IDENTIFYING FACTORS DRIVING LOAN DEFAULTS THROUGH DATA MINING IN STATA

1. Background and Problem Statement

A regional microfinance institution faced increased loan defaults despite stricter credit checks. Although customer profiles, repayment histories, and loan metadata were maintained, there was no structured data mining approach to extract insights on borrower risk. The finance team needed a reliable method to analyze historical loan data using Stata to uncover the behavioral and demographic factors influencing default risk.

The aim was to use Stata to mine the dataset and identify key predictors of loan default, build a logistic regression model, and produce a segmented risk matrix to guide lending decisions.

2. Objectives

- To analyze loan repayment data and determine the main causes of default
- To develop a logistic regression model in Stata to predict default probability
- To use segmentation to categorize borrowers by risk level
- To generate a comprehensive Stata-based report for internal use and audit readiness

3. Methodology

3.1 Data Summary

- Source: Loan database (2019–2023)
- Sample Size: 8,200 loans issued to 6,000 clients
- Variables Used:
    - LoanID, ClientID, LoanAmount, RepaymentTerm, InterestRate, IncomeLevel, Age, EmploymentType, PastDefault (binary), Defaulted (binary)

3.2 Data Preparation in Stata

- Missing data handled using mvdecode and imputation (mean/median for numeric; mode for categorical)
- Dummy variables created for employment categories using tabulate with gen()
- Outlier detection in LoanAmount and IncomeLevel through boxplots

3.3 Exploratory Data Analysis

- Summary statistics with tabstat and summarize
- Frequency tables for categorical predictors
- Visualizations using graph bar, histogram, and scatter plots

3.4 Modeling and Segmentation

- Logistic regression model specified as:
- logit Defaulted LoanAmount RepaymentTerm InterestRate IncomeLevel i.EmploymentType Age PastDefault
- ROC curve generated with lroc
- Multicollinearity checked using vif
- Used predicted values (predict) and xtile for borrower segmentation

4. Results

- Key Predictors ($p < 0.01$):
    - High InterestRate (+), Low IncomeLevel (+), PastDefault (+), Short RepaymentTerm (+)
- ROC AUC = 0.76
- Marginal effects showed past default history increased the probability of default by 22%
- Segmentation Output:
    - High Risk: 17% of borrowers, 65% default rate
    - Medium Risk: 42% of borrowers, 24% default rate
    - Low Risk: 41% of borrowers, 4% default rate

5. Interpretation and Insights

- Repeat defaulters and low-income borrowers need special scrutiny
- Default probability reduces significantly with longer repayment periods and lower interest
- Employment type had moderate effect; government-employed borrowers had lower default rates

6. Recommendations

- Adjust credit policies by tightening checks for high-risk borrowers
- Offer longer repayment terms with flexible installment plans for new borrowers
- Introduce a “second look” review mechanism for borrowers with prior defaults

7. Deliverables

- Stata .do file for full analysis and modeling
- Final report in APA format with regression output, interpretation, and recommendations
- Segmentation spreadsheet with borrower risk categories

8. Stakeholder Relevance

Academic:

- Demonstrates real-world application of logistic regression, marginal effects, and risk segmentation in Stata
- Useful for coursework in development economics, finance, and predictive modeling

Corporate:

- Applicable to credit risk modeling in banks, NBFCs, and MFIs
- Enhances evidence-based decision-making for loan officers and compliance teams