Lending Club Loan Default Prediction

Prepared by Hasnath Unnisa

1. Objective

The objective of this analysis is to build and evaluate classification models to predict loan default probability using LendingClub's borrower-level data. The analysis involves data preprocessing, exploratory analysis, model development, performance evaluation, and business recommendations.

2. Data Preprocessing

The dataset underwent cleaning, missing value handling, encoding of categorical variables, and scaling of numeric features. Outliers were reviewed and redundant variables removed to improve model quality.

3. Exploratory and Bivariate Analysis

Bivariate analysis examined how borrower and loan characteristics relate to loan default risk. Median values for numeric variables and default rates for categorical variables were compared between defaulters and non-defaulters.

Key findings:
- Higher interest rates and loan amounts are associated with greater default risk.
- Defaulters generally have lower annual incomes and higher debt-to-income ratios (DTI).
- Higher revolving utilization and installments correlate with increased default likelihood.
- Renters and small business borrowers show elevated default tendencies.

4. Model Development

Two models were built: Logistic Regression (baseline) and Random Forest (ensemble). Logistic Regression provides interpretability, while Random Forest improves predictive performance.

5. Model Evaluation Summary

Performance metrics on the test dataset:

Model | Accuracy | Precision | Recall | F1-score | ROC-AUC

--- | --- | --- | --- | --- | ---
Logistic Regression | 0.640 | 0.221 | 0.626 | 0.327 | 0.685
Random Forest | 0.665 | 0.231 | 0.602 | 0.334 | 0.685

Interpretation:
- Random Forest outperforms Logistic Regression in accuracy and F1-score.
- Both achieve ROC-AUC ≈ 0.68, indicating moderate discrimination.
- Recall (~0.60) captures most defaulters, while lower precision (~0.23) is acceptable in risk contexts.

6. Key Predictors of Loan Default

Top 5 features influencing loan default probability:
1. Interest Rate (int_rate) : Higher rates increase default likelihood.
2. Loan Term (term_months): Longer tenures raise repayment risk.
3. Revolving Utilization (revol_util) : High utilization reflects financial stress.
4. Annual Income (annual_inc) : Lower income correlates with higher risk.
5. Debt-to-Income Ratio (dti) : High DTI indicates over-leverage.

7. Business Insights

- Risk-based pricing is effective: higher rates correspond to riskier borrowers.
- Borrower leverage and utilization are critical drivers of default.
- Income-based segmentation enhances credit scoring.
- Loan purpose and state variations reveal structural risks.

8. Recommendations

- Include behavioral indicators (revol_util, dti) in scoring models.
- Tighten approval for high-interest or long-term loans.
- Develop early-warning systems for high-risk borrowers.
- Reward low-risk borrowers with differentiated pricing.
- Use a hybrid model combining Logistic Regression and Random Forest.

9. Conclusion

Machine learning effectively predicts borrower default risk using LendingClub data. Key drivers include interest rate, term, utilization, income, and DTI. Random Forest provides higher predictive power, while Logistic Regression offers interpretability together forming a strong credit risk assessment framework.