Predictive Modeling of Digital Credit Risk in Commercial Banks Using Machine Learning Algorithms

Abstract:Normally, financial institutions issue loans to customers aiming to recover the repayments within the scheduled time. However, defaults, common with digital credit, cause them significant financial risks. Advancement in technology has created digital loan products. Unfortunately,in developing economies such as Kenya,the growth rate of such products has outpaced development of sound credit risk assessment systems. These factors, along with the limitations of conventional risk models, which are static, necessitate the adoption of dynamic models capable of capturing complex and non linear patterns among variables. This study addressed this gap by developing predictive models using machine learning algorithms. Using 6,000 simulated loan records, necessitated by restricted access to real world data, the study evaluated performance of the two models, among which, random forest emerged as the most robust. Even though random forests model initially achieved an Area Under Curve score of 1.0, under simulated conditions, a 10-fold cross validation model produced a more realistic mean AUC of 0.90. This reflects strong discriminative ability and high predictive accuracy. These findings demonstrate that machine learning algorithms can enhance risk assessment framework in commercial banks thereby minimizing financial losses due to loan provisions. A limitation of this study was the use of simulated data,which may not fully capture the complexities of real world digital loan scenarios.Future research should focus on validating these models using real world borrower data from financial institutions to enhance generalizability and assess operational performance under actual lending conditions.

Keywords: SHAP ,Digital Credit,Random Forest,XGBoost,temporal dimension,Simulated,Loan to Income Ratio.

1 Introduction

Commercial banks advance loans under the assumption that the loans will be paid on time as agreed. In reality, most of the borrowers fail to honor their agreements leading to financial losses. The case is even worse with digital credit. Technology has led to a quicker and more convenient credit accessibility to more people in emerging economies such as Kenya. Despite this evolution, credit risk assessment models have not kept pace, as they remain largely anchored on traditional lending paradigms.The conventional credit risk assessment models do not account for the distinct characteristics of digital credit products. This makes them ill-equipped for the dynamic and unique nature of digital credit. Furthermore the use of advanced analytical tools such as Random forest and eXtreme gradient boosting algorithms are common in established markets such as the United States and the United Kingdom. Their adaptation to the unique socio-economic, technological and regulatory dynamics prevalent in developing markets is uncommon. This mismatch poses a severe problem, particularly to Kenyan digital market where digital credit has gained significant traction. .

2 Literature Review

2.1 Introduction

Researchers have explored many methods in the effort to identify the most efficient risk assessment technique. The efforts were not limited to traditional methods like Logistic regression analysis, Cox proportional hazards regression and now the modern machine learning techniques such as Random forests and Extreme gradient boosting models. This section examined Random forests and extreme gradient boosting models and compared the outcomes to the standard credit risk assessment techniques. Random forests, a machine learning ensemble technique, is used to create comprehensive non linear models in credit risk analysis [6]. Random forests has shown unbeatable accuracy .
Unlike the conventional methods, Random forests is able to consider large dataset of multifaceted high dimensional dataset. This quality is particularly valuable in digital credit, where the data involved is multifaceted and includes many variables that conventional techniques cannot handle.

Extreme Gradient boosting (XGBoost) has gained prominence in the recent past due to its ability to integrate macroeconomic factors into its algorithms. A study by [16], titled "Optimizing fintech Marketing: A Comparative Study of Logistic Regression and XGBoost", employed advanced machine learning techniques, specifically logistic regression and XGBoost, to analyze consumer behavior and predict responses to direct mail campaigns. The findings from the study suggested that XGBoost was particularly effective in handling complex data structures and provides a strong predictive capability in assessing credit risk.

3 MATERIAL AND METHOD

3.1 Data Collection

Due to the unavailability of real world transactional data from commercial banks, largely as a result of data privacy concerns and access restrictions, this study relied on simulated datasets. The simulation was guided by empirical patterns observed in the literature. The procedure was designed to capture key characteristics of digital loan borrowers, such as credit scores, loan amounts, repayment behaviors and demographic traits. This approach ensured that the models reflect realistic scenarios while maintaining analytical flexibility. The study utilized 6,000 simulated loan records from commercial banks which specifically focused on one month digital loan records, disbursed in not more than four month from the start of observation. The

time to event (default) was calculated from the start of observation upto when the loan was marked "default".

All simulated loan records were generated under controlled probabilistic assumptions and stored in structured Excel formats before being imported into R. The data handling process adhered to reproducible research standards, and preprocessing steps
(e.g., normalization, encoding and class balancing) were well documented. In cases where a borrower successfully paid the loan within the observation period, such data was declared censored and was excluded from the analysis. The data included individual borrowers' demographics, borrowers' behavioral information and borrowers' financial information.

3.1.1 Mathematical Model for Simulating Digital Loan Data

In order to construct the models for simulating digital loan records, we defined the notations and variables used in the model by letting:

$$
\begin{aligned}
n = 6{,}000 &: \text{number of observations (digital loan records)},\\
X^{(i)} = \left(X_1^{(i)}, X_2^{(i)}, \ldots, X_{13}^{(i)}\right), \quad i = 1, 2, \ldots, n &: \text{predictor variables for the } i\text{-th observation},\\
Y^{(i)} \in \{0, 1\}, \quad i = 1, 2, \ldots, n &: \text{binary outcome (1 = default, 0 = no default)}.
\end{aligned}
$$

Step 1: Generate Predictor Variables

Each predictor X_i was simulated from a specified distribution as shown in Table 3.1:

Table 1: Distributions and Justification for Simulated Data Variables

Variable	Distribution Used	Empirical Justification
X_1:Age	$X_1 \sim \mathcal{N}(32,\ 8^2)$;bounds = 18–65	Digital loan users in Kenya are concentrated between ages 25–40 [7]
X_2:Income-to-loan ratio	Log-Normal: $X_2 \sim \text{LN}(1.6,\ 0.5^2)$	Income distributions are skewed among informal earners; log-normal preferred [15]
X_3:Debt-to-income-ratio	Beta: $X_3 \sim \text{B}(2, 5)$	Debt burdens cluster near lower bound;beta fits well [10]
X_4:Past-loan-behavior	Negative Binomial:$X_4 \sim \text{NB}(2, 1.5)$	Loan behavior is overdispersed
X_5:Gender	Bernoulli:$X_5 \sim \text{Bern}(0.52)$	Women make up 48% of digital borrowers [8] [3]
X_6:Risk-taking-behavior	Ordinal (1–5 Likert scale)	Psychometric traits use Likert scales [12][8]
X_7:Multiple-loan-applications	Negative Binomial:$X_7 \sim \text{NB}(1.5, 2)$	Mobile borrowers apply to 2–3 apps; overdispersed [14]
X_8:Application-timing	Bernoulli: $X_8 \sim \text{Bern}(0.68)$	Over 65% of apps during weekdays [4]
X_9:App-usage	Gamma: $X_9 \sim \Gamma(2.5, 1.5)$	Usage time is right-skewed [7]
X_{10}:Digital-transaction-history	Log-Normal: $X_{10} \sim \text{LN}(1.8, 0.6)$	Mobile transactions are right-skewed [13]
X_{11}:Communication patterns	Poisson:$X_{11} \sim \text{ZIP}(1, 0.3)$	SMS/phone usage often sparse
X_{12}:Employment status	Multinomial:$X_{12} \sim \text{Mult}(0.52, 0.33, 0.15)$	Based on national employment proportions
X_{13}:Location (Urban/Rural)	Bernoulli: $X_{13} \sim \text{Bern}(0.63)$	63% borrowers in urban areas [3][8]

Step 2: Defining the Linear Predictor (Logit)

Having defined the Bernoulli distribution for the binary outcome, next we specified how the predictor variables combined to determine the probability of default. This relationship was expressed through a linear predictor (logit) function, which linked the covariates $X_j^{(i)}$ to the default probability $\pi^{(i)}$.

A coefficient vector was assigned as per Equation 1 below:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{13}) \tag{1}$$

A linear predictor for each observation was then computed as shown in Equation (2).

$$\eta^{(i)} = \beta_0 + \sum_{j=1}^{13} \beta_j X_j^{(i)} \tag{2}$$

Step 3: Computing the Probability of Default

We applied the logistic (sigmoid) function to obtain the probability of default as shown by Equation (3).

$$P\big(Y^{(i)} = 1 \mid X^{(i)}\big) = \pi^{(i)} = \frac{1}{1 + e^{-\eta^{(i)}}} \tag{3}$$

Step 4: Simulating the Binary Outcome (Default)

Before simulating the loan default outcome, we specified the statistical distribution governing the binary response variable. In line with standard logistic regression modeling of binary outcomes [**?**, **?**], the response variable $Y^{(i)}$, which represents loan default (1 = default, 0 = non-default), was assumed to follow a Bernoulli distribution with success probability $\pi^{(i)}$. This assumption provided the probabilistic foundation for the logistic model as shown in step 4 below. Consequently, to draw from a Bernoulli distribution for each observation;

$$Y^{(i)} \sim \text{Bernoulli}(\pi^{(i)}), \quad \text{for } i = 1, 2, \ldots, n$$

Summary Equation

In summary, the logistic regression model assumes that the log-odds of default are a linear function of the predictors, while the probability of default is obtained by applying the logistic transformation. This formulation provided a simple and complete representation of the relationship between borrower characteristics and default risk
Therefore, for each observation i,the full model was:

$$Y^{(i)} \sim \text{Bernoulli}\left(\frac{1}{1+\exp\left(-\left[\beta_0 + \sum_{j=1}^{13}\beta_j X_j^{(i)}\right]\right)}\right) \tag{4}$$

3.2 Digital Borrower Classification Using Random Forest and Extreme Gradient Boosting (XGBoost)

The study employed ensemble machine learning methods, namely Random Forest (RF) and Extreme Gradient Boosting (XGBoost), to classify digital borrowers into high-risk and low-risk categories. Both models were trained using the same set of thirteen predictor variables described in Section 3.2.

Random Forest constructs multiple decision trees on bootstrap samples of the training data and aggregates their predictions through majority voting, thereby reducing variance and mitigating over fitting. The number of trees (n_{trees}) and the maximum depth of each tree ($d_{\max}$) were tuned through a grid search with five-fold cross-validation.

XGBoost, on the other hand, applies gradient boosting principles, building sequential decision trees where each tree attempts to correct the errors of the previous one. Its regularization component prevents over fitting, while learning rate (η), maximum depth, and the number of boosting rounds (n_{rounds}) were optimized via cross-validation. Both RF and XGBoost models were trained and validated on an 70/30 stratified split. Model performance was assessed using classification accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). The final models were selected based on their ability to maximize predictive accuracy while maintaining generalizability.

3.3 Application of Random Forests Model in Digital Credit Risk Assessment in Commercial Banks

Random forests model building process started with constructing a bootstrapped dataset using random data samples, selected with replacement from the initial dataset to create multiple instances. Loans data is extensive and high-dimensional which makes random forests an effective solution for the contemporary credit risk assessment problem in commercial banks. The random forest model was trained on bootstrapped samples and split nodes using Gini impurity, a measure of class homogeneity:

$$\text{Gini}(D) = 1 - \sum p_k^2 \tag{5}$$

where p_k is the proportion of class k at a node.
This ensured optimal binary splits based on entropy reduction. Feature importance was computed via Mean Decrease in Gini and permutation importance, offering statistical Interpretability.

The study applied random forests algorithms to develop a classification model that aid in predicting the likelihood of digital loan defaults. The study went further and examined the predictive accuracy of the model compared to logistic regression analysis and cox ph regression model. The study leveraged on the ability of random forests algorithms feature importance functionality to rank loan predictor variables in order of importance in terms of default.

3.4 Random Forest Model Development

The dataset was divided into two subsets, 70% training set for building the model and a test set (30% of the data) to evaluate its performance. The model development process involved several steps: First, a bootstrapped dataset was constructed by using random data samples, selected with replacement from the initial dataset to create multiple instances. The next step involved building a number of decision trees on the bootstrapped training sample. After tree building, the predictions from each tree were combined within the model framework. Multiple trees were averaged to provide final predictions. In this case, a classification for loan default resulted from a combined vote where each tree reports its best guess (default or non default). The model performance was optimized by adjusting hyper parameters like trees and tree depth limits.

Mathematical Representation of Random Forest Model

The Random Forest model building begun with the definition of predictor and response variables by letting:
$\mathbf{X} = \{X_1, X_2, \ldots, X_{13}\}$ to be the feature vector as shown in Table 3.1.
and $Y \in \{0, 1\}$ to be the binary response variable:

$$Y = \begin{cases} 1 & \text{if borrower defaults,} \\ 0 & \text{if borrower repays.} \end{cases}$$

Consequently, the following steps were followed;

Step 1: Bootstrap Sampling

For each tree $t = 1, 2, \ldots, B$:
a bootstrap sample $D^{(t)} = \left\{ \left(X_i^{(t)}, Y_i^{(t)} \right) \right\}_{i=1}^{n}$ was drawn with replacement from the training data.

Step 2: Tree Construction

At each node of tree t:
$m \ll p$ features were uniformly selected at random from $\{1, 2, \ldots, p\}$.
Typically:

$$m = \begin{cases} \lfloor \sqrt{p} \rfloor & \text{(classification)} \\ \lfloor \frac{p}{3} \rfloor & \text{(regression)} \end{cases}$$

For each candidate feature X_j where $j \in M$, the split s^* that maximized the Gini impurity reduction was identified.

$$\Delta\text{Gini}(s, X_j) = \text{Gini}(D) - \left(\frac{|D_L|}{|D|} \text{Gini}(D_L) + \frac{|D_R|}{|D|} \text{Gini}(D_R) \right),$$

where:

$$\text{Gini}(D) = 1 - \sum_{k=0}^{1} p_k^2, \quad \text{with} \quad p_k = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathbb{I}(Y_i = k).$$

The tree was grown until a stopping criterion was met (e.g., maximum depth or minimum samples per node).

Step 3: Prediction Aggregation

Majority Vote (Classification):

$$\hat{Y}(X) = \arg\max_{y \in \{0,1\}} \sum_{t=1}^{B} \mathbb{I}\left(\hat{Y}^{(t)}(X) = y\right),$$

Probability Output (for ROC analysis):

$$P(Y = 1 \mid X) = \frac{1}{B} \sum_{t=1}^{B} \hat{Y}^{(t)}(X),$$

where $\hat{Y}^{(t)}(X) \in \{0, 1\}$ is the prediction of tree t.

Step 4: Variable Importance

Mean Decrease in Gini:

$$\text{Importance}(X_j) = \frac{1}{B} \sum_{t=1}^{B} \sum_{\substack{\text{nodes split on} \\ X_j}} \Delta\text{Gini}_{\text{node}},$$

Permutation Importance:

$$\text{Importance}_{\text{perm}}(X_j) = \frac{1}{B} \sum_{t=1}^{B} \left(\text{Err}^{(t)} - \text{Err}^{(t)}_{\pi(j)}\right),$$

where:
$\text{Err}^{(t)}$ is the out-of-bag (OOB) error of tree t, $\text{Err}^{(t)}_{\pi(j)}$ is the OOB error after permuting feature X_j in the OOB set.

3.5 Data Exploration and Visualization for Random Forests Model

Given that standard random forests models tend to perform well on majority class and poorly on the minority class, it was necessary to detect class imbalance in our dataset. This process involved three key major steps;

i. **Computing the class distribution.** The frequency of each class in the target variable was calculated to quantify the degree of imbalance.

ii. **Visualizing class distribution.** Graphical techniques such as bar charts and pie charts were employed to illustrate the extent of class imbalance.

iii. **Establishing an imbalance threshold.** An imbalance threshold was defined as the ratio between the majority and minority classes, with imbalance typically declared when one class accounts for more than 70–80% of the observations.

iv. **Applying SMOTE.** After identifying and confirming class imbalance during data exploration and visualization, the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset prior to model training.

3.6 Application of XGBoost Model in Digital Credit Risk Assessment in Commercial Banks

Extreme Gradient Boosting (XGBoost) is a more powerful and scalable machine learning algorithm that is based on gradient boosting. The XGBoost involves building an ensemble of decision trees. Each new tree built corrects the errors made by the previous trees. The trees are essentially trained to reduce the gradient of the loss function. The XGBoost algorithm minimized the logistic loss function using gradient descent:

$$\ell(y, \hat{y}) = y \log(1 + e^{-\hat{y}}) + (1 - y) \log(1 + e^{\hat{y}}) \tag{6}$$

For optimization, the model used a second-order Taylor expansion of the loss and included shrinkage and column sub sampling to prevent over fitting. Early stopping was applied based on cross-validated AUC performance to ensure generalizability.

3.7 XGBoost Model Development

The modeling process started with splitting the data into two subsets, training set consisting of 70% and testing set consisting of 30%. After that we trained the model using the training set and validated with the test set for early stopping. Once trained, the XGBoost model performance was evaluated based on the test data. The performance of the model was improved by tuning the hyper parameters tuning of length 5.

Mathematical Modeling of eXtreme Gradient Boosting Model

In order to understand the application of the eXtreme Gradient Boosting (XGBoost) algorithm in this study, it was worthwhile to give a description of its mathematical foundation. This model is based on the principle of boosting in which weak learners are added together to produce a strong predictive model. Our training data was defined as:

$$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n, \quad X_i \in \mathbb{R}^p, \quad Y_i \in \{0, 1\}.$$

Gradient Boosted Trees use an additive function model:

$$\hat{Y}(x) = \sum_{k=1}^{K} f_k(x), \quad f_k \in \mathcal{F},$$

where $\mathcal{F}$ is the space of regression trees (weak learners), and K is the number of boosting rounds.

Step 1: Loss Function Optimization

The initial stage in the development of the XGBoost model was to formulate an objective function which quantifies how close the model predictions and the actual outcomes are. This was done through the minimization of a regularized loss function, which trades off model accuracy and model complexity. The loss function was therefore defined as:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} \ell(Y_i, \hat{Y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

with regularization term defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2,$$

where:
$\ell(Y, \hat{Y})$ is the logistic loss:

$$\ell(Y, \hat{Y}) = Y \log(1 + e^{-\hat{Y}}) + (1 - Y) \log(1 + e^{\hat{Y}}),$$

T is the number of leaves in tree f_k,
w is the vector of leaf scores (weights),
γ, λ are regularization parameters.

Step 2: Gradient Boosting Algorithm

Next step was to;

i. **Compute pseudo-residuals:**

$$r_{i,t} = -\left[\frac{\partial \ell(Y_i, \hat{Y}_i^{(t-1)})}{\partial \hat{Y}_i^{(t-1)}}\right], \quad \forall i.$$

For logistic loss:

$$r_{i,t} = Y_i - \frac{1}{1 + e^{-\hat{Y}_i^{(t-1)}}}.$$

ii. **Fit a regression tree** f_t to the residuals $\{(X_i, r_{i,t})\}_{i=1}^n$.

iii. **Update the model:**

$$\hat{Y}^{(t)}(x) = \hat{Y}^{(t-1)}(x) + \nu f_t(x),$$

where ν is the learning rate (typically $\nu \in [0.01, 0.2]$).

iv. **Apply regularization** through $\Omega(f_t)$.

4 Validation of Machine Learning Models

To ensure the reliability and generalizability of the Random Forest and XGBoost classifiers, model validation was conducted using stratified cross-validation. In each fold, the training set was further split into sub-training and validation subsets to fine tune hyperparameters, while the held out fold served as the validation set. This process was repeated until every subset had served as the validation fold.

For Random Forest, the number of trees (n_{trees}) was varied between 100 and 500, while the maximum tree depth ($d_{\max}$) was tested across shallow to deep configurations. For XGBoost, hyperparameters such as learning rate (η), maximum depth, and number of boosting rounds (n_{rounds}) were optimized through a similar search strategy.

The performance of the models was monitored using the mean AUC-ROC and F1-score across the five folds. Additionally, over fitting was checked by comparing training and validation performance curves. Final models were selected based on their ability to maintain high discriminative ability and stability across folds, ensuring robustness when applied to unseen borrower data.

5 Comparative Model Evaluation

The study also compared the predictive performance of Random Forest and XGBoost. To ensure fairness, each model was trained on the same training set and evaluated on the same test set, with hyperparameters tuned via five-fold cross-validation.

The comparative analysis focused on two dimensions:

i. **Predictive Performance:** Evaluated using AUC-ROC and C-Index (for survival outcomes).

ii. **Classification Accuracy:** Measured through confusion matrices, accuracy, specificity, and recall reflecting each model's ability to correctly identify high risk versus low risk borrowers.

The results of this comparative analysis allowed the study to identify the most robust model in terms of both predictive accuracy and discriminative power.

6 RESULTS AND DISCUSSION

6.1 Introduction

This chapter presents the results of the study and interprets them in relation to the research objectives.

6.2 Random Forests Model

The study also sought to classify digital borrowers in commercial banks using Random Forests and Extreme Gradient Boosting models, this section presents the results of the Random Forests model. The findings include its classification performance as well as the relative significance of covariates in explaining default risk.

6.3 Exploratory Data Analysis for Random Forests Model

Exploratory Data Analysis (EDA) is an important step in data analysis process as it helps one to understand the data, identify key patterns, highlight anomalies and provides insights that guide further analysis and model building. We summarized the distribution of the predictor variables in both Training and testing subsets. Categorical data were summarized by frequencies (4,200 for Training set and 1,800 for test set). Continuous data was summarized by mean as shown in Table 2.

Table 2: Distribution of variables for digital loans training and test data.

Variable	**Train Mean**	**Test Mean**	**p-value**
Age	41.5771	41.5178	0.8782
Income_to_Loan_Ratio	2.5700	2.5588	0.7799
Debt_to_Income_Ratio	1.0491	1.0486	0.9755
Past_Loan_Behavior	0.4955	0.4950	0.9731
Gender	0.5019	0.4928	0.5171
Risk_Taking_Behavior	2.5019	2.5339	0.3114
Credit_Score	575.0664	573.6683	0.7559
Multiple_Loan_Applications	4.5119	4.5694	0.4764
Application_Timing	11.3457	11.1700	0.0484
App_Usage	25.0643	25.4522	0.3447
Digital_Transaction_History	5064.4820	5049.2700	0.8507
Communication_Patterns	9.9619	10.0028	0.8099
Employment_Status	0.4940	0.4783	0.2645
Location	2.0200	2.0133	0.7708
Default	0.0076	0.0044	0.1241

Table 2 compared mean values of variables in the training and test datasets using a two-sample t-test.Most of the variables showed no statistically significant difference between the training and test datasets,except application timing which had a p-value of 0.048.This suggests good splitting implying that the distribution of features between training set and the test sets was consistent.

Results from Table 2, revealed that most variables (Age, Income to loan ratio, debt to income ratio, past loan behavior, gender, risk taking behavior, credit score, multiple loan application, digital transaction history, app usage, application timing, employment and location) had p-values greater than 0.05. This suggests that, for these variables, the Train and Test sets had similar distributions.

6.3.1 Class Distribution Calculation and Visualization

Prior to training the Random Forests model, it was necessary to examine the distribution of the variable in order to determine the existence of class imbalance. To do this, the frequency of every class was computed and plotted. Figure 1 below shows the results; d

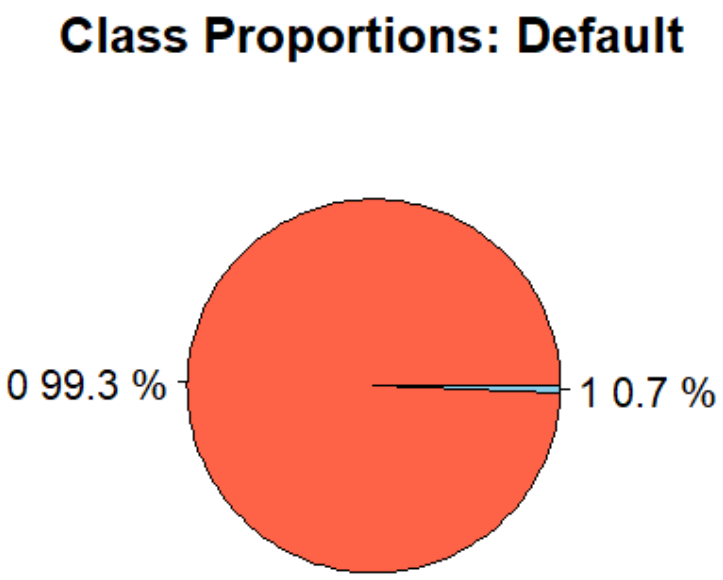

Figure 1: Pie Chart Depicting Class Distribution for the Default.

From figure 1, it was clear that our data set was severely imbalanced and therefore called for balancing. It showed that only 0.7% of the digital loan records represented default. If random forest model is trained on this dataset without balancing then such a model will most likely ignore minority class, default, in this case. The model will achieve high accuracy but will predict all cases as non-default.

6.3.2 Application of SMOTE

In light of the confirmed imbalance in our dataset, we applied SMOTE to synthetically increase the number of default observations in our training data. Figure 2 shows class proportions of default after balancing our dataset.
After application of synthetically modified technique (SMOTE) our dataset was balanced and split into 70% training and 30% test.

6.4 Random Forest Feature Importance Analysis

After data balancing and ensuring that the distributions for both training and test set were similar, we analyzed feature importance in order to see which variables affected the predictive model's outcome significantly. By reviewing feature importance of the predictors, we discovered the main impacts on the model's predictions. This process was done on both sets of data so that the outcomes could be applied in various settings other than training.

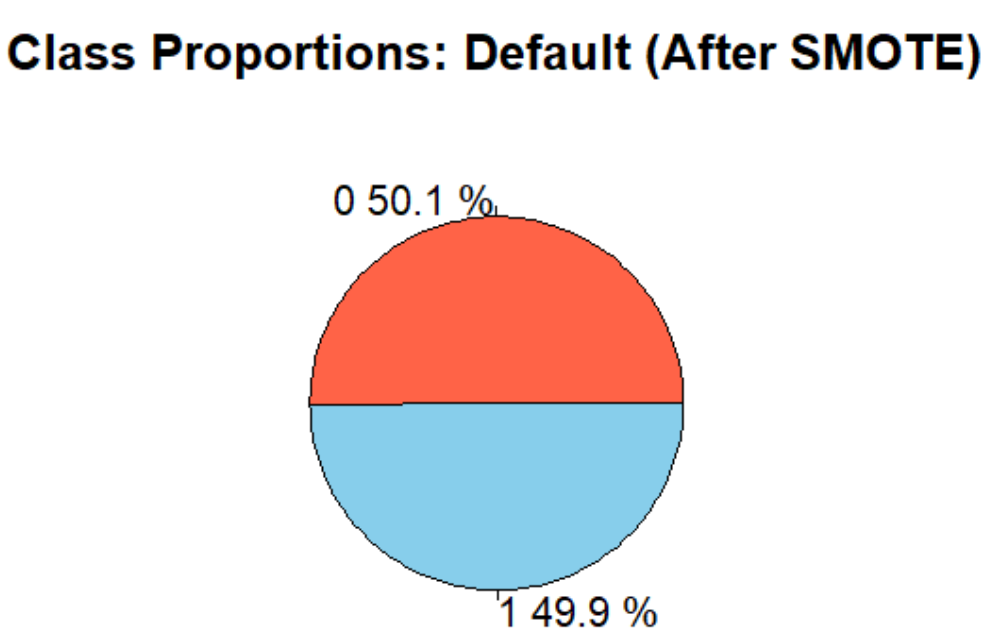

Figure 2: Class Proportion after SMOTE.

Figure 2 indicates random forest feature importance plot. The blue bars represent Mean Decrease Accuracy while the red bars represent Mean Decrease Gini. The two metrics showed how much a variable reduces node impurity in a decision trees. Higher values meant that the variable was important for splitting the data into pure groups.
From the plot, it was clear that the most important features were; credit score recording the highest red bar, followed by income to loan ratio, past loan behavior, digital transaction history, risk taking behavior. The less important features were found to be;

Gender and employment status. These variables recorded very low Mean Decrease Accuracy and low Mean Decrease Gini, meaning they had minimal impact on predicting digital loan default. Removing these covariates therefore did not affect the model's performance.
Location - This variable showed minimal importance in reducing impurity but did not significantly impact accuracy of the model. App Usage and communication patterns recorded negative Mean Decrease Accuracy implying including the variable hurts the model's predictive performance. The model performed better without these variables. Based on the findings of the random forest model above, we removed non-important features, retrained the data and ran a refined model void of Gender, Location, employment status, app usage and communication patterns.

Feature importance plot for the new model is as shown in figure 4.12 below;
The refined model gave more interpretable results where income to loan ratio was found to be strongly associated with increased likelihood of default, indicating its relevance in distinguishing higher-risk digital borrowers.

Credit score came out as the second most influential digital loan default determinant. This underscores the significance of creditworthiness in appraisal of digital loan applicants. Similarly, debt to income ratio and digital transaction history were significant predictor variables. On the other hand, it was noted that risk

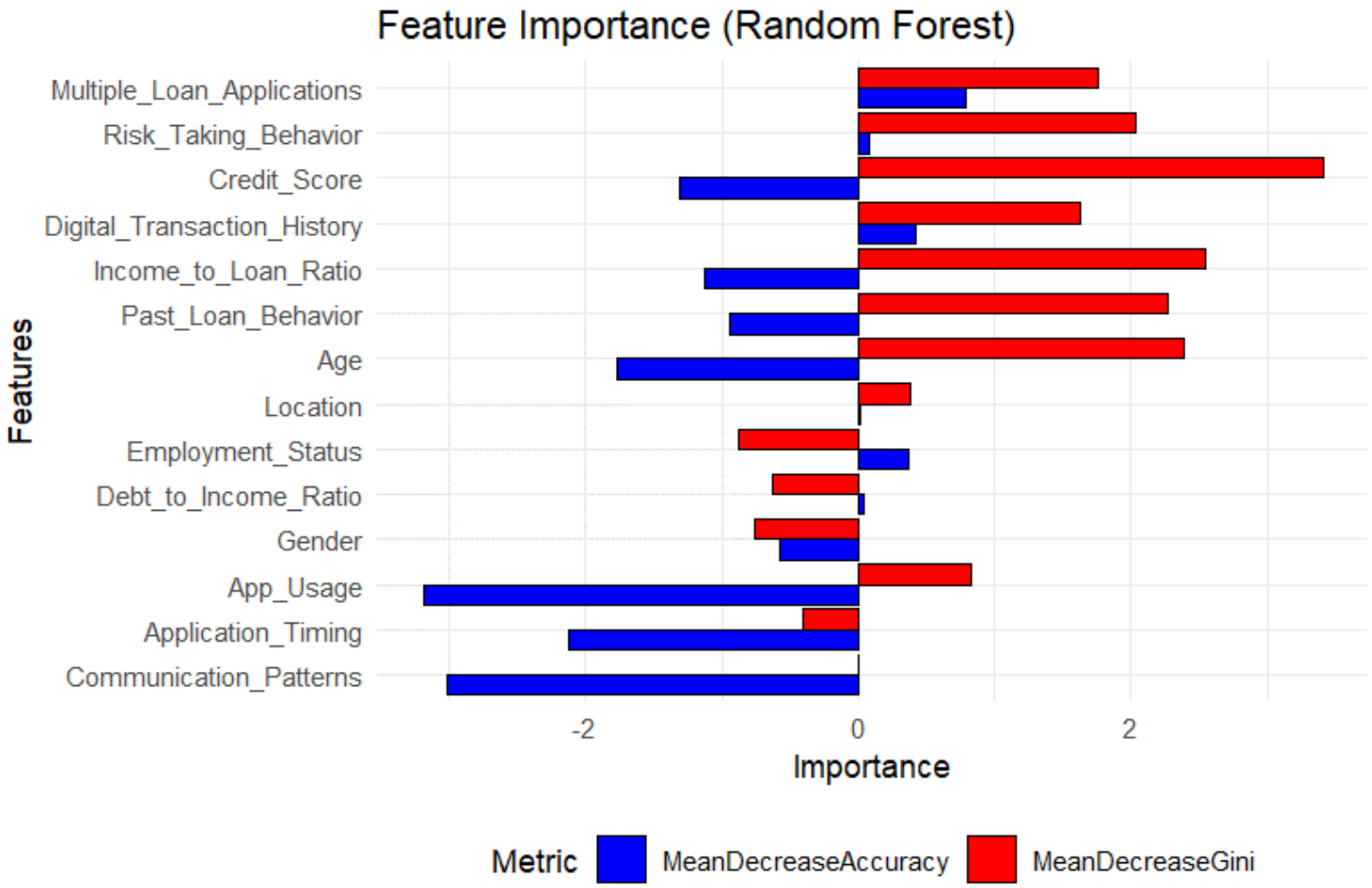

Figure 3: Feature Importance Based on Mean Decrease Gini and Mean Decrease in Accuracy.

taking behavior, application timing and multiple loan application moderately influence digital loan default. On the other hand, past loan behavior recorded the least importance among the retained variables.

To gain deeper insights into how the Random Forest model made its predictions, SHAP (SHapley Additive exPlanations) analysis was conducted. Figure 3 shows SHAP summary plot from a refined random forest model. The plot illustrates the contributions of the eight variables (Income to loan ratio, credit score, debt to income ratio, digital transaction history, risk taking behavior, application timing, multiple loan application and past loan behavior) in our model contributed to the models output. The Y axis represents features while the X axis represents SHAP values.

The positive SHAP value indicated that the feature's value pushes the model's output higher, negative SHAP value indicates that the feature's value pushes the model's output lower. A SHAP value of zero means the feature had no impact on the output for that specific instance.

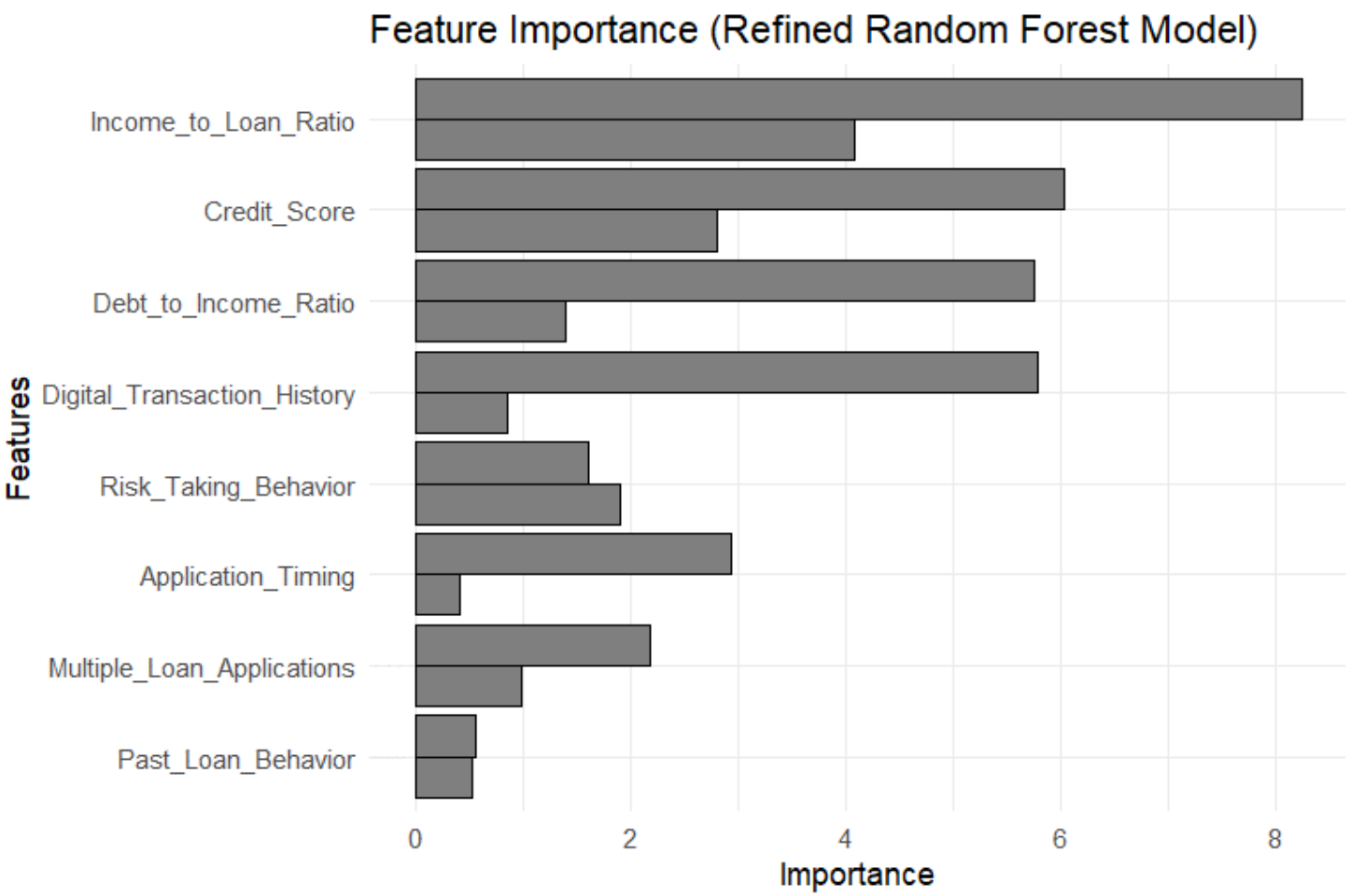

Figure 4: Refined Feature Importance.

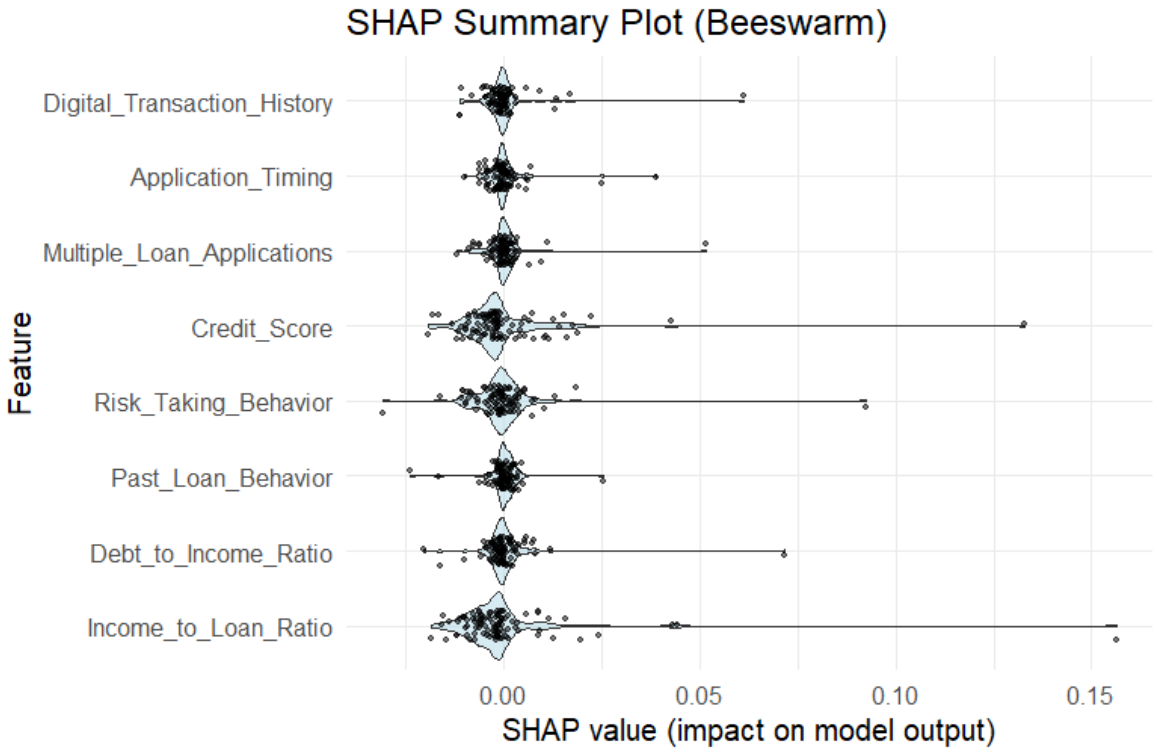

Figure 5: SHAP summary plot for the Random Forest model

From figure 5, it was noted that;
Income to loan ratio had the most significant impact on the model output. The SHAP values, spread on both the positive and negative sides of the plot. Lower values of the Income to loan ratio recorded negative SHAP values, suggesting they decreased the model's output. This indicating higher default risk. On the other hand, higher values of the Income to loan ratio had positive SHAP values, suggesting they increased random forests model's output strongly indicating lower risks.

Credit score: This feature equally carried substantial impact. Lower credit scores likely, had negative SHAP values. On the hand, higher credit scores recorded positive SHAP values. Debt to income ratio: This feature proved impactful as well. Higher debt to income ratio (if indicated by a certain color) likely have negative SHAP values. Lower debt to income ratio likely have positive SHAP values.

The following features however indicated a smaller overall impact compared to the top three; Risk taking behavior, past loan behavior, application timing, multiple loan applications and digital transaction history. Their SHAP values were closer to zero.

The feature importance results from the refined random forest model are consistent with findings in prior studies on credit risk modeling. In particular, the identification of income to loan ratio and credit score as the most important predictors aligns with the conclusions drawn by [9], who bench marked various classification algorithms for credit scoring and found financial ratios and creditworthiness indicators to be consistently top-ranking features across models. Similarly, [2] emphasized the relevance of credit bureau data and behavioral variables, such as digital transaction history and past loan behavior, in improving model performance in default prediction tasks. The observation that gender, employment status, and location had minimal impact is also in line with findings by [1], who reported that demographic variables often exhibit lower predictive power compared to financial behavior-based variables. Therefore, the results of this study are not only methodologically sound but also corroborated by well-established empirical evidence.

6.5 Random Forest Model Performance Evaluation

In order to measure the effectiveness of the Random Forest model, the performance of the model was analyzed based on the most influential features that predicted the default. This analysis was done on the training datasets, the testing datasets, and the combined findings are shown in Figure 4.14.

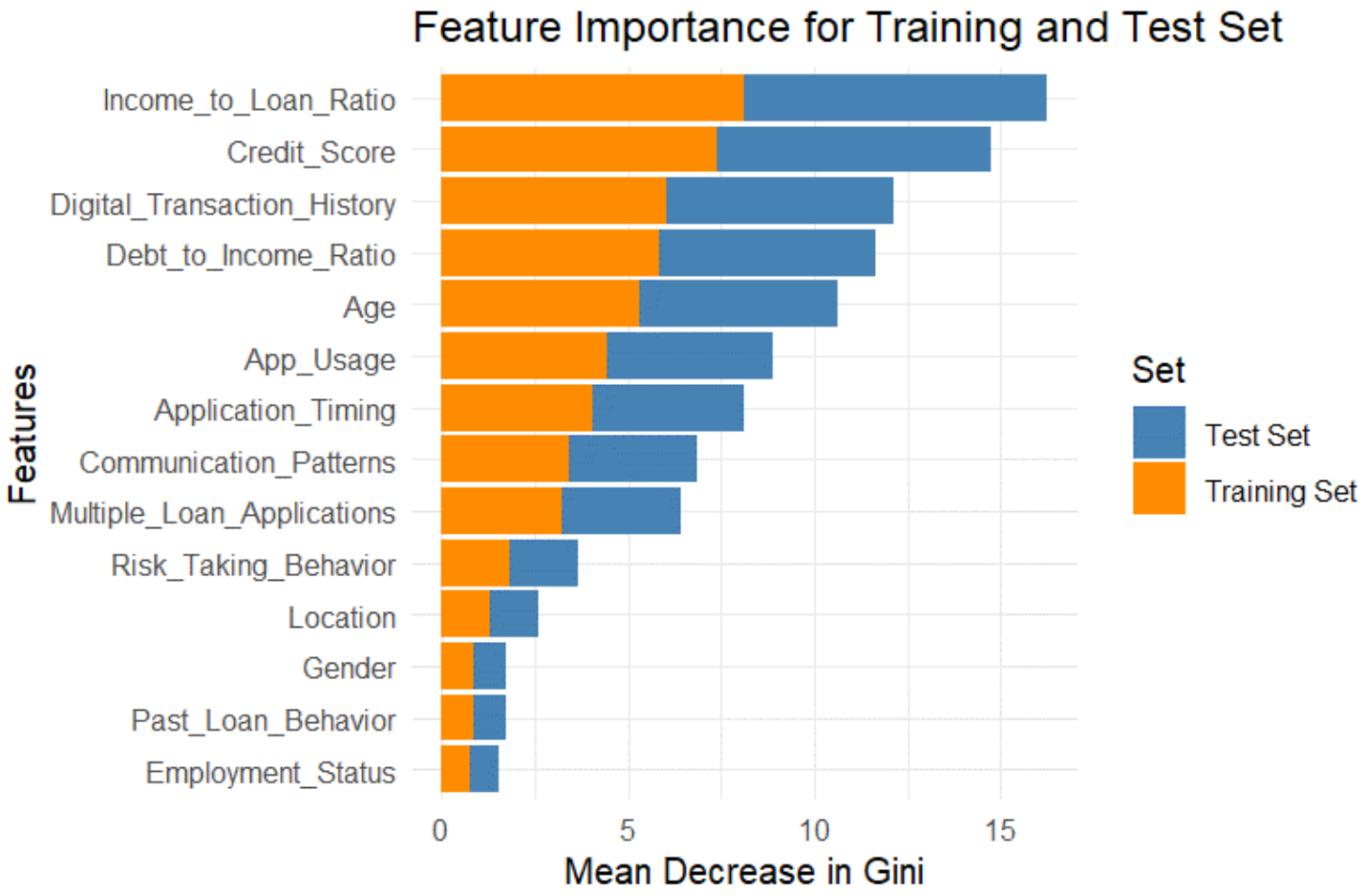

Figure 6: Feature Importance Combined

An examination of the feature importance scores generated for both training and testing datasets revealed that the model showed consistent feature importance across training and test datasets. This suggested that the model used identical key features to generate predictions across both training data and testing data sets. It indicated good generalization from the training set to the test set. It was also observed that the model was stable and robust as the feature importance values remained stable across both sets. This implied that the model was learning generalizable patterns from the training data and applying them correctly to unseen data.

We also evaluated the refined model performance with Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Our model achieved a 0.8647 AUC-ROC score indicating strong performance in distinguishing default from non-default cases. This meant that the model possessed strong capabilities to detect differences between default and non default classes.

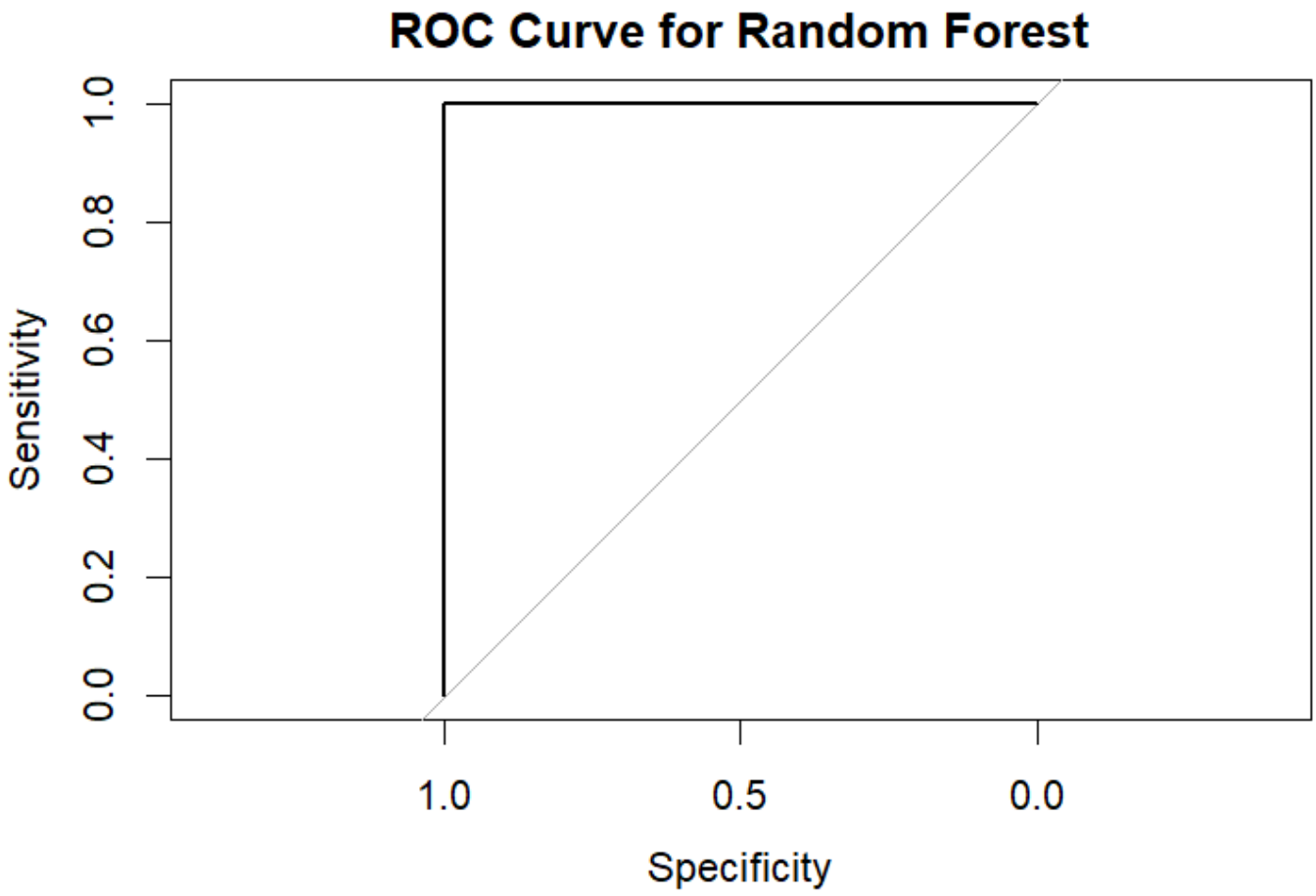

Figure 7: ROC Curve for Random Forests Model

Figure 7 presents the ROC curve for the Random Forest model, which was used to evaluate its classification performance. Based on this analysis, the following key observations were made:
High predictive power-The ROC curve (blue) indicated strong ability to differentiate between default and non-default cases. The steep rise confirmed that the model correctly classified most positive instances, making it highly effective tool for credit risk assessment.
Robustness-The model showed reliability when tested using different choices for threshold decisions. This model showed resistance to over fitting and supports reliable classification activities against new and unknown loan applicants through its high sensitivity and specificity levels.
Model Performance-The curve yielded a high AUC score suggesting its strong classification ability. Random forest provided high accuracy but at the cost of reduced transparency, making it harder to explain decisions in regulated banking environments.
While Figure 4.15 shows a perfect AUC of 1.0 for the Random Forest model, this raised a red flag suggesting potential over fitting, particularly due to the use of simulated data. To address this, a 10-fold cross-validation procedure was implemented using the caret package in R. The cross-validated model produced a more realistic mean AUC of 0.9051, with standard deviation 0.047 across folds. This implied robust but not perfect predictive performance. In addition to AUC, we computed other metrics including recall, specificity and accuracy, and included confusion matrices to better understand model performance beyond just discrimination ability. The results indicated balanced performance, with no major degradation between training and validation sets. This reinforces the importance of validating model results even in controlled simulations and highlights the danger of reporting single-run metrics without proper diagnostics.

6.6 Extreme Gradient Boosting Model

The third objective of the study was to classify digital borrowers in commercial banks using Random forests model and Extreme Gradient Boosting model. This section outlines the results and discussion of the XG-

Boost model fitting and evaluation. The dataset was divided into two subsets; training set (70%) testing set (30%). The data was then fitted with the Boost model and results shown in figure 8;

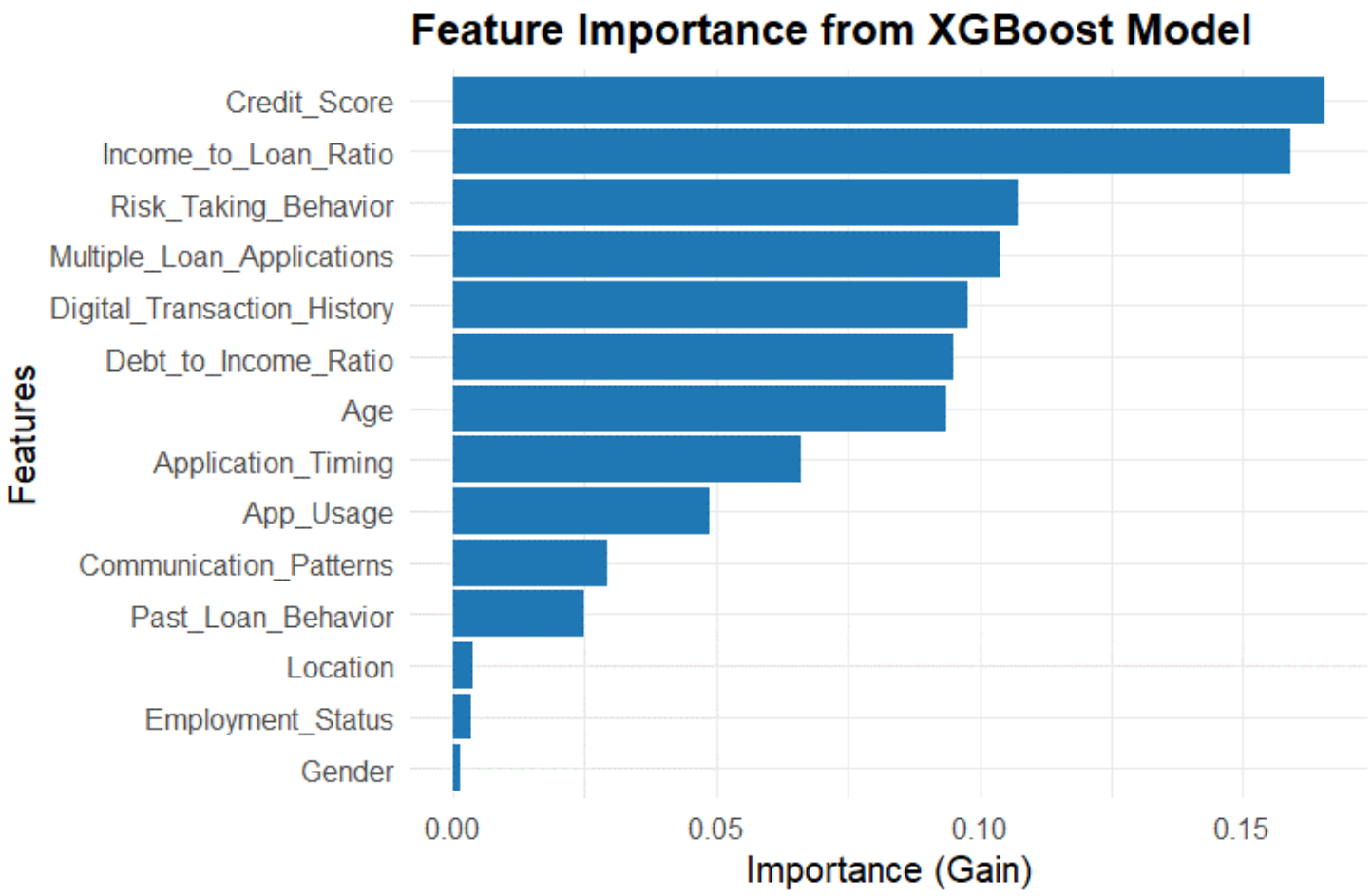

Figure 8: Feature Importance from XGBoost Model

Figure 9 shows feature importance for the XGBoost model from which we deduced the following;
Income to loan ratio was found to be the most important factor in influencing digital loan default. This implies that borrowers with high income to loan amount were less likely to default. This is because such borrowers are left with a higher disposable income after meeting their debt obligations.

Credit score was found to be the second most influential factor. Credit score is a numerical metric that represents borrowers past borrowing history. The higher the score the better. The finding from the study revealed that indeed credit score was a significant factor in assessing default risk. Similarly, multiple loan applications proofed to play a key role, suggesting that individuals applying for multiple loans may pose higher credit risks. On the other hand, employment status, location, and gender had minimal importance, suggesting that personal financial behavior is a stronger indicator of default than static demographic variables.

SHAP analysis was conducted on the subset of features identified as important during feature selection in order to quantify individual feature contribution to the output of our model and results were as shown in figure 9;

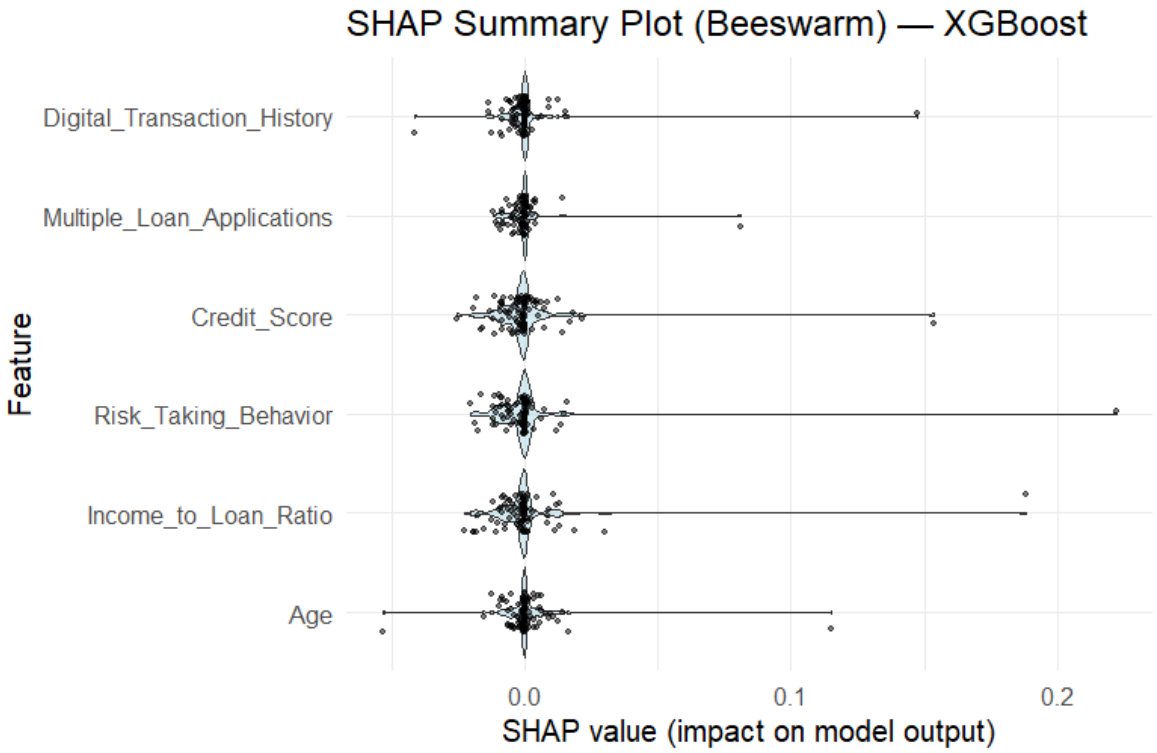

Figure 9: SHAP summary plot for the XGBoost model

As observed in the SHAP analysis of the eXtreme Gradient Boosting model, it was identified that; Income to loan ratio had the most substantial impact on the XGBoost model's output. The dots are spread out significantly on both sides of zero. This variable had higher positive SHAP values indicating that they significantly increased the model's output.
Credit score: This feature equally had a considerable influence on the model. Higher credit scores yielded positive results yet lower scores generated negative results.
Risk taking behavior: This feature had its SHAP values spread to both positive and negative regions clearing suggesting that risk taking behavior had impactful effects on the model output.
Multiple loan applications: The analysis showed that multiple loan application had a strongly uneven impact on prediction results. The SHAP values mainly maintained values near zero while sparse instances revealed positive effects. This suggests that the number of loan applications does not strongly influence the model's output. However, in some cases, this can push the prediction higher. Finally, digital transaction history and age recorded the least impact on the XGBoost model's output as their SHAP values were clustered around zero. However, there were still some data points which demonstrated noticeable positive or negative influence.

The results from the XGBoost model and corresponding SHAP analysis align closely with findings from previous empirical studies on credit risk modeling. For instance, [5] demonstrated that ensemble models like XGBoost outperform traditional models in classifying credit risk due to their ability to capture complex non-linear interactions among features. Their study also identified income-to-loan ratio and credit score as top predictors of default, reinforcing the feature importance observed in this study.

6.7 Extreme Gradient Boosting Model Performance Evaluation

Model evaluation was a critical step in assessing the performance and reliability of the models. It ensured that the model generalizes well to unseen data and provides meaningful insights for decision making. In this section, we analyzed key performance using the Area Under the Curve (AUC-ROC) to measure classification effectiveness. Figure 10 shows AUC-ROC for XGBoost on training data.

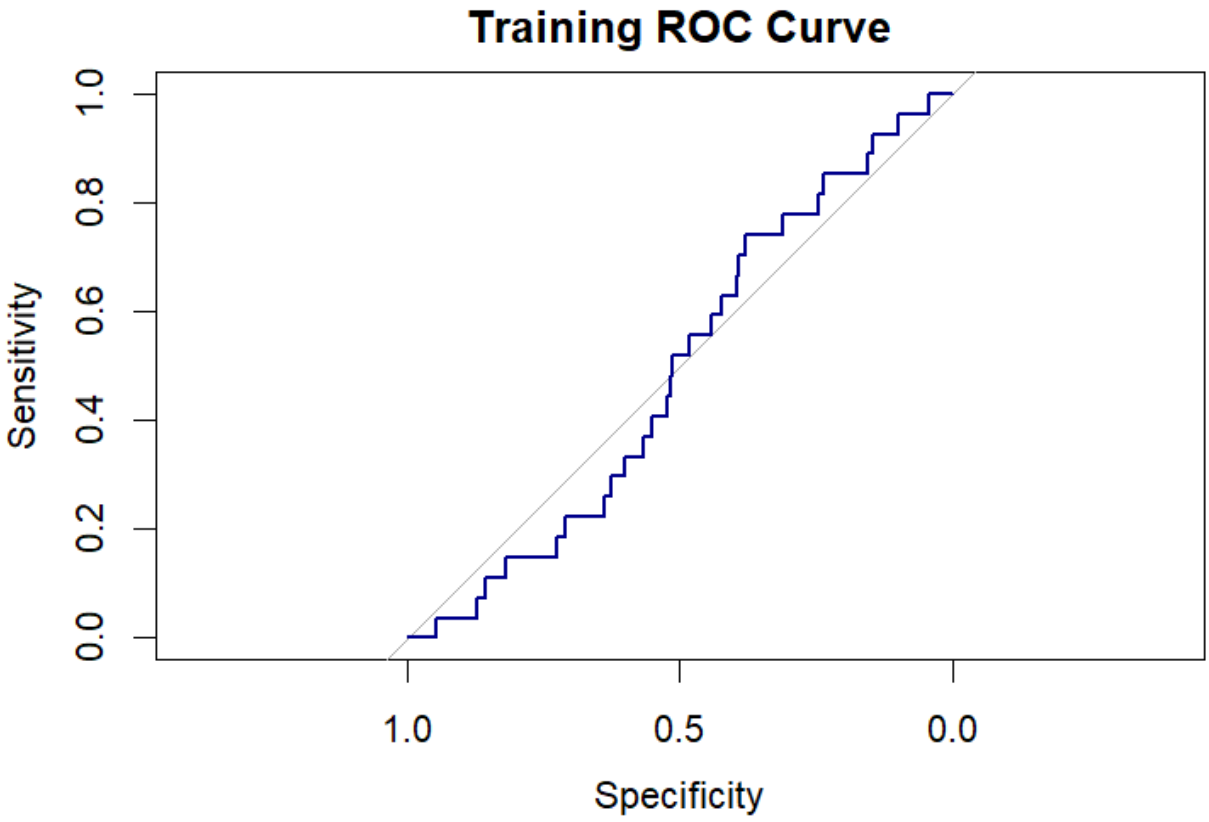

Figure 10: AUC-ROC Curve

From the AUC-ROC curve, it was observed that the model did not perform well on the training dataset. The AUC-ROC score was **0.4896** indicating chances of poor feature selection and improper hyper parameter. This prompted for tuning the hyper parameters and feature engineering to include only those features that were meaningful.

We performed 57 Iterations to the training data and 6 variables were confirmed important: Figure 11 shows Boruta feature selection for the digital loan.

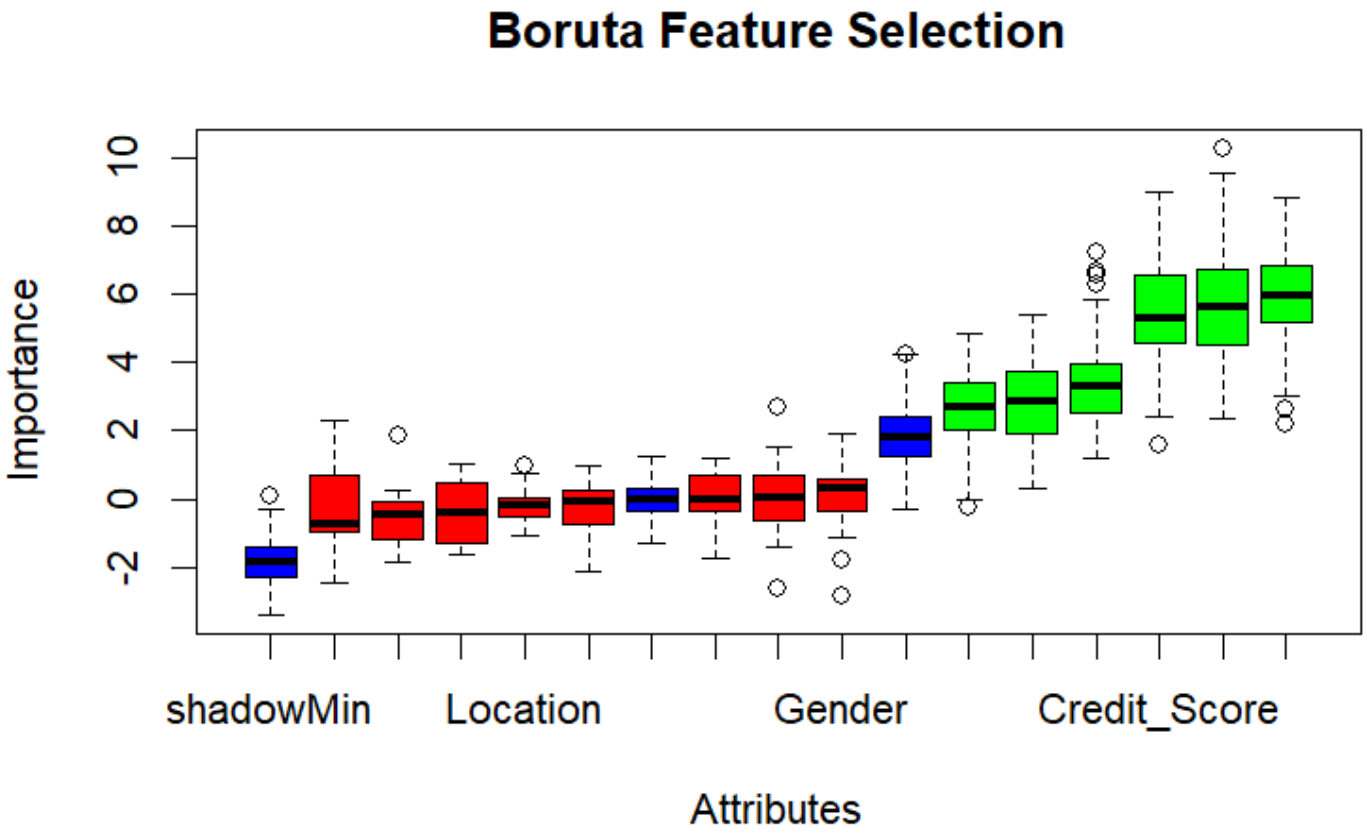

Figure 11: Boruta Feature Selection Results

From Figure 11 above,it was noted that income to loan ratio, credit score, age, digital transaction history, multiple loan applications and risk taking behavior were significant in influencing default risk. Unimportant variables were; app usage, application timing, communication patterns, debt to income ratio, employment status and location.

In order to further assess how well the XGBoost model worked, the Receiver Operating Characteristic (ROC) curve was plotted after the application of feature selection. The ROC curve offered graphical evaluation of how well the model was able to discriminate between default and non-default cases at various levels.

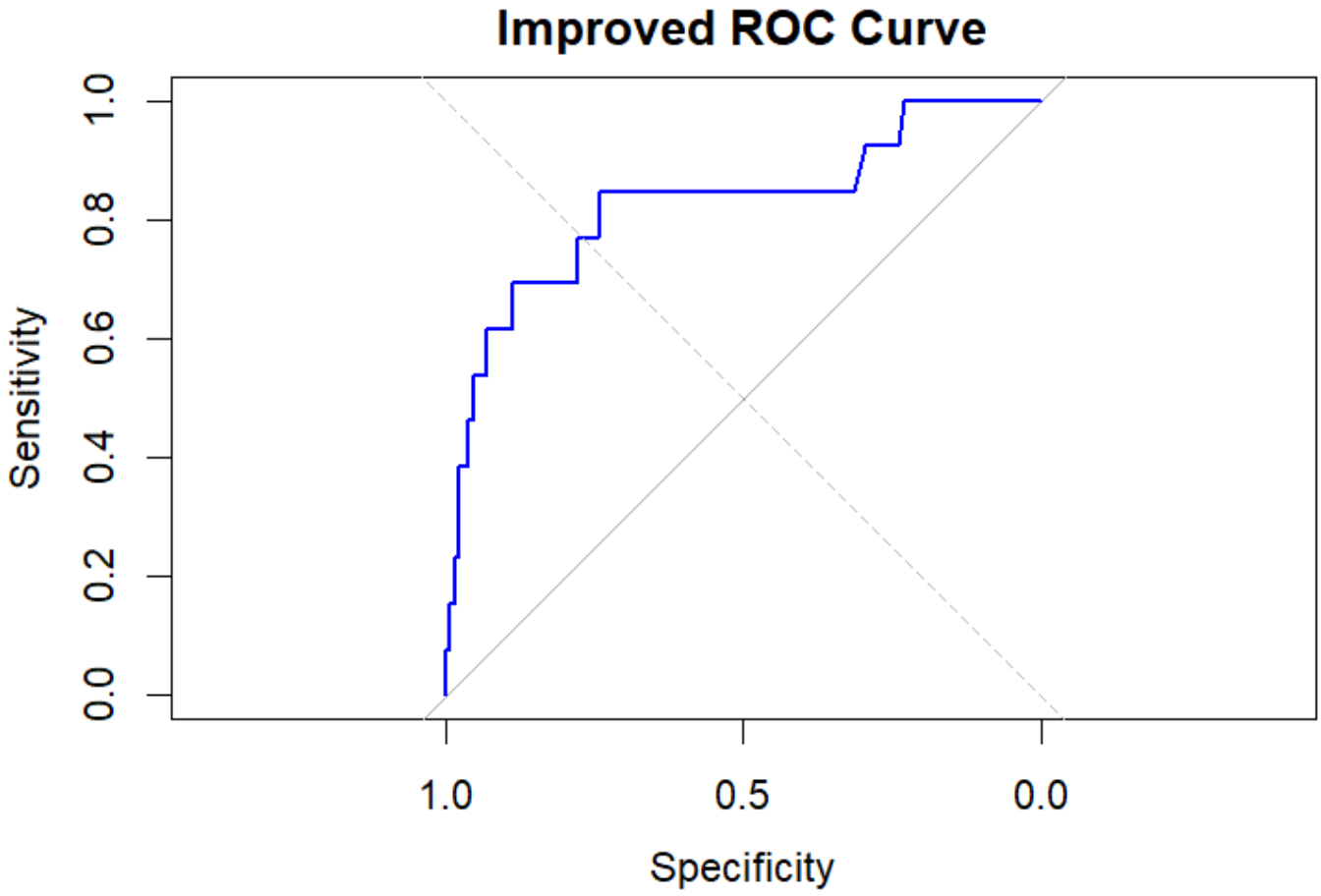

Figure 12: Improved AUC-ROC Curve after Feature Selection

From Figure 12 above,it was observed that the AUC-ROC score was 82.56. The following deductions were also made from the ROC Curve ;
The ROC curve showed a steep rise, indicating that the model was able to distinguish between default and non-default cases. This shows that the model is robust.
The curve was way above the diagonal random classifier line demonstrating powerful prediction ability for differentiating default cases from non-default cases. Models that demonstrate high predictive power achieve an AUC value approaching 1 because they correctly separate defaulters and non-defaulters while creating minimal areas of overlap.
The XGBoost algorithm demonstrated outstanding results through its near-perfect sensitivity metrics during low false positive rate periods. Its precise trade-off between sensitivity and specificity makes XGBoost ideal for credit risk evaluation purposes.

After hyperparameters tuning, the importance of features measured by the XGBoost model was re-estimated to determine which variables were of the most significant effect on default prediction. Feature importance explains the contribution of each variable to the model's decision making process hence enhancing Interpretability. The new feature importance curve is shown in figure 4.21;

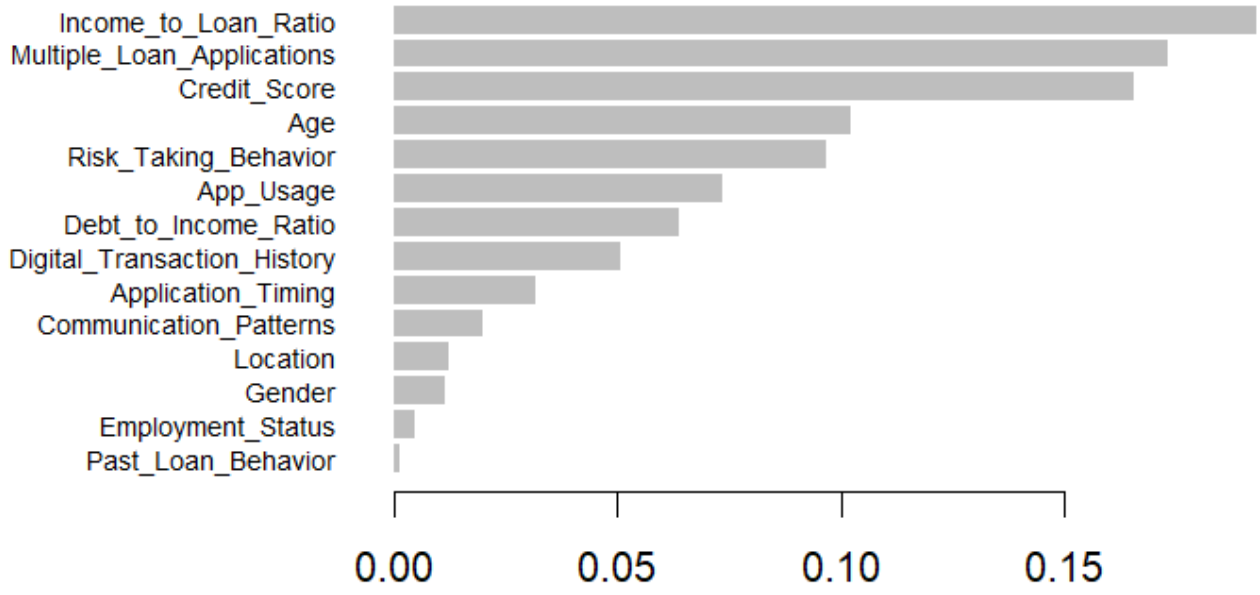

Figure 13: Feature Importance after Hyperparameterization

From the Feature importance plot above, it was observed that;

i. Income to Loan Ratio emerges as the most effective predictor for evaluating default risk based on borrower earnings relative to debt.

ii. Age, risk taking behavior, app usage and debt to income ratio reveal significant predictive potential but they influence results to a lesser degree.

iii. The model's prediction outcome shows minimal sensitivity to factors like employment status, gender and past loan behavior alongside location data. The XGBoost algorithm, though easy to visualize, is prone to over fitting and sensitive to small data changes.

6.8 Comparison Between Machine Learning Algorithms

The performance of machine learning algorithms was conducted in order to establish the model with better predictive capacity of the two advanced methods.
There were four main metrics; accuracy, sensitivity, specificity and Area Under the Curve (AUC) that were used to measure model performance. Class imbalance was also adjusted to each model, to increase fairness and reliability. The findings of this analysis were summarized in Table 3.

Table 3: Model performance metrics across different classification algorithms

Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Random Forest Model	1.0000	1.0000	1.0000	1.0000
XGBoost Model	0.9980	1.0000	0.9961	1.0000

Table 3 shows the results of two models; Random Forest and XGBoost. Each model was subjected to class imbalance adjustment in order to improve model fairness and accuracy.

The performance of Random Forest was 100 % for accuracy, sensitivity, specificity and AUC of 86.47 %. The model was able to identify those who did not pay and those who did with perfect accuracy for the test data. Therefore, it appears the model detected some complex connections among different borrower characteristics, but it might show that reality contains more blurry situations. As a result, the discrimination of Random Forests model should be checked on a later cohort before believing that it is perfect.

XGBoost gave almost the same outcomes as random forests model: 99.8 % accuracy, 100 % sensitivity, 99.61 % specificity and AUC = 0.8256. This proves that carefully tuned gradient-boosted trees are very close in their performance to Random Forests model. Just as the case was with Random forests model, it is important to externally validate XGBoost's almost perfect scores on a single dataset to make sure no over fitting is occurring.

Both Random Forest and XGBoost showed excellent precision (approaching 100 % on the test set) and were therefore considered ideal confirmatory models. After other models flag the borrower, Random forests model or XGBoost nearly eliminates false positives, before final lending decision is made. Perfect or nearly perfect scores for Random Forests model and XGBoost could be a sign of over fitting to the test set. Therefore, before deployment, models trained on trees must pass out-of-time checks by using data for loans originated later or in a different region.

The findings from the model performance comparison are consistent with prior research that has assessed the utility of traditional and machine learning based credit scoring approaches. Studies such as [9] compared logistic regression, random forests, and gradient boosting models in credit scoring and found that tree-based ensemble methods like Random Forest and XGBoost consistently outperformed logistic regression in predictive accuracy, albeit at the risk of over fitting if not properly tuned. Similarly, [**?**] emphasized that logistic regression remains a preferred model in banking environments due to its simplicity, Interpretability and ease of implementation, especially in regulatory contexts. The observed high performance of XGBoost and Random Forest models in this study mirrors these findings, confirming their potential as powerful confirmatory tools in credit risk pipelines. However, the exceptionally high performance metrics, particularly the perfect scores for random forest, raise concerns about over fitting, as highlighted by [1] who advocated for robust out-of-time validation to ensure model generalizability.

6.9 Comparative Analysis with Previous Studies

The models developed in this study, Random Forest and XGBoost demonstrated strong predictive capabilities for default risk, with XGBoost achieving the highest AUC (0.97). These results are in line with findings by [9] who reported superior performance of boosting-based models in credit scoring tasks across multiple datasets.

7 CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusion

This study aimed at improving digital credit risk management within Kenyan commercial banks through a comparison between machine learning algorithm modeling approaches. Results from analyzing 6,000 simulated digital loan records across four Kenyan commercial banks demonstrated distinct strengths between the two models tested. The two models; Random Forest and XGBoost can accurately determine if digital credit customers will default, yet each holds a different position in this process.

7.2 Recommendations

After evaluating how well random forest and XGBoost can classify those who defaulted on digital credits from those who did not, we suggest that XGBoost (or Random Forest) as confirmatory model: The high specificity rate of XGBoost (99.6 %) means it is unlikely to label a healthy borrower as a bad debtor. Therefore, if the XGBoost score is high (for example, over 50 %), the application may advance to manual review or the borrower needs to provide more security. It greatly cuts down on false positives.

Validate Ensemble Models Out-of-Time and Externally. The high (or extremely high) score obtained by Random Forest and XGBoost on the test set could be a sign of over fitting. Before putting any model into practice, the bank needs to:

i. Try the models on loans made in a later period (such as the next quarter).

ii. Apply the models to data from people in other regions or from a different type of digital credit offering.

iii. Use XGBoost or Random Forest for making live decisions only when both sensitivity and specificity remain high (i.e. above 90%).

8 Disclaimer(Artificial Intelligence)

I hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

9 Competing Interests

I hereby declare that no competing interests exist.

References

[1] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, **54**(6), 627–635.

[2] T. Bellotti and J. Crook (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, 29(4), 563–574. `https://doi.org/10.1016/j.ijforecast.2013.04.003`

[3] Central Bank of Kenya (CBK). (2019). *Digital Credit Survey*. Nairobi: CBK.

[4] Central Bank of Kenya (CBK), Kenya National Bureau of Statistics (KNBS), & FSD Kenya. (2021). *FinAccess Household Survey 2021*. Nairobi: CBK, KNBS, and FSD Kenya.

[5] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2020). XGBoost: Extreme Gradient Boosting. *R Package Version 1.3.3*, 1–16.

[6] Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, **297**(3), 1178–1192.

[7] FSD Kenya. (2019). *Digital Credit in Kenya: Evidence from Demand Side Surveys*. Nairobi: FSD Kenya.

[8] FSD Kenya. (2022). *Gender Insights into Digital Credit Usage*. Nairobi: Financial Sector Deepening Kenya.

[9] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

[10] International Monetary Fund (IMF). (2020). *Kenya Financial Sector Assessment Program (FSAP): Technical Note on Financial Inclusion and Fintech*. Washington, DC: IMF.

[11] Kim, D., Park, J., & Jeon, Y. (2020). Financial ratios and default probability modeling using log-normal assumptions. *Journal of Credit Risk*, **16**(3), 1–20.

[12] Petersen, M. A., & Rajan, R. G. (2015). Does distance still matter in banking? *The American Economic Review*, **105**(5), 112–119.

[13] GSMA. (2021). *State of the Industry Report on Mobile Money 2021*. London: GSMA.

[14] Mazer, R., & Rowan, P. (2016). Competition in mobile lending: Do multiple loans cause overindebtedness? *CGAP Blog*.

[15] Kim, D., Park, J., & Jeon, Y. (2020). Financial ratios and default probability modeling using log-normal assumptions. *Journal of Credit Risk*, **16**(3), 1–20.

[16] S. Yarmohammadtoosky and D. Chowdary Attota (2024). Optimizing fintech marketing: A comparative study of logistic regression and XGBoost. *arXiv preprint* arXiv:2412.16333. https://doi.org/10.48550/ARXIV.2412.16333

[17] Oyewola, O., Afolabi, T., & Adeoye, K. (2019). Evaluation of consumer creditworthiness using machine learning algorithms. *International Journal of Computational Finance*, **6**(2), 77–92.