EVALUATING LOAN RISK SCORES USING MATLAB REGRESSION AND CONFIDENCE INTERVALS

1. Overview

Client:

A financial services startup based in the United States offering personal loans to mid-income segments

Objective:

To develop and validate a data-driven loan risk model using MATLAB. The model aimed to predict borrower default probability based on financial, demographic, and behavioral indicators.

2. Background

The client maintained a borrower database but lacked an analytical model to assess risk objectively. The need was for a transparent, explainable regression-based model that would allow underwriters to improve screening accuracy without depending on external credit scores.

3. Data Summary

Dataset:

Loan application data from 3,000 customers

Key Variables:

Variable	Type	Description
Default_Status	Binary	1 = Default, 0 = No Default
Monthly_Income	Continuous	In USD
Credit_History_Length	Continuous	In months
Number_of_Open_Loans	Integer	Active loan accounts
Debt_to_Income_Ratio	Continuous	% of income committed to debt
Education_Level	Categorical	High School, College, Postgrad
Age	Continuous	In years

Employment_Status	Categorical	Employed, Unemployed, Self-Employed

4. Methodology

Software Used:

MATLAB R2023b

Workflow:

1. **Data Import & Preprocessing:**
    - Imported .csv file using readtable()
    - Encoded categorical variables using dummyvar()
    - Scaled income and ratio variables to prevent coefficient distortion
2. **Modeling:**
    - Used fitglm() to build a logistic regression model:
      Default_Status ~ Monthly_Income + Credit_History_Length + Debt_to_Income_Ratio + Education_Level + Age + Employment_Status
    - Verified model assumptions using residuals() and plotDiagnostics()
3. **Statistical Analysis:**
    - Estimated odds ratios and 95% confidence intervals
    - Evaluated model significance via anova()
    - Calculated classification accuracy, precision, and recall on a test split

5. Key Results

Metric	**Value**
Overall Classification Accuracy	84.3%
Statistically Significant Predictors	Income, Debt-to-Income, Education
Odds Ratio: College vs. High School	0.67 (95% CI: 0.55–0.82)
Model p-value	< 0.001
ROC AUC	0.89

Notable Insight:

- Applicants with **monthly income < $3,000** and **debt-to-income > 0.5** had the highest predicted default risk
- Higher education consistently associated with lower default risk

6. Visual Outputs (MATLAB):

- ROC Curve: True Positive Rate vs. False Positive Rate
- Coefficient bar plot with confidence intervals
- Predicted risk heatmap by income and age
- Residual plot to check model fit

7. Deliverables

- .m scripts for data preprocessing, modeling, and evaluation
- Clean dataset with encoded variables
- 14-page technical report:
    - Logistic regression table
    - Predictive analysis summary
    - Visual risk stratification
- Business deck (4 slides):
    - Predictive model performance
    - Key borrower segments flagged
    - Integration plan with underwriting team

8. Application & Outcome

- Model integrated into the client's loan application dashboard
- Credit team used outputs to request secondary documentation for flagged high-risk applicants
- Reduced early-stage delinquencies by 22% in the first quarter post-implementation
- MATLAB workflow retained internally for continuous risk model refinement

statssy

9. Strategic Value Delivered

- Enabled **predictive screening** using internal data
- Provided **statistically explainable risk modeling**
- Offered **reusable and modifiable MATLAB codebase**
- Increased confidence in loan underwriting decisions

Statssy