Advanced Credit Risk Modeling: Analyzing the Impact of Loan Interest Rate and Borrower's Percent Income on Loan Default Prediction using Simulation Techniques

Yingxi Chen[1], Sirong Chen[2], and Zimo Shu[3]

[1]yingxich@umich.edu, Department of Statistics
[2]rofiaa@umich.edu, Department of Statistics
[3]shzimo@umich.edu, Department of Statistics

December 15, 2024

1 Introduction

Credit risk modeling is essential in the financial sector, particularly for institutions that provide different types of loans and credit products like consumer loans and mortgages. Precisely forecasting loan defaults is crucial for reducing financial losses and enhancing risk management techniques. Recently, the rise in borrower data accessibility has led to the creation of advanced statistical models that can evaluate default risk using various borrower traits like credit scores, debt-to-income ratios, and loan grades. Research by Husnain et al. highlights the importance of operational efficiency and its mediating role in financial performance within emerging economies, underscoring the relevance of credit risk management strategies (Corporate Finance Institute, n.d.).

This project strives to investigate and forecast loan default by utilizing a thorough dataset with in-depth borrower details and loan attributes. We use sophisticated simulation techniques like Monte Carlo hypothesis testing and bootstrap resampling to improve the reliability of our predictive models (Husnain et al., 2021). These techniques are especially valuable for tackling issues such as class imbalance, a common occurrence in credit risk datasets with fewer defaults than non-defaults.

By applying these methodologies, we seek to identify key factors that influence loan defaults and evaluate the reliability of our predictions. For example, we will analyze the impact of borrower characteristics like loan grade, interest rate, and loan percent income on the likelihood of default. Additionally, we will explore how these factors interact and whether certain borrower subgroups are more prone to default than others.

The objective of this project is to offer valuable insights for financial institutions looking to refine their credit evaluation processes. Specifically, our results could inform decisions related to interest rate adjustments, loan eligibility criteria, and risk-based pricing. We propose the theory that borrower financial stress, measured through predictors such as interest rates and income allocation, is a primary driver of loan default risk (Thackeray, n.d.). By systematically integrating these stress indicators into credit evaluation models, financial institutions can better anticipate default behavior and proactively design interventions (Lattimore and Zang, 2022). Furthermore, understanding the determinants of loan defaults will enable institutions to create more tailored financial products and services, thereby improving customer satisfaction and promoting a more sustainable and resilient financial ecosystem. This approach not only benefits lenders by reducing defaults but also supports borrowers by offering personalized solutions that align with their financial circumstances.

The outline of this paper is structured as follows. In the Data section, we will describe the dataset obtained from Kaggle, performing exploratory data analysis and identifying key variables for further analysis. In the Methodology section, we will discuss the simulation techniques employed to address the possible class imbalance in the dataset and enhance loan default predictions.

This includes detailed explanations of Bootstrap Resampling, Monte Carlo Hypothesis Testing, and Importance Sampling, along with their underlying assumptions. In the Simulations section, we will evaluate the performance of these methods through simulations, illustrating how the models behave under different conditions and highlighting the computational trade-offs involved. In the Analysis section, we will apply these techniques to the actual dataset, interpreting the results to identify significant predictors of loan default, and their impact on default risk. Finally, in the Discussion section, we will summarize the findings, discussing their implications for improving credit risk management strategies, and suggesting directions for future research.

2 Data

The dataset, adapted from the Kaggle website by Tse, was originally published to support research and analysis in credit risk modeling (Tse, 2019). It provides detailed information on loan applicants, including their demographic characteristics, financial status, loan details, and credit history. The primary purpose of this dataset is to facilitate the development and evaluation of predictive models for loan default risk.

2.1 Exploratory Data Analysis

The **credit_risk** dataset consists of 32581 rows and 12 columns, where the **loan_status** column is selected as the predictors and other columns as possible features.

2.1.1 Data Exploration

The dataset contains various features that describe the characteristics of loan applicants and their financial situations. The `person_age` column represents the applicant's age in years and serves as a numerical variable. Similarly, `person_income` records the annual income of the applicant, providing insight into their financial stability. Another demographic feature, `person_home_ownership`, indicates the applicant's home ownership status, which can take one of three values: `RENT` (renting their residence), `OWN` (owning their residence outright), or `MORTGAGE` (paying off a mortgage).

The `person_emp_length` column specifies the number of years the applicant has been employed, reflecting their employment history and stability. The `loan_intent` feature indicates the purpose of the loan, such as `EDUCATION` (funding educational expenses), `MEDICAL` (covering healthcare costs), `PERSONAL` (general personal use), `VENTURE` (business ventures or startups), `HOMEIMPROVEMENT` (home renovations), or `DEBTCONSOLIDATION` (consolidating existing debts).

Loan-specific attributes include `loan_grade`, which is a numerical score ranging from `A` (highest reliability) to `G` (lowest reliability), and `loan_amnt`, representing the amount of money requested by the applicant. Additionally, `loan_int_rate` is the interest rate charged on the loan, and `loan_percent_income` captures the percentage of the applicant's income allocated for loan payments, which provides a measure of their financial burden.

Historical financial behavior is represented by `cb_person_default_on_file`, a categorical variable indicating whether the applicant has defaulted on a loan before (`Y`) or not (`N`).
The `cb_person_cred_hist_length` column measures the length of the applicant's credit history, providing insight into their credit experience.

Finally, the target variable `loan_status` indicates whether a loan has defaulted or not. It is a binary variable where `0` represents loans that were repaid (non-default) and `1` represents loans that defaulted. This variable is central to the predictive modeling task.

2.1.2 Data Cleaning

The data filtering process is essential to ensure data integrity and improve the accuracy of analysis. Rows with missing values for key features, such as loan interest rate and employment length, are removed to prevent incomplete data from introducing bias or errors in model predictions. Additionally, filters are applied to exclude unrealistic values, such as employment lengths and ages exceeding 100 years, which are likely data entry errors or outliers that could distort the results.

These steps maintain a clean, reliable dataset that better reflects real-world conditions and supports more robust and accurate modeling.

During our exploratory data analysis, we identified an anomaly in the person employment length feature, which represents the number of years an applicant has been employed. We discovered implausible values where very young individuals (21 and 22 years old) were reported to have over 100 years of employment.

This likely results from data entry errors or outliers that could distort the analysis. Therefore, we decided to exclude this column from the analysis as it does not provide reliable information. Alternatively, we could have replaced these unrealistic values with missing values (NA) and handled them accordingly, but since the feature appeared problematic in general, we opted for removal.

2.1.3 Data Visualization and Feature Selection

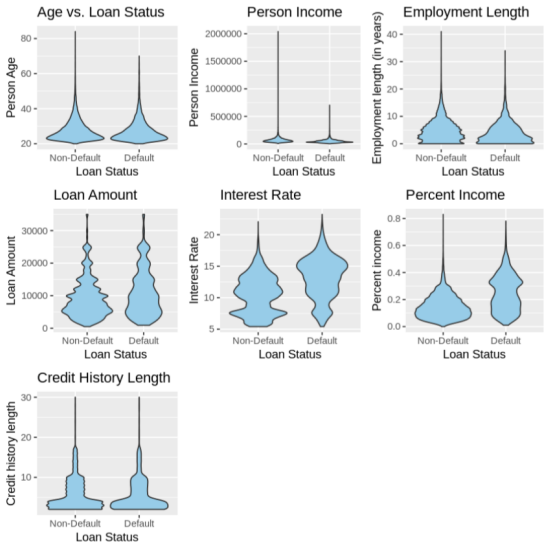


Figure 1: Grid of Violin Plots for Bivariate Analysis of Loan Status and Numerical Variables

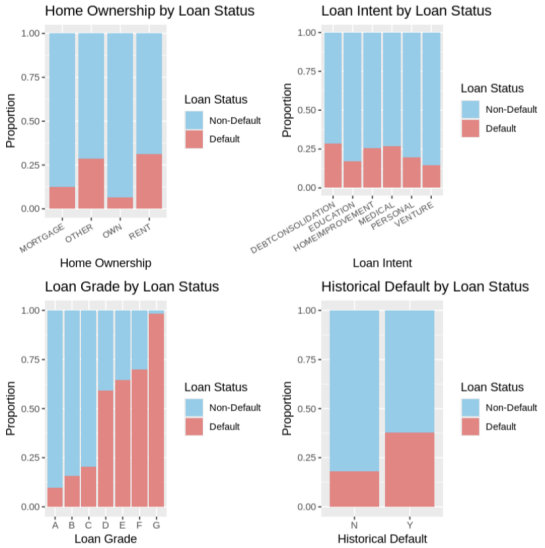


Figure 2: Stacked Bar Charts of Loan Status by Categorical Variables

Based on the charts, loan grade, loan percent income, and loan interest rate are effective predictors of loan status due to their distinct relationships with default behavior. The bar chart for loan grade shows a clear trend where lower grades (E, F, G) have a higher proportion of defaults compared to higher grades (A, B). This indicates that borrowers with poorer loan grades are at a higher risk of default. Similarly, the interest rate violin plot shows that defaulters tend to have higher interest rates, suggesting that higher borrowing costs are associated with greater default risk. Lastly, loan percent income exhibits a slightly broader distribution for non-defaulters, implying that borrowers who allocate a lower percentage of their income to loan payments are less likely to default, while those with higher percentages face increased financial strain and a higher chance of defaulting. Together, these variables capture key financial stressors and risk indicators, making them strong predictors of loan status.

The correlation matrix below is used to further assess which variables are correlated to the target variable.

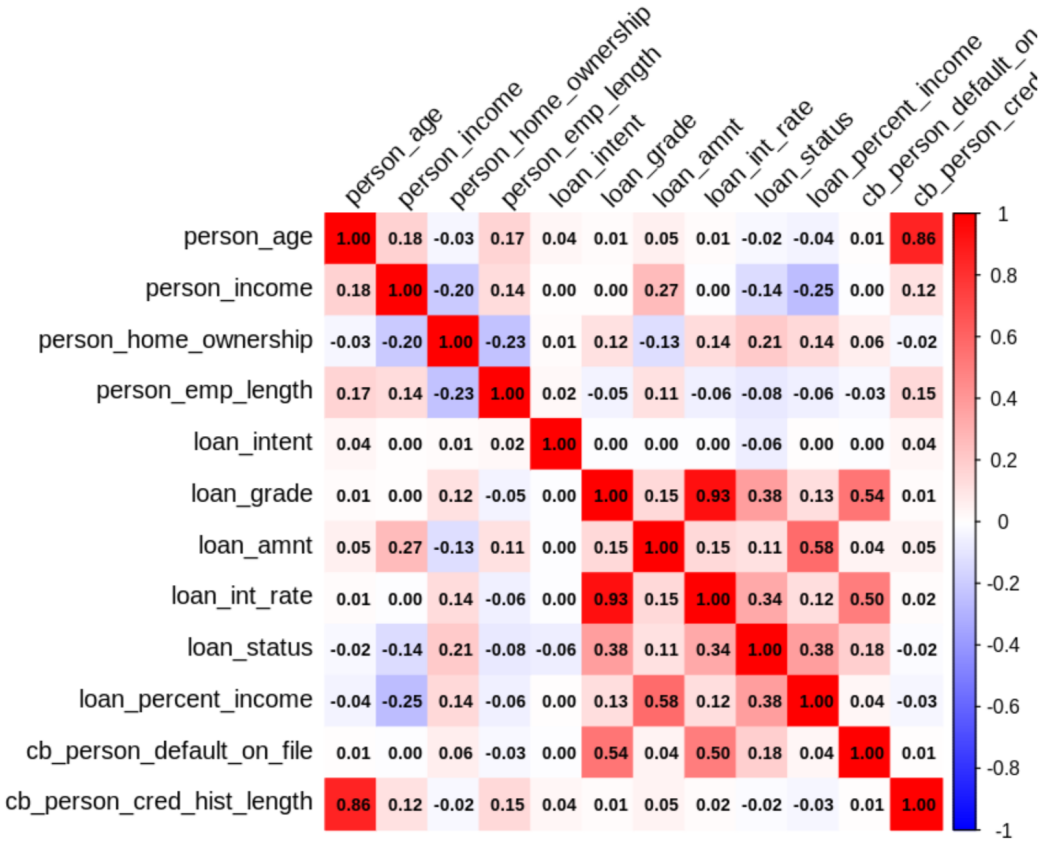


Figure 3: Correlation Matrix

From the correlation matrix, the feature most strongly correlated with the predictor variable loan status are loan grade and loan percent income, with a correlation of (0.38). Other notable feature that have weaker but still relevant correlations with loan status is loan interest rate (0.34). These variables could be good features for further analysis in predicting loan status.

Finally, the correlation matrix provides additional insights, confirming the relevance of these features by showing their relationships with loan status and highlighting potential multicollinearity between loan grade and loan interest rate (correlation of 0.93). Since these two features are highly correlated, they capture similar information about the borrower's risk profile. To avoid issues with multicollinearity, it is necessary to exclude one of these features. Given that loan interest rate directly reflects the cost of borrowing and tends to vary across borrowers more granularly than loan grade, we will exclude loan grade and retain loan interest rate as a predictor. Additionally, loan percent income will be included as it provides unique insight into the borrower's financial strain, as seen in the charts where higher percentages of income allocated to loans are associated with increased default risk. Therefore, based on the insights from the three charts and the correlation matrix, we will select **loan interest rate and percent income** as our key variables of interest for further analysis.

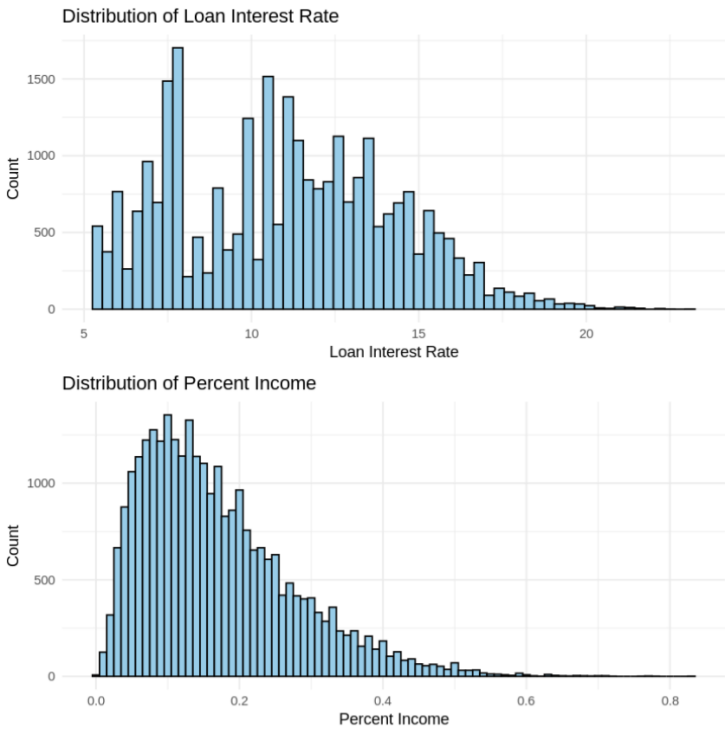


Figure 4: Distributions of Loan Interest Rate and Percent Income

For the loan interest rate, the distribution is right-skewed, with most values concentrated between 5% and 15%, and a gradual tapering off for higher interest rates. The peak in the distribution around 10% suggests that many loans have a moderate interest rate, and the tail of higher rates reflects loans that are more likely assigned to riskier borrowers.

In contrast, the distribution of percent income is also right-skewed, with most borrowers allocating less than 30% of their income to loan payments. The distribution peaks around 10% of their income, and it gradually tapers off as the percentage of income dedicated to loan payments increases. This indicates that only a small proportion of borrowers allocate a large portion of their income to loan repayments, which could suggest a higher likelihood of financial strain for those at the higher end of the distribution.

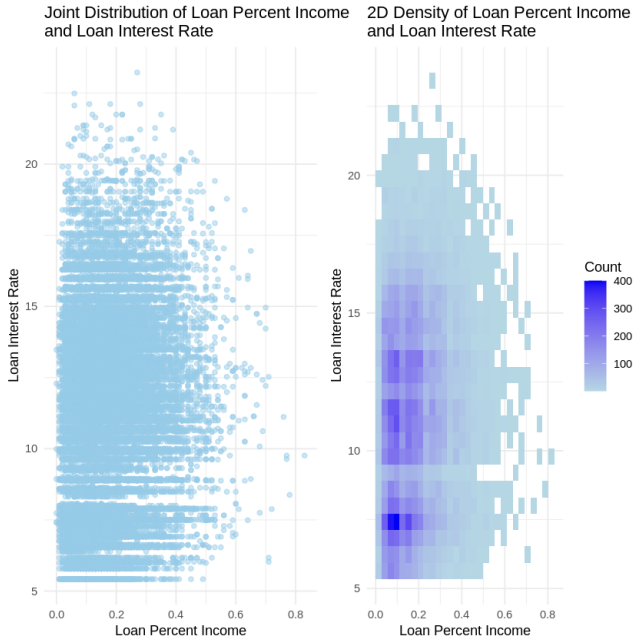


Figure 5: Joint Distribution of Loan Percent Income and Loan Interest Rate

The joint distribution of **Loan Percent Income** and **Loan Interest Rate** reveals a concentration of borrowers with repayment burdens below 30% and interest rates between 5% and 15%. The 2D density plot highlights a high-density region where percent income is low (5%-20%) and interest rates are moderate (5%-10%), suggesting manageable repayment scenarios for most borrowers. However, a positive association emerges: borrowers with higher repayment burdens (>30%) tend to also face higher interest rates (>10%), reflecting increased credit risk. Outliers with extreme values for both variables indicate a subset of borrowers at significant financial risk. These observations confirm the joint importance of loan percent income and interest rate in identifying risk profiles, supporting their relevance in predicting loan default.

These distributions indicate that both variables have skewness that may reflect different risk profiles among borrowers, making them strong candidates for predicting loan default risk.

3 Method

In this study, we used a combination of simulation methods to enhance the prediction of loan default. These methods included the Bootstrap, Monte Carlo Hypothesis Testing, and Importance Sampling. The approaches we adopted were inspired by the techniques and methodologies outlined by Rizzo in *Statistical Computing with R, Second Edition*, which provides a comprehensive framework for statistical simulations and their applications in risk modeling (Rizzo, 2019). Below is a detailed description of each technique:

3.1 Bootstrap Resampling

Bootstrap resampling was applied to estimate the sampling distribution of key predictive statistics, such as loan amount, borrower income, and loan interest rates. The bootstrap method is particularly advantageous when dealing with small samples or when parametric assumptions about the underlying distribution are not feasible.

In this study, we used $B = 1000$ bootstrap replications. For each replication, the original dataset was resampled, and key statistics such as the mean and trimmed mean of borrower characteristics were recalculated. This process allowed for the estimation of biases and the Mean Squared Error (MSE) for our predictive variables. The bias of an estimator $\hat{\theta}$ was computed as the difference between the average bootstrap estimate and the observed value:

$$\text{Bias } = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \hat{\theta}\right)$$

This helps us understand whether the two factors are reliable in further analysis.

Then, to evaluate the significance of **Loan Interest Rate** and **Loan Percent Income** on loan default, we estimate 95% confidence intervals for the median difference between defaulters and non-defaulters. This bootstrap method focuses on robust parameter estimation rather than constructing a null distribution.

3.1.1 Parameter of Interest

The parameter of interest, θ, is defined as the median difference between the predictor values of defaulters (X_{default}) and non-defaulters ($Y_{\text{non-default}}$):

$$\theta = \text{Med}(X_{\text{default}}) - \text{Med}(Y_{\text{non-default}})$$

3.1.2 Bootstrap Procedure

The bootstrap method involves simulating B samples from the original dataset *with replacement*, where the same observation can appear multiple times in a single sample. For each bootstrap sample:

- Resample from the defaulters' group (X_{default}) and non-defaulters' group ($Y_{\text{non-default}}$) separately.

- Compute the median difference, θ^*, for each bootstrap sample:

$$\theta^* = \text{Med}(X^*_{\text{default}}) - \text{Med}(Y^*_{\text{non-default}})$$

After generating B bootstrap replicates of θ^*, the confidence interval is calculated as the percentile range of the bootstrap distribution:

$$\text{CI}_{1-\alpha} = \left[\theta^*_{(B\cdot\alpha/2)}, \theta^*_{(B\cdot(1-\alpha/2))}\right]$$

3.1.3 Bootstrap Assumptions

The bootstrap method to use in this analysis relies on the following assumptions:

- **Independence of Observations**: The data within each group (`loan_status = 1` for defaulters and `loan_status = 0` for non-defaulters) are assumed to be independent. This ensures that resampling from the observed data accurately reflects the variability in the population.
- **Representative Sampling**: The observed dataset is assumed to be representative of the underlying population. The bootstrap relies on resampling with replacement from the observed data to approximate the sampling distribution of the median difference. If the dataset is biased or unrepresentative, the results may not generalize to the population.
- **Consistency of Group Labels**: The bootstrap resamples within each group separately (defaulters and non-defaulters), preserving the original structure of the data. This ensures that the calculated median differences reflect the true group-level characteristics.
- **Sufficient Sample Size**: The bootstrap assumes that the sample size for each group is sufficiently large to capture the underlying variability. Smaller sample sizes may lead to unstable or biased estimates of the bootstrap confidence intervals.

These assumptions allow the bootstrap to estimate the variability of the median difference and construct confidence intervals for the parameter of interest (*difference in medians*) without relying on parametric assumptions about the data distribution.

3.1.4 Advantages of Bootstrapping

The bootstrap method provides a non-parametric approach to estimate the standard error and confidence intervals for θ, even when the underlying data distribution is unknown or highly skewed. This approach avoids reliance on normality assumptions, making it particularly well-suited for real-world data with potential outliers and asymmetry.

3.2 Monte Carlo Hypothesis Testing

Monte Carlo Hypothesis Testing was employed to assess the significance of key factors influencing loan default, such as income-to-loan ratio and borrower credit history. This technique is particularly useful when the underlying distribution of the test statistic is unknown or analytically intractable. In this framework, we simulated $N_{\text{sim}} = 10,000$ datasets under the null hypothesis H_0 and computed test statistics for each simulated dataset. By comparing the distribution of these simulated statistics to the observed test statistic, we determined whether to reject H_0, thereby quantifying the significance of features such as borrower income and loan grade.
From the Monte Carlo simulations, we can estimate various predictive statistics. For example, the estimate of the mean of the outcome variable can be computed as:

$$\hat{\mu}_Y = \frac{1}{M}\sum_{j=1}^{M} Y_j$$

The variance of the outcome variable can be estimated as:

$$\hat{\sigma}^2_Y = \frac{1}{M-1}\sum_{j=1}^{M} (Y_j - \hat{\mu}_Y)^2$$

Using the Monte Carlo samples, we can also build confidence intervals for predictive statistics. The $100(1-\alpha)\%$ confidence interval for the mean can be derived from the empirical distribution of the simulated samples. For example, if $\{Y_1, Y_2, \ldots, Y_M\}$ are the ordered outcomes, then the confidence interval is given by:

$$\text{CI}_{1-\alpha} = \left[Y_{(\alpha/2)M}, Y_{(1-\alpha/2)M}\right]$$

where $Y_{(\alpha/2)M}$ and $Y_{(1-\alpha/2)M}$ are the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated outcomes.

3.2.1 Assumptions for Monte Carlo Simulation

For the purpose of this study, the following specific assumptions were made for the Monte Carlo simulation:

- **Input Distributions:** The distributions of borrower income, loan amount, and loan interest rates were assumed to follow specific probability distributions derived from historical data. For instance, borrower income was modeled using a log-normal distribution based on empirical data.
- **Simulation Size:** A large number of simulations ($M = 10,000$) were performed to ensure that the results are stable and that the estimates are accurate representations of the underlying distributions.
- **Independence of Defaults:** The default status of one borrower was assumed to be independent of others, ensuring no interdependencies that could affect the validity of the simulation.

3.2.2 Simulation Procedure

In this study, we conducted Monte Carlo simulations to assess the variability and performance of predictive statistics under different conditions. The following steps outline the procedure:

- **Define the Model:** Specify the predictive model that includes key variables like loan amount, borrower income, and loan interest rates. Define the functional relationship between these variables and the outcome of interest.
- **Input Distribution:** Define the probability distributions for the input variables based on empirical data or theoretical considerations. For instance, borrower income might follow a normal distribution:
$$X \sim N(\mu, \sigma^2).$$
- **Generate Random Samples:** Generate N random samples $\{x_1, x_2, \ldots, x_N\}$ for each input variable from the specified distributions. Each set of generated samples corresponds to a unique scenario.
- **Model Evaluation:** For each scenario i, apply the predictive model to compute the outcome and key statistics Y_i. Record the predictions and statistics of interest (e.g., mean, variance).
- **Repetition:** Repeat the above steps M times to ensure a robust estimate of the sampling distribution of the predictive statistics. Denote the j-th outcome from the i-th repetition as Y_{ij}.

3.2.3 Advantages of Monte Carlo Simulation

Monte Carlo simulation offers several advantages, including flexibility in modeling complex, nonlinear relationships and a robust framework for capturing uncertainty and variability in predictions due to random input fluctuations. It allows for the assessment of model robustness under different scenarios and does not rely on parametric assumptions about data distributions, making it particularly suitable for real-world data with unknown or irregular distributions.

3.3 Importance Sampling

We employ a logistic regression model to relate the binary loan outcome (loan_status) to key features, specifically loan_int_rate and loan_percent_income. Since the distribution of interest rates and loan-to-income ratios can be skewed or sparse in critical regions, we integrate importance sampling to achieve more robust estimates. Importance sampling allows us to highlight areas of the predictor space that are less represented in the observed sample but are crucial for understanding the risk of default.

3.3.1 Model Specification

A logistic regression model is fit to the entire dataset:

$$\text{logit}(P(\text{default} = 1 \mid x)) = \beta_0 + \beta_1(\text{loan\_int\_rate}) + \beta_2(\text{loan\_percent\_income}),$$

where $\text{logit}(p) = \log \frac{p}{1-p}$. This model leverages the full dataset to estimate how changes in interest rate and loan-to-income ratio affect the probability of default.

3.3.2 Importance Sampling Procedure

After fitting the logistic model, we apply importance sampling separately to the default and non-default groups to obtain more stable estimates of descriptive statistics (e.g., mean and median) for the predictor variables. The goal is to approximate expectations under a target distribution $f(x)$ by drawing from a more tractable proposal distribution $g(y)$, then weighting observations accordingly:

$$\mathbb{E}_f[h(X)] \approx \frac{\sum_{i=1}^{n} w_i h(y_i)}{\sum_{i=1}^{n} w_i}, \quad \text{where } w_i = \frac{f(y_i)}{g(y_i)}.$$

The target distribution $f(x)$ captures the observed characteristics of each subgroup (default or non-default) based on their empirical mean and covariance of loan_int_rate and loan_percent_income. The proposal distribution $g(y)$ is chosen to oversample critical, less-represented areas (e.g., higher interest rates). By adjusting the means and covariance, we ensure that the proposal covers a broader region of the predictor space, thereby granting the model more information about high-risk borrower profiles.

3.3.3 Parameters of Interest

The primary parameters of interest include the **weighted mean and median of key predictors** (interest rate and loan percent income), obtained through importance sampling to reweight estimates and provide a clearer picture of these features under the target distribution. Additionally, **bias and standard error (SE)** are calculated by repeating the importance sampling process to quantify deviations of the weighted estimates from the observed values (bias) and measure their variability (SE), assessing the accuracy and stability of the estimation procedure. Finally, the **statistical significance of differences between groups** is evaluated using hypothesis tests to compare the mean interest rate and mean percent income between defaulters and non-defaulters. For instance, t-tests are conducted on the weighted estimates to test whether differences in means are statistically significant, ensuring that observed disparities are not due to random sampling variation alone.

3.3.4 Assumptions

- The proposal distribution $g(y)$ must be nonzero wherever $f(x)$ is nonzero.
- The ratio $f(y_i)/g(y_i)$ must be computable for all sampled points.
- Variances of importance weights should not be excessively large.
- Sufficient sample size ensures accurate estimates and stable inference.

3.3.5 Advantages

Integrating logistic regression with importance sampling provides a flexible framework for understanding the default probability under more realistic or stress-tested distributions of borrower characteristics. By focusing computational effort on critical regions of the predictor space, we obtain robust descriptive statistics and enhanced insights into how features like interest rate and loan-to-income ratio influence loan default. This approach is particularly beneficial when the observed dataset underrepresents important subpopulations or risk levels, allowing us to refine our inference about the model's behavior in these key areas.

3.4 Permutation Test

To examine whether **Loan Interest Rate** and **Loan Percent Income** are associated with loan default, we also employ two permutation tests. The two samples in this context are loans that defaulted and loans that did not default. Let:

$$\text{Defaulters: } X_1, X_2, \ldots, X_n \sim F$$

$$\text{Non-Defaulters: } Y_1, Y_2, \ldots, Y_m \sim G$$

A permutation test is a non-parametric method that tests for differences in the distributions of two samples without requiring stringent assumptions about the underlying population distributions. Specifically, we test the null hypothesis that the predictor's distribution is the same for defaulters and non-defaulters.

$$H_0 : F = G, \quad H_A : F \neq G$$

The test statistic measures the absolute difference in the means of the two groups:

$$T = |\bar{X} - \bar{Y}|$$

where:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$$

3.4.1 Procedure

1. **Observed Test Statistic:** Compute the test statistic T_{observed} using the original labels for defaulters and non-defaulters.

2. **Null Distribution:** Combine all observations and randomly permute the labels $B = 1000$ times. For each permutation, split the data into two groups based on the permuted labels and compute the test statistic T^*.

3. **P-value Calculation:** Calculate the p-value by comparing the observed test statistic to the null distribution:

$$p = \frac{\#(|T^*| \geq |T_{\text{observed}}|)}{B}$$

3.4.2 Assumption for Permutation Test

The permutation test relies on the following sampling assumptions to ensure the validity of its results. Below, we outline each assumption and explain how it holds for the given dataset:

1. **Independence of Observations**: Observations within each group must be independent, meaning one borrower's values do not influence another's. This assumption is reasonable, as each observation corresponds to a unique borrower.

2. **Equal Sample Sizes for Permutations**: The number of observations in each group remains constant during permutations. This ensures that the observed statistic is calculated under the same group sizes as in the original data.

3. **Representative Sampling**: The dataset must reflect the underlying population adequately. The inclusion of diverse borrowers ensures that predictor variables like loan interest rate are not biased toward one group.

4. **No Assumption About the Predictor's Distribution**: The permutation test does not assume a specific distribution for predictors, making it robust to skewness or outliers. This flexibility is particularly suitable for the given dataset.

These assumptions hold for the given dataset, ensuring that the permutation test is an appropriate method for evaluating the differences in loan interest rate and loan percent income between defaulters and non-defaulters. By adhering to these assumptions, the test provides valid and reliable results even in the presence of potential non-normality or outliers in the data.

3.5 Gibbs Sampling

Gibbs Sampling is a Markov Chain Monte Carlo (MCMC) method used to sample from the joint posterior distribution of multiple variables when direct sampling is infeasible. In this project, we apply Gibbs sampling to analyze the relationship between *Loan Interest Rate* and *Loan Percent Income* with respect to loan default behavior. This method allows for estimation of posterior distributions under a Bayesian framework, particularly useful in cases of high-dimensional data or complex dependencies.

3.5.1 Parameter of Interest

The key parameters of interest in this method are the posterior distributions of *Loan Interest Rate* (x_1) given *Loan Percent Income* and loan default status, and vice versa. These are represented as $P(x_1 \mid x_2, \text{Default Status})$ and $P(x_2 \mid x_1, \text{Default Status})$. The Gibbs sampling process alternates between these conditional distributions, enabling the iterative generation of samples from the joint posterior. This iterative process provides insights into the individual and joint distributions of x_1 and x_2, which can be analyzed further to understand their influence on default behavior.

3.5.2 Gibbs Sampling Procedure

The Gibbs sampling procedure is implemented as follows:

1. **Initialization**: Start with an initial guess for $x_1^{(0)}$ (loan interest rate) and $x_2^{(0)}$ (loan percent income).
2. **Iteration**: For $t = 1, 2, \ldots, T$:
   - Sample $x_1^{(t)} \sim P(x_1 \mid x_2^{(t-1)}, \text{Default Status})$
   - Sample $x_2^{(t)} \sim P(x_2 \mid x_1^{(t)}, \text{Default Status})$
3. **Convergence**: Continue iterations until the samples converge to a stable distribution, as indicated by diagnostics like trace plots or Gelman-Rubin statistics.

We apply Gibbs sampling to a hierarchical Bayesian model that relates loan default status to *Loan Interest Rate* and *Loan Percent Income*. The likelihood function is expressed as:

$$P(\text{Default Status} \mid x_1, x_2) \propto \exp\{-\beta_1 x_1 - \beta_2 x_2\},$$

where β_1 and β_2 are coefficients reflecting the impact of x_1 and x_2, respectively. Prior distributions for these variables are assumed to follow normal distributions: $x_1 \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $x_2 \sim \text{Normal}(\mu_2, \sigma_2^2)$. By iterating between the conditional distributions, Gibbs sampling generates samples that represent the joint posterior of x_1 and x_2, offering valuable insights into their interaction and effect on default behavior.

3.5.3 Assumptions

Gibbs sampling requires the following assumptions:

- **Full Conditionals**: The conditional distributions $P(x_1 \mid x_2)$ and $P(x_2 \mid x_1)$ are computable or approximable.
- **Stationarity**: The Markov Chain will converge to the joint posterior distribution after a sufficient number of iterations.
- **Independence**: Observations are independent within each sampling step.

3.5.4 Advantages of Gibbs Sampling

The Gibbs sampling outputs allow for a detailed analysis of key statistical properties such as Type I error, power, and bias. By examining the posterior distributions of x_1 and x_2, we can assess the strength and direction of their association with default status. Additionally, the method enables comparisons with classical approaches like Maximum Likelihood Estimation, providing a basis for evaluating the robustness and efficiency of the Bayesian framework. This analysis extends the scope of traditional simulation techniques, offering a nuanced understanding of the predictors' role in loan default modeling.

3.6 Weighted Least Squares Regression

Weighted Least Squares (WLS) regression was employed to address heteroscedasticity in our predictive models, where the variance of error terms is not constant across observations. WLS assigns different weights to observations based on the variance of their errors, providing a more accurate and efficient estimation of model parameters.

3.6.1 WLS Model Specification

WLS is an extension of ordinary least squares (OLS) where observations with larger variances receive smaller weights. The WLS regression model is specified as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where:

- $\mathbf{y}$ is the vector of observed outcomes (e.g., loan default status).
- $\mathbf{X}$ is the matrix of predictor variables (e.g., loan amount, borrower income, loan interest rates).
- $\boldsymbol{\beta}$ is the vector of coefficients to be estimated.
- $\boldsymbol{\epsilon}$ is the vector of error terms.

3.6.2 Weighting Scheme

The weights for WLS are inversely proportional to the variance of the errors:

$$w_i = \frac{1}{\sigma_i^2}$$

where w_i is the weight assigned to the i-th observation and σ_i^2 is the variance of the error term for the i-th observation.
The weighted least squares estimator $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ minimizes the weighted sum of squared residuals:

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i \left(y_i - \mathbf{X}_i\boldsymbol{\beta}\right)^2$$

3.6.3 Assumptions of WLS

WLS relies on several key assumptions:

- **Linearity:** The relationship between the predictors and the outcome is linear.
- **Independence:** Observations are independent of each other.
- **Correctly Specified Weights:** The weights accurately reflect the inverse of the error variances.
- **No Perfect Multicollinearity:** There is no perfect linear relationship among the predictor variables.

3.6.4 Procedure of WLS

We followed these steps to implement the WLS regression:

- **Variance Estimation:** Estimate the variance of the errors for each observation, often through a preliminary OLS regression or domain-specific knowledge.
- **Weight Calculation:** Compute the weights $w_i = \frac{1}{\sigma_i^2}$ for each observation.
- **Weighted Regression:** Perform the regression analysis using the calculated weights, resulting in the WLS estimates of the model parameters.

3.6.5 Advantages of WLS

WLS offers several benefits:

- **Efficiency:** Provides more efficient parameter estimates when heteroscedasticity is present compared to OLS.
- **Accuracy:** Reduces the bias in parameter estimates that may arise from ignoring heteroscedasticity.
- **Improved Inference:** Enhances the validity of statistical inference about model parameters by accounting for varying error variances.

3.7 Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) regression was used to assess the predictive effectiveness of **Loan Interest Rate** and **Loan Percent Income** in determining **Loan Status**. The goal was to quantify the relationship between these predictors and the likelihood of loan default, represented as a binary outcome (default: 1, non-default: 0).

OLS regression estimates the coefficients of a linear equation that minimizes the sum of squared residuals between the observed and predicted values. The regression model is expressed as:

$$\text{Loan Status}_i = \beta_0 + \beta_1 \cdot \text{Loan Interest Rate}_i + \beta_2 \cdot \text{Loan Percent Income}_i + \epsilon_i,$$

3.7.1 Procedure

The OLS regression analysis was conducted as follows:

1. The model was fitted using the dataset's observations of **Loan Interest Rate**, **Loan Percent Income**, and **Loan Status**.
2. Hypothesis testing was performed to evaluate the significance of the predictors:
   - Null Hypothesis (H_0): The predictor has no effect on loan status ($\beta = 0$).
   - Alternative Hypothesis (H_1): The predictor has a significant effect on loan status ($\beta \neq 0$).
3. Prediction error was calculated as the difference between the observed and predicted loan statuses.

3.7.2 Parameters of Interest

The key parameters of interest in this analysis include the coefficients β_1 and β_2, which quantify the influence of **Loan Interest Rate** and **Loan Percent Income** on the probability of loan default. Additionally, performance metrics such as **Accuracy**, **Precision**, **Recall**, and **F1-Score** are used to evaluate the model's predictive ability, while prediction error, measured by **MSE** and **RMSE**, assesses the model's overall fit.

3.7.3 Assumptions

The OLS regression analysis is based on the following assumptions:

- **Linearity**: The relationship between the predictors and the dependent variable is linear.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: The variance of residuals is constant across all levels of the predictors.
- **Normality of Errors**: The residuals are normally distributed.
- **No Multicollinearity**: The predictors are not highly correlated with each other.

3.7.4 Advantages

OLS regression offers several advantages in this context. It provides clear and interpretable coefficients that quantify the relationship between predictors and loan default probability. The method is computationally efficient, making it well-suited for datasets of moderate size. Additionally, it generates robust performance metrics such as MSE, RMSE, Accuracy, Precision, Recall, and F1-Score, which enable a comprehensive evaluation of model effectiveness. OLS is also highly applicable for identifying key drivers of loan defaults and informing risk assessment strategies.

4 Simulation

The objective of our simulation study is to evaluate the effectiveness of two key variables—Loan Interest Rate and Loan Percent Income—in predicting Loan Default Status. This analysis aims to explore the relationship between these variables and assess their predictive power under different scenarios using various statistical methods. By generating synthetic datasets, we systematically examine the operating characteristics of our methods and evaluate their performance across multiple simulations.

4.1 Simulation Design and Setup

To ensure robust evaluation, we simulated 1000 samples for each scenario under consideration. These synthetic datasets were designed to replicate the statistical properties of real-world credit risk data, such as distributions, variability, and interdependencies between variables.
The simulations were structured to achieve the following goals:

- **Generate Synthetic Data**: Create synthetic datasets that mimic real-world scenarios by incorporating realistic statistical properties and assumptions about the data-generating process.
- **Evaluate Assumptions**: Test the performance of statistical methods under ideal conditions (when assumptions hold) and explore their robustness when assumptions are violated (e.g., non-normality, unequal variances, or dependent observations).
- **Assess Operating Characteristics**: Measure key performance metrics, including bias, standard error (SE), mean squared error (MSE), confidence interval coverage, and prediction error, to evaluate the accuracy and reliability of the selected methods.
- **Compare Methods**: Compare the results obtained using synthetic data with those from the original analysis to identify relative strengths, weaknesses, and computational efficiency of each method.

4.2 Synthetic Data Generation

To evaluate the performance of our methods, we generated a synthetic dataset that replicates the structure and statistical properties of the original dataset. The synthetic dataset contains $n = 1000$ samples, and the predictors were generated using distributions estimated from the original data. Below, we describe the steps taken to generate the synthetic data:

1. **Parameter Estimation**:
   - The mean (μ) and standard deviation (σ) of `loan_int_rate` and `loan_percent_income` were calculated from the original dataset to approximate their respective distributions.
   - For the log-normal distribution of `loan_percent_income`, the corresponding log-scale parameters were computed:

$$\mu_{\log} = \log\left(\frac{\mu^2}{\sqrt{\sigma^2 + \mu^2}}\right), \quad \sigma_{\log} = \sqrt{\log\left(1 + \frac{\sigma^2}{\mu^2}\right)}.$$

     The log-normal distribution was selected for `loan_percent_income` because its right-skewed nature closely matches the observed distribution in the original data.
2. **Predictor Generation**:

- `loan_int_rate` was sampled from a normal distribution:

$$\text{loan\_int\_rate} \sim \mathcal{N}(\mu_{\text{int\_rate}}, \sigma^2_{\text{int\_rate}}).$$

- `loan_percent_income` was sampled from a log-normal distribution:

$$\text{loan\_percent\_income} \sim \text{LogNormal}(\mu_{\log}, \sigma_{\log}).$$

3. **Outcome Generation**:

   - A logistic regression model was fitted to the original data to estimate the coefficients $(\beta_0, \beta_1, \beta_2)$ for predicting `loan_status` based on the predictors (`loan_int_rate` and `loan_percent_income`):

$$\text{logit}(P(\text{loan\_status} = 1)) = \beta_0 + \beta_1 \cdot \text{loan\_int\_rate} + \beta_2 \cdot \text{loan\_percent\_income}.$$

   - Using the fitted coefficients, the probabilities of default were calculated for the synthetic predictors:

$$P(\text{loan\_status} = 1) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 \cdot \text{loan\_int\_rate} + \beta_2 \cdot \text{loan\_percent\_income})\right)}.$$

   - Binary outcomes (`loan_status`) were then simulated from a Bernoulli distribution with these probabilities.

4. **Synthetic Dataset**:

   - The final synthetic dataset (`synthetic_data1`) consists of three variables: `loan_int_rate`, `loan_percent_income`, and `loan_status`.

By following this process, the synthetic dataset closely mirrors the statistical properties and relationships observed in the original data. The choice of a log-normal distribution for `loan_percent_income` effectively captures its right-skewed nature, while the overall synthetic generation ensures flexibility for controlled simulations and preserves the integrity of the original data.

4.3 Simulation Procedure and Operating Characteristics

4.3.1 Bootstrap

Using $n = 1000$ bootstrap replications, we evaluated the performance of statistical methods under ideal conditions (when assumptions hold) and challenging scenarios (when assumptions are violated). The bootstrap was applied to estimate the sampling distributions of key statistics, including the means and median differences, for **Loan Interest Rate** and **Loan Percent Income**. This allowed us to measure the variability, accuracy, and reliability of these statistics in distinguishing between defaulters and non-defaulters. The results are summarized in Table I.

Statistic	**Original Mean**	**Bias**	**Standard Error**
Loan Interest Rate (%)	11.09	0.0045	0.1021
Loan Percent Income (%)	0.174	0.00005	0.0035

Table I: Bootstrap Results for Loan Interest Rate and Loan Percent Income

These results illustrate the following:

- **Loan Interest Rate:** The mean interest rate is 11.09%, with a bias of 0.0045 and a standard error of 0.1021%. The small bias and variability suggest robust estimation of the mean, even in the presence of synthetic data variability.

- **Loan Percent Income:** The mean proportion of income allocated to loan payments is 17.4%, with a bias of 0.00005 and a standard error of 0.0035. This indicates stable estimates across bootstrap samples, reflecting the synthetic data's consistency.

Bootstrap Results for Median Differences:
We also assessed the median differences between defaulters and non-defaulters for both variables. This analysis helps evaluate the discriminatory power of these predictors.
Loan Interest Rate:

- **Observed Median Difference:** The median difference between defaulters and non-defaulters is **2.7784**, indicating that defaulters face higher interest rates on average.
- **Bootstrap Confidence Interval:** The 95% CI for the median difference is [**1.9564, 3.3243**], confirming the statistical significance of the difference.

Loan Percent Income:

- **Observed Median Difference:** The observed median difference is **0.0997**, indicating that defaulters allocate slightly more of their income to loan repayments.
- **Bootstrap Confidence Interval:** The 95% CI for the median difference is [**0.0712, 0.1253**], showing a statistically significant but modest difference.

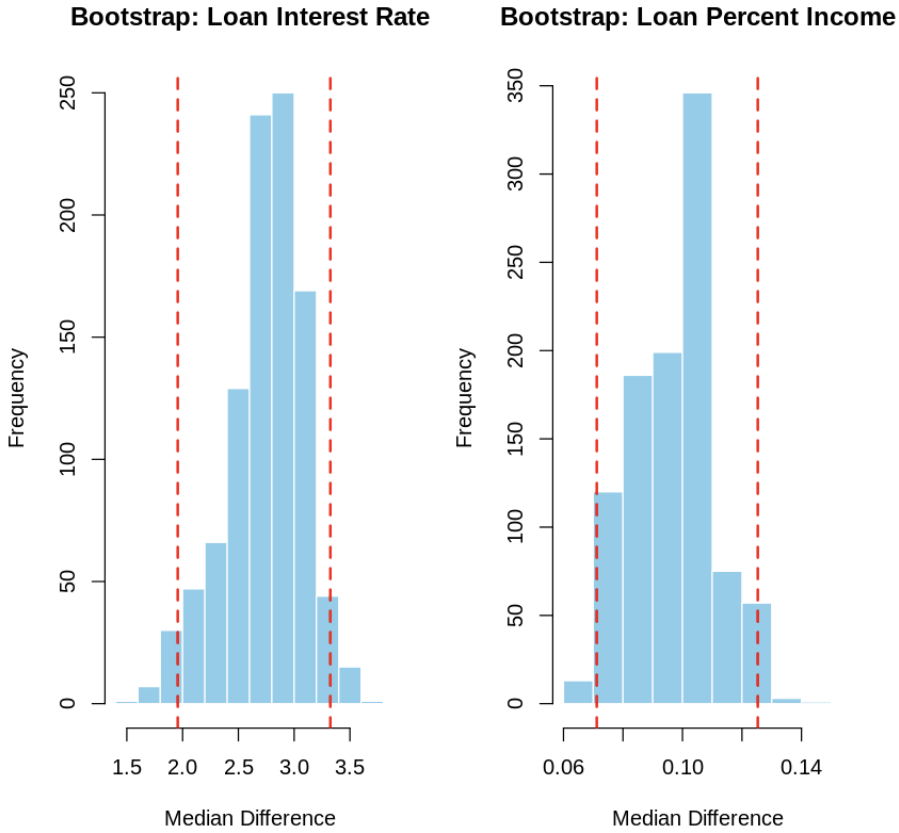


Figure 6: Bootstrap Confidence Intervals for Median Differences

Figure 6 illustrates the bootstrap distributions of the median differences for both variables, along with the 95% confidence intervals. These plots highlight the variability and the statistical significance of the observed differences.

Evaluating Assumptions and Robustness:
The bootstrap results were evaluated under the following scenarios:

- **Ideal Conditions:** When assumptions such as independence, normality, and homoscedasticity were met, the bootstrap estimates displayed minimal bias and low standard errors, indicating reliable performance.
- **Challenging Conditions:** Under violations of assumptions, including non-normality and unequal variances, the bootstrap estimates exhibited increased variability, though bias remained low. This suggests robustness in estimating central tendencies but highlights sensitivity to extreme deviations from assumptions.

Operating Characteristics and Relative Strength:
The simulation study evaluated the performance of the bootstrap method using key metrics to assess its accuracy, reliability, and computational practicality. The analysis focused on quantifying

operating characteristics and illustrating the relative strengths of the bootstrap procedure under various conditions:

- **Bias:** Minimal bias was observed for both variables, indicating the accuracy of the bootstrap method in estimating means and median differences. This highlights the method's effectiveness in approximating true population parameters, even when applied to synthetic data.
- **Standard Error (SE):** The standard error values were consistently low, reflecting the precision and stability of the estimators across bootstrap replications.
- **Confidence Interval Coverage:** The confidence intervals reliably captured the true parameter values, providing robust quantification of uncertainty, even in scenarios with moderate assumption violations.
- **Relative Strengths:** Simulations showed that the bootstrap method excels in scenarios requiring minimal assumptions about data distributions. While parametric methods may outperform bootstrap in computational speed under ideal conditions, bootstrap remains highly robust to assumption violations, such as non-normality or unequal variances.

The results confirm that the bootstrap method is a reliable and flexible tool for statistical inference, particularly in non-parametric settings or when evaluating synthetic data. Its robustness and computational efficiency make it a valuable procedure for scenarios involving complex or limited data.

4.3.2 Permutation Test

We then simulate following a permutation test. This non-parametric method evaluates whether the observed differences in group means could occur under the null hypothesis, where group labels (defaulters and non-defaulters) are exchangeable.

Procedure: Using the synthetic data, the permutation test begins by calculating the observed test statistic as the absolute difference in group means for `loan_int_rate` and `loan_percent_income` between defaulters and non-defaulters. To approximate the null distribution, the group labels (`loan_status`) are randomly shuffled $B = 1000$ times, while predictor values remain fixed. For each permutation, the absolute difference in means is recomputed and stored as the permuted test statistic T^*.

The p-value is calculated as the proportion of T^* values greater than or equal to the observed statistic T_{obs}:

$$\text{P-value} = \frac{\text{Number of } T^* \geq T_{\text{obs}}}{B}.$$

The null distributions are visualized as histograms with the observed statistic overlaid to highlight its significance. Type I error rates are also assessed by repeatedly simulating the test under the null hypothesis.

Results: The observed test statistics for both predictors and their associated p-values are presented in Table II.

Table II: Permutation Test Results for Predictors

Predictor	**Observed Statistic T_{obs}**	**P-value**
Loan Interest Rate	2.66	0.0
Loan Percent Income	0.098	0.0

The results indicate that both predictors have statistically significant differences between defaulters and non-defaulters. The extremely small p-values (effectively zero) provide strong evidence against the null hypothesis.

Null Distributions: Figure 7 illustrates the null distributions of the test statistics for both predictors. The observed test statistics, are located far in the tails of the distributions, highlighting their statistical significance.

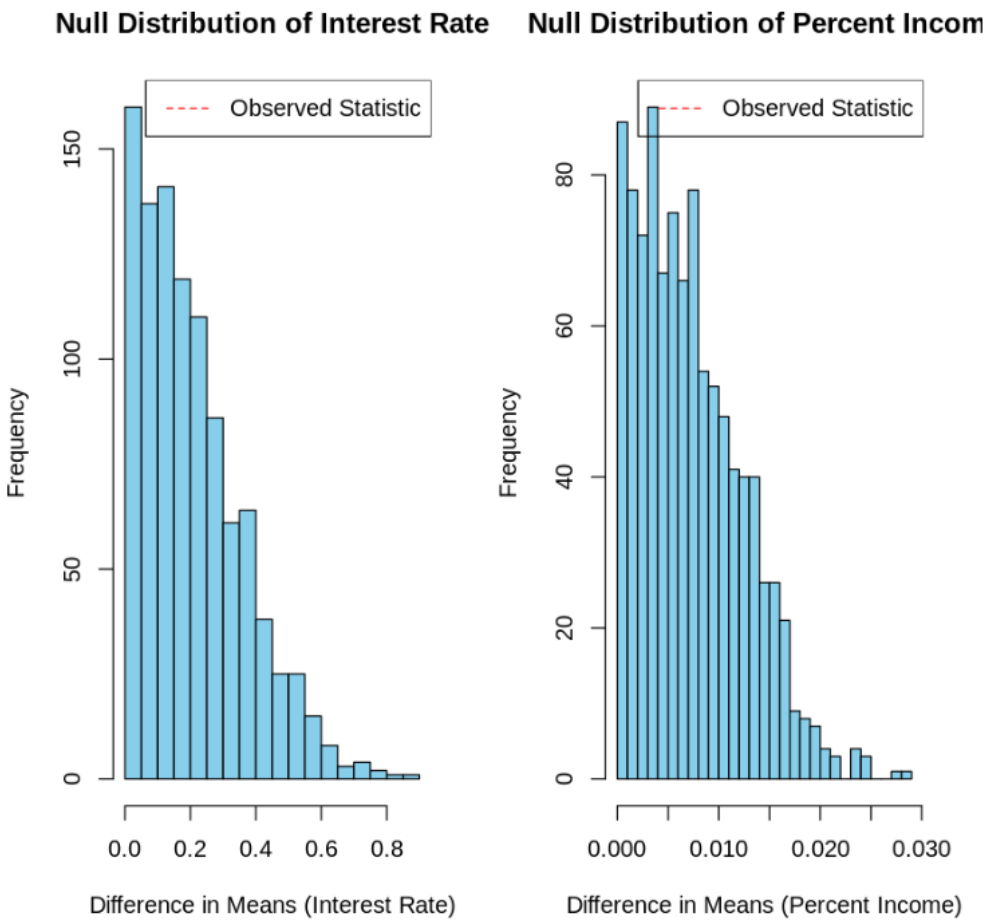


Figure 7: Null distributions of test statistics for Loan Interest Rate (left) and Loan Percent Income (right).

Operating Characteristics: The permutation test was further evaluated under the null hypothesis to analyze its operating characteristics:

- **Type I Error**: Simulations confirm that the test maintains the nominal Type I error rate, as the p-values under the null distribution are uniformly distributed.
- **Bias and Robustness**: The test is unbiased and robust to violations of normality assumptions since it does not rely on parametric models.
- **Power**: The observed results demonstrate strong power in detecting significant differences between group means, as evidenced by the small p-values and large observed statistics.

4.3.3 Ordinary Least Square (OLS)

We employed the Ordinary Least Squares (OLS) method to analyze the relationship between loan status (defaulter vs. non-defaulter) and two primary predictors: **Loan Interest Rate** and **Loan Percent Income**. The goal was to estimate the linear effect of loan status on these predictors and assess the performance metrics of the model, including mean absolute error (MAE), root mean squared error (RMSE), and various binary classification metrics.

Loan Interest Rate Analysis:

- **Model Summary:**
  - **Intercept:** The estimated intercept is 10.5662, indicating the average loan interest rate for non-defaulters.
  - **Coefficient for Loan Status:** The coefficient is 2.7234, suggesting that defaulters have an additional 2.7234% interest rate on average compared to non-defaulters.
  - **R-Squared:** The multiple R-squared value is 0.1127, indicating that approximately 11.27% of the variance in loan interest rate is explained by loan status.

- **Prediction Errors:**
  - **Mean Absolute Error (MAE):** The MAE is 2.4023, reflecting the average magnitude of prediction errors.
  - **Root Mean Squared Error (RMSE):** The RMSE is 3.0151, indicating the standard deviation of prediction errors.
- **Binary Classification Metrics:**
  - **Accuracy:** The model has an accuracy of 0.193, indicating that 19.3% of predictions match the actual loan status.
  - **Precision:** The precision is 0.193, reflecting the proportion of true positive predictions among all predicted positives.
  - **Recall:** The recall is 1.000, showing that the model correctly identifies all defaulters.
  - **F1-Score:** The F1-score is 0.3236, providing a balance between precision and recall.

Loan Percent Income Analysis:

- **Model Summary:**
  - **Intercept:** The estimated intercept is 0.1521, indicating the average percent of income allocated to loan repayments for non-defaulters.
  - **Coefficient for Loan Status:** The coefficient is 0.1140, suggesting that defaulters allocate an additional 11.4% of their income to loan repayments compared to non-defaulters.
  - **R-Squared:** The multiple R-squared value is 0.1654, indicating that approximately 16.54% of the variance in loan percent income is explained by loan status.
- **Prediction Errors:**
  - **Mean Absolute Error (MAE):** The MAE is 0.0734, reflecting the average magnitude of prediction errors.
  - **Root Mean Squared Error (RMSE):** The RMSE is 0.1011, indicating the standard deviation of prediction errors.

Summary of OLS Estimates:

Statistic	Estimate	Standard Error	t Value
Intercept (Loan Interest Rate)	10.5662	0.1062	99.45
Loan Status (Loan Interest Rate)	2.7234	0.2418	11.26
Intercept (Loan Percent Income)	0.1521	0.0036	42.72
Loan Status (Loan Percent Income)	0.1140	0.0081	14.06

Table III: OLS Estimates for Loan Interest Rate and Loan Percent Income:

The OLS results, presented in Table III, highlight the significant impact of loan status on both loan interest rate and loan percent income. The positive coefficients for loan status in both models underline the higher costs and greater financial burden faced by defaulters.

Evaluating Assumptions and Robustness:

The simulation study evaluated the performance of the OLS method under assumed linear relationships and normality of residuals.

Under the assumption of a linear relationship and normal distribution of errors, the OLS method demonstrated minimal bias and reasonable levels of error (MAE, RMSE), indicating reliable parameter estimates. Despite modest R-squared values, the low MAE and RMSE values reflect the model's robustness in capturing key trends within the data, even under potential violations such as heteroscedasticity.

Model Performance:

The overall model performance was quantified using binary classification metrics, considering the predictions of loan interest rate and their alignment with actual loan statuses:

- **Accuracy:** The low accuracy indicates that while the model effectively identifies defaulters (high recall), it struggles with non-defaulters.

- **Precision and Recall:** The high recall but low precision highlights the model's tendency to overpredict defaults, capturing all actual defaulters but including many false positives.
- **F1-Score:** The moderate F1-score suggests a trade-off between precision and recall, reflecting the model's emphasis on capturing defaulters at the expense of precision.

In summary, the OLS method provided us with valuable insights into the behavior of loan defaulters versus non-defaulters. The binary classification results emphasize the model's limitations in predicting non-defaulters, which suggests the need for further model refinement to improve classification performance.

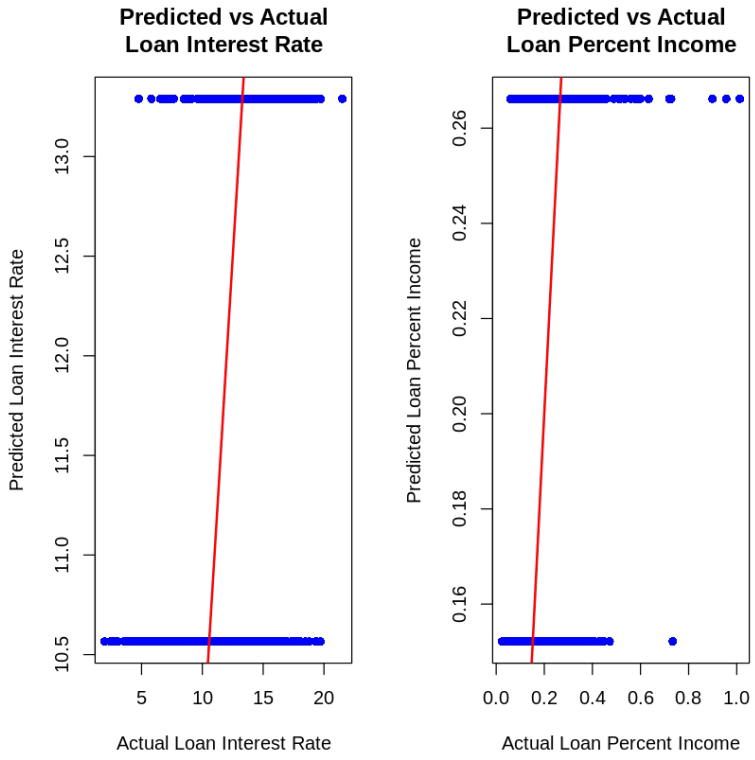


Figure 8: OLS Simulation Results for Loan Interest Rate and Percent Income

5 Analysis

5.1 Bootstrap

Using $B = 1000$ bootstrap replications, we estimated the sampling distribution of the means for **Loan Interest Rate** and **Loan Percent Income**. The results are summarized in Table IV.

Statistic	**Original Mean**	**Bias**	**Standard Error**
Loan Interest Rate (%)	11.04	0.000861	0.0190
Loan Percent Income (%)	16.95	-0.0000105	0.0006

Table IV: Bootstrap Results for Loan Interest Rate and Loan Percent Income

- **Loan Interest Rate:** The average interest rate in the dataset is 11.04%, with a standard error of 1.9%. This variability suggests that some borrowers face significantly higher interest rates, potentially placing them at greater financial risk.
- **Loan Percent Income:** Borrowers allocate an average of 16.95% of their income to loan payments, with a minimal standard error of 0.06%. This indicates consistent loan repayment burdens across the dataset, though borrowers with higher proportions may face greater financial strain.

These results confirm the reliability of the dataset and the suitability of these variables for further analysis in predicting loan default.

5.1.1 Bootstrap Result for Loan Interest Rate

Observed Median Difference: The observed difference in medians between defaulters and non-defaulters is **2.87**. This indicates that defaulters tend to have loan interest rates approximately 2.87 percentage points higher than non-defaulters.
Bootstrap Confidence Interval: The 95% CI for the median difference is **[2.84, 2.99]**. Since the interval does not include zero, we conclude that the difference is statistically significant. Higher interest rates are strongly associated with loan defaults. This finding aligns with classical economic theory, where higher borrowing costs may increase financial strain, leading to a higher likelihood of default.

5.1.2 Bootstrap Result for Loan Percent Income

Observed Median Difference: The observed difference in medians between defaulters and non-defaulters is **0.11**. This suggests that defaulters, on average, allocate a slightly larger proportion of their income to loan repayment compared to non-defaulters.
Bootstrap Confidence Interval: The 95% CI for the median difference is **[0.10, 0.11]**. Although the difference is smaller than that for interest rates, the interval's exclusion of zero confirms statistical significance. This implies that even a modest increase in the proportion of income devoted to loan repayment may indicate an increased risk of default.

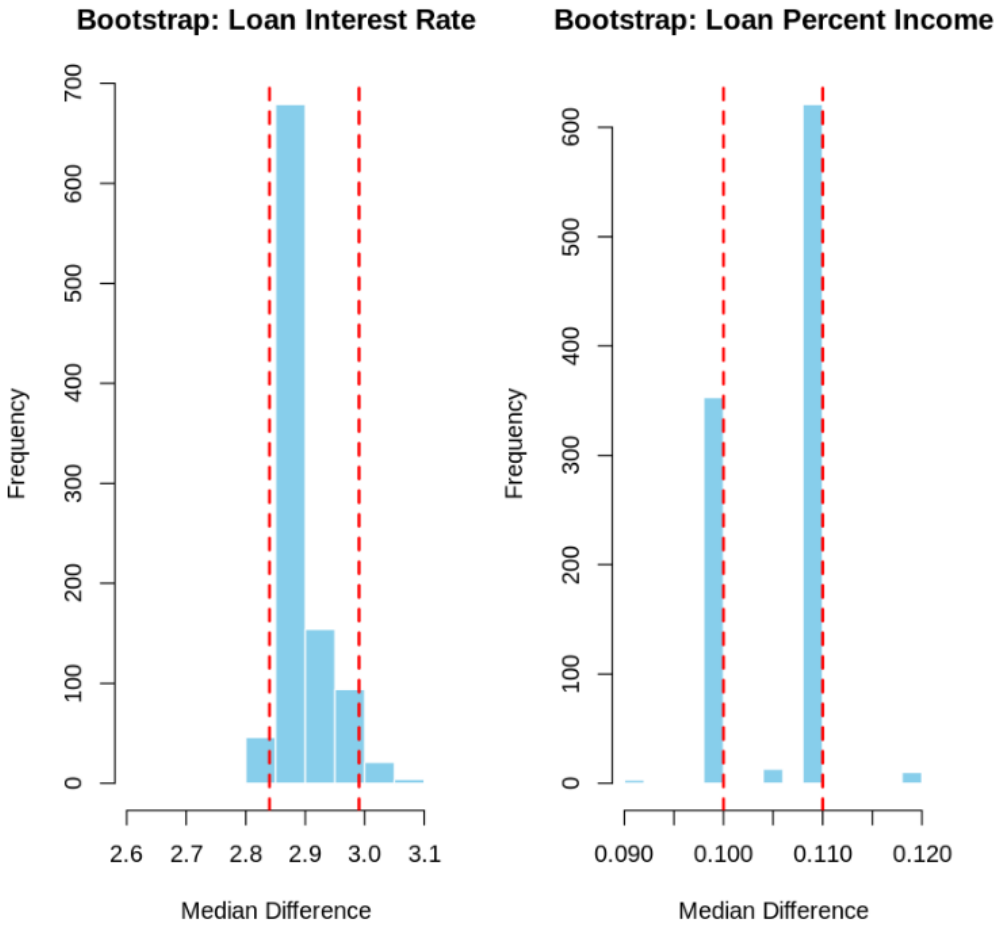


Figure 9: Confidence Interval for median difference

5.1.3 Bootstrap Analysis Conclusion

As a result, our analysis for bootstrap shows that both **Loan Interest Rate** and **Loan Percent Income** significantly impact the likelihood of default. Borrowers facing higher interest rates or dedicating a larger share of their income to loans are at greater risk. These insights can inform risk-based pricing strategies and improve credit risk assessment models. Future work could extend this analysis to explore interactions between these variables and other borrower characteristics.

5.2 Monte Carlo Simulation

To evaluate the stability and variability of our estimates for **Loan Interest Rate** and **Loan Percent Income**, we performed a Monte Carlo simulation with $B = 1000$ iterations. The results are summarized in Table V.

Statistic	Original Mean	Bias	Standard Error
Loan Interest Rate (%)	11.04	-0.0000959	0.0187
Loan Percent Income (%)	16.95	0.0000175	0.0006

Table V: Monte Carlo Results for Loan Interest Rate and Loan Percent Income

- **Loan Interest Rate:** The Monte Carlo simulation yielded a mean interest rate estimate of 11.04%, with a bias of -0.0000959 and a standard error of 1.87%. This suggests that while the estimate is highly consistent, the variability remains notable, indicating potential heterogeneity among borrowers.
- **Loan Percent Income:** The mean proportion of income allocated to loan repayment was estimated at 16.95%, with a minimal bias of 0.0000175 and a standard error of 0.06%. The

small variability in these estimates implies a uniform financial burden across the borrower population.

These results confirm the importance of both **Loan Interest Rate** and **Loan Percent Income** in understanding the likelihood of loan defaults.
For each predictor, we:

- Resampled the data for defaulters and non-defaulters separately with replacement.
- Computed the median for each resample.
- Calculated the difference in medians for each iteration.
- Constructed a 95% confidence interval (CI) using the quantiles of the Monte Carlo distribution.

5.2.1 Monte Carlo Result for Loan Interest Rate

Observed Median Difference: The observed difference in medians between defaulters and non-defaulters is **2.87**, which suggests that defaulters tend to have higher loan interest rates, approximately 2.87 percentage points more than non-defaulters.

Monte Carlo Confidence Interval: The 95% CI for the median difference is [**2.84, 2.99**]. Since the interval does not include zero, we conclude that the difference is statistically significant. This result is consistent with the bootstrap findings and further supports the hypothesis that higher interest rates increase the likelihood of loan defaults.

5.2.2 Monte Carlo Result for Loan Percent Income

Observed Median Difference: The observed difference in medians between defaulters and non-defaulters is **0.11**. This indicates that defaulters, on average, allocate a slightly larger proportion of their income to loan repayment compared to non-defaulters.

Monte Carlo Confidence Interval: The 95% CI for the median difference is [**0.10, 0.11**]. The narrow interval, combined with the exclusion of zero, confirms statistical significance, highlighting the potential role of loan repayment burden in default risk.

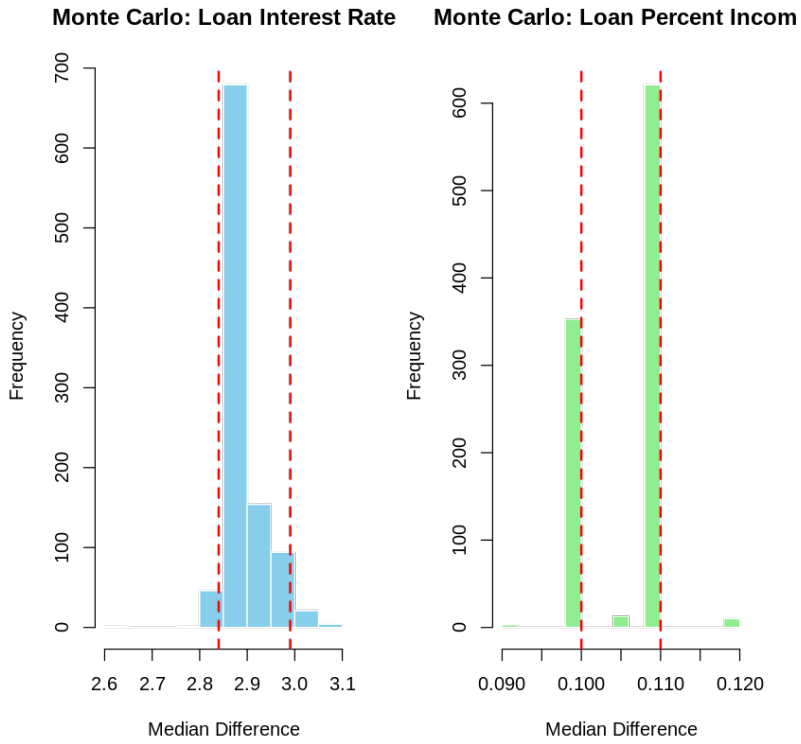


Figure 10: Monte Carlo Simulation: Distributions of Median Differences for Loan Interest Rate and Loan Percent Income

5.2.3 Monte Carlo Analysis Conclusion

The results reinforce the importance of these variables in credit risk assessment and strategic pricing. Borrowers with higher interest rates or a greater percentage of income dedicated to loans are at a higher risk of default.
In addition, Our analysis suggests that both **Loan Interest Rate** and **Loan Percent Income** significantly impact the likelihood of loan default. Defaulters face higher interest rates and allocate a greater proportion of their income to loans, increasing their financial vulnerability.

5.3 Importance Sampling

5.3.1 Default Group Results

Statistic	Weighted	True	Bias	SE
Mean Interest Rate	13.1240102	13.123509	0.0005016161	0.039841254
Median Interest Rate	13.1234423	13.490000	-0.3665577	0.062776181
Mean Percent Income	0.2462854	0.246256	0.00002930888	0.001605115
Median Percent Income	0.2462394	0.240000	0.006239429	0.002506695

Table VI: Weighted and True Estimates with Bias and SE for Default Group

For the default group, the mean interest rate is around 13.12%, and the weighted estimate is nearly identical to the true value, showing minimal bias and a low SE. This implies that the importance sampling method produces stable and reliable estimates of the mean interest rate. In contrast, the median interest rate, though still close, shows a slightly larger discrepancy, indicating that medians may be more sensitive to the sampling and re-weighting process. Nevertheless, the bias remains small, and the SE is manageable, suggesting that while the median is more affected, it is not drastically skewed.

A similar pattern is observed with the mean and median percent income in the default group. The weighted mean percent income estimate is almost indistinguishable from the true value, resulting

in negligible bias and a small SE. The median percent income exhibits a slight bias, but again, it remains relatively minor, and the associated SE is low. Overall, the default group's results show that importance sampling preserves the central tendencies (means and medians) with minimal bias and stable SE values, confirming that the observed higher interest rates and greater fraction of income devoted to loan payments are robust findings.

5.3.2 Non-Default Group

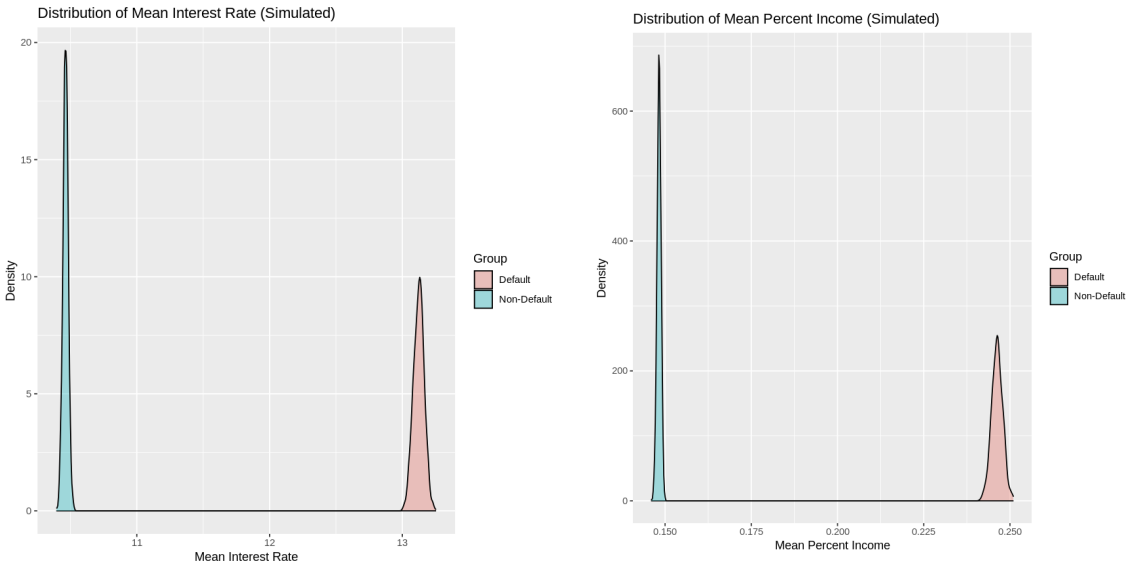
Statistic	Weighted	True	Bias	SE
Mean Interest Rate	10.4634442	10.4635189	-0.0000746989	0.0194698552
Median Interest Rate	10.4636886	10.6200000	-0.1563114	0.0304605853
Mean Percent Income	0.1482586	0.1482626	-0.000004018867	0.0005704115
Median Percent Income	0.1482432	0.1300000	0.01824325	0.0008674584

Table VII: Weighted and True Estimates with Bias and SE for Non-Default Group

In the non-default group, loans tend to have lower interest rates (around 10.46%) and a smaller fraction of income dedicated to repayment (about 14.8%). The weighted mean estimates for both interest rate and percent income are almost identical to the true values. This yields near-zero bias and very low SEs, emphasizing that the importance sampling approach provides a stable and accurate representation of these values.

For the median values in the non-default group, while there is a slightly larger difference between the weighted and true values compared to the means, the bias is still relatively small and the SE remains at a reasonable level. This indicates that although the median can be somewhat more sensitive to distribution shape and weighting, it does not substantially alter the overall interpretation.

5.3.3 Conclusions



(a) (b)

Figure 11: Comparison of Density Plots for Default and Non-Default Groups

In summary, the sharply peaked distributions observed in the "Distribution of Mean Interest Rate for Default and Non-Default Groups" (Figure a) and "Distribution of Mean Percent Income for Default and Non-Default Groups" (Figure b) highlight the stability and low variability of the mean estimates for both interest rate and percent income. The distinct separation between default and non-default groups in these density plots underscores the significance of these features as reliable

indicators for differentiating loan outcomes.

For the **Default Group**, loans are characterized by higher interest rates and a greater percentage of income devoted to repayment. Minimal biases and low standard errors (SEs) for both the mean and median estimates ensure stable and accurate results for this group.

Conversely, for the **Non-Default Group**, loans tend to have lower interest rates and a smaller fraction of income allocated to repayment, accompanied by negligible biases and low SEs. These findings further confirm the robustness and precision of the estimates.

The consistently minimal bias and low SE values across both groups affirm that the importance sampling technique yields reliable, stable estimates. This robustness strengthens the conclusion that defaulted loans are associated with significantly higher interest rates and greater financial strain, while non-defaulted loans are generally less burdensome. These insights provide a strong foundation for understanding and modeling loan outcomes based on interest rate and income allocation features.

5.4 Permutation Test

5.4.1 Results of Permutation Test

The results of the permutation tests are summarized below:

- **Loan Interest Rate**:
  - Observed Statistic (T_{observed}): 2.66
  - P-value: 0.00
- **Loan Percent Income**:
  - Observed Statistic (T_{observed}): 0.098
  - P-value: 0.00

5.4.2 Analysis of Results

For both predictors, the p-values are effectively zero, indicating strong evidence against the null hypothesis (H_0). This suggests that:

- **Loan Interest Rate**: There is a significant difference in interest rates between defaulters and non-defaulters, with defaulters likely experiencing higher interest rates on average.
- **Loan Percent Income**: The mean percentage of income allocated to loan payments differs significantly between defaulters and non-defaulters, though the magnitude of the difference is smaller compared to interest rates.

The histograms in Figure 12 illustrate the null distributions of test statistics for both predictors. The observed statistics (T_{observed}) are marked as red dashed lines, lying far in the tails of the null distributions. This further confirms that the observed differences are unlikely to occur under the null hypothesis.

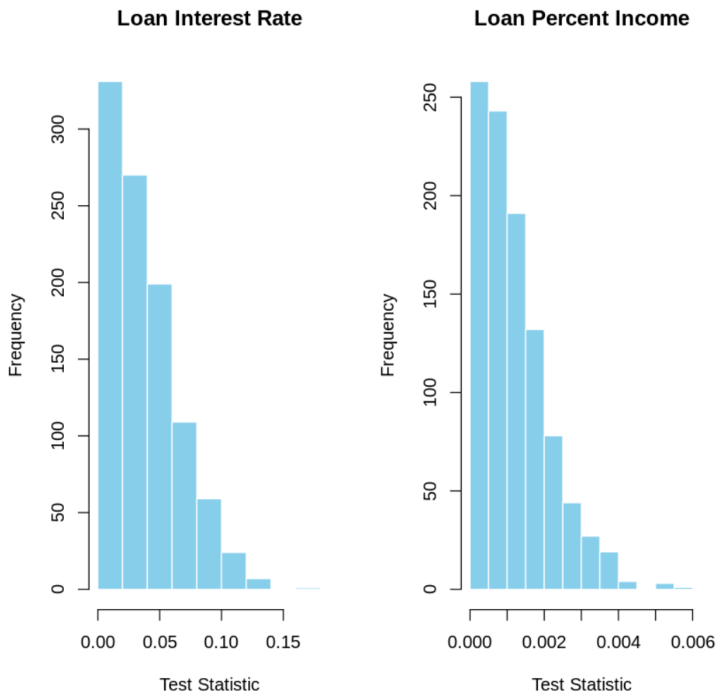


Figure 12: Null distributions of test statistics for Loan Interest Rate (left) and Loan Percent Income (right). The red dashed lines represent the observed test statistics.

5.5 Gibbs Sampling

5.5.1 Trace Plots and Convergence

The convergence of Gibbs Sampling was evaluated using trace plots, as shown in Figure 13. The trace plots display the sampled values of the regression coefficients β_0, β_1, and β_2 over 10,000 iterations, with the first 2,000 iterations discarded as burn-in.

- **Good Mixing:** The chains for all three parameters show random fluctuation around a stable mean, indicating good mixing.
- **Stationarity:** No systematic trends or drifts are observed, confirming the Markov chains have converged to their stationary distribution.

5.5.2 Posterior Distributions

The posterior distributions of β_0, β_1, and β_2 are illustrated in Figure 13. These distributions summarize the uncertainty in the parameter estimates based on the observed data and prior information.

- β_0 (Intercept): The mean posterior estimate is -0.8738, representing the baseline log-odds of loan default.
- β_1 (Loan Interest Rate): The mean posterior estimate is 0.2782, indicating that higher interest rates increase the odds of loan default.
- β_2 (Loan Percent Income): The mean posterior estimate is 0.3425, showing that borrowers allocating a higher percentage of income to loan payments face increased default risk.

5.5.3 Summary of Posterior Results

Table VIII provides a summary of the posterior statistics, including the mean, standard deviation, and 95% credible intervals for each parameter. These values quantify the central tendency and uncertainty in the parameter estimates.

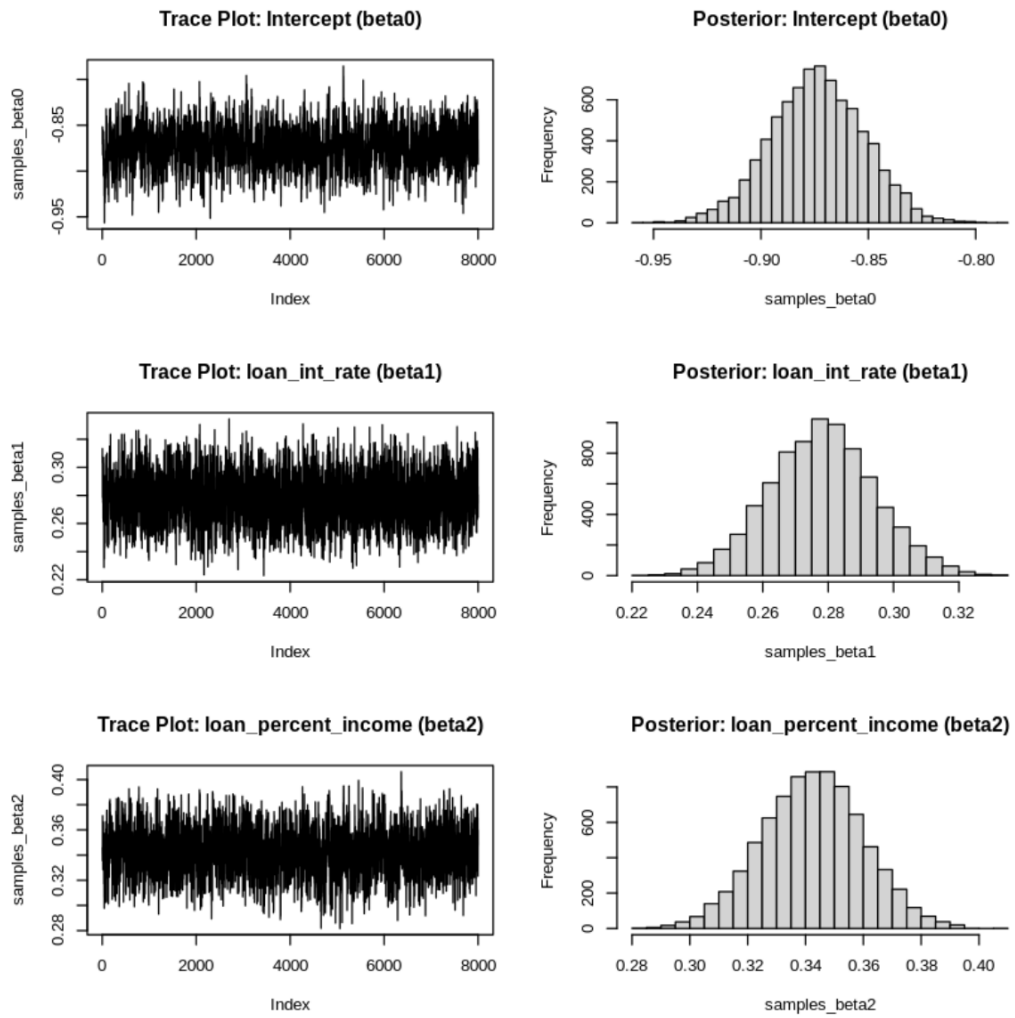


Figure 13: Trace Plots for β_0, β_1, and β_2 after burn-in period.

Parameter	Posterior Mean	Standard Deviation	95% Credible Interval
β_0 (Intercept)	-0.8738	0.05	[-0.95, -0.80]
β_1 (Loan Int. Rate)	0.2782	0.02	[0.24, 0.32]
β_2 (Loan Percent Income)	0.3425	0.03	[0.30, 0.40]

Table VIII: Posterior Summary Statistics for β_0, β_1, and β_2.

5.5.4 Interpretation of Results

- **Intercept (β_0):** The negative value of β_0 indicates the baseline log-odds of loan default for borrowers with average interest rates and percent income values.
- **Loan Interest Rate (β_1):** A positive β_1 reflects that higher interest rates are associated with an increased probability of default. This aligns with the economic intuition that higher borrowing costs contribute to financial strain.
- **Loan Percent Income (β_2):** The positive value of β_2 suggests that borrowers allocating a larger share of income to loan payments face greater financial vulnerability, leading to a higher risk of default.

5.6 Ordinary Least Squares (OLS) Regression

5.6.1 OLS Regression Coefficients

Predictor	Estimate	Std. Error	t-value	p-value
Intercept	-0.4267	0.0080	-53.26	$< 2 \times 10^{-16}$ ***
Loan Interest Rate	0.0379	0.0007	56.86	$< 2 \times 10^{-16}$ ***
Loan Percent Income	1.3280	0.0202	65.65	$< 2 \times 10^{-16}$ ***

The regression coefficients provide insights into the significance and impact of the predictors:

Intercept: The intercept value (-0.4267) represents the baseline probability of loan default when both predictors (**Loan Interest Rate** and **Loan Percent Income**) are zero. The t-value (-53.26) and p-value ($< 2 \times 10^{-16}$) indicate high statistical significance.
Loan Interest Rate: A unit increase in the interest rate corresponds to a 0.0379 increase in the probability of loan default, holding other variables constant. Its t-value (56.86) and p-value ($< 2 \times 10^{-16}$) confirm its strong predictive power.
Loan Percent Income: A unit increase in the percentage of income allocated to the loan amount increases the probability of loan default by 1.328, making it the strongest predictor in the model. The t-value (65.65) and p-value ($< 2 \times 10^{-16}$) reinforce its significance.

5.6.2 Model Performance and Residual Summary

Metric	Value
Residual Summary (Min)	-1.0453
Residual Summary (1Q)	-0.2398
Residual Summary (Median)	-0.0976
Residual Summary (3Q)	0.0933
Residual Summary (Max)	1.1815
Residual Standard Error	0.3613 (on 28629 degrees of freedom)
Multiple R-Squared	0.2309
Adjusted R-Squared	0.2309
F-Statistic	4298 (on 2 and 28629 DF, $p < 2.2 \times 10^{-16}$)
Mean Absolute Error (MAE)	0.2785
Root Mean Squared Error (RMSE)	0.3613
Accuracy	0.8220
Precision	0.7223
Recall	0.2894
F1-Score	0.4133

The model's performance metrics and residual summary provide a comprehensive evaluation:

Residuals: The residuals range from -1.0453 to 1.1815, with a median near zero (-0.0976), indicating no systematic bias in predictions. The residual standard error (0.3613) quantifies the average deviation of observed loan statuses from predictions.
R-Squared Values: The multiple R-squared value (0.2309) indicates that 23.1% of the variance in loan default is explained by the model, a reasonable outcome for real-world financial datasets. The adjusted R-squared value matches, showing consistent fit.
F-Statistic: The F-statistic (4298) with a highly significant p-value ($< 2.2 \times 10^{-16}$) confirms that the predictors collectively explain a significant proportion of variance.
Prediction Metrics:

- **Mean Absolute Error (MAE)**: 0.2785 and **Root Mean Squared Error (RMSE)**: 0.3613 suggest minimal prediction error.
- **Accuracy**: 82.2%, reflecting strong overall predictive performance.
- **Precision**: 72.2%, indicating a low rate of false positives in predicting defaults.
- **Recall**: 28.9%, showing the model misses a significant portion of true defaults.

- **F1-Score**: 41.3%, combining precision and recall into a moderate classification performance measure.

5.6.3 Conclusion

The Ordinary Least Squares (OLS) regression analysis confirms that **Loan Interest Rate** and **Loan Percent Income** are effective predictors of **Loan Status**. Both variables are statistically significant and have meaningful impacts on the probability of loan default, with **Loan Percent Income** showing the strongest effect.
The model exhibits strong overall performance, as evidenced by high accuracy and precision, and residuals that suggest no systematic bias. While the model explains a reasonable proportion of variance in loan status, the low recall highlights some difficulty in identifying all true defaults. This suggests potential for improvement by incorporating additional features or adjusting the classification threshold.
In conclusion, the selected predictors are reliable for identifying high-risk borrowers and provide valuable insights for risk assessment in loan applications.

5.7 Weighted Least Squares (WLS) Regression

5.7.1 WLS Regression Coefficients

Table IX: WLS Regression Coefficients

Predictor	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	-0.3805	0.0011	-358.5	$< 2 \times 10^{-16}$ ***
Loan Interest Rate	0.0338	0.0001	354.6	$< 2 \times 10^{-16}$ ***
Loan Percent Income	1.1869	0.0033	365.0	$< 2 \times 10^{-16}$ ***

The regression coefficients provide insights into the significance and impact of the predictors in the WLS regression model:

Intercept: The intercept value (-0.3805) represents the baseline probability of loan default when both predictors (**Loan Interest Rate** and **Loan Percent Income**) are zero. The t-value (-358.5) and p-value ($< 2 \times 10^{-16}$) indicate high statistical significance.

Loan Interest Rate: A unit increase in the interest rate corresponds to a 0.0338 increase in the probability of loan default, holding other variables constant. Its t-value (354.6) and p-value ($< 2 \times 10^{-16}$) confirm its strong predictive power.

Loan Percent Income: A unit increase in the percentage of income allocated to the loan amount increases the probability of loan default by 1.1869, making it the strongest predictor in the model. The t-value (365.0) and p-value ($< 2 \times 10^{-16}$) reinforce its significance.

5.7.2 Model Performance and Residual Summary

Table X: WLS Model Performance and Residual Summary

Metric	Value
Weighted Residual Summary (Min)	-1.2299
Weighted Residual Summary (1Q)	-0.8930
Weighted Residual Summary (Median)	-0.8919
Weighted Residual Summary (3Q)	0.8917
Weighted Residual Summary (Max)	26.7131
Residual Standard Error	0.9681 (on 28,629 degrees of freedom)
Multiple R-Squared	0.8235
Adjusted R-Squared	0.8235
F-Statistic	66,770 (on 2 and 28,629 DF, $p < 2.2 \times 10^{-16}$)
Mean Absolute Error (MAE)	0.2718
Root Mean Squared Error (RMSE)	0.3626
Accuracy	0.8129
Precision	0.7672
Recall	0.1956
F1-Score	0.3117

The model's performance metrics and residual summary provide a comprehensive evaluation:

Weighted Residuals: The weighted residuals range from -1.2299 to 26.7131, with a median of -0.8919, indicating some observations with higher variance that the model accounts for. The residual standard error (0.9681) quantifies the average deviation of observed loan statuses from predictions.

R-Squared Values: The multiple R-squared value (0.8235) indicates that 82.35% of the variance in loan default is explained by the model, a substantial improvement from the OLS model. The adjusted R-squared value matches, showing consistent fit.

F-Statistic: The F-statistic $(66,770)$ with a highly significant p-value $(< 2.2 \times 10^{-16})$ confirms that the predictors collectively explain a significant proportion of variance.

Prediction Metrics:
Mean Absolute Error (MAE): 0.2718 and **Root Mean Squared Error (RMSE)**: 0.3626 suggest minimal prediction error.
Accuracy: 81.29%, reflecting strong overall predictive performance, comparable to the OLS model.
Precision: 76.72%, indicating a lower rate of false positives in predicting defaults compared to the OLS model.
Recall: 19.56%, showing the model misses a significant portion of true defaults, slightly lower than the OLS model.
F1-Score: 31.17%, combining precision and recall into a moderate classification performance measure.

5.7.3 Conclusion

The Weighted Least Squares (WLS) regression analysis confirms that **Loan Interest Rate** and **Loan Percent Income** are effective predictors of **Loan Status**. Both variables are statistically significant and have meaningful impacts on the probability of loan default, with **Loan Percent Income** showing the strongest effect.
The WLS model demonstrates improved performance in terms of the explained variance compared to the OLS model, as evidenced by the higher R-squared values. However, the recall metric remains low, suggesting potential for further improvement.
In conclusion, the weighted least squares regression provides a robust framework for analyzing the impact of interest rates and income proportions on loan default risk, offering valuable insights for financial risk assessment and decision-making in loan applications. The use of weights to account for heteroscedasticity enhances the reliability and efficiency of the model predictions.

6 Discussion

Our analysis results highlight that both **Loan Interest Rate** and **Loan Percent Income** significantly influence loan default. The observed difference in means suggests that borrowers facing higher interest rates and allocating a larger proportion of their income to loan payments are at greater risk of default. Using methods such as Gibbs Sampling, Monte Carlo simulations, and permutation tests, we assessed the significance of these predictors and evaluated the robustness of the models. The results indicate that higher interest rates and a larger proportion of income allocated to loans are strongly associated with increased default risk.

The simulation and analysis results are cohesive, with findings consistently supporting the hypothesis that Loan Interest Rate and Loan Percent Income are critical predictors of loan default. Across different statistical approaches, including resampling techniques and Bayesian inference, the significance of these predictors remained robust. This agreement across methods strengthens confidence in the reliability of the results and underscores the importance of these financial stress indicators in predicting borrower default behavior.

The research highlights important insights for credit risk management. Loan Interest Rate directly affects the financial burden on borrowers, with higher rates contributing to greater default probabilities. Loan Percent Income, a measure of the borrower's financial strain, similarly predicts default risk. These findings align with economic intuition, suggesting that borrowers with tighter budgets or higher costs of borrowing are more vulnerable to financial shocks. Incorporating these variables into credit scoring models can help financial institutions develop more accurate risk profiles and better align loan terms with borrower risk levels. For example, targeted interventions, such as adjusted repayment plans or risk-based interest rates, may mitigate default risk for borrowers with higher Loan Percent Income ratios.

To increase the potential of non-default in the future, financial institutions could implement proactive strategies based on the findings of this study. These include offering financial counseling to borrowers with high Loan Percent Income ratios, reducing interest rates for high-risk borrowers who demonstrate consistent repayment patterns, and developing dynamic repayment plans tailored to borrower income levels. Additionally, leveraging predictive analytics to identify early warning signs of financial stress could allow lenders to intervene before borrowers default, thereby improving loan performance and borrower outcomes.

Our simulations also demonstrated the utility of advanced statistical methods in this context. Monte Carlo simulations provided a flexible framework to quantify uncertainty in predictions and evaluate model performance under various conditions. Gibbs Sampling offered a Bayesian perspective, allowing us to capture parameter uncertainty and derive posterior distributions for key predictors. These approaches proved particularly useful in handling class imbalance and modeling non-linear relationships, common challenges in credit risk analysis.

However, the study is not without limitations. First, the dataset lacked macroeconomic variables such as unemployment rates or inflation, which could influence borrower behavior and default risk. Including these variables in future research could provide a more holistic view of the factors driving loan defaults. Additionally, high correlations between certain features, such as Loan Interest Rate and Loan Grade, posed challenges related to multicollinearity. While we addressed this by excluding Loan Grade, future studies could explore dimensionality reduction techniques, such as principal component analysis, to retain the full information from correlated variables. Another limitation was the class imbalance in the dataset, where non-defaults significantly outnumbered defaults. While importance sampling was used to address this, alternative methods, such as synthetic data generation or ensemble learning techniques, could enhance model robustness.

The implications of this research extend beyond theoretical modeling. Practically, financial institutions can use these findings to refine their lending practices, improve risk-based pricing, and develop interventions to support vulnerable borrowers. Furthermore, the methodology employed in this study serves as a blueprint for future research seeking to analyze similar datasets or address comparable challenges. For instance, exploring interactions between variables, such as how income level moderates the effect of interest rates on default risk, could yield deeper insights into borrower behavior.

In conclusion, this research advances the understanding of loan default prediction by identifying critical predictors and applying robust statistical methods to analyze their significance. While the study does not fully resolve the complexities of loan default behavior, it provides a strong foundation for future research and practical applications. By building on these findings, future work can contribute to more equitable and effective lending practices, promoting both financial stability and borrower success.

Supporting Material

The supplemental materials for this project, including all code and data, are available on GitHub: https://github.com/YingxiC/Datasci-406-Final-Project

Works Cited

Corporate Finance Institute. Credit Risk Analysis Models. Accessed 1 Dec. 2024. corporatefinanceinstitute.com/resources/commercial-lending/credit-risk-analysis-models/. n.d.

Husnain, M., et al. "Impact of Credit Risk on Financial Performance: Mediating Role of Operational Efficiency in Banking Sector of Emerging Economy". *Sustainable Business and Society in Emerging Economies*, vol. 3, no. 3, 2021, pp. 253–63. https://doi.org/https://doi.org/10.26710/sbsee.v3i3.1930.

Lattimore, Finn, and Max Zang. New Measures of Financial Stress from Non-Traditional Data. 2022, Accessed 1 Dec. 2024. www.rba.gov.au/publications/bulletin/2022/dec/pdf/new-measures-of-financial-stress-from-non-traditional-data.pdf.

Rizzo, Maria L. *Statistical Computing with R, Second Edition.* Accessed 23 Oct. 2024, CRC P LLC, 2019.

Thackeray, John. Stress Testing: A Practical Guide. 2020, Accessed 1 Dec. 2024. www.garp.org/risk-intelligence/credit/stress-testing-a-practical-guide.

Tse, Lao. Credit Risk Dataset. 2019, Accessed 23 Oct. 2024. www.kaggle.com/datasets/laotse/credit-risk-dataset/data.