### **Final Report: Modernizing Credit Risk Management Using Advanced Data Analytics**

**Prepared by:** Jean C. Quinones

**Course:** CIS 731

**Date:** December 15, 2024

### **1. Introduction**

In the banking sector, effective risk management is critical to maintaining financial stability. With the growing complexity of financial products and increased loan applications, understanding the factors that contribute to loan defaults is essential for both risk mitigation and profitability. This project analyzes a credit risk dataset to identify key factors influencing loan defaults and provide actionable insights for improving loan approval processes.

The primary objectives of this analysis include:

* Understanding the distribution of loan defaults across the applicant pool.
* Identifying significant predictors of loan defaults, such as income, loan amount, employment length, and interest rates.
* Evaluating correlations between various numerical and categorical features.
* Providing data-driven recommendations to improve risk assessment and decision-making.

### **2. Data Overview**

The dataset consists of **32,581 records** with **12 features**, including:

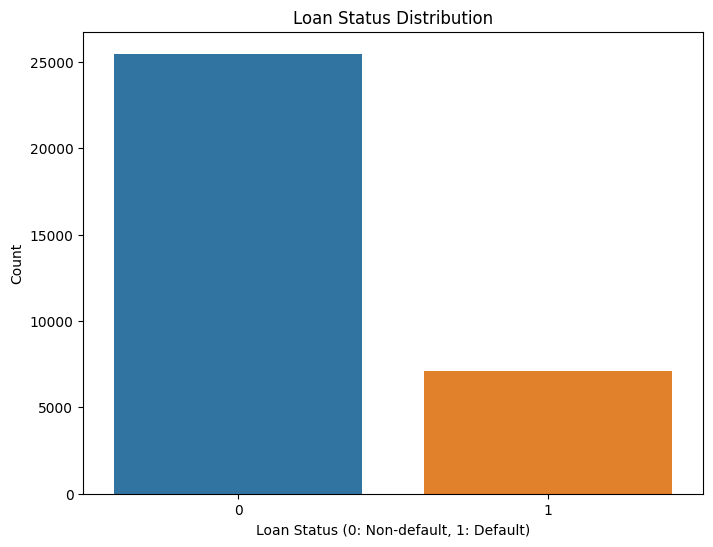
* **Numerical Variables**:
  + person\_age: Age of the applicant.
  + person\_income: Annual income of the applicant.
  + loan\_amnt: Loan amount requested by the applicant.
  + loan\_int\_rate: Interest rate applied to the loan.
  + loan\_percent\_income: Loan amount as a percentage of the applicant's income.
  + cb\_person\_cred\_hist\_length: Credit history length in years.
* **Categorical Variables**:
  + person\_home\_ownership: Home ownership status (e.g., RENT, OWN, MORTGAGE).
  + loan\_intent: Purpose of the loan (e.g., EDUCATION, PERSONAL, MEDICAL).
  + loan\_grade: Loan grade assigned based on creditworthiness.
  + cb\_person\_default\_on\_file: Historical default status (Y/N).
* **Target Variable**:
  + loan\_status: Loan repayment status (1 = Default, 0 = Non-default).

#### **Missing Values**

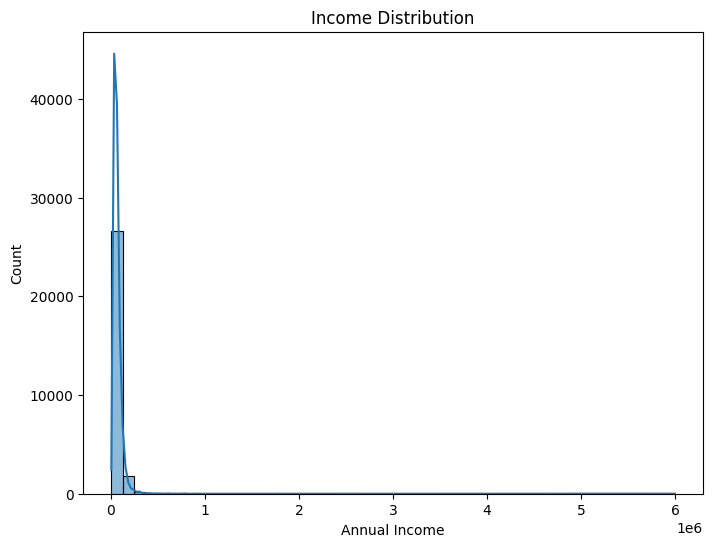
* person\_emp\_length: 2.7% missing values.
* loan\_int\_rate: 9.6% missing values.
* **Action Taken**: Rows with missing values were removed for cleaner analysis.

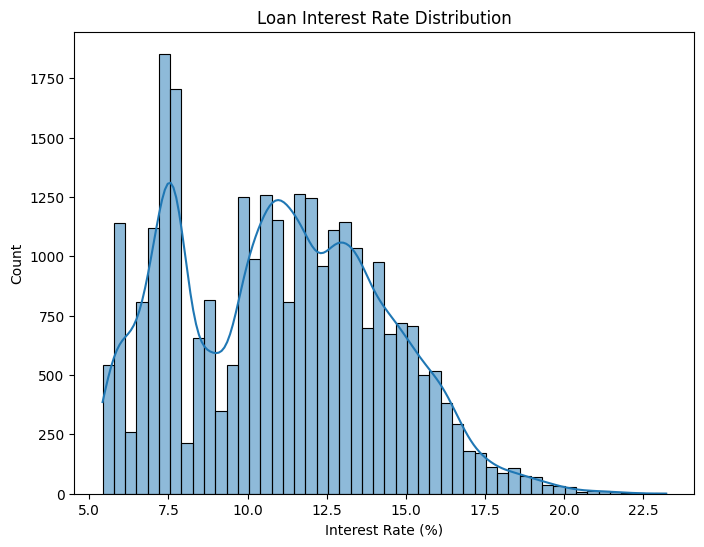
### **3. Key Findings**

#### **3.1 Loan Status Distribution**

* **Observation**: The majority of loans fall under the **non-default** category, indicating a significant class imbalance.
* **Impact**: Class imbalance can lead to bias in predictive modeling, favoring the majority class (non-defaults). Addressing this imbalance through resampling techniques or adjusting class weights is essential for improving model performance.
* 

#### **3.2 Income and Loan Interest Rate Analysis**



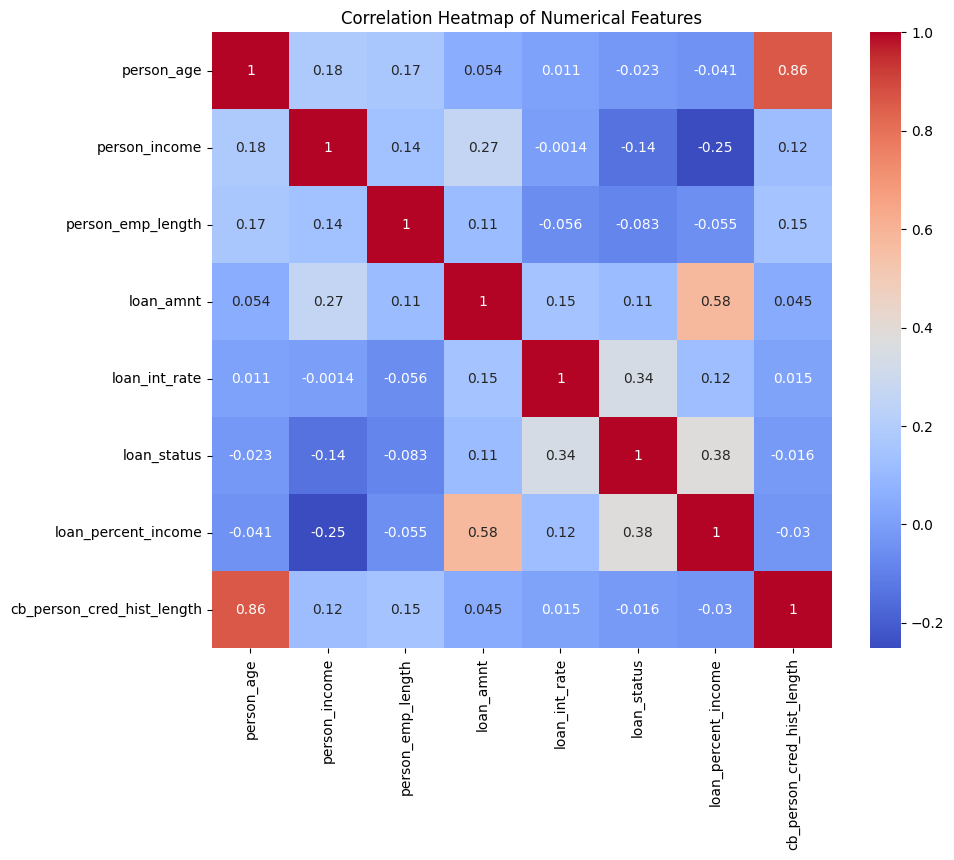
* **Income Distribution**:
  + The distribution shows that most applicants fall into **lower-income brackets**.
  + Applicants with annual incomes below $50,000 represent a significant portion of the dataset, making them more susceptible to loan defaults due to financial strain.
* **Interest Rate Distribution**:
  + Interest rates are skewed toward higher values, with many loans carrying rates between **12% and 18%**.
  + Higher interest rates often correlate with higher credit risk, indicating a need for tailored risk management strategies for high-rate borrowers.
  + 

#### **3.3 Correlation Analysis**

The correlation heatmap of numerical features reveals the following key relationships:

* **Income vs. Loan Status**:
  + A **negative correlation** indicates that applicants with lower incomes are more likely to default.
* **Loan Percent Income**:
  + Positive correlation with loan\_amnt highlights the increased financial burden for higher loan amounts relative to income.
* **Credit History Length**:
  + Longer credit history is associated with better loan repayment behavior.

These insights highlight the need to carefully evaluate loan requests from low-income applicants and those with shorter credit histories.



### **4. Detailed Analysis of Key Features**

#### **4.1 Loan Purpose (Intent)**

Analyzing the purpose of the loan reveals significant differences in default rates:

* **Personal Loans** and **Education Loans** have higher default rates compared to other categories like **Home Improvement** or **Debt Consolidation**.
* **Insight**: Applicants seeking loans for personal or educational purposes may lack stable financial backing, leading to higher risk.

#### **4.2 Employment Length**

* Applicants with shorter employment lengths (less than 5 years) show a higher likelihood of default.
* **Observation**: Stability in employment is a critical factor in determining an applicant’s ability to repay loans.
* **Recommendation**: Loans to applicants with limited employment history should undergo additional scrutiny.

#### **4.3 Loan Grades**

Loan grades (A to G) are assigned based on creditworthiness:

* Grades **D, E, F, and G** show disproportionately high default rates.
* **Insight**: Lower-grade loans correlate with higher interest rates, which further increases the financial burden and default risk.

### **5. Recommendations**

Based on the findings, the following recommendations are proposed to improve credit risk management:

1. **Enhanced Screening for Low-Income Applicants**:
   1. Develop stricter risk assessment models for applicants with annual incomes below $50,000.
   2. Introduce financial counseling programs to educate borrowers on managing debt effectively.
2. **Interest Rate Adjustments**:
   1. Consider offering lower interest rates to financially stable applicants with higher incomes and longer credit histories.
   2. Implement dynamic interest rate models that factor in employment stability and loan purpose.
3. **Improved Loan Grading System**:
   1. Refine loan grades to include additional risk factors such as employment length and loan purpose.
   2. Monitor loans graded **D and below** more closely to minimize default rates.
4. **Address Class Imbalance**:
   1. Use resampling techniques (e.g., SMOTE) or adjust model class weights to improve the predictive performance of machine learning models.
5. **Predictive Modeling**:
   1. Develop machine learning models using features such as income, loan amount, employment length, and loan grade to predict default probabilities.
   2. Implement real-time risk monitoring systems to identify high-risk loans early and take preventive actions.

### **6. Conclusion**

This analysis highlights the critical role of income, loan amount, interest rates, and employment length in predicting loan defaults. Key findings include the vulnerability of low-income borrowers, the risk associated with high-interest loans, and the importance of employment stability. By implementing data-driven recommendations, financial institutions can improve risk assessment, minimize loan defaults, and ensure sustainable growth.

Future work includes:

* Developing and deploying predictive machine learning models for loan default prediction.
* Incorporating additional variables such as borrower education level and industry type.
* Implementing real-time dashboards for dynamic credit risk monitoring.
* Exploring external economic factors, such as unemployment rates, to enhance risk models.

### **7. Acknowledgment**

Special thanks to Professor Hsu for guidance and support in completing this analysis. Appreciation is also extended to peers and collaborators who provided valuable feedback throughout the project.