Credit Risk Assessment Using Data Mining Techniques

Aditya Shah (as5069), Kris Patel (ksp177), Manan Parikh (mpp150)
16:958:588:02 - Data Mining, Rutgers New Brunswick, NJ - 08901, USA

***Abstract*—This project develops a data-driven model for assessing credit risk using data mining techniques. Our goal is to identify key factors affecting loan approval and default risk, thus enabling financial institutions to enhance decision-making processes, mitigate financial risks, and optimize loan strategies.**

I. Introduction

Credit risk assessment is crucial for banks to minimize loan defaults while ensuring deserving applicants receive necessary financial support. Effective analysis can significantly impact banks by reducing potential losses and enhancing operational efficiency.

II. Data Description

The dataset comprises 32,581 records with detailed financial and demographic data. Key variables include:

- **person_age**: Applicant's age (20-144 years).
- **person_income**: Annual income (4,000 USD - 6,000,000USD).
- **person_home_ownership**: Ownership status (RENT, OWN, MORTGAGE, OTHER).
- **person_emp_length**: Employment length (0-123 years).
- **loan_intent**: Purpose of the loan.
- **loan_grade**: Loan grading (A-G).
- **loan_amnt**: Loan amount (500 USD - 35,000 USD).
- **loan_int_rate**: Interest rate (5.42 % - 23.22 %).
- **loan_status**: Default status (0=non-default, 1=default).
- **loan_percent_income**: Ratio of loan amount to income.
- **cb_person_default_on_file**: Historical default record (Y/N).
- **cb_person_cred_hist_length**: Credit history duration (2-30 years).

Plot: Home Ownership Status Distribution

The majority of loan applicants are renters, followed by those with mortgages, and then homeowners. A minimal number of applicants fall into the 'OTHER' category.

What we can interpret from this is that renters represent the largest group seeking loans, possibly due to less financial stability or equity compared to homeowners. Banks might consider renters as a higher-risk segment and may need to scrutinize their financial stability more thoroughly. The prevalence of mortgage holders also indicates significant financial commitments, potentially affecting their capacity to manage additional debt.

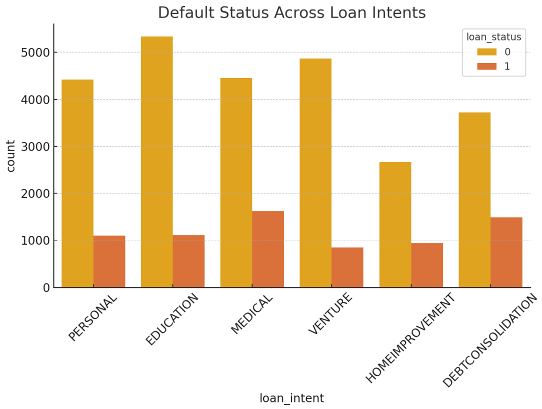

Fig. 1. Home Ownership Status Distribution

Plot: Default Status Across Loan Intents

What we can interpret from this is that applicants requesting education and medical loans pose a higher risk of default, likely due to uncertain financial returns from these expenses or higher financial stress. Loans for debt consolidation or home improvement appear less risky, possibly indicating better financial planning or existing financial stability among these applicants. Banks might prioritize these intents when allocating loans or establish stricter requirements for high-risk categories like education and medical loans.

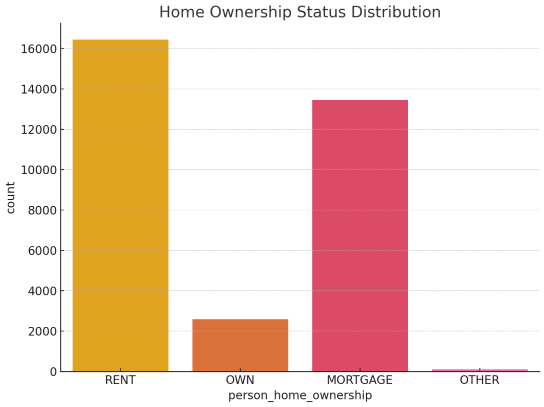

Fig. 2. Default Status Across Loan Intents

III. Preliminary Analysis

Detailed preliminary analysis identified the following key insights:

- Applicant ages primarily cluster between 20-40 years, but extreme values exist, requiring careful treatment.
- Income distribution exhibits significant right skewness, with several exceptionally high-income individuals indicating potential outliers.
- The loan default rate at approximately 22% suggests an imbalanced dataset, necessitating specialized methods to handle class imbalance effectively.
- Missing data is relatively minimal yet significant, particularly for loan interest rates (9.56%) and employment length (2.75%), necessitating appropriate imputation strategies.

IV. Proposed Methodologies

A. Data Preprocessing

Comprehensive preprocessing will include:

- Imputation using median values for numeric variables and mode for categorical variables to maintain data integrity.
- Outlier treatment via the Interquartile Range (IQR) and Z-score methods to prevent distortion of predictive models.
- Normalization of skewed data (income, loan amount) through transformations (logarithmic scaling).

B. Exploratory Data Analysis (EDA)

EDA will delve deeper by:

- Evaluating distributions and densities using boxplots, histograms, and violin plots to understand variable characteristics and interactions.
- Investigating categorical variable relationships with loan default through grouped bar plots.
- Conducting correlation analyses using heatmaps to identify relationships and guide subsequent modeling.

Plot: Loan amount by loan grade

Higher-grade loans (grades A and B) typically have smaller loan amounts compared to lower-grade loans (grades D, E, F, G), which show a higher median loan amount. Loan grades A and B exhibit less variation, indicating more conservative lending practices for higher-rated applicants. Lower grades show significantly higher variability in loan amounts, possibly indicating riskier lending.

What we can interpret from this is that financial institutions tend to limit loan amounts for high-quality borrowers (low-risk individuals), whereas for riskier loans, larger amounts are requested and possibly granted, suggesting that higher-risk borrowers seek or are approved larger loans.

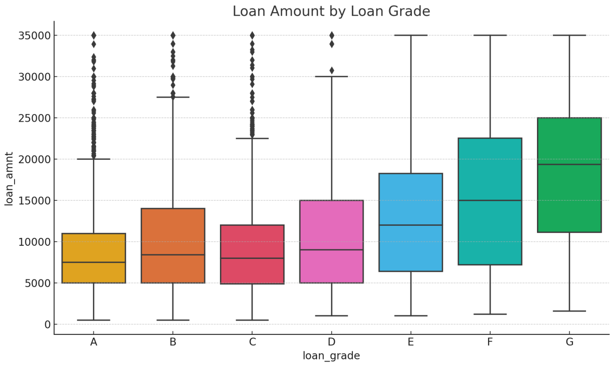

Fig. 3. Loan amount by loan grade

Plot: Interest Rate Distribution by Default Status

Defaulters have a noticeably higher median interest rate compared to non-defaulters. The density of higher interest rates is visibly more substantial for the defaulter group.

What we can interpret from this is that individuals with higher interest rates tend to default more often, indicating that elevated interest rates might correspond with higher credit risks.

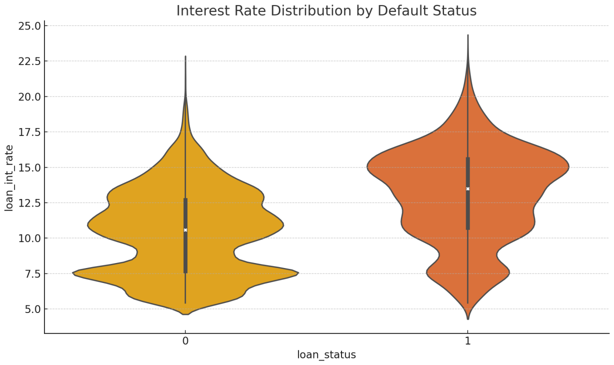

Fig. 4. Interest Rate Distribution by Default Status

Plot: Income vs. Loan Amount by Default Status

Default occurrences tend to cluster at lower income levels with relatively higher loan amounts. Non-defaulters span a broader range of incomes and loan amounts, with less clustering at the extremes.

What we can interpret from this is that borrowers with lower incomes seeking high loan amounts are potentially riskier. Financial institutions might consider implementing stricter lending criteria for low-income applicants seeking substantial loans.

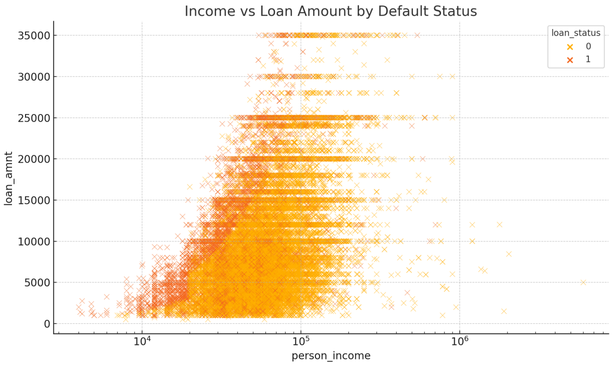

Fig. 5. Income vs. Loan Amount by Default Status

Plot: Correlation Matrix Heatmap

Strong positive correlations exist between loan amount and loan percent income, indicating loans frequently represent substantial proportions of borrower income. Interest rate correlates positively with loan status (default), suggesting higher interest loans default more often. A negative correlation between credit history length and default status is observed, implying that borrowers with longer credit histories tend to default less.

What we can interpret from this is that these relationships guide feature selection, highlighting essential predictors such as loan amount proportion, interest rate, and credit history length. Financial institutions can use these factors as significant indicators for their credit risk assessment models.

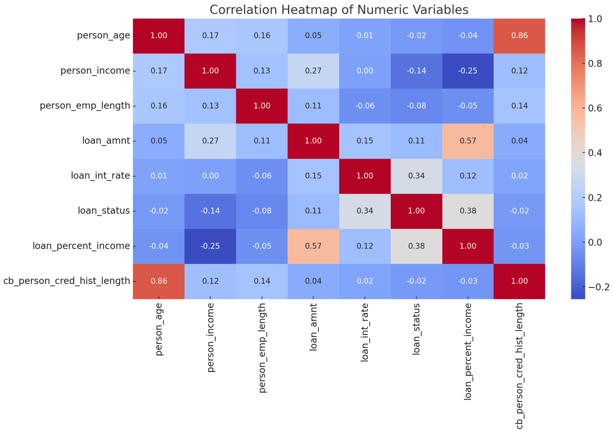

Fig. 6. Correlation Matrix Heatmap

C. Feature Engineering

We enhance predictive power through the following engineered features:

- **Debt-to-income ratio**: Captures the applicant's capacity to handle additional debt relative to income.
- **Employment stability indicators**: Reflect consistency and duration of employment, directly influencing repayment capability.
- **Interaction terms**: Between income and loan amount, as well as income and interest rate, to capture complex financial interactions.
- **Encoding categorical variables**: Applying One-Hot Encoding for nominal categorical variables (loan intent, home ownership) and Label Encoding for ordinal variables (loan grades).

Plot: Top 10 Feature Importances - Random Forest Classifier

Financial institutions should prioritize these features during the credit evaluation process. The loan interest rate emerges as a highly predictive feature, indicating borrowers with higher rates are significantly more prone to default. Similarly, the loan amount relative to income highlights borrowers' repayment ability, suggesting that careful analysis of these factors could drastically improve loan approval accuracy and risk mitigation strategies.

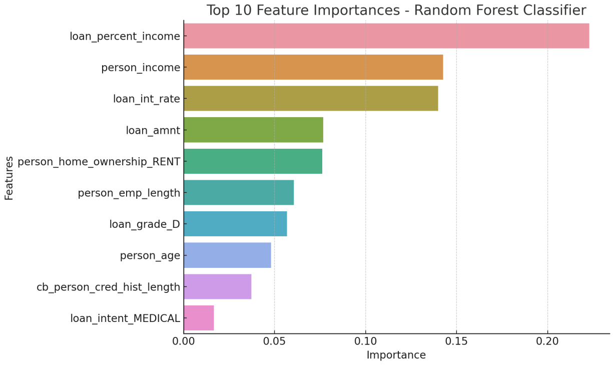

Fig. 7. Top 10 Feature Importances - Random Forest Classifier

D. Modeling and Classification

We employ multiple algorithms, each with distinctive strengths:

- **Logistic Regression**: Provides baseline performance and interpretability of feature impacts.
- **Decision Trees & Random Forests**: Allow the handling of complex interactions and offer high interpretability through feature importance metrics, robust against outliers and missing data.
- **Support Vector Machines (SVM)**: Effectively manage non-linear relationships and offer high accuracy through kernel functions, though at the expense of interpretability.

E. Model Evaluation

To robustly evaluate the models, we utilize:

- **Accuracy, Precision, Recall, and F1-score**: Providing comprehensive metrics beyond mere accuracy to handle class imbalance effectively.
- **ROC-AUC**: Evaluating overall model performance at various classification thresholds, critical for imbalanced datasets.
- **K-Fold Cross-validation**: Enhancing reliability by testing model performance across multiple folds, reducing

the likelihood of overfitting and ensuring generalization capability.

- **Confusion Matrix Analysis**: Providing insights into types of prediction errors, essential for financial institutions that weigh false positives differently than false negatives.

Plot: ROC Curve - Random Forest Classifier

An AUC close to 1.0 (observed around 0.93) indicates the Random Forest model effectively differentiates defaulters from non-defaulters. This performance means that the model can reliably assign higher risk scores to actual defaulters, which is highly valuable for financial institutions aiming to minimize financial risk and strategically manage loan approvals.

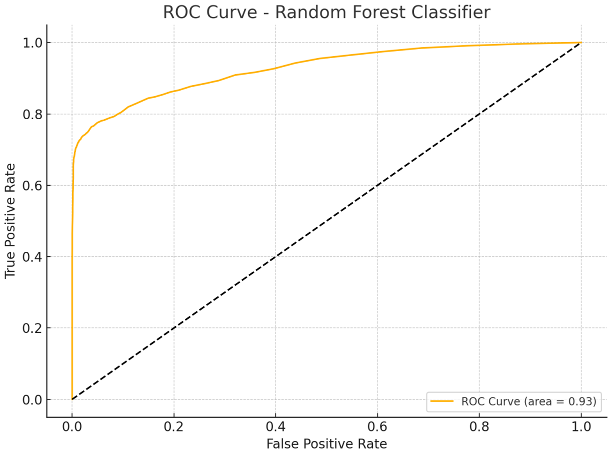

Fig. 8. ROC Curve - Random Forest Classifier

Plot: Confusion Matrix - Random Forest Classifier

The model performs strongly overall but reveals the necessity for financial institutions to carefully balance the costs of false positives and false negatives. In lending scenarios, reducing false negatives (incorrectly classifying high-risk applicants as safe) might be prioritized, as these errors directly result in financial loss. The confusion matrix guides banks on refining classification thresholds to align predictions with their specific risk management strategies.

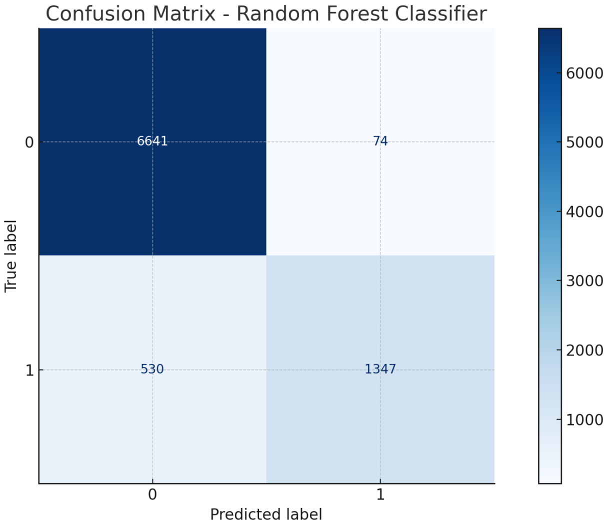

Fig. 9. Confusion Matrix - Random Forest Classifier

V. Expected Results

We anticipate:

- Identification of key predictive factors (income, loan amount, credit history).
- Development of accurate predictive models for classifying defaulters and non-defaulters.
- Actionable insights aiding financial institutions in improving lending practices and reducing default risks.

VI. Conclusion

This study aims to significantly enhance the accuracy and fairness of credit assessments, thereby supporting financial institutions in minimizing default risks, refining loan approval strategies, and improving overall financial stability.