Identifying Key Predictors of Loan Default in Kenyan Microfinance Institutions: A Backward Logistic Regression Analysis in Makueni County, Kenya

Jackson Musau[1*], Dr. Martin Kasina[2] & Dr. Ayubu Anapapa[3]

[1,2]Department of Mathematics and Statistics, Machakos University, Machakos, Kenya

[3]Department of Mathematics and Actuarial Science, Murang'a University of Technology, Murang'a, Kenya

***Correspondence Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100600375

Received: 08 June 2026; Accepted: 13 June 2026; Published: 24 June 2026

ABSTRACT

Microfinance institutions (MFIs) play a critical role in improving financial inclusion in Kenya; however, high loan default rates continue to threaten their financial sustainability. This study aimed to determine the predictors of loan default using logistic regression among microfinance institution clients in Makueni County, Kenya. The study adopted a quantitative research design and used secondary data comprising 4,592 borrower records obtained from selected MFIs operating within the county. Data analysis was done using python programing. Borrower socio-economic and financial attributes were extracted from loan records and preprocessed through data cleaning, normalization and encoding procedures. Prior to model fitting, logistic regression assumptions, including multicollinearity, independence of observations and model goodness-of-fit, were assessed and found to be satisfactory. The dataset was split into training and testing sets using an 80:20 ratio. A binary logistic model was fitted with all predictor variables and was found to be significant (χ^2 (19)= 2695.70, $p < 0.001$) at α=0.05. Backward logistic regression was then run to identify the key predictors of loan default. The findings revealed that key predictors of loan default included loan amount, interest amount, outstanding loan balance, loan term, number of serviced loans, employment status, loan purpose and borrower type (new or repeat clients). In conclusion, these results indicate that loan default is primarily driven by financial exposure (loan amount, outstanding balance, repayment term) and borrower characteristics (employment status and repayment history). These findings demonstrate the effectiveness of data-driven credit risk modelling in improving loan approval decisions and enhancing financial sustainability of MFIs in rural Kenyan contexts.

Key Word: Loan default, Microfinance institutions, Logistic regression, Credit risk.

INTRODUCTION

Microfinance institutions (MFIs) are the key instruments for increasing financial inclusion in developing economies. This is because they provide seamless access to credit services to the poor households and micro-enterprises. [9] argues that entrepreneurship and equitable financial growth become possible when there is access to credit facilities. However, sustainability of MFIs is based on their capacity to manage credit risk, which is a quantitative issue of estimating the likelihood of a loan defaulter, a problem that lies at the core of applied statistical modelling.

In Kenya, microfinance industry has increased financial accessibility but there is growing pressure on credit-risk. In 2024, the Central Bank of Kenya [1] reported a non-performing loan (NPL) ratio of 16.4%. In 2025, the State of the Banking Industry Report [2] noted a slow credit growth and poor transmission of the monetary policy. According to an article by [5], the overall MFI assets decreased by almost ten percent in 2024 with the gross NPLs increasing 5 times between 2014 and 2024 taking the ratio to over 35 percent. These statistics measure systemic vulnerability requiring statistical research. This quantitative view is supported by empirical research by scholars in Kenya. Through multiple regression, [6] demonstrated that the capital adequacy is significantly inversely related to Non-Performing Loans (NPLs) among microfinance banks. Using logistic regression to analyze digital-loan data of commercial banks, [7] verified the model strength in estimating probabilities of default, the income-to-loan ratio and credit score were found to be determinants of loan default. [12] evaluated the performance of logistic regression, XGBoost and random-forest algorithms to predict credit risk using 6,771 loan accounts of Machakos County MFIs. The study found out that XG Boosting was the best performer with 83.3% accuracy. Findings from these studies indicate the necessity of strict comparative analysis of statistical models of classification in Kenyan credit-risk environments such as in MFIs.

In Makueni County, the main credit providers among rural households and smallholders are MFIs, but repayment problems are very acute. [10] found that in Nzaui/Kilili/Kalamba Ward the default rate was 61 percent attributed to the literacy of the borrowers, off-farm earnings, frequency and level of supervision by the loan officers. Such socio-economic variables are quantifiable and easy to model statistically and are often seldom included in forecasting models. Reliance on rain-based agriculture and a lack of access to banking services increases the exposure to repayment risk thereby undermining objectivity through subjective lending practices.

Although the study by [10] offers valuable descriptive information on the loan repayment issues in Nzaui/Kilili/Kalamba Ward, the study methodology and the level of analysis of the current study are of a different nature. The main methods applied in the work of Mwaka were the traditional methods of inferential identification of socio-economic factors that affect the repayment behavior with the focus on the borrower literacy, off-farm earnings and loan supervision. The analysis was not aimed at creating predictive models and its results were much context-based and descriptive.

The fact that the loan default rates are still on the increase in Makueni County is an indication that there may be some structural and methodological flaws in the credit risk assessment practice. This has been mostly due to the over dependence on the use of single models, mostly logistic regression, without a systematic comparison with other classification algorithms that may be able to capture nonlinear relationship and complex behavior of borrowers. The lack of localized and empirically validated comparative modeling frameworks has thus limited the timeliness, precision and objectivity in credit decisions made by MFIs leading to ongoing increases in non-performing loans and endangering the sustainability of the institutions.

The specific aim of this study was to determine the key socio-economic and financial predictors of loan default using logistic regression among microfinance institution clients in Makueni County, Kenya. This study may improve credit risk assessment, reduce non-performing losses and become more sustainable by making the credit decisions based on the objective and data-driven lending decisions.

MATERIAL AND METHODS

This study adopted a quantitative research design with a comparative approach to evaluate and compare the predictive accuracy and effectiveness of the selected models. The design was appropriate because the study aimed at developing and comparing statistical classification models in predicting the probability of loan default. There were 9 MFIs with 22 branches operating within Makueni County [2]. This study adopted a purposive sampling technique to select nine microfinance institutions (MFIs) for inclusion in the analysis. Each MFI was represented by one branch to form a sample of nine institutions. The institutions were chosen based on the availability of complete and relevant loan-level records necessary for developing and evaluating predictive models for loan default. Data was analyzed using python programming.

Preprocessing

Before data was subjected to analysis, it was first prepared to accommodate the predictive algorithms. Data was cleaned to address any missing entries by imputation. Before fitting the models, the dataset was prepared by handling outliers through capping, eliminating duplicates and scaling continuous variables to a 0 to 1 range through normalization. Categorical variables were encoded as shown in Table 1.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

Table 1: Coding of Categorical Variables

Variable Name	Description	Coding
Gender	Sex of borrower	1 = Male, 2 = Female
Marital_Status	Marital status	1 = Single, 2 = Married, 3 = Others
Education_Level	Highest education attained	1 = Primary, 2 = Secondary, 3 = Tertiary, 4 = University
Loan_Purpose	Purpose of loan	1 = Business, 2 = Education, 3 = Personal, 4 = Other
Collateral	Loan security status	1 = Yes, 0 = No
Employment_Status	Employment status	1 = Employed, 2 = Self-employed, 3 = Unemployed
New_Repeat	Loan history status	1 = New, 2 = Repeat
Loan_Status (Target)	Indicates loan default status	1 = Defaulted, 0 = Not Defaulted

To address class imbalance, class weighting was applied using the class_weight=’balanced’ parameter in Python’s scikit-learn library. This automatically adjusted penalties such that misclassification of the minority class carries higher weight, effectively balancing model sensitivity without altering the original dataset.

THEORETICAL RESULTS

Logistic Regression

Logistic Regression is a classification model that is used for classification tasks. In this study, binary logistic regression was used to predict the probability of a client defaulting on a loan, which is a binary outcome, i.e., default (1) or no default (0). The choice of binary logistic regression was motivated by several considerations specific to this study. Binary logistic regression directly models the probability of a binary outcome (default vs. non-default) using the sigmoid function which produces default likelihoods. As opposed to advanced machine learning algorithms, logistic regression provides transparent coefficient estimates and odds ratios, which are easily interpretable, enabling loan officers in Makueni County MFIs to understand and explain which borrower characteristics drive default risk. In addition, logistic regression handles class imbalance through class weighting, which was applied in this study to ensure the model does not bias predictions toward the majority (non-default) class. These characteristics establish logistic regression as both an appropriate baseline model and a practically useful tool for MFI credit risk assessment in Makueni County.

Logistic regression model uses sigmoid function to transform linear combination of input features into a probability. Suppose we take the input borrower’s features, such as income, age, employment status and other relevant predictor variables to be represented by a vector $X = x_1, x_2, \dots, x_n$ and the dependent variable Y with binary values 0 and 1 such that:

$$y = \begin{cases} 0 & \text{if No Default} \\ 1 & \text{if Default} \end{cases} \qquad (1)$$

We first compute a linear combination z of input features such that:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad (2)$$

where β_0 is the intercept and $\beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients corresponding to the predictor variables $x_1, x_2, \dots, x_n$. We now take the log odds of a defaulter such that:

$$\log\left(\frac{p}{1-p}\right) = z \quad (3)$$

where p is the likelihood of a loan applicant defaulting. The value z is a continuous value from a linear model. To transform z into a probability function, we use the sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (4)$$

where x is the input value to the sigmoid function such that;

$$\lim_{x\to\infty} \sigma(x) = 1, \qquad \lim_{x\to-\infty} \sigma(x) = 0. \quad (5)$$

resulting in the characteristic S-shaped curve shown in Figure 1.

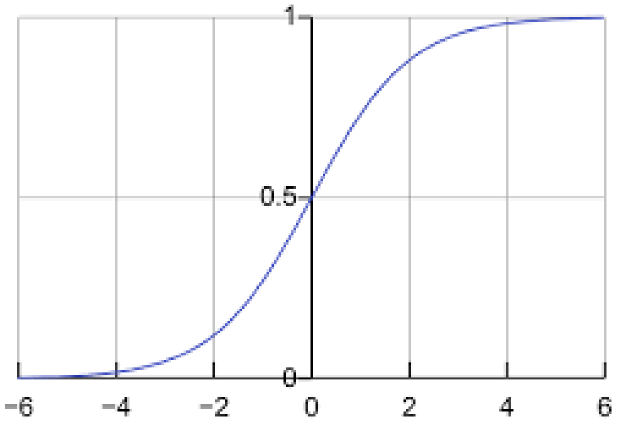

Figure 1. Sigmoid function curve.

In our case, x is the value of z. The sigmoid function then becomes as follows:

$$\sigma(z) = p(y = 1|X) = \frac{1}{1+e^{-z}} \quad (6)$$

Taking the value of z from equation (2), we obtain a probability function:

$$p(y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}} \quad (7)$$

where $p(y = 1|X)$ represents the probability that a client defaults. The output is a probability, hence lies between 0 and 1. To classify a new loan client we compute the probability of default $\hat{p} = P(y = 1 \mid X)$ and classify the client such that:

$$\hat{y} = \begin{cases} 1, & \text{if } \hat{p} \geq 0.5 \\ 0, & \text{if } \hat{p} < 0.5 \end{cases} \quad (8)$$

Parameter Estimation

To obtain the best model for predicting loan default, we estimate the logistic regression parameters to find the optimal coefficients β and intercept β_0 that maximize the probability of observing our data. In this study, we used the method of Maximum Likelihood Estimation. For each observation i, the binary outcome is:

$$p(y_i = 1 \mid x_i) = p(x_i), \qquad p(y_i = 0 \mid x_i) = 1 - p(x_i) \quad (9)$$

The likelihood of n independent observations then becomes:

$$L(\beta,\beta_0) = \prod_{i=1}^{n}[p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i} \quad (10)$$

Taking the natural logarithm of equation (10) we obtain:

$$\ln L(\beta,\beta_0) = \sum_{i=1}^{n}[y_i \ln p(x_i) + (1-y_i)\ln(1-p(x_i))] \quad (11)$$

Substituting equation (11) to the sigmoid function:

$$p(x_i) = \frac{1}{1+e^{-(\beta\cdot x_i+\beta_0)}} \quad (12)$$

we obtain a simplified log-likelihood function such that:

$$\ln L(\beta,\beta_0) = \sum_{i=1}^{n}\left[y_i(\beta\cdot x_i+\beta_0) - \ln\left(1+e^{\,\beta\cdot x_i+\beta_0}\right)\right] \quad (13)$$

This represents the maximized logistic regression function. To obtain the optimal parameters β and β_0, we apply the gradient ascent method to the log-likelihood function. The gradient with respect to coefficient β_j is:

$$\frac{\partial \ln L(\beta,\beta_0)}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - p(x_i;\beta,\beta_0))\,x_{ij} \quad (14)$$

Similarly, the gradient with respect to the intercept β_0 is:

$$\frac{\partial \ln L(\beta,\beta_0)}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - p(x_i;\beta,\beta_0)) \quad (15)$$

In practice, the model learns iteratively updating the parameters using the gradient until no significant change to the parameter values occurs(convergence), such that:

$$\beta_j^{(t+1)} = \beta_j^{(t)} + \alpha\frac{\partial \ln L}{\partial \beta_j}, \qquad \beta_0^{(t+1)} = \beta_0^{(t)} + \alpha\frac{\partial \ln L}{\partial \beta_0} \quad (16)$$

where α is the learning rate and t is the current iteration number. At convergence, we say the model is optimal and hence can be used to give the best prediction of the probability of loan default for a new client.

Result

The sample dataset was made up of 4,592 records drawn from the 9 selected MFIs. The adequacy of the dataset size (n = 4,592) was evaluated using learning curve generated from Logistic Regression Model as shown in Figure 2

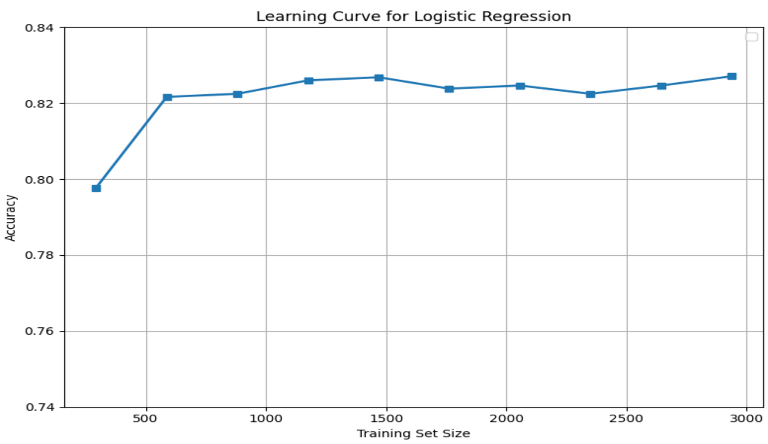

Figure 2. Learning Curves

The results of the learning curve analysis across sample sizes of 500 to 3,500 shows that Logistic Regression

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

plateaued at 1,500-sample size. These findings establish 1,500-sample size as the optimal training size for maximum accuracy. This indicates that the current dataset size (n=4,592) was sufficient for stable model performance.

Exploratory Data Analysis

Target Distribution

An analysis of the distribution of the target variable was conducted to evaluate the proportion of defaulted and non-defaulted loans. The results are presented in Figure 3.

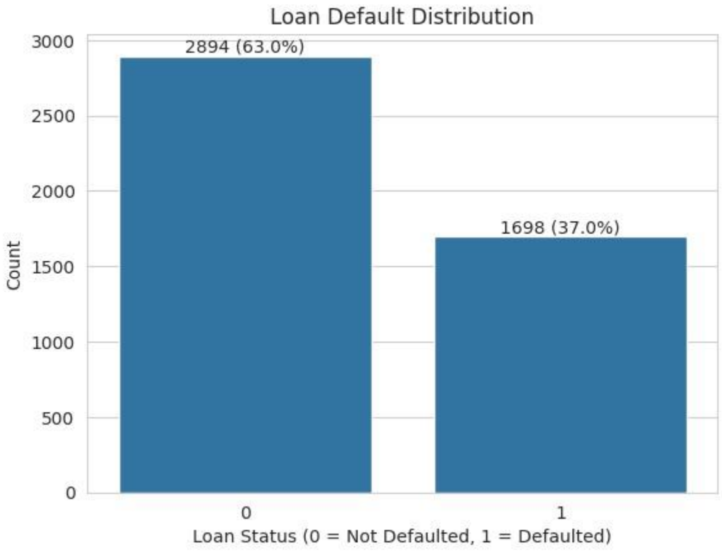

Figure 3. Distribution of the Target Variable

The analysis shows that 2894 (63.0%) of borrowers did not default, while 1698 (37.0%) defaulted. This indicates that although the majority of borrowers repay their loans successfully, a significant proportion still default. This finding are similar to the study by [11], which reported that loan datasets are often characterized by class imbalance with a higher proportion of non-defaulters. This imbalanced dataset hence call for application of classification methods that are robust to class imbalance.

Descriptive Analysis of Continuous Variables

A descriptive analysis for continuous variables was done to summarize their key feature characteristics, thereby providing an overview of the dataset structure. The results are presented in Table 2.

Table 2. Descriptive Statistics of the Continuous Variables

Variable	*N*	Mean	SD	Min	Max
Age	4592	36.97	8.78	18.00	73.00
Loan Amount	4592	22150.70	12089.75	5000	40000
Interest Amount	4592	6255.16	5040.37	1000	26000
Loan Term	4592	15.41	9.89	4	35
Interest Rate	4592	31.62	21.09	8	60
OLB	4592	11134.26	9492.17	0	36000

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

Monthly Income	4592	27776.79	12293.02	3000	50000
Serviced Loans	4592	3.29	2.04	0	10

The dataset contains 4,592 observations across all continuous variables. Loan amounts average Ksh 22,150, with a wide dispersion (SD = 12,089), indicating variation in borrowing capacity among clients. Interest rates average was at 31.62%, with a maximum of 60%, suggesting risk-based pricing across borrowers. Monthly income averages Ksh 27,776, indicating that most borrowers fall within the low-to-middle income category typical of microfinance clients. The outstanding loan balance (OLB) shows a mean of Ksh 11,134, but ranges up to Ksh 36,000, indicating that some borrowers still carried significant repayment burdens.

Descriptive analysis of Categorical Variables

A visualization of how the distribution of loan default varied across different categorical variables was done and results presented in Figure 4.

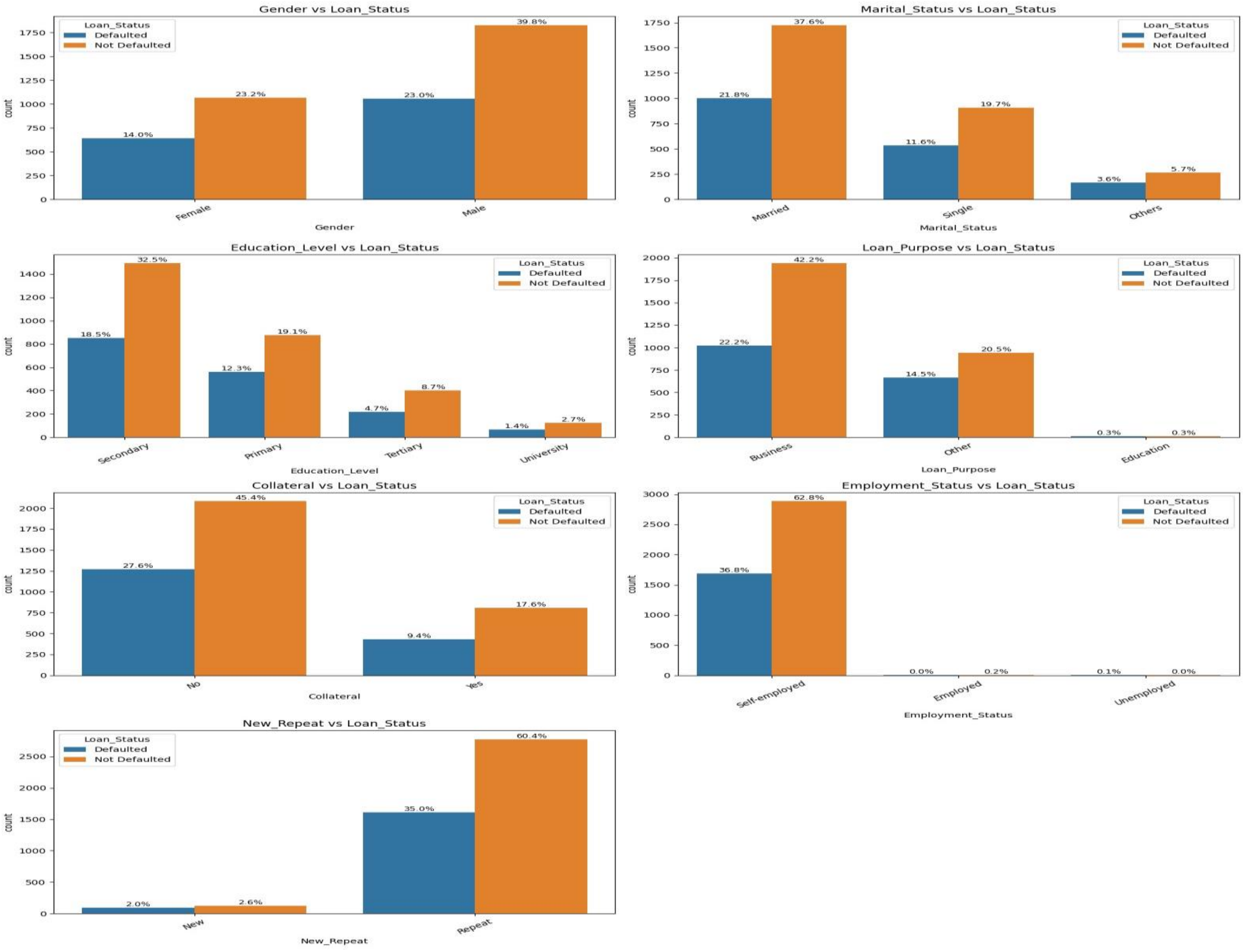

Figure 4. Categorical Variables and Loan Default Status

The series of bar charts shows relationship between borrower characteristics and loan default status. The findings are summarized below.

Gender and Loan Status

Male borrowers exhibited higher frequencies in both defaulted (23.0%) and non-defaulted loans (39.8%) compared to female borrowers (14.0% defaulted; 23.2% not defaulted). This suggests that males constitute a larger proportion of borrowers overall, though default patterns appear proportional across gender.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

Marital Status and Loan Status

Married individuals accounted for the highest proportion of both defaulted (21.8%) and non-defaulted loans (37.6%), followed by single borrowers (11.6% defaulted; 19.7% not defaulted). Borrowers classified as "others" represented the smallest share (3.6% defaulted; 5.7% not defaulted). This indicates that marital status may be associated with loan participation but not necessarily with disproportionate default risk.

Education Level and Loan Status

Borrowers with secondary education formed the largest group (18.5% defaulted; 32.5% not defaulted), followed by primary education (12.3% defaulted; 19.1% not defaulted). Tertiary and university-educated borrowers had considerably lower representation in both categories. This pattern suggests that individuals with lower education levels dominate the borrower population.

Loan Purpose and Loan Status

Loans taken for business purposes had the highest proportions (22.2% defaulted; 42.2% not defaulted), followed by other purposes (14.5% defaulted; 20.5% not defaulted). Loans for education purposes were negligible (0.3% for both categories). This indicates that business loans are the most common and may carry relatively higher exposure to default simply due to volume.

Collateral and Loan Status

Borrowers without collateral showed higher default rates (27.6%) compared to those with collateral (9.4%). Similarly, non-default rates were higher among those without collateral (45.4%) than those with collateral (17.6%). These findings suggest that lack of collateral is associated with increased loan uptake and higher default counts.

Employment Status and Loan Status

Self-employed individuals constituted the majority of borrowers (36.8% defaulted; 62.8% not defaulted), while employed and unemployed categories had negligible representation. This indicates that the dataset is heavily skewed toward self-employed individuals.

Loan History and Loan Status

Repeat borrowers accounted for the majority of both defaulted (35.0%) and non-defaulted loans (60.4%), whereas new borrowers represented a very small proportion (2.0% defaulted; 2.6% not defaulted). This suggests that repeat borrowing is common and may be associated with sustained loan access regardless of default status.

These findings are consistent with [5], who observed that borrower characteristics such as business engagement, collateral status and loan usage patterns play an important role in determining the repayment outcomes in microfinance settings. The dominance of business-related loans and self-employed borrowers reveals the typical characteristics of microfinance client base, where access to credit is closely tied to informal economic activities.

Correlation Analysis

A correlation analysis was computed to assess relationships and multicollinearity among the independent variables before model fitting was done. The results were visualized using a correlation heatmap as shown on Figure 5.

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

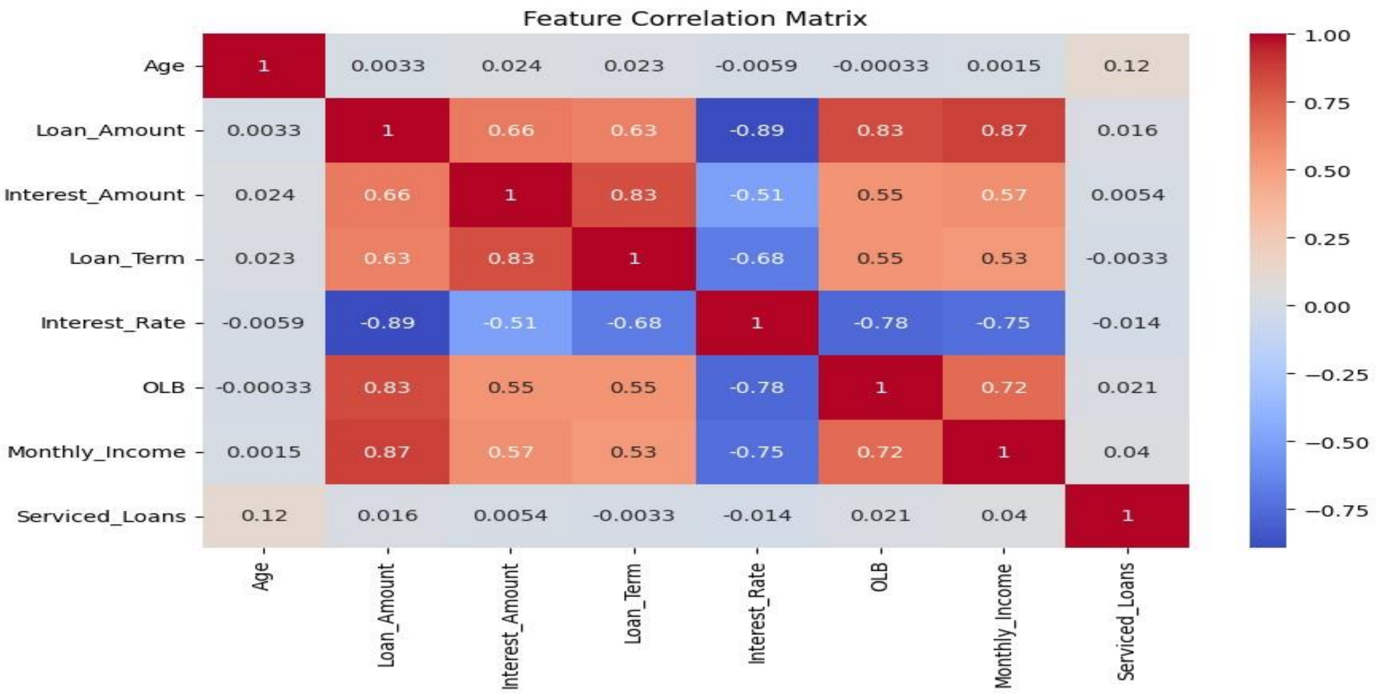

Figure 5. Correlation Heatmap

The correlation matrix revealed several strong relationships among the financial variables. Loan amount was strongly positively correlated with outstanding loan balance (OLB) (r = .83) and monthly income (r = .87), indicating that borrowers with higher incomes tend to receive larger loans and maintain higher outstanding balances. Loan amount was also moderately correlated with interest amount (r = .66) and loan term (r = .63).

Interest amount showed a strong positive correlation with loan term (r = .83), suggesting that longer loan durations are associated with higher interest charges. Additionally, OLB was strongly correlated with monthly income (r = .72), indicating a relationship between income levels and remaining loan obligations.

Notably, interest rate exhibited strong negative correlations with loan amount ($r = -.89$), OLB (r = −.78) and monthly income (r = −.75), suggesting that borrowers with higher loan amounts and incomes tend to receive lower interest rates. Age and serviced loans showed weak or negligible correlations with most variables, indicating minimal linear association with the other predictors.

It is worth noting that, the presence of several strong correlations (|r|) among key predictor variables suggests potential multicollinearity. Since Logistic Regression, is highly sensitive to multicollinearity, it was therefore necessary to further assess multicollinearity using the Variance Inflation Factor (VIF) prior to model fitting.

From the heatmap results in Figure 5, Interest Rate shows high correlation across all other variable. The researcher then drew attention to this particular variable. A Variance Inflation Factor (VIF) analysis was hence conducted before and after dropping Interest Rate variable. A VIF value greater than 10 indicates severe multicollinearity, while values between 5 and 10 suggest moderate multicollinearity. The results are presented in Table 3.

Table 3. Comparison of VIF Values Before and After Removing Interest Rate

Variable	VIF (Before)	VIF (After)
Loan Amount	17.97	7.42
Interest Rate	12.48	–

Loan Term	7.74	3.40
Interest Amount	7.61	3.57
Monthly Income	4.30	4.23
OLB	3.33	3.30
Serviced Loans	1.02	1.02
Age	1.02	1.02

This results indicate that prior to removal of Interest Rate variable, Loan Amount exhibited severe multicollinearity (VIF = 17.97), while Interest Rate also showed high collinearity (VIF = 12.48). Loan Term (VIF = 7.74) and Interest Amount (VIF = 7.61) indicated moderate to high multicollinearity. In contrast, Monthly Income (VIF = 4.30), OLB (VIF = 3.33), Serviced Loans (VIF = 1.02) and Age (VIF = 1.02) showed acceptable levels of multicollinearity. After removing Interest Rate, the VIF values significantly decreased. Loan Amount reduced to 7.42, Loan Term to 3.40 and Interest Amount to 3.57. Monthly Income (VIF = 4.23), OLB (VIF = 3.30), Serviced Loans (VIF = 1.02) and Age (VIF = 1.02) remained largely unchanged. Overall, the results confirm that multicollinearity was sufficiently reduced and the dataset was suitable for subsequent model fitting.

Model Fit

A binary Logistic Regression model was fitted to predict loan default among microfinance institution clients in Makueni County, Kenya. The model incorporated socio-economic and financial features including age, loan amount, loan term, outstanding loan balance, income and borrower characteristics.

Table 4 summarizes the fit statistics of the logistic regression model. These measures, including the chi square value, model significance, R^2 and model accuracy indicate the overall adequacy and explanatory strength of the model in predicting loan default.

Table 4. Logistic Regression Model Summary

Statistic	**Value**
Number of Observations (Weighted)	3,673
Model Chi-square (χ^2, df = 19)	2695.70
Model Significance	$p < .001$
Cox & Snell R^2	0.479
Classification Accuracy	0.84

The model shows a significantly good fit ($\chi^2(19)$ = 2695.70, p < 0.001) at α=0.05. The R^2 value of 0.479 indicates that the predictor features explain approximately 47.9% of the variation in loan default. The model achieved good predictive accuracy, correctly classifying 84% of cases in the test data set.

Model Regression Coefficients

Table 5 presents the estimated coefficients of the logistic regression model used to identify the determinants of loan default among microfinance institution clients. The results include parameter estimates, standard errors, z-values and p-values, which are used to assess the statistical significance and direction of influence of each predictor variable.

Table 5. Weighted Binary Logistic Regression Results

Variable	**Coef.**	**Std. Err**	**z**	**p-value**	**2.5%**	**97.5%**
Constant	-1.4459	0.086	-16.753	< .001	-1.615	-1.277
Age	0.0863	0.049	1.768	.077	-0.009	0.182
Loan Amount	-6.3741	0.260	-24.531	< .001*	-6.883	-5.865
Interest Amount	0.7533	0.112	6.737	< .001*	0.534	0.972
Loan Term	-0.9904	0.105	-9.434	< .001*	-1.196	-0.785
OLB	5.1987	0.204	25.515	< .001*	4.799	5.598
Monthly Income	-0.1435	0.102	-1.413	.158	-0.343	0.056
Serviced Loans	0.4099	0.059	7.005	< .001*	0.295	0.525
Gender (2)	0.0140	0.048	0.293	.770	-0.080	0.108
Marital Status (2)	-0.0060	0.053	-0.114	.910	-0.110	0.098
Marital Status (3)	-0.0009	0.051	-0.018	.986	-0.101	0.100
Education Level (2)	-0.0706	0.055	-1.287	.198	-0.178	0.037
Education Level (3)	-0.0179	0.053	-0.336	.737	-0.122	0.087
Education Level (4)	-0.0060	0.053	-0.112	.911	-0.111	0.099
Loan Purpose (2)	-0.0159	0.053	-0.299	.765	-0.120	0.088
Loan Purpose (4)	0.3209	0.054	5.927	< .001*	0.215	0.427
Collateral (1)	-0.0290	0.049	-0.590	.555	-0.125	0.067
Employment Status (2)	0.1178	0.057	2.064	.039*	0.006	0.230
Employment Status (3)	0.3240	0.064	5.027	< .001*	0.198	0.450
New Repeat (2)	-0.2036	0.055	-3.692	< .001*	-0.312	-0.096

Note. * indicates statistically significant variables at $p < .05$.

As shown in Table 5, a unit increase in age increased the log odds of defaulting by 0.0863, meaning that older borrowers were slightly more likely to default compared to younger borrowers, holding other variables constant. For every additional unit increase in loan amount, the log odds of defaulting decreased by 6.3741, indicating that higher loan amounts were associated with a lower likelihood of default, holding other factors constant. For every additional unit increase in Interest Amount, the log odds of defaulting increased by 0.7533 suggesting that an increase in the interest amount increased the chances of defaulting. For every additional unit increase in loan term, the log odds of defaulting decreased by 0.9904, suggesting that longer repayment periods reduce the probability of default. A unit increase in outstanding loan balance (OLB) increased the log odds of defaulting by 5.1987, indicating that borrowers with higher outstanding balances were significantly more likely to default, holding all other variables constant. For every additional unit increase in monthly income, the log odds of

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

defaulting decreased by 0.1435, although this effect was not statistically significant at $\alpha = 0.05$. For every additional unit increase in loans serviced, the log odds of defaulting increased by 0.4099, indicating that borrowers with more prior loan engagements were more likely to default. A male applicant changed the log odds of defaulting by 0.0140, indicating a very small increase in the likelihood of default compared to female applicants, holding all other variables constant. Applicants marital status and education level did not have a significant effect ($p > 0.05$) to loan defaulting at $\alpha = 0.05$. Borrowers whose loan purpose category was 4(other loan uses) had their log odds of defaulting increased by 0.3209 compared to the reference category (Business). Collateral category 1(loan with collateral) did not have a significant effect to the loan defaulting. Being in employment status category 3(unemployed) increased the log odds of defaulting by 0.3240, while employment status category 2(self-employed) increased the log odds of defaulting by 0.1178.. Being a repeat borrower (New/Repeat = 2) reduced the log odds of defaulting by 0.2036, indicating that repeat clients were less likely to default compared to new borrowers.

Results show that Loan Amount, Interest Amount, Loan Term, OLB, Serviced Loans, Loan Purpose, Employment Status and New/Repeat (loan history) predictors were statistically significant ($p < 0.05$) at $\alpha = 0.05$. This means that they significantly affected the applicants' ability to repay their loans. Using these statistically significant factors, the researcher constructed the parsimonious model as follows.

From Equation [6] , the logistic regression model is expressed in terms of probability as :

$$p(y = 1|X) = \frac{1}{1+e^{-z}} \tag{17}$$

where z is the linear predictor given by:

$$\begin{aligned} z = {} & 2.3573 - 0.0005(LoanAmount) + 0.0001(InterestAmount) - 0.0992(LoanTerm) \\ & +0.0005(OLB) + 0.1975(ServicedLoans) + 0.6951(LoanPurpose\_4) \\ & +1.7667(EmploymentStatus\_2) + 7.6876(EmploymentStatus\_3) \\ & -0.9194(New/Repeat\_2) \end{aligned} \tag{18}$$

Thus, the model estimates the probability that the outcome variable equals one given the explanatory variables included in the model($p(y = 1|X)$). If $p(y = 1|X) > 0.5$, the client is classified as a defaulter otherwise a non-defaulter. This is the best binary logistic model that can be used to predict the probability of default.

Key Predictors of Loan Default

Backward elimination logistic regression method was conducted to identify the key features that can be used to predict loan default. This method starts with all the features included in the model and subsequently eliminates weak features until the model performance improves significantly.

The initial model included 18 predictors, which were progressively reduced based on statistical significance (p < .05), resulting in a final parsimonious model with improved performance as presented in Table 6

Table 6. Comparison of Full and Reduced Logistic Regression Models

Model	No. of Predictors	Accuracy	AUC-ROC
Full Model	18	0.835	0.923
Reduced Model	9	0.850	0.925

The full model, which included 18 predictors, achieved an accuracy of 83.5% and an AUC-ROC of 0.923. In comparison, the reduced model obtained through backward elimination included only 9 predictors with a significant improvement in performance, achieving an improved accuracy of 85.0% and an AUC-ROC of 0.925. These results indicate that reducing the number of predictors did not compromise model performance but demonstrated improvements in both classification accuracy and discriminatory power.

All the weak predictors were removed and only the key predictors were included in the final model as shown in Table 7.

Table 7. Logistic Regression Coefficients for Final Reduced Model

Predictor	Coef.	Std. Err.	z	p-value
Constant	2.3573	0.9440	2.496	0.013
Loan Amount	-0.0005	0.0000	-28.366	< 0.001
Interest Amount	0.0001	0.0000	6.938	< 0.001
Loan Term	-0.0938	0.0100	-9.617	< 0.001
OLB	0.0005	0.0000	27.842	< 0.001
Serviced Loans	0.2248	0.0250	9.029	< 0.001
Loan Purpose (Cat 4)	0.8423	0.1010	8.374	< 0.001
Employment Status (Cat 2)	1.7874	0.9080	1.969	0.049
Employment Status (Cat 3)	6.6938	1.5460	4.331	< 0.001
New/Repeat (Repeat)	-1.0192	0.2270	-4.489	< 0.001

All the remaining features were significant ($p < 0.05$) at $\alpha = 0.05$ meaning they were significantly predicting loan default. The key predictors identified were loan amount, interest amount, loan term, outstanding loan balance, number of serviced loans, loan purpose, employment status and the type of client (New or Repeat). These results indicate that both financial capacity and borrower behaviour play a crucial role in determining default risk. It is then worthy noting that, loan default is driven by a combination of borrower financial exposure and socio-economic characteristics, which should be carefully considered in credit risk assessment models.

DISCUSSION

The study identified several statistically significant predictors of loan default using backward elimination Logistic Regression analysis. The final reduced model retained nine significant predictors, which were, loan amount, interest amount, loan term, outstanding loan balance (OLB), serviced loans, loan purpose, employment status and borrower type (new or repeat client). The reduced model achieved strong predictive performance with an accuracy of 85.0% and an AUC of 0.925, indicating that the selected variables significantly explained variations in loan default behavior.

Loan amount emerged as one of the most influential predictors across both full and in the reduced Logistic Regression model. This finding is consistent with [8], who identified total loan amount as a major determinant of personal loan default in Tanzania. Similarly, a study by [11] found that loan size significantly influenced credit risk among microfinance institutions in Kenya. Large loan amounts may increase repayment burden and financial exposure, thereby increasing the likelihood of borrower default. Outstanding loan balance (OLB) also demonstrated strong predictive influence in the present study. This finding supports the observations of [6], who established that loan portfolio characteristics significantly affect non-performing loans in Kenyan microfinance banks. High outstanding balances may indicate over-indebtedness and repayment difficulties among borrowers, which consequently increases the default risk.

Interest amount and loan term were also found to significantly influence loan default. These findings agree with [4], who reported that interest rates and loan repayment conditions significantly affect borrower repayment behavior among small business owners in Rwanda. Higher interest obligations and unfavorable repayment periods may place financial pressure on borrowers and increase the probability of default. The number of serviced loans was another important predictor retained in the final model. This suggests that previous borrowing experience and loan management history significantly influence repayment performance. These findings are consistent with [10], who established that borrower monitoring and repayment behavior significantly affect loan repayment among microfinance clients in Makueni County.

Employment status and borrower type (new or repeat client) were also statistically significant predictors in the final model. Repeat borrowers were less likely to default compared to new borrowers, suggesting that borrower familiarity and repayment history contribute positively to creditworthiness assessment. These findings confirm that loan default among microfinance institution clients is influenced by a combination of borrower financial characteristics, loan conditions and behavioral factors

CONCLUSION

This study concludes that loan default is primarily driven by financial exposure (loan amount, outstanding balance, repayment term) and borrower characteristics (employment status and repayment history). Notably, repeat borrowers were less likely to default, suggesting that credit history is a strong protective factor. Employment status emerged as the most influential categorical predictor, showing that income stability significantly affects repayment behavior.

Future researchers should incorporate additional behavioral and digital transaction variables (e.g., mobile money usage) to improve prediction accuracy.

REFERENCES

1. Association of Micro-finance Institutions–Kenya, Microfinance Sector Report 2025. AMFI-K, 2025. [Online]. Available: https://www.amfikenya.com.
2. Central Bank of Kenya, Bank Supervision Annual Report 2024. Central Bank of Kenya, 2024.
3. Central Bank of Kenya, State of the Banking Industry Report 2025. Central Bank of Kenya, 2025.
4. D. B. Uphade, A. A. Muley, and S. V. Chalwadi, “Identification of Most Preferable Machine Learning Technique for Prediction of Bank Loan Defaulters,” Indian Journal of Science & Technology, vol. 17, no. 4, pp. 343–351, 2024. doi: 10.17485/IJST/v17i4.2978.
5. D. Twesige, A. Uwamahoro, P. Ndikubwimana, F. Gasheja, I. K. Misago, and U. Hategikimana, “Causes of Loan Defaults Within Microfinance Institutions: Learning from Micro and Small Business Owners in Rwanda: A Case of MSEs in Kigali,” Rwanda Journal of Social Sciences, Humanities and Business, vol. 2, no. 1, pp. 27–49, 2021. doi: 10.4314/rjsshb.v2i1.3.
6. H. Njuguna, “Kenya’s Microfinance Banks: A Decade of Declining Fortunes,” 2025. [Online]. Available: https://kenyanwallstreet.com/kenyas-microfinance-banks-a-decade-of-declining-fortunes.
7. H. R. Boye and F. Mithi, “Firm Characteristics and Non-Performing Loans of Microfinance Banks in Kenya,” The Strategic Journal of Business & Change Management, vol. 10, no. 4, pp. 368–381, 2023. doi: 10.61426/sjbcm.v10i4.2762.
8. I.N. Barasa, S. W. Wanyonyi, and M. M. Kololi, “Application of Logistic Regression in Enhancing Digital Credit Risk Management in Commercial Banks,” Asian Journal of Probability and Statistics, vol. 27, no. 2, pp. 13–26, 2025. doi: 10.9734/ajpas/2025/v27i2710.
9. J. E. Mwende, “Effect of Machine Learning in Early Prediction of Personal Loan Defaulting on Commercial Banks in Tanzania (2009–2020),” Master’s dissertation, University of Rwanda, College of Business and Economics, 2022.
10. M. A. Rahman, “Financial Inclusion and Economic Development,” International Journal of Economics and Management Intellectuals (IJEMI), 2024. doi: 10.63665/ijemi-y1f1a002. [Ahead of print].
11. M. K. Mwaka, “Factors Influencing Repayment Among Microfinance Loan Consumers in Makueni County: A Case of Nzaui/Kilili/Kalamba Ward, Makueni County, Kenya,” Master’s thesis, University

ISSN No. 2454-6186 | DOI: 10.47772/IJRISS | Volume X Issue VI June 2026

of Nairobi, 2017.

12. M. M. Kasina, J. M. Kihoro, and A. Kibet, “Credit Risk Indicators for Microfinance Institutions Within Machakos County, Kenya,” Indian Journal of Computer Science and Technology, vol. 4, no. 3, pp. 111–120, 2025a.
13. M. M. Kasina, J. M. Kihoro, and A. Kibet, “Performance of Different Machine Learning Algorithms for Credit Risk Classification,” Journal of Data Analysis and Information Processing, vol. 13, pp. 504–519, 2025b. doi: 10.4236/jdaip.2025.134029.