

Loan Default Prediction in Microfinance: A Comparative Analysis of Logistic Regression and Decision Forest at Grameen Bank

Risshit Sikka

12th Grade IB

American International School, Dhaka

DOI:10.37648/ijtbm.v15i04.003

¹Received: 04/07/2025; Accepted: 14/07/2025; Published: 17/10/2025

Abstract

Loan Defaulting has become a prevalent issue in financial systems worldwide, as it threatens institutional stability and increases credit risks. When borrowers fail to meet repayment obligations, financial institutions face increasing risk. In Bangladesh the problem of loan defaulting has become especially severe with default rates crossing over 20.20% and even higher in some commercial owned banks at 45.79% [1]. The aim of this study was to understand what differentiates microfinance institutes like Grameen Bank that boast a relatively low default rate at 4.32% as of July 2025 compared to the majority of Banks in Bangladesh [2]. To evaluate the characteristics of loan defaulters, this study analyses the predictive performance of two-class logistic regression and two-class decision forest models in forecasting common loan default characteristics among micro-borrowers of Grameen Bank in Bangladesh. This study contributes by applying a primary survey of 106 Grameen Borrowers. The key findings of this study reveals that Monthly Household Income, Total Number of Microloans, Age and People in Household are the key characteristics of loan defaulters. Additionally, the decision tree model overall outperformed the logistic regression model with a higher F1 score and less error.

Keywords: *Loan Defaulting; Credit Risk; Grameen Bank; Logistic Regression; Decision Forest; Borrowers*

1. Introduction

Microfinance institutes (MFIs) is a type of banking service which provides banking service to low-income individuals who otherwise would not have access to financial services [7]. However, one of the most persistent challenges followed by MFIs is the problem with loan default rate. In Bangladesh, the microfinance sector is one of the most important due to its high proportion of rural population and small entrepreneurs, who otherwise would not get access to banking service.

Despite the success of institutions like Grameen Bank, a leading figure in the microfinance industry, in maintaining recovery rates above 95%, the majority of commercial banks boast a recovery rate of only 54.21%. [1]

Understanding which borrower characteristics are most strongly affected with default can help Grameen Bank and other institutes in Bangladesh design more effective and fairer lending policy. Therefore the aim of the study is twofold: first, to identify which borrower characteristics most strongly influence default risk; and second, to evaluate which model offers superior predictive performance between logistic regression and decision forest model.

¹ How To Cite The Article: Sikka R (October 2025); Loan Default Prediction in Microfinance: A Comparative Analysis of Logistic Regression and Decision Forest at Grameen Bank; *International Journal Of Transformations In Business Management*, Vol 15, Issue 4, 40-55, Doi: <http://doi.org/10.37648/ijtbm.v15i04.003>

Several variables may influence the likelihood of loan defaulting. In this study, the borrower characteristics are separated into two categories: quantitative and qualitative characteristics. Quantitative variables include age, income level, loan amount, interest rate, repayment term, and household size, which can be measured numerically. While, qualitative factors on the other hand, include gender, marital status, education level, occupation type, and loan purpose, which represent descriptive characteristics which might have an influence.

To achieve the research objectives, two models are applied and compared. Logistic regression is a parametric statistical model that is widely used for binary classification issues, such as forecasting whether a borrower would default (1) or not (0) [4]. The model calculates the chance of default using a logistic function, stated as follows:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

where p represents the probability of default and X_i denotes the independent variables (quantitative and qualitative characteristics) and the resulting coefficients β_i measure how much each X_i affects default probability [5].

The Decision Forest is a non-parametric ensemble learning algorithm. It operates by building a large number of individual decision trees, each trained on a random subset of the data and features, and then combining their predictions through majority voting to produce a more accurate classification result [3] [6].

Using these two models we can evaluate them using other models like, Odds ratio, p-value, Test Accuracy, OOB Score, OOB Error Rate, Feature Gini Importance, Accuracy, and F1 Score.

2. Related Work

A. Comparative Performance Analysis between Conventional and Islamic Banks in Bangladesh- An Application of Binary Logistic Regression [8]

This study by Khandani, Kim, and Lo investigates the application of binary logistic regression in predicting consumer credit risk. This study was conducted using real financial data from 30 banks between 2003 and 2013. They analysed more quantitative variables like profitability, credit risk, capitalization, liquidity, efficiency, and bank size to assess their models. The results showed that Islamic banks generally maintained higher capitalization and lower credit risk, while conventional banks exhibited greater profitability

B. A Comparative Assessment of Credit Risk Model Based on Machine Learning —a case study of bank loan data [10]

Wang et al. conducted a comparative study on credit risk prediction models using real commercial bank loan data to evaluate five classifiers: KNN, Decision Tree, Random Forest, Naive Bayes, Logistic Regression. Their experimental results revealed that the Random Forest model achieved the highest precision, recall, AUC, and accuracy, outperforming the other methods.

3. Implementation

The two main models that were used in this study were Two Class Logistic Regression and Two Class Decision Forest. The study is focused on the specific characteristics that loan defaulters portray, and how accurate are Logistic Regression and Decision Forest at predicting these characteristics. We have chosen to focus on this study in characteristics of loan defaulters due to the high proportion of defaulted loans in private commercial banks of 20.16% as of March 2025 [1]. Meanwhile, Grameen Bank consistently boasts a loan default rate of less than 95% [2].

To analyse we used a convenience sample from different villages in Chittagong, Bangladesh and distributed physical copies of the survey directly. A total of 106 samples were taken over the research period. Participants were informed

at the beginning of the survey that their responses would be used solely for academic research purposes. This was done to ensure transparency and allow them to provide informed consent. They were also assured that their answers would remain confidential, which helped safeguard their privacy and encouraged openness when responding to more sensitive questions, such as those regarding age or estimated annual household income. The following disclaimer was presented to them:

4. Disclaimer: Your responses will not affect your access to banking services or loan approval in any way.

Before the two participants were asked to answer questions in regard to defaulting. We asked them to answer personal questions, where we narrowed down to 13 different variables Age, Gender, Marital Status, Education, People in Household, Monthly Household Income, Monthly Household Expenses, Home Ownership Status, Repayment of Last Loan, Total Amount Borrowed, Total Number of Microloans, Purpose of Loan, Loan Duration (Months), Monthly Installment.

A. Survey Results

Figure 1. Age Distribution by Range
Age Distribution by Range

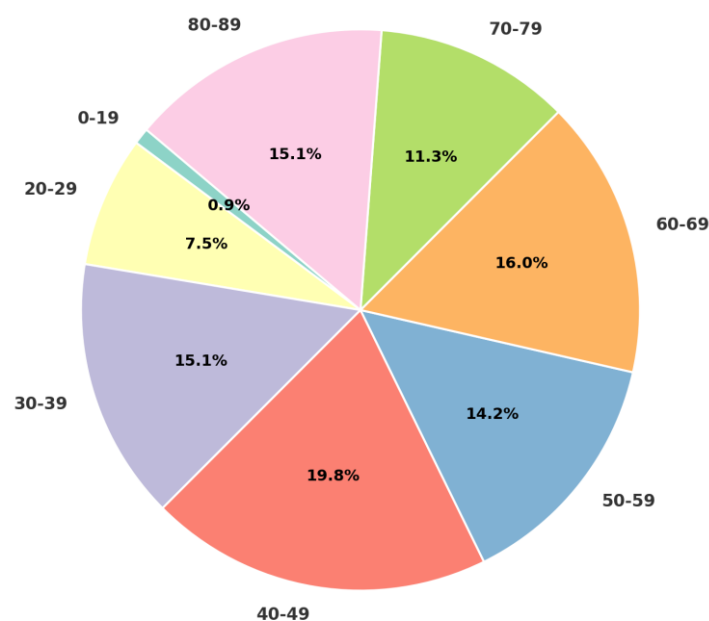


Figure 1 illustrates the age distribution where the largest group is ages 40–49, making up 19.8% of the population. This is followed by 60–69 at 16.0%, both 30–39 and 80–89 at 15.1% each, and 50–59 at 14.2%. The 70–79 range accounts for 11.3%, while 20–29 makes up 7.5%. The smallest proportion is the 0–19 group, representing only 0.9%. This indicates a population skewed more toward middle-aged and older age groups than toward the youngest.

Figure 2. Gender Distribution

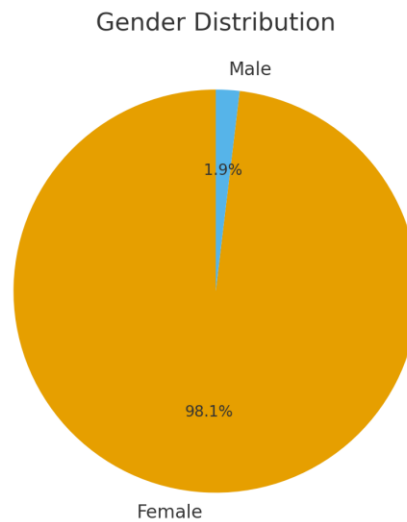


Figure 2 represents the overall gender distribution of the borrowers. Since the majority of the individuals that Grameen Bank focuses on are females. An overwhelming 98.1% are female, while only 1.9% are male. This indicates that the group is predominantly female, with males representing only a very small minority.

Figure 3. Marital Status Distribution

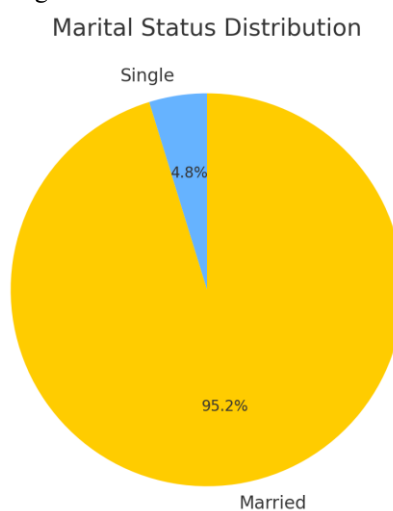


Figure 3 illustrates the overall Marital Status of the borrowers. Where the vast majority of respondents are married (95.2%), while only a small proportion are single (4.8%).

Figure 4. Education level distribution

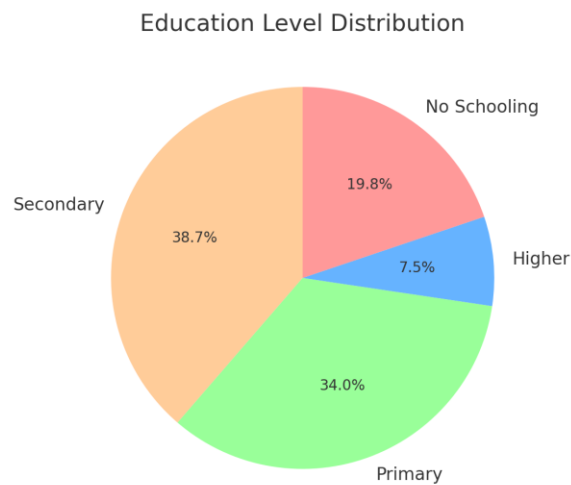


Figure 4 shows that most respondents have a secondary education (38.7%) or primary education (34.0%), while 19.8% have no schooling and only 7.5% attained higher education.

Figure 5. Household Size Distribution

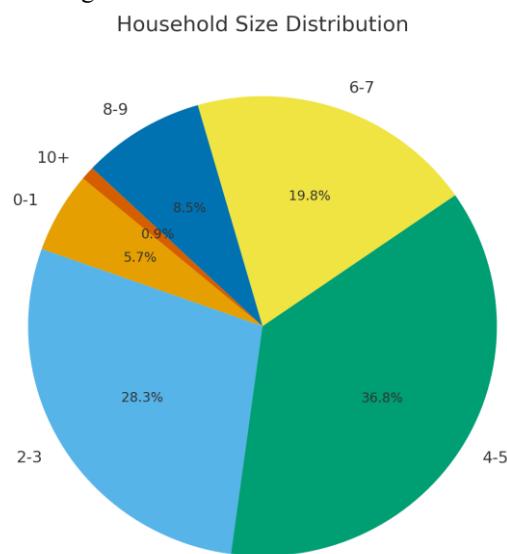


Figure 5 highlights that the largest portion (36.8%) belongs to households with 4-5 members. The smallest group is households with 10 or more members at just 0.9%. The other categories are households with 2-3 members (28.3%), 6-7 members (19.8%), 8-9 members (8.5%), and 0-1 members (5.7%).

Figure 6. Monthly Household Income

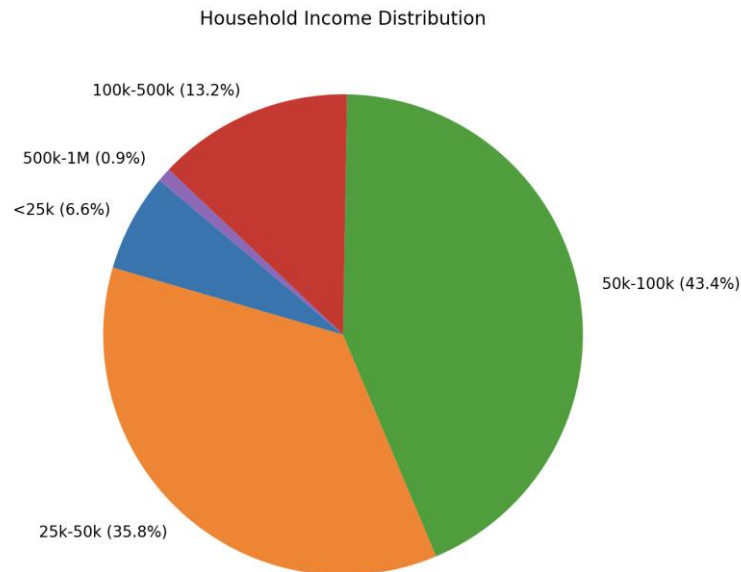


Figure 6 illustrates that the largest income bracket is 50k-100k, accounting for 43.4% of households. Following that is the 25k-50k bracket at 35.8%. Households earning 100k-500k make up 13.2%, while those earning less than 25k represent 6.6%. The smallest segment is households earning 500k-1M, at only 0.9%. All of these values are in the local Bangladeshi currency: Taka.

Figure 7. Household Expenses Distribution

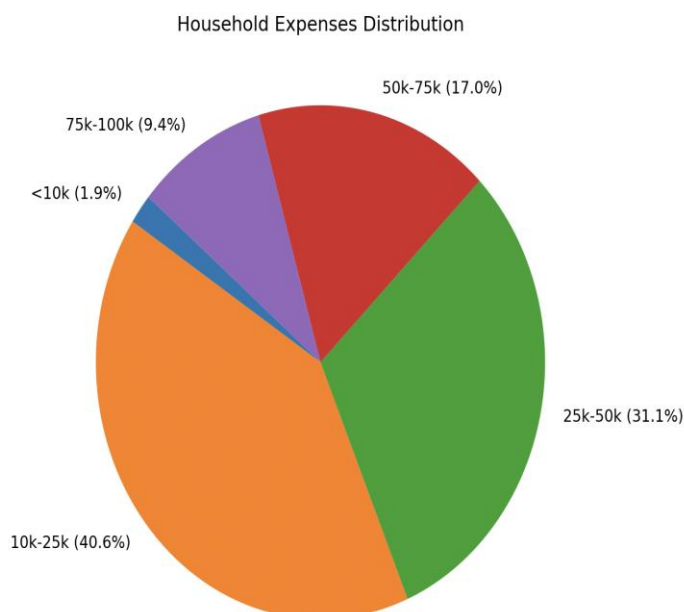


Figure 7 shows that the largest expense category is 10k-25k, representing 40.6% of households. The next largest is 25k-50k at 31.1%. Following these are 50k-75k (17.0%) and 75k-100k (9.4%). The smallest expense category is households spending less than 10k, at only 1.9%.

Figure 8. Home Ownership Status

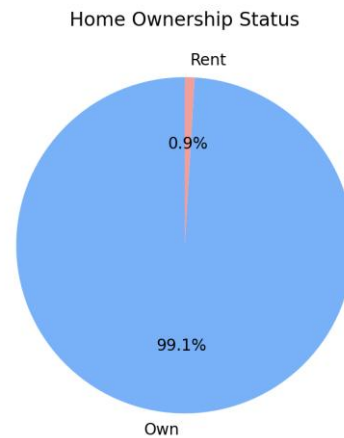


Figure 8 illustrates that the vast majority of households own their homes, accounting for 99.1%. Only a very small fraction, 0.9%, are in households that rent.

Figure 9. Loan Amount Borrowed Distribution

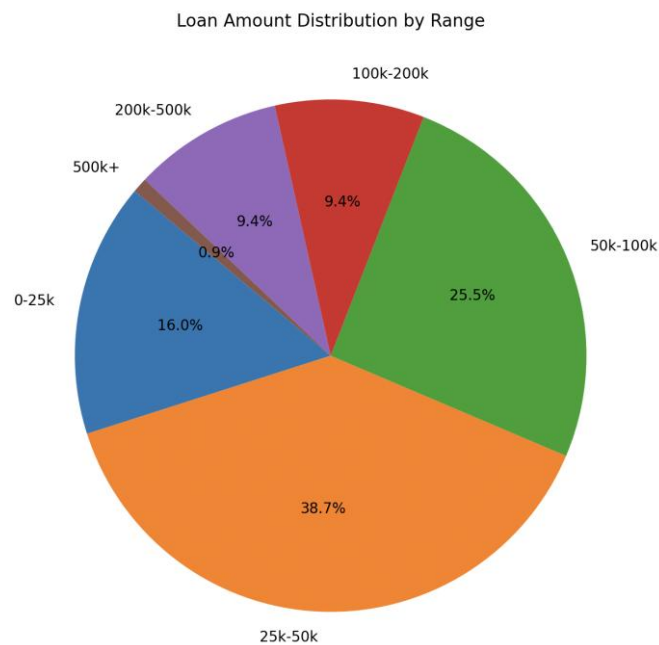


Figure 9 illustrates that the most common loan amount range is 25k-50k, representing 38.7% of loans. This is followed by 50k-100k at 25.5%. The 0-25k range accounts for 16.0% of loans, while 100k-200k and 200k-500k each make up 9.4%. The smallest category is loans of 500k+, at just 0.9%.

Figure 10. Number of Microloans Taken
Distribution of Microloans by Range

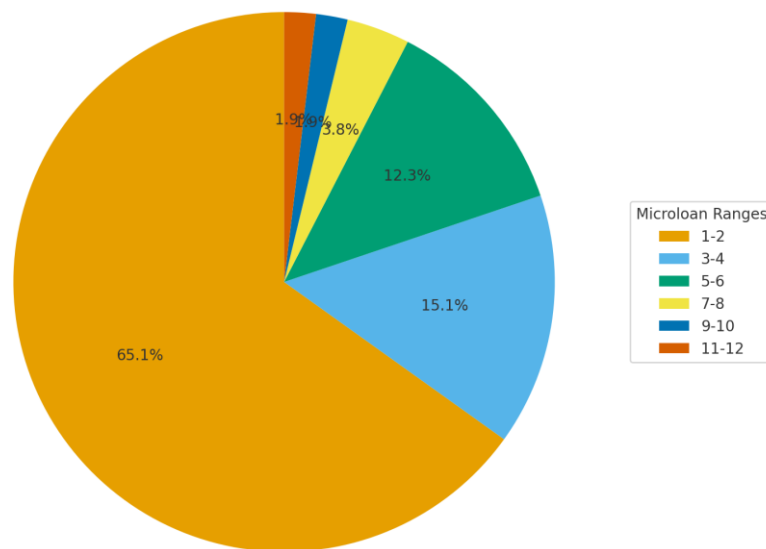


Figure 10 highlights that the largest proportion of individuals have taken 1-2 microloans, making up 65.1%. The next largest group has taken 3-4 microloans (15.1%), followed by 5-6 microloans (12.3%). A smaller percentage have taken 7-8 microloans (3.8%), and the smallest categories are 9-10 microloans (1.9%) and 11-12 microloans (1.9%).

Figure 11. Purpose of Borrowing Loan
Distribution of Loan Purposes

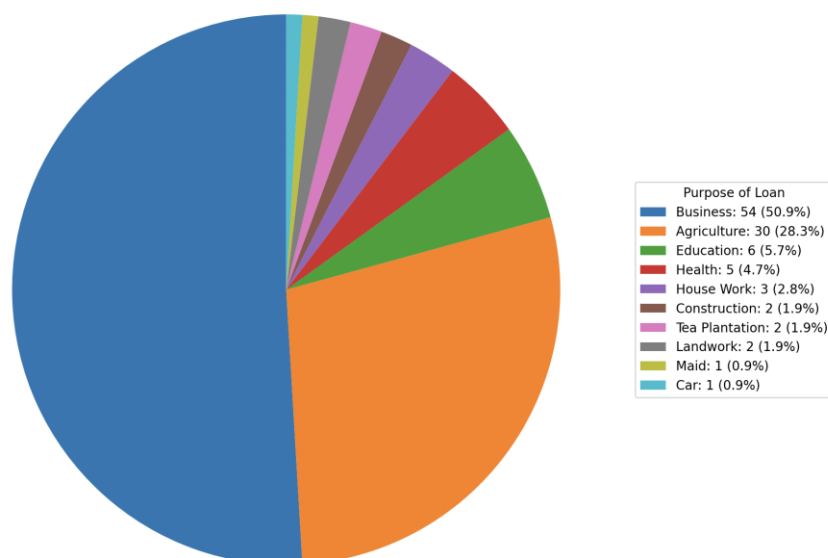


Figure 11 signifies that the most common purpose for borrowing a loan is for Business, accounting for 50.9% of loans. This is followed by Agriculture at 28.3%. Other purposes include Education (5.7%), Health (4.7%), and House Work (2.8%). The remaining purposes, Construction, Tea Plantation, Landwork, Maid, and Car, each represent a very small percentage of the total.

Figure 12. Total Loan(s) Duration
Distribution of Loan Durations (Months)

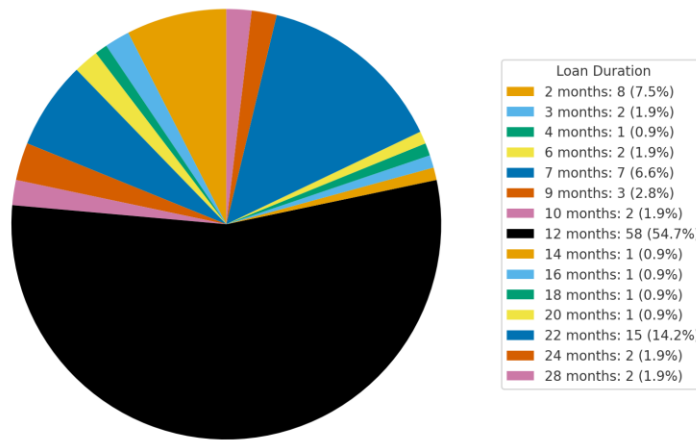


Figure 12 illustrates that the most significant category is 12 months, with a duration of 58 (54.7%). This is followed by 22 months at 14.2%, and 2 months at 7.5%.

Figure 13. Monthly Instalment of the loan
Monthly Installments Distribution by Range

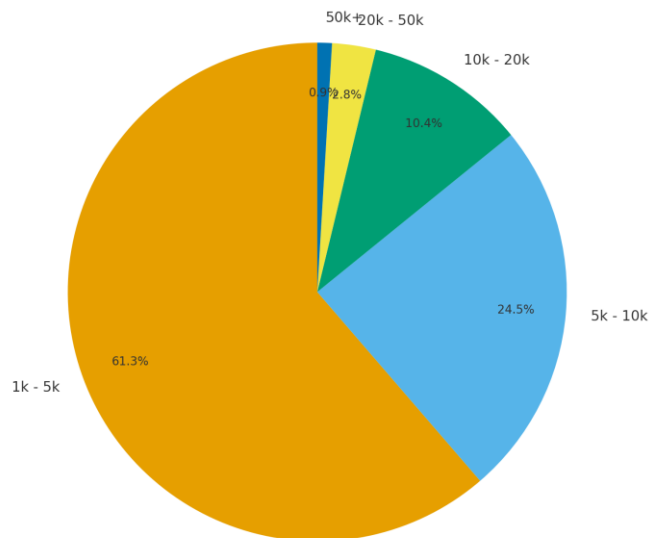


Figure 13 illustrates that the largest portion of monthly installments falls within the 1k - 5k range, accounting for a significant 61.3%. The next largest category is 5k - 10k, representing 24.5% of installments.

B. Analysis

The analysis section begins with the results of the Logistic regression model, which was used to identify significant borrower characteristics that influence likelihood of default. Logistic regression estimates a relationship between a binary dependent variable and multiple independent variables. The core equation behind the logistic regression model is:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

From this logistic regression results two metrics are derived: Odds Ratio and p-value. The odds ratio in this case shows how the independent variable affects the odds of an event happening in logistic regression and indicates how a one-unit increase in a variable affects the odds of default. Where,

$$\text{Odds} = \frac{p}{1-p}$$

[8] Using the logistic regression model we can isolate the term:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}$$

For a specific independent variable the calculation for a specific odds ratio will be the following:

$$\text{Odds ratio} = e^{\beta_i}$$

- If $OR > 1$, the dependent variable increases the likelihood of defaulting
- If $OR < 1$, the dependent variable decreases the likelihood of defaulting
- If $OR = 1$, the variable has little or no effect.

The p-value is based on the null hypothesis (H_0) and alternative hypothesis (H_1). The null hypothesis (H_0) in the context of this study is that the independent variables (Income, expenses, education, etc) do not influence the likelihood of loan default. The alternative hypothesis is that borrower characteristics do influence the likelihood of loan default. Therefore if the $p < 0.05$ reject H_0 and the variable is statistically significant and if the $p > 0.05$ it fails to reject H_0 and the variable is not statistically significant.

Logistic Regression Results (Odds Ratio & p-value)		
Variable	Odds Ratio	P-value
Monthly Household Income	0.912	0.0286
Monthly Household Expenses	1.055	0.0992
Loan Duration (Months)	0.865	0.1667
Age	1.037	0.2643
Education	0.625	0.4528
Total Amount Borrowed	1.004	0.4828
Purpose of Loan	0.761	0.6125
Monthly Instalment	0.937	0.6728
Total Number of Microloans	0.923	0.7989
People in Household	0.972	0.9036
Gender	0.000	1.0000

Martial Status	0.000	1.0000
Home Ownership Status	0.000	1.0000

Table 1. Logistic Regression Results

Borrowers with higher monthly household expenses tend to have slightly higher odds of defaulting on their loans. However the effect is only marginally significant since $0.0992 > 0.05$ and thus the relationship is not statistically significant at the 5% level. This case is similar with age and the total amount borrowed of the borrower with p-value being significantly higher 0.2643 and 0.4828 with odds ratio being close to 1, highlighting a neutral relationship or no effect.

All variables with odds ratio below 1, indicating that increases in these factors reduce the likelihood of loan default. Monthly Household Income (OR = 0.912, $p=0.0286$) and total number of microloans (OR= 0.923, $p=0.0450$) are statistically significant in this case, showing an inverse relationship where the higher the monthly household income and number of microloans the lower the chance of them defaulting on the loan.

Finally, Gender, Martial Status and Home ownership status all had OR = 0 and $p=1$. These combinations indicate that logistic regression models could not detect any meaningful effect in these variables. This is mainly because of the lack of variation in the dataset, 98% of participants were female, 95% are married, and 99% own their home. Because of the lack of variation between the values, the model cannot effectively calculate the slope or estimate any reliable effect for these characteristics.

To complement the findings from the logistic regression model, a decision forest model was applied to this dataset. The feature importance model is a technique used for calculating a score for all independent variables in a model. The score represents the respective “importance” of each feature. A higher score of feature importance is going to indicate the feature having a larger effect on the model [9].

Variable	Feature Importance
Age	0.244824
People in Household	0.146407
Monthly Household Expenses	0.123790
Total Amount Borrowed	0.123256
Monthly Household Income	0.101702
Total Number of Microloans	0.089105
Monthly Instalment	0.073140
Education	0.052810

Purpose of Loan	0.049072
Loan Duration (Months)	0.040628
Gender	0.000266
Marital Status	0.000000
Home Ownership Status	0.000000

Table 4. Decision Forest Model

Feature importance is one of the key evaluative metrics to determine whether a variable has an influence on the other. In comparison to logistic regression, age emerged as the most important predictor (0.244824). Other factors like people in household (0.146407) and Total Amount Borrowed (0.123256) are also significant influential predictors of loan defaults. Followed by monthly household income (0.101702) and total number of microloans (0.089105) which was similar to the logistic regression model. Finally, monthly instalment (0.073140), Education (0.052810), Purpose of Loan (0.049072), Loan Duration (0.040628), Gender (0.000266), Marital Status (0.000000) and Home Ownership Status (0.000000) showed little to no importance.

The model accuracy and reliability of the decision tree model were assessed using several performance metrics, including test accuracy, OOB score, and OOB error rate

Metric	Value
Test Accuracy	0.9062
OOB Score	0.9459
OOB Error Rate	0.0541

Table 5. Test Error and Accuracy

The test accuracy refers to how many predictions made by the model were correct on test data. The value of 0.9062 means that the model predicted 90.62% of borrowers correctly. It can be calculated using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP: True positive (correctly predicted defaulters)
- TN: True negative (correctly predicted non-defaulters)
- FP: False positive (predicted as defaulters but did not default)
- FN: False Negative (failed to predict the defaulter)

OOB score is a validation technique used for Random Forest models. Each tree in the forest is trained on a random subset. Some percent of the data is not used to train that tree. Those OOB samples act as “unseen data” to test the model internally. An OOB score of 0.9459 means that the model correctly predicted 94.59% of OOB samples, making it an unbiased source of estimate for the model accuracy. The OOB error rate is simply:

$$OOB\ Error\ Rate = 1 - OOB\ score$$

This means that the model misclassified 5.41% of the total data.

The confusion matrix evaluates how well each model classifies borrowers as either defaulters or non-defaulters

- True Positive: correctly predicted defaulters
- True Negative: correctly predicted non-defaulters
- False Positive: predicted as defaulters but did not default
- False Negatives: failed to predict actual defaulters

Figure 16. Confusion Matrix Logistic Regression

		Actual	
		1	0
Predicted	1	3	0
	0	4	99

The confusion matrix in Figure 16 shows the classification results for the logistic regression model. Out of all observations in this study, 3 were predicted correctly (True Positives) with 99 being correctly identified as non defaults (True Negatives). However the model missed 4 actual defaulters (False Negatives) and did not have any False positives.

Figure 17. Confusion Matrix Decision Forest

		Actual	
		1	0
Predicted	1	5	0
	0	2	99

The confusion matrix in Figure 17 shows the classification results for the decision forest model. Out of all observations in this study, 5 were predicted correctly (True Positives) with 99 being correctly identified as non defaults (True Negatives). However the model missed 2 actual defaulters (False Negatives) and did not have any False positives.

Table 6 summarizes the overall performance of both models using two evaluation methods: Accuracy and F1 score. These metrics help to assess the balance between correctly identifying defaulters in the study and avoiding false alarms.

Metric	Logistic Regression	Decision Forest
Accuracy	0.962	0.906
F1 Score	0.600	0.833

Table 6. Evaluation of Logistic Regression and Decision Forest

Both models perform relatively well, but the logistic regression is slightly higher at 96.2%, compared to decision forests 90.6%. Accuracy measures the proportion of correct predictions and is based on the results from the confusion matrix:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The F1 score combines both precision and recall into one model which is a more balanced approach, reflecting the ability of the model to both correctly identify defaulters and avoid false classifications. The F1 score can be mathematically defined as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The Decision forest model (83.3%) performed better overall compared to the logistic regression model (60.0%).

C. Discussion

The evaluation of the Two-Class Logistic Regression and Two-Class Decision Forest models reveals important insights into both model performance and borrower characteristics that most strongly influence loan default at Grameen Bank. The Logistic Regression model achieved an overall accuracy of 96.2%, slightly higher than the Decision Forest model's 90.6%. In contrast, the Decision Forest model achieved a higher recall of 71.4% and a stronger F1 score of 0.833, compared to Logistic Regression's F1 score of 0.600. The Decision Forest's superior recall indicates that it was more successful in detecting borrowers who were genuinely at risk of default [6].

In the Logistic Regression model, two borrower characteristics stood out as statistically significant predictors:

- 1) Monthly Household Income ($p = 0.0286$, OR = 0.912)
- 2) Total Number of Microloans ($p = 0.0450$, OR = 0.923)

This suggests that Grameen should generally target individuals with higher monthly income and total number of microloans, they remain significant because of factors like loyalty and security which acts as a safety net. This furthermore reinforces the Grameen Bank lending culture which is instilled with discipline and long-lasting established trust with the institution.

In the Decision Forest model, feature importance scores highlighted the variables with the greatest influence on predictive performance. The most influential characteristics were:

- 1) Age (Feature Importance = 0.2448)
- 2) People in Household (Feature Importance = 0.1464)

Older individuals were found to have a higher likelihood of defaulting on their loan. This possibly is due to limited employment opportunities or dependence on seasonal livelihoods, which accounts mostly for agriculture.

Furthermore, larger household sizes were associated with higher default risk, suggesting greater financial dependency and less available income per household member.

While this study provides valuable insights into the predictive modeling of loan defaults in the microfinance sector, several limitations should be acknowledged to ensure an accurate interpretation of the results.

The dataset used in this study was a relatively small number of 106 samples and drawn from a limited number of villages in Chittagong, Bangladesh. Small datasets can increase the risk of overfitting the data particularly in the case of decision forest and logistic regression which illustrate a precision and AUC of 1.000, which is unrealistic. Some of the data points were heavily imbalanced with 98% of participants being female, 95% being married, and 99% owning their homes. This lack of diversity along with the small number of participants limits the model's ability to evaluate how these characteristics influence loan default. Finally the models were evaluated using internal validation techniques such as test accuracy and Out-of-Bag (OOB) error [10], but external validation using cross-validation or out-of-sample testing was not possible due to dataset size limitations. Future studies might benefit from a larger sample size and diversity in the data, along with exploring advanced algorithms like Gradient Boosting or Neural Networks to strengthen model reliability.

D. Conclusion

Together, these results demonstrate that Logistic Regression offers strong interpretive clarity with higher accuracy, while the Decision Forest captures more complex relationships among variables like age, household composition, and total loan amount with higher F1 and recall score.

This study also followed a survey driven approach, where the dependent variables of the characteristics were entirely self reported. As a result, there is a potential for response bias or inaccurate reporting, which may have led to minor inconsistencies in the dataset. Future studies should address this limitation by imposing a larger sample size to strengthen the reliability of the study.

Grameen Bank's low default rates may be the result of lending to borrowers with steady incomes and established repayment histories, both of which are highly related with non-defaulting behaviour. However, the Decision Forest's identification of risk indicators such as age and household size implies that even financially secure borrowers may fail their repayment if household reliance or loan size exceeds their earning capability.

Thus from a policy perspective the findings recommend a hybrid modelling approach where both models are used jointly for both understanding borrower dynamics and early detection.

Acknowledgment

This paper is an independent work, separate from any school curriculum or qualification. It was done in the pursuit of higher education. I would like to acknowledge the use of Open AI's ChatGPT as a feedback tool for improving structure and clarity in the paper.

References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- GeeksforGeeks. (2017, May 9). *Logistic regression in machine learning*. <https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/>
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. John Wiley & Sons, Inc.

Kagan, J. (2024, July 31). *Microfinance definition: Benefits, history, and how it works*. Investopedia. <https://www.investopedia.com/terms/m/microfinance.asp>

Khandker, S., Khalily, B., & Khan, Z. (1994). *Is Grameen Bank sustainable?* (Human Resources Development and Operations Policy Working Papers). The World Bank. <https://documents1.worldbank.org/curated/en/658601468768006874/pdf/multi-page.pdf>

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787. <https://doi.org/10.1016/j.jbankfin.2010.06.001>

Noman, A. H. M., Pervin, S., Chowdhury, M. M., & Hossain, M. A. (2015). Comparative performance analysis between conventional and Islamic banks in Bangladesh- an application of binary logistic regression. *Asian Social Science*, 11(21), 248–259. <https://doi.org/10.5539/ass.v11n21p248>

Shin, T. (2023, November 7). *Understanding feature importance in machine learning*. Built In. <https://builtin.com/data-science/feature-importance>

Staff Correspondent. (2025, June 15). Defaulted loan rises to Tk 4.2 trillion. *Prothomalo*. <https://en.prothomalo.com/business/local/6jwomf618g>

Wang, Y., Wang, T., & Liu, J. (2020). A comparative assessment of credit risk model based on machine learning—a case study of bank loan data. *Procedia Computer Science*, 174, 141–149. <https://doi.org/10.1016/j.procs.2020.06.069>