Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250403021
Volume 4, Issue3 (September-December 2025), PP: 111-120.
www.indjcst.com
ISSN No: 2583-5300

Credit Risk Indicators for Microfinance Institutions within Machakos County, Kenya

Martin M. Kasina*[1], John M. Kihoro[2], Dr. Alex Kibet[3]
[1,2] *Department of Information Technology and Computer Science, The Cooperative University of Kenya, Kenya.*
[3]*Lecturer, Department of Computing and Informatics, School of Science and Applied Technology, Laikipia University, Kenya.*

To Cite this Article: *Martin M. Kasina*[1], John M. Kihoro[2], Dr. Alex Kibet[3], "Credit Risk Indicators for Microfinance Institutions within Machakos County, Kenya", Indian Journal of Computer Science and Technology, Volume 04, Issue 03 (September-December 2025), PP: 111-120.*

Abstract: *Microfinance institutions in Kenya play a unique role in promoting financial inclusion, loans, and savings provision, especially to low-income individuals and small-scale entrepreneurs. However, despite their benefits, most of their products and programs in Machakos County have been reducing due to repayment challenges, threatening their financial ability to extend further credit. The objective of the research was to establish key credit risk indicators for microfinance institutions operating within Machakos County, Kenya. The study adopted a mixed research design using supervised machine learning approach. It randomly sampled 6771 loan appli- cation account records and repayment history.Rstudio and Python programming languages were deployed for data pre-processing and analysis. The logistic regression algorithm, XG Boosting and the random forest ensemble method were used to rank the feature importance. Based on the study findings; The amount of loan required, the income level, the gender and the age of the applicant were the main features that influenced loan default rate. Integration of the hard and soft data into machine learning for better credit risk assessment outcome was recommended. Similar research but using different target population and institutions, to as- certain the validity, reliability and the generalizability of the study findings was recommended for further research.*

Key Words: *Feature importance; Loan payment default; Classification modeling.*

I.INTRODUCTION

Default refers to the failure of the borrower to meet the legal obligation of the loan payment. Although Kenyan SMEs have gained a lot of attention in the recent past, most of them do not live long enough to celebrate their first birthday, while others stagnate in growth for so too long due to non-performing loans [1]. The study also noted that, the evidence of the default pattern of loans to individuals and SMEs in a fast growing emerging economies was still missing. Financial institutions are facing a shocking number of increased loan defaulters all over the world. According to [2] and [3] most of the MFIs products and programs in Machakos County were falling due to repayment problems, threatening their financial sustainability and the ability to extend additional credit. Having a good understanding of the signs and characteristics of willful defaulters can help financial institutions and other lenders take preventive measures. Therefore, this research paper sought to identify key risk indicators among loan applicants using machine learning techniques in MFIs operating within Machakos, Kenya.

II.MATERIAL AND METHODS

The amount of money borrowed, the cost of the loan and the reason of borrowing the money may influence the default rate. The study by [4] on factors which affect the loan repayment rate among the farmers in Ethiopia addressed the impact of many factors, that included; loan ratings and the cost of borrowing, on default rate. The study observed that, the higher the cost of the loan charged, the higher the chances of defaulting on the payment. The study further noted that, the borrower's demographic characteristics also significantly influenced the default rate whereby, the default rate was higher among the young applicants compared to the aged ones. The research conducted by [5] on determining factors of default risk in Savings and Credit Groups found that, the probability of having problems in loan repayment was higher for male borrowers than for females counterparts. The study by [6] which focused on understanding the impact of borrowers' behavioural and psychological traits on credit default revealed that, some non-financial factors, that can be assessed and looked into, then quantify and verify their soft information (character and reliability) of debtors while granting a loan to a borrower could help in safeguarding against bad debts among the MFIs. This was helpful for applicants without an appropriate advanced credit history and reliable documented financial statements. The research by [7] on institutional factors and efficiency performance in the global microfinance industry MFIs, noted that the location of a MFI affects the default rate and performance significantly whereby the default rate was higher in urban set up than rural areas. The study also noted that the MFIs managed by women had less defaults compared to those administered by their male counterparts, thus bringing out the key and central role of women in the business achievement of MFIs. The interrelationships between the variables of the study is illustrated in the conceptual framework given in figure 1

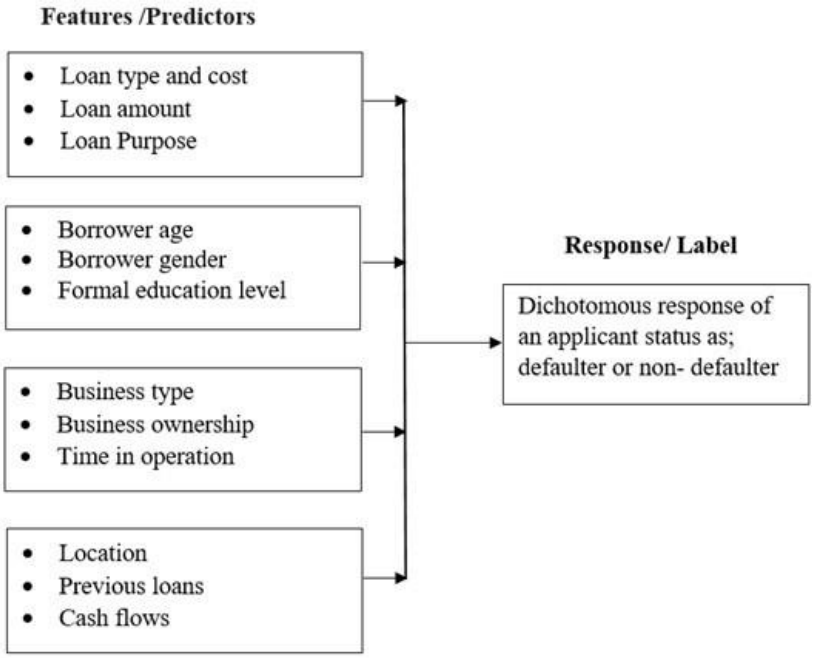

Figure 1: The Conceptual Framework

Individual loan applicant characteristics included the sex category, the number of years lived, marriage status, residence, and the highest formal training level attained. The details of SMEs business unit included, the business location, time in operation, previous loan payment behavior, and cash flow. The response variable was a credit risk status, categorical label of a loan applicant, categorized as either a default or non- default.

III.METHODOLOGY

The study adopted a mixed research design under supervised machine learning approach. The target population for this project were the loaned individuals and SMEs within the period 2020 to 2024 by the 21 registered MFIs operating in Machakos County, Kenya.

3.1 Preprocessing

To ensure the data structure conformed to the machine learning formats, a preprocessing analysis was carried out. In order to ensure that there was no vital information unnecessarily got lost, mean imputation was performed for numeric features to address the missing data values while mode imputation was used for categorical features. Capping off the most extreme data values, categorical encoding, erasing all the duplicates was done before scaling the continuous variables into a zero to one range using normalization techniques so as to accommodate most of the ML algorithms.

IV.MODELS

The study used; Logistic Regression algorithm, Extreme Gradient Boosting and Random forests ensemble method to analyze the data for prediction and classification of loan applicants as default- ers or non-defaulters. Rstudio, Excel and Python programming languages were deployed for data pre-processing, manipulation and classification.

4.1 Logistic Regression

Logistic regression method for classification was used whereby the dichotomous response variable of whether the applicant's status was non-defaulter (0) or defaulter (1) such that; by letting the independent design matrix be X, then the response variable y with the binary values was;

$$Y = \begin{cases} 0 & \text{if non-defaulter} \\ 1 & \text{if a defaulter} \end{cases}$$

Then the $log(Odds(defaulter)) = logit\pi$ such that;

$$log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{4.1}$$

where π is the likelihood of the credit risk score status of a loan applicant. The sigmoid function is most appropriate for performing logistic regression because it maps the linear combination of input features to a probability. The sigmoid function is given by equation 5.2.

$$S(x) = \frac{1}{1 + e^{-x}} \tag{4.2}$$

where x is the input variable such that; as $x \to \infty$, the function approaches 1 and as $x \to -\infty$, the function approaches 0 resulting in the sigmoid shape illustrated in Figure 2 The sigmoid function takes any real input log-odds and gives output probability. Such that $\sigma : \mathsf{R} \to (0, 1)$. where x is a linear function of multiple explanatory variables such that.

$$S(x) = p(x) = \frac{1}{1 + e^{\beta_0 + \beta_i x_i}}$$

In the logistic model, $p(x)$ is interpreted as the probability of a binary response variable y whereby, the normal threshhold for the event being successful ranges between 0.5 to 1, otherwise it is deemed not successful. That is in this case, the probability of failing to pay ranged between 0.5 to 1. Next, the log odds inverse is explained as

$$\text{logit}\, p(x) = In \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_i x_i \tag{4.3}$$

Then equivalently after exponentiation in both sides, we get the odds such that,

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_i x_i} \tag{4.4}$$

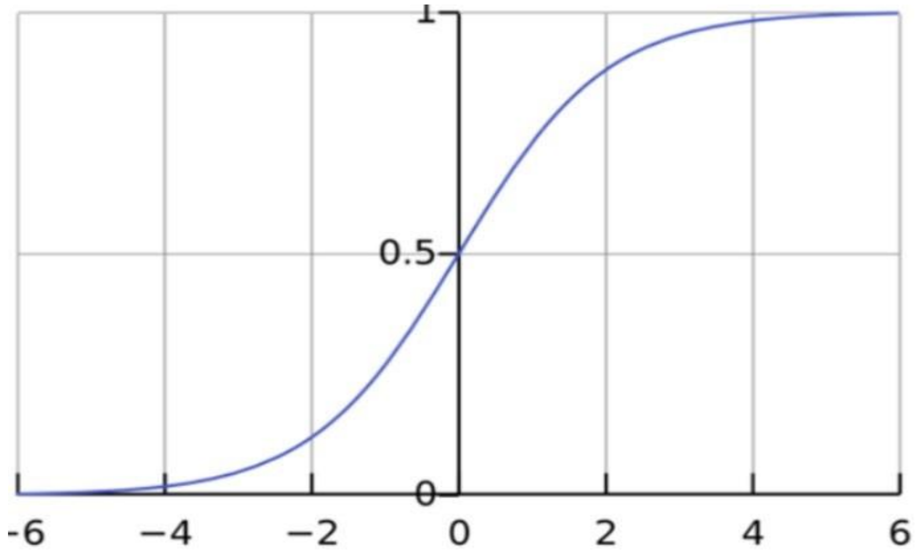

Figure 2: The Sigmoid function.

Whereby

$p(x)$ is the probability that the loan is not paid given some linear combination of the features.β_0 is the intercept from linear regression when the numeric feature is equal to zero and the cate- gorical feature is equal to the referenced feature. The study used both the regression coefficients and the odds ratio to establish the feature contribution and importance in the model.

4.2 Random Forest

Decision tree which is a simple model that divides data into different categories based on certain characteristics, is the foundation of the random forest. Decision trees mostly use entropy and information gain to decide where to split the data. Such that by letting

$entropy = H$, then $H(S) = \sum_i^n pi \log_2 pi$.

where pi is the probability of each class in the dataset and n represents the number of classes.Main idea is to select the feature that leads to the greatest reduction in entropy (informationgain). According to [8] ”A Random Forest model is a robust supervised ML algorithm composedof a tree-like structure that uses a set of decision trees to make predictions.

Whereby each tree is trained on a random subset (with replacement) and a random subset of features". This concept is well illustrated in Figure 3

The final prediction for classification task is decided upon by a majority vote, while for the regression ones, it is the arithmetic mean of the predictions from all the trees.

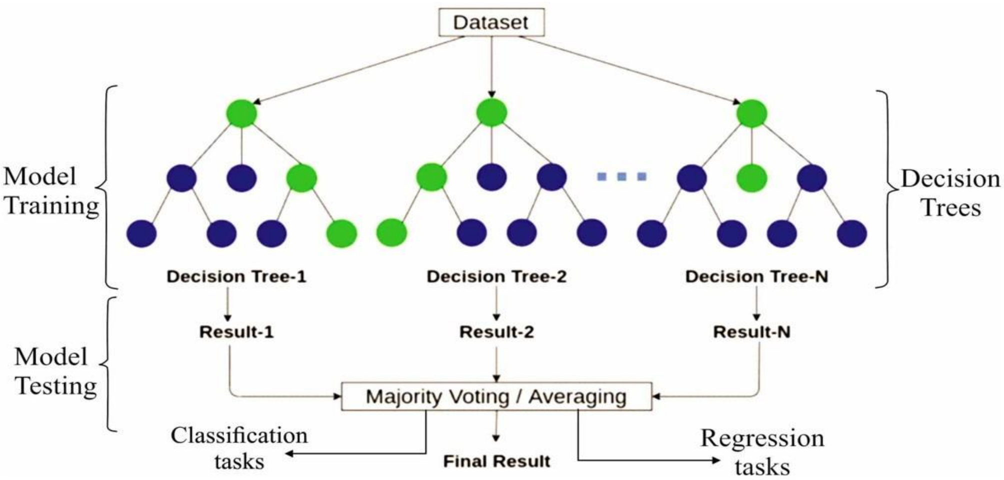

Figure 3: The Random Forest Structure.

4.3Extreme Gradient Boosting

The Extreme Gradient Boosting, also referred to as XGBoost, is a ML model used for supervised learning challenges. It uses training data to predict a target variable. Its main strength is that it is a scalable and optimized algorithm that does improve the speed and prediction performance of Gradient Boosting Machines (GMB). It does either classification or regression based on the explanatory variables which could either be quantitative and, or qualitative. The XGBoost relies on the intuition that the best possible succeeding model, when combined with preceding models, minimizes the overall prediction error. The errors for the classification tasks are minimized through the random forest process illustrated in Figure 4

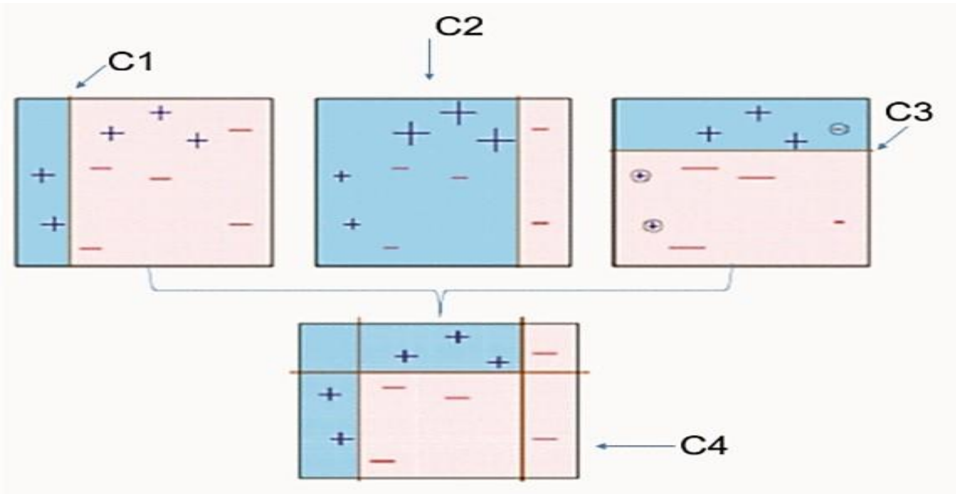

Figure 4: The Xgboost for classification.

Each succeeding model majors on misclassifications or errors of the preceding model and tries to reduce it in the succeeding iterations. XG boost gradually and iteratively improves the predictions by concentrating on where the previous model made errors. As demonstrated in figure 4, by drawing strait lines similar to building to correct the residual errors made by the combined predictions of all previous trees, the model is able to separate the negatives and positives at classifier 4

V.RESULTS

The study used learning curve to choose a sample size appropriate for model training. Typically, the ideal sample was the saturation point where the curve no longer rises but plateau, indicating that there was no further significant performance improvement by adding more data. learning models may stabilize at slightly different levels of sample size as shown in Figure 5 and the other models used in this study, plateau started at 3000 training sample size.

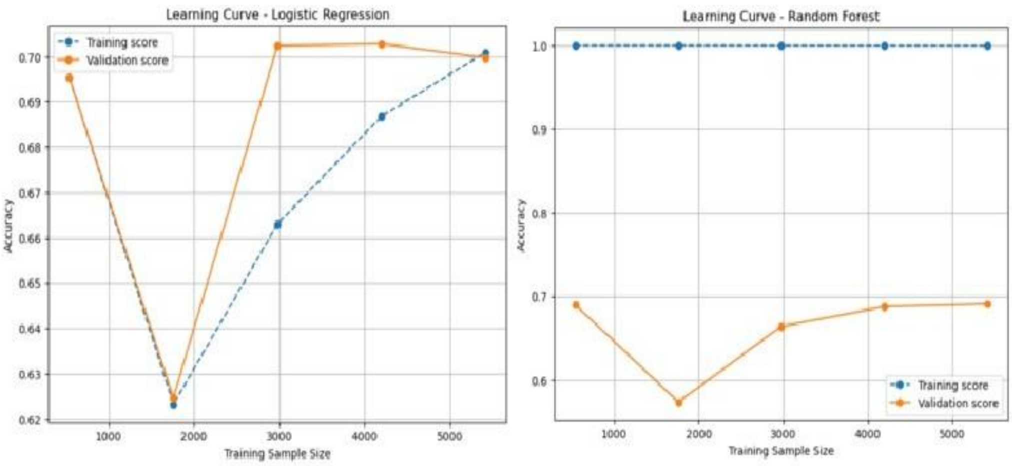

Figure 5: The Learning Curve

Therefore, based on the learning curves from the three models, the ideal sample size for train- ing data ranged between 3000 to 6000. The study used 6,770 individual and small business loan accounts from the 21 registered microfinance institutions operating within Machakos County. The training, validation and testing datasets were portioned in the ratios; 80% to 10% to 10% re- spectively. The stratified sampling was performed to ensure a structured representation of the perspectives in the sector, as well as to present the research questions being investigated by the study. Each loan was identified by the unique loan ID with the labeled information on default payments status. The demographic factors such as marital status, gender, age of the applicant, loan amount , credit history of the applicant, operating area and the date of disbursement between 2020 to 2024 were captured.

5.1. Data Exploratory

The study explored both the target and the features to get an overview of the target and features distribution and their general relationships. The findings are discussed in the following sections.

5.2. Target Distribution

The distribution of the response variable among the sampled loan accounts was performed. About a third of the loan applicants defaulted on payment, as shown in figure 6 Based on the findings displayed in figure 6, non-defaulters outnumbered the defaulters implying an imbalanced dataset. Further visualization was performed as displayed in the following sections.

5.3. Target and Features Distribution

This section related the target to the features of the study such as sex, marital status, credit history of the applicants and loan purpose as summarized in Figure 7

As visualized in 7, proportionately the higher default was by loan applicants with secondary school certificates compared to the primary school level applicants and post secondary school applicants. There was 18% default by business loan applicants compared to 15.1% defaults by loan applicants for personal reasons. Least default was by those without any credit history followed by those with more than three years' experience. Male applicants were more likely to default compared to their female counterpart applicants.

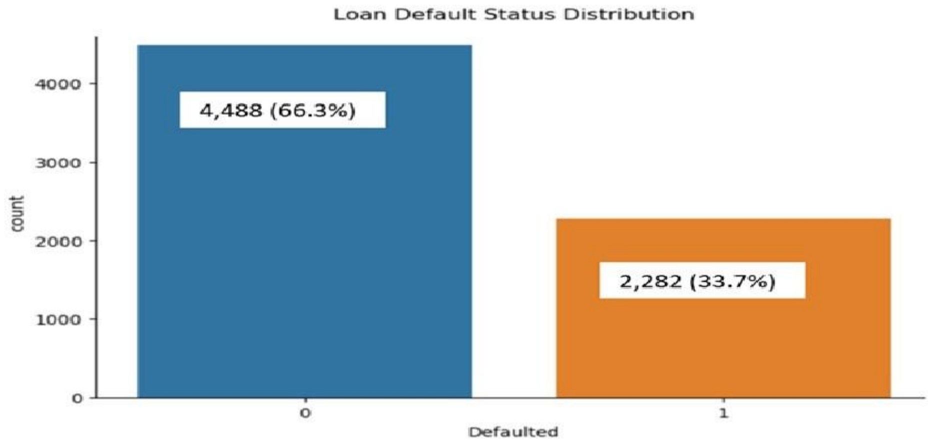

Figure 6: The Loan Default Status

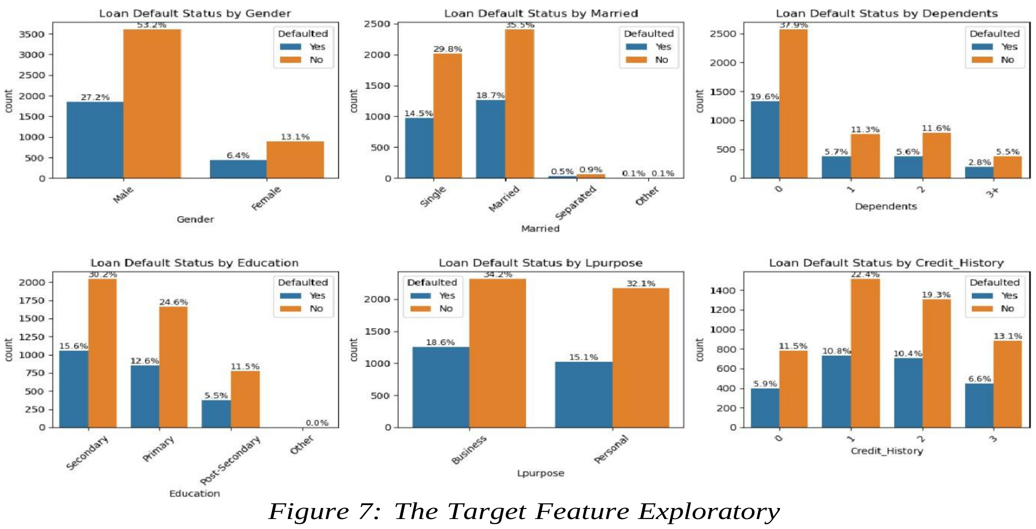

Figure 7: The Target Feature Exploratory

5.4 Logistic Regression Coefficients

Among the logistic regression analysis outputs is the coefficients of the features summarized in Table 1 The significant factors in $\alpha = 0.05$ were sex, the number of dependents, the amount of

Table 1: Logistic Regression Coefficients

Features	Estimates	Std. Errors	Z values	$Pr(>\ \|Z\|)$
(Intercept)	$-1.025e+00$	$1.187e+00$	-0.864	0.3878
Married2	$7.775e-01$	$1.164e+00$	0.668	0.5042
Married3	$8.119e-01$	$1.213e+00$	0.669	0.5032
Education2	$2.182e-02$	$8.421e-02$	0.259	0.7956
Education3	$-3.699e-02$	$1.141e-01$	-0.3245	0.7458
GenderMale	$1.884e-01$	$9.388e-01$	1.904	0.0185*
LoanAmount	$2.611e-06$	$2.886e-07$	9.048	$<$ 2e-16***
Dependant1	$-1.816e-02$	$1.037e-01$	-0.175	0.8610
Dependant2	$-2.065e-01$	$1.051e-01$	-1.965	0.0494*
Dependant3+	$3.316e-02$	$1.172e-01$	0.254	0.7992
CreditHistory2	$-1.430e-02$	$1.388e-01$	-0.103	0.9460
CreditHistory2	$-2.430e-02$	$2.458e-02$	-0.258	0.7961
CreditHistory3+	$-6.353e-03$	$1.228e-01$	-0.198	0.8430
ApplicantInc	$-1.523e-05$	$2.055e-06$	-7.411	$1.25e-13$ ***
PurposePersonal	$3.316e-03$	$7.590e-02$	0.044	0.9652
ApplicantAge	$-6.057e-03$	$4.771e-02$	-1.270	0.2043
AreaSemiurban	$1.278e-01$	$9.302e-02$	1.374	0.1694
AreaUrban	$3.263e-02$	$9.696e-02$	0.336	0.7365

loan applied for, and the income level of the applicant. Male applicants changed the log odds of defaulting by 0.1884, having two dependents changed the log odds of defaulting by -0.2065 compared to having no dependent. For every additional unit increase in loan amount applied for, the log odds of defaulting increased by $2.611e-06$ while for every additional unit increase in applicant's income, the log odds of defaulting reduced by $1.523e-05$ keeping all the other factors constant. The following factors were not significant at $\alpha = 0.05$ however, they affected the defaulting rate as discussed. By being married, i.e Married2 changed the log odds of defaulting by 0.7775 compared to being single (married1), while being separated (Married3) changed the log odds of defaulting by 0.8199 compared to being single (married1).

The odds ratio for the output is summarized in Table 2 General interpretation of the Odds Ratio (OR) is such that; when $OR = 1$, there is no effect, if $OR > 1$, then the predictor increases odds of outcome and if the $OR < 1$, then the predictor decreases odds of outcome. Based on the result in table 4.2 the following was observed. Marital status was not statistically significant factor in predicting the likelihood of defaulting at $\alpha = 0.05$. However, married applicants were associated with 4% lower default odds compared to the singles (referenced category) and the separated were associated with 1.5% higher odds of defaulting compared to the singles. Being male is associated with 19.6% higher odds of the defaulting compared to being female (reference) and because the confidence interval doesn't include 1 the effect is statistically significant at $\alpha = 0.05$. For each unit increase in loan amount applied for, the odds of defaulting increase by a very small fraction (0.0003%) holding all the other factors constant. However, since the loans were measured in thousands of Kenya shillings the changes were sizable amount and it was statistically significant at alpha=0.05. Having one, two or more than three dependents was associated with about 4%, 8% and 5% respectively of lower odds of defaulting compared to the referenced category of having no dependents. The number of dependents a loan applicant had was not statistically significant factor to consider at alpha=0.05 when determining the chances of a loan applicant defaulting. Having acquired a loan before was not statistically a significant factor in predicting the loan applicant's defaulting status. However, the second time applicants were associated with about 9% lower odds of defaulting compared to the first time applicants (referenced category), while the third time

Table 2: The Odds Ratio and 95% CI

Features	OR	2.5%	97.5%
(Intercept)	0.6504458	0.1203122	2.9905843
Married2	0.9612165	0.2225860	4.9449928
Married3	1.0052797	0.2050571	5.7269593
Education2	1.0335128	0.8759444	1.2198533
Education3	1.0153713	0.8092205	1.2716776
GenderMale	1.1961345	1.0054109	1.4560660
LoanAmount	1.0000031	1.0000025	1.0000037
Dependant1	0.9569769	0.7749066	1.1788576
Dependant2	0.9228192	0.7510539	1.1311421
Dependant3+	0.9507465	0.7194672	1.2490680
CreditHistory1	0.9198977	0.7363393	1.1507100
CreditHistory2	0.9835436	0.7870306	1.2307224
CreditHistory3	0.9343430	0.7316321	1.1935832
ApplicantInc	1.0009816	1.0009775	1.0009856
PurposePersonal	0.8525753	0.7329552	0.9914890
ApplicantAge	0.9939611	0.98467244	1.0032685
AreaSemiurban	0.9783631	0.8139986	1.1763193
AreaUrban	0.9497023	0.7854475	1.1483931

and more than three time applicants were associated with 2% and % lower odds of defaulting compared to the first time applicants. Age of the applicant was not statistically a significant factor in predicting the loan applicant's defaulting status. However for each unit increase in age, the odds of defaulting decreased by marginally by (0.6%) holding all the other factors constant. The residence of the loan applicant was classified as either rural, urban, or semi-urban and was not statistically a significant factor in predicting the loan applicant defaulting status. However, loan applicants who resided in urban areas were associated with 6% lower chances of default compared to applicants who resided in rural areas (referenced category), while loan applicants who reside in semi-urban areas were associated with about 2% lower odds of default compared to applicants who reside in rural areas.

A step analysis for variable selection was done and settled at gender, loan amount requested, loan applicant's income and then lastly, the loan applicant's age features that would be effective, precise and still achieve a good fit to the data as summarized in figure. 8 The parsimonious

```
Call:  glm(formula = Defaulted ~ Gender + LoanAmount + ApplicantInc + Applicant_Age,
    family = binomial, data = train_data)

Coefficients:
  (Intercept)    GenderMale      LoanAmount
   -1.799e-01     1.598e-01     2.612e-06
 ApplicantInc  Applicant_Age
   -1.518e-05    -7.313e-03

Degrees of Freedom: 3363 Total (i.e. Null);  3359 Residual
Null Deviance:      4292
Residual Deviance: 4121       AIC: 4131
```

Figure 8: The Parsimonious Model

model balanced simplicity with features' power, under minimal assumptions and complexities while maximizing forecast. Based on its output; the few significant factors that could explain the model as accurately as the full model were gender of the applicant, the amount of loan requested, the applicants income and the age of the applicant.

A decision tree classifier was carried out to establish the key factors which influenced the loan payment classification. The summary findings are shown in Figure 9 Based on the decision tree classifier the most significant factors in classifying the loan applicants as either defaulters or non- defaulters were the applicant's income level and the amount of loan applied. Loan applicants with income greater than 0.21 applying for more than 0.18 were likely to default. However, applicant with the same level of income but applying for less than 0.18 were not likely to default. Based on the loan amount applied the classifier categorized majority (89%) as non-defaulters while placing 2% (applicants with less than 0.21 income) and 9% of the applicants with more than 0.21 income as defaulters.

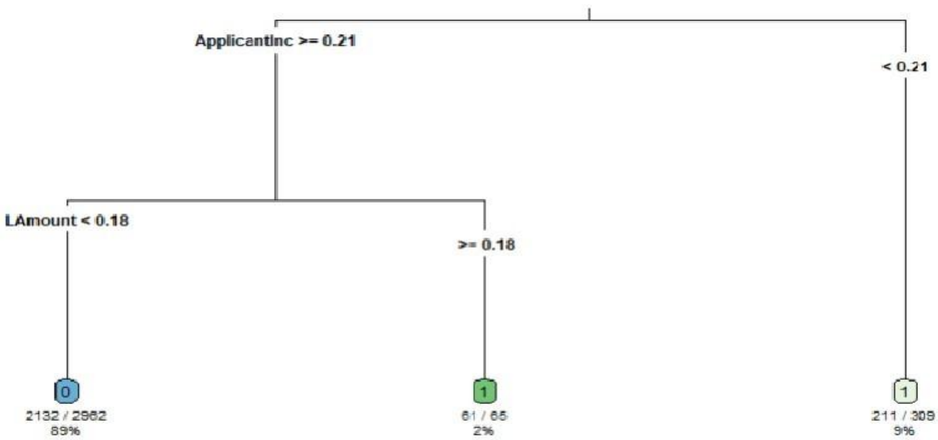

Figure 9: The Decision Tree Classifier

5.5Random Forest Model

Before applying the random forest algorithm, the necessary libraries were imported for data pre- processing and data analysis. Data were split into training, validation and testing sets (80% train, 10% validation and 10% test) prior to training random forest model. In order to better understand which features were most useful for making predictions, feature importance analysis was carried out, and its findings are summarized in Table 3

Table 3: Feature Importance Random Forest

Serial	Feature	Importance
1	Applicant Income	0.333623
2	Loan Amount	0.166662
3	Credit History	0.156935
4	Gender	0.107458

5	Month dispersed	0.101123
6	Area	0.041931
7	Dependents	0.036239
8	Education	0.027593
9	Marriage	0.015382
10	Loan Purpose	0.013054

The income of the loan applicant was the most important factor in predicting the loan default status, contributing 33% of model's decision power. The amount of loan requested by the applicant strongly influenced the decision, also with 17% contribution. The credit history of the applicant also a played a critical role as well in determining the outcome at 15.7%. The demographic factors of the applicant such as sex, number of dependents, highest formal education level, marital status, and loan purpose had much less influence on default status.

5.6XG Boost Algorithm

The XG Boost findings and the contribution of each feature to the improvement splits in the trees are summarized in Table 4 The higher values possess more predictive power. The top two features

Table 4: Feature Importance XG Boost

Serial	Feature	Importance
1	Loan Amount	0.213766
2	Applicant Income	0.165173
3	Gender	0.097111
4	Applicant Age	0.077745
5	Education	0.076715
6	Credit History	0.076523
7	Loan Purpose	0.074176
8	Dependents	0.074053
9	Marriage	0.072813
10	Area	0.071925

which are financial in nature, explains the large share of predictive power. The loan amount requested and the applicant's financial capacity strongly affect the model's decision. Although loan purpose (7.4%), dependents (7.4%), marriage (7.3%), and area (7.2%) add some predictive value, they were not major drivers. Loan approval or risk assessment in this model is primarily driven by financial capacity (loan size and income), while demographic characteristics (gender, age, marital status, area) play a secondary role.

VI.DISCUSSION

The study finding were inline with the study findings by [4] who observed that, the default rate was higher among the young applicants compared to the aged applicants. The borrower's location influence on default rate findings differed from the research findings by [7] whose research findings noted that, the default rate was higher in urban set up than rural areas. The differences could be attributed to different target population and research timings. Regarding the client's history influence the study findings were in agreement with [9] research that established, understanding the client's thorough credit history and assessment of client's literacy level could reduce the MFIs delinquency significantly. The study exposed male applicants as risk clients which supports the previous findings among them, the observation by [5] whose research noted that, the probabil- ity of having problems in loan repayment was higher for male borrowers than it was for female counterparts.

VII.RECOMMENDATIONS

Microfinance institutions must invest in capacity building for handling big data, modeling, and out- put interpretation for evidence-based decision making. County and national governments should develop a legal framework to govern ethical and open use of machine learning assessment while ob- serving the data protection Act. Explore more different machine learning techniques and compare their performance with the traditional credit score assessment performance.

VIII.AUTHOR CONTRIBUTION

Conceptualization: MMK and JMK, methodology: MMK, JMK and AK, software: MMK and JMK, Data Analysis MMK, MMK, AK and writing original draft preparation: MMK, JMK and AK supervision JMK and AK.

IX.CONFLICTS OF INTREST

All authors have read and agreed to the published version of the manuscript.

X.FUNDING

This research did not receive external funding.

References

1. Steve Gakuhi Karanja and Eddie Mungami Simiyu. Credit management practices and loan performance of microfinance banks in kenya. Journal of Finance and Accounting, 6(1):108– 139, 2022.
2. Alex Kimani, Robert Arasa, and James Karau. Resource allocation and performance of microfi-nance institutions in machakos county, kenya. International Journal of Finance and Accounting,3(1):28–35, 2024.
3. Jeniffer Muendi Muthini and Fredrick WS Ndede. Socio-economic factors and table banking loans default levels among women groups in machakos county, kenya. Int J Curr Aspects Fin Banking Acc, 4(1):91–103, 2022.
4. Andualem Kassegn and Ebrahim Endris. Factors affecting loan repayment rate among small- holder farmers got loans from the amhara credit and saving institution: in the case of habru district, amhara regional state, ethiopia. International Area Studies Review, 25(1):73–96, 2022.
5. Pilar Lopez-Sanchez, Elena Urquia-Grande, Cristina Del Campo, and Andr´es L Cancer. Delving into the determinants of default risk in savings groups: empirical evidence from ecuador. The European Journal of Development Research, 34(6):2625, 2021.
6. Akanksha Goel and Shailesh Rastogi. Understanding the impact of borrowers' behavioural and psychological traits on credit default: review and conceptual model. Review of Behavioral Finance, 15(2):205–223, 2023.
7. Md Aslam Mia, Gary John Rangel, Mohammad Nourani, and Rajesh Kumar. Institutional factors and efficiency performance in the global microfinance industry. Benchmarking: An International Journal, 30(2):433–459, 2023.
8. Hasan Ahmed Salman, Ali Kalakech, and Amani Steiti. Random forest algorithm overview. Babylonian Journal of Machine Learning, 2024:69–79, 2024.
9. Basu Dev Lamichhane. Loan delinquency in microfinance institutions (mfis): Ways to overcome the problem. Nepalese Journal of Management Research, 2(1):37–43, 2022.