def make_predictions(pred_data):
    model = joblib.load('catboost_model.pkl')
    pred = model.predict(pred_data)
    return pred


prediction = make_predictions(df_pred)
if st.button('Predict'):
    if prediction[0] == 0:
        st.write('<p class="big-font">Your Credit Score is High.You can apply
for a Loan.:thumbsup:</p>',
                 unsafe_allow_html=True)
    else:
        st.write('<p class="big-font">Your Credit Score is Low.You are Likely
to Default.:thumbsdown:</p>',
                 unsafe_allow_html=True)
```