October 2024

Optimization of Bank Loan Approvals

Enhancing Lending Practices Through Data Analysis and Predictive Modeling

Joao Pedro Jacomossi

Table of Contents

Table of Figures

Table of Tables

Executive Summary

This project was undertaken to address a bank's high loan default rate of 20% by implementing a data-driven approach to optimize the home improvement loan approval process. Leveraging predictive modeling, the initiative aimed to reduce defaults, mitigate credit risks, and ensure compliance with the Equal Credit Opportunity Act (ECOA) while supporting profitability and sustainable growth.

The dataset consisted of 5,960 records and 13 variables. After addressing missing data and inconsistencies 5,572 records were retained for analysis. Significant predictors included debt-to-income ratio, delinquency history, derogatory records, credit line age, number of credit lines and loan amount. A credit scoring system inspired by the FICO methodology was also developed to categorize applicants into risk groups such as "Poor," "Fair," "Good," and "Excellent.", providing additional insights into borrower behavior and lending patterns. Analysis of this scoring system revealed that default rates decreased significantly as credit scores improved, dropping below 5% for scores above 732. Borrowers in the "Excellent" category exhibited the lowest default rates (4.4%) and the highest average loan amounts (over $21,000), while those in the "Poor" category had default rates exceeding 50% and received lower average loan amounts. These insights helped inform lending strategies and risk segmentation.

After evaluating multiple machine learning algorithms, the Random Forest model demonstrated the best performance, achieving a test set accuracy of 92.11%, sensitivity of 91.51%, and precision of 73.48%.

The model achieved a significant reduction in the estimated default rate, decreasing it from 20% to 2.11%. Comparisons between historical and model-driven scenarios highlighted reductions in defaulted loan amounts from $18,860,200 to $491,100, definitive losses by $1,321,568 (98%), and recoverable losses by $17,047,532 (97%). The total loan issuance adjusted accordingly, reflecting more prudent lending decisions. These results underscore the model's potential to enhance credit risk management and improve portfolio quality.

By integrating this model into its approval process, alongside insights from the credit scoring system, the bank can strengthen its risk management capabilities, minimize financial losses, and offer more competitive lending terms to reliable borrowers. This initiative positions the bank for long-term growth, improved profitability, and enhanced customer trust.

Section 1: Introduction

The Consumer Credit Department of a bank aims to enhance its home improvement loan approval process by automating decision-making through a data-driven model. This initiative seeks to optimize lending decisions while adhering to the principles of the Equal Credit Opportunity Act (ECOA), ensuring fairness and compliance.

The proposed model will be developed using historical data from recent loan applicants, enabling the bank to predict the likelihood of default or financial delinquency. By leveraging predictive analytics, the bank intends to identify high-risk applicants proactively while maintaining transparency in its decision-making process. This ensures that, in cases of loan denial, clear and understandable explanations can be provided.

The dataset comprises information on 5,960 bank loans, of which 1,189 applicants (approximately 20%) defaulted or experienced significant payment difficulties. The data includes 12 variables for each applicant, offering a foundation for building a model capable of identifying potential risks with precision and reliability.

This predictive approach will enable the bank to improve its lending decisions, mitigate financial risks, and responsibly extend credit to those in need, all while maintaining a balance between profitability and social responsibility.

Section 2: Business Problem

A business problem represents a challenge or obstacle that prevents an organization from achieving its objectives or operating efficiently.

Problem Statement:

- The bank is experiencing a **high loan default rate of nearly 20%**, which significantly exceeds the typical default rates of 1–5% observed in the industry.
- This elevated default rate adversely impacts the bank's profitability and reputation, posing a serious challenge to its long-term success.

Section 3: Business Goal

A business goal is defined as an accomplishment or target an organization aims to achieve to drive its overall success and growth.

The goal of this project is to optimize the bank's loan approval process to address the high default rate. By improving risk assessment and ensuring transparency in decision-making, the bank aims to reduce loan defaults and mitigate credit risk.

This optimization will enhance the bank's profitability, minimize financial losses, and strengthen its reputation, supporting sustainable growth and competitive advantage in the industry.

Section 4: Analytics Goal

The analytics goal refers to the outcome expected to be achieved using analytics. This is the guiding purpose behind the analytical efforts and ensures the project remains focused on supporting the business goal.

Analytics Goals:

1. **Exploratory Data Analysis (EDA):**
    - Gain insights into the data, identify patterns, and understand the key factors influencing loan defaults and customer behavior.
2. **Classification Model Development:**
    - Build a predictive model to classify the likelihood of customer default.
    - Support the bank in assessing credit risk and improving decision-making for loan approvals.

Section 5: Data Preprocessing

The aim of data preprocessing is to prepare the dataset for accurate and reliable analysis or modeling by ensuring it is clean, consistent, and well-structured. Key steps include data cleaning, transformation, feature engineering, and defining attributes.

Data Cleaning:

- Verifying that data types, formats, and variable names are correct throughout the dataset.
- Detecting and handling outliers to avoid distorting the results.

Data Transformation:

- Adjusting variables to the correct data types for analysis and modeling purposes.

Feature Engineering:

- Creating new variables from existing data to capture meaningful patterns or offer additional insights for analysis.

Attributes Definition:

- Providing clear definitions for each variable, including its meaning, type, unit of measurement, and expected value range.

5.1 Dataset

```
'data.frame':   5960 obs. of  13 variables:
$ DEFAULT: Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 2 2 ...
$ LOAN   : int  1100 1300 1500 1500 1700 1700 1800 1800 2000 2000 ...
$ MORTDUE: num  25860 70053 13500 NA 97800 ...
$ VALUE  : num  39025 68400 16700 NA 112000 ...
$ REASON : Factor w/ 3 levels "","DebtCon","HomeImp": 3 3 3 1 3 3 3 3 3 3 ...
$ JOB    : Factor w/ 7 levels "","Mgr","Office",..: 4 4 4 1 3 4 4 4 4 6 ...
$ YOJ    : num  10.5 7 4 NA 3 9 5 11 3 16 ...
$ DEROG  : int  0 0 0 NA 0 0 3 0 0 0 ...
$ DELINQ : int  0 2 0 NA 0 0 2 0 2 0 ...
$ CLAGE  : num  94.4 121.8 149.5 NA 93.3 ...
$ NINQ   : int  1 0 1 NA 0 1 1 0 1 0 ...
$ CLNO   : int  9 14 10 NA 14 8 17 8 12 13 ...
$ DEBTINC: num  NA NA NA NA NA ...
```

Figure 5.1: Original Dataset Structure

The original dataset contains 5960 observations and 13 variables. For a better interpretation the variable “BAD”, was renamed “DEFAULT”, and blank values were replaced with NA values. Additionally, the categorical and binary variables "DEFAULT," "REASON," and "JOB" were converted into factor variables. Converting categorical and binary variables into factors is essential because it allows statistical models and algorithms to properly interpret and handle these variables.

5.2 Variable Definition

Variable Name	Definition	Type
DEFAULT	Whether or not an applicant defaulted on their loan. (Defaulted =1 or Not Defaulted = 0)	Categorical/Binary
LOAN	The amount of the loan requested by the applicant.	Integer
MORTDUE	The amount due on an existing mortgage.	Numeric
VALUE	The value of the applicant's current property.	Numeric
REASON	The reason for the loan: either DebtCon (debt consolidation) or HomeImp (home improvement).	Categorical/Binary
JOB	The occupational category of the applicant.	Categorical
YOJ	Years at present job.	Numeric
DEROG	The number of major derogatory reports (e.g., bankruptcies, liens).	Integer
DELINQ	The number of delinquent credit lines the applicant currently has.	Integer
CLAGE	The age of the oldest credit line in months.	Numeric
NINQ	The number of recent credit inquiries.	Integer
CLNO	The total number of credit lines the applicant has.	Integer
DEBTINC	The debt-to-income ratio, which measures the applicant's monthly debt obligations relative to their monthly income.	Numeric
EQUITY	The amount of equity the applicant has in their property (calculated as property value minus mortgage due).	Numeric

UNSECURED LOAN	Portion of a loan not covered by the collateral's equity.	Numeric
DEFINITIVE LOSS	Actual financial loss incurred by the lender for defaulted applicants.	Numeric
LTV RATIO	Loan-to-Value ratio, calculated as loan amount divided by the property value.	Numeric
LTE RATIO	Loan-to-Equity ratio, calculated as loan amount divided by the equity in the property (measures how much of the loan is covered by the applicant's equity).	Numeric
OWNERSHIP	The percentage of the property the applicant owns (similar to equity but expressed as a percentage).	Numeric
CREDIT_SCORE	A score reflecting the applicant's creditworthiness	Numeric
SCORE_CATEGORY	The category in which the applicant's credit score falls (Excellent, Good, Fair, or Poor) based on score ranges.	Categorical

Table 5.1: Variable Definition

5.3 Missing Values

Handling missing values is a critical step that can significantly impact the quality and reliability of the results. Missing data can arise from different sources such as human error, system failures, or incomplete data collection. These gaps, if not handled appropriately, can introduce bias, reduce the statistical power of the analysis, or lead to inaccurate conclusions. This subsection will explore the nature and patterns of the missing and its potential impact on the analysis. Lastly, it will discuss the method employed to handle these missing values effectively.

5.3.1 Exploration of Missing Values

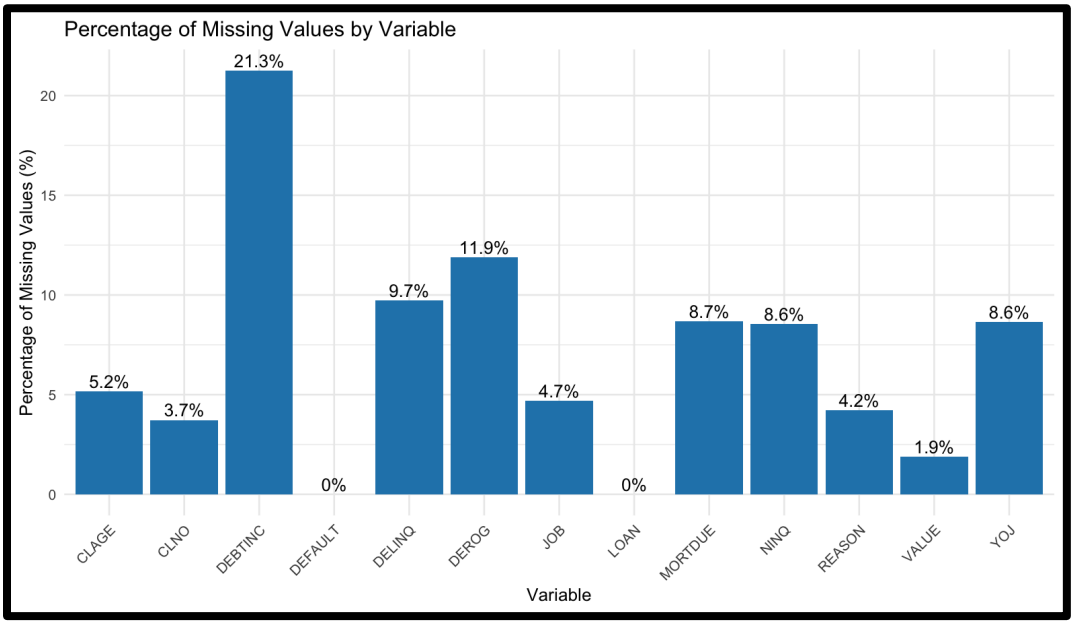

Figure 5.2: Bar Plot of Percentage of Missing Values by Variable

Figure 5.2 shows the percentages of missing values across all variables in the dataset. It is possible to observe that 11 out of 13 variables contain missing values and most of them account for more than 5% of the number of observations. Therefore, it requires careful consideration because selecting an inappropriate method can compromise the validity of the entire analysis.

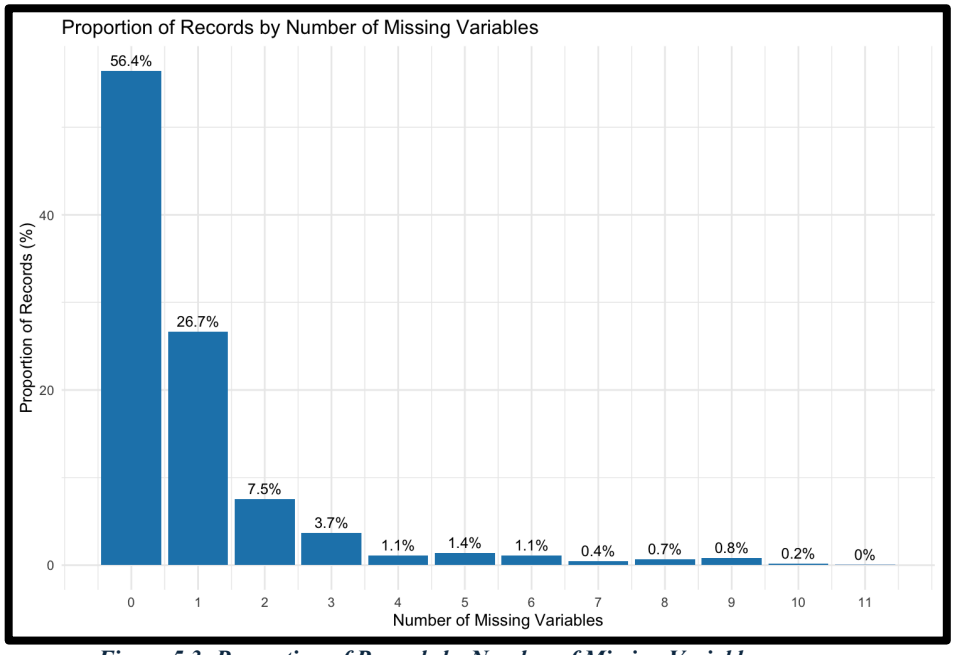

Figure 5.3: Proportion of Records by Number of Missing Variables

Figure 5.3 shows that 56.4% of the records are complete cases, meaning they contain values for all 13 variables. Using only these complete cases would reduce the dataset to 3,364 observations, including 300 defaults and 3,064 non-defaults. However, due to the significant loss of data, this approach is not ideal.

5.3.2 Removing Rows with Excessive Missing Data

Given that the dataset contains 13 variables, rows with more than four missing values were removed to ensure data quality and maintain the integrity of the analysis. This approach preserves a substantial portion of the dataset while minimizing the potential impact of missing data on the results.

```
> summary(loan_subset)
 DEFAULT       LOAN          MORTDUE          VALUE            REASON        JOB           YOJ             DEROG            DELINQ
 0:4496   Min.   : 1100   Min.   :  2063   Min.   :  8000          :   0   Other  :2267   Min.   : 0.000   Min.   : 0.0000   Min.   : 0.0000
 1:1125   1st Qu.:11300   1st Qu.: 46385   1st Qu.: 66922   DebtCon:3801   ProfExe:1251   1st Qu.: 3.000   1st Qu.: 0.0000   1st Qu.: 0.0000
          Median :16500   Median : 65000   Median : 90008   HomeImp:1682   Office : 938   Median : 7.000   Median : 0.0000   Median : 0.0000
          Mean   :18846   Mean   : 73977   Mean   :103025   NA's   : 138   Mgr    : 740   Mean   : 9.004   Mean   : 0.2393   Mean   : 0.4487
          3rd Qu.:23500   3rd Qu.: 91989   3rd Qu.:120724                  Self   : 189   3rd Qu.:13.000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
          Max.   :89900   Max.   :399550   Max.   :855909                  (Other): 109   Max.   :41.000   Max.   :10.0000   Max.   :15.0000
                          NA's   :343      NA's   :84                      NA's   : 127   NA's   :327      NA's   :415       NA's   :265
     CLAGE            NINQ            CLNO          DEBTINC
 Min.   :   0.0   Min.   : 0.000   Min.   : 0.00   Min.   :  0.5245
 1st Qu.: 115.6   1st Qu.: 0.000   1st Qu.:15.00   1st Qu.: 29.4313
 Median : 173.6   Median : 1.000   Median :20.00   Median : 35.0209
 Mean   : 179.8   Mean   : 1.188   Mean   :21.45   Mean   : 34.0705
 3rd Qu.: 230.7   3rd Qu.: 2.000   3rd Qu.:26.00   3rd Qu.: 39.1434
 Max.   :1168.2   Max.   :17.000   Max.   :71.00   Max.   :203.3121
 NA's   :72       NA's   :199                      NA's   :1174
```

Figure 5.4: Summary Statistics of New Dataset Containing Rows with a Maximum of 3 Missing Values per Row

The summary in ***Figure 5.4*** shows that the updated dataset includes 4,496 non-default observations and 1,125 default observations, totaling 5,621 records. This represents a loss of 339 rows (approximately 5.7% of the original dataset), which is relatively small given that these rows had substantial missing data.

It was observed earlier that removing all rows with missing values would leave only 8.91% default observations (DEFAULT = 1). In contrast, with this last approach, the percentage of defaults is 20.01%, which is much closer to the original dataset's percentage of 19.95%. This indicates that this method has successfully preserved a balanced ratio of non-default and default observations compared to the original dataset.

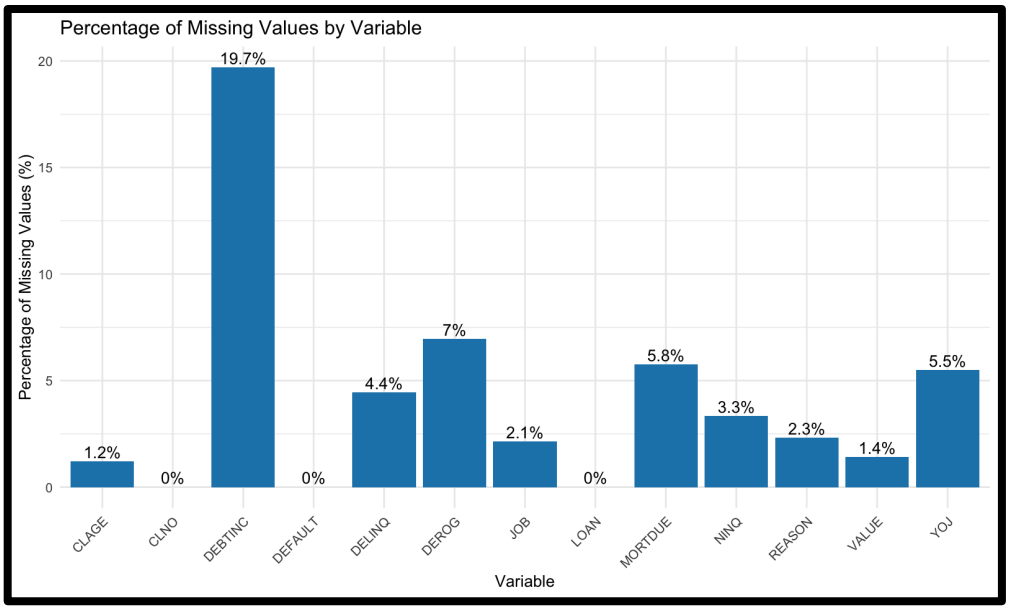

Figure 5.5: Percentage of Missing Values by Variable on After Variable Deletion

Figure 5.5 shows that the number of variables with missing values has reduced to 10, as the **CLNO** variable now has no missing data. Additionally, the percentage of missing values for several variables, including **CLAGE**, **DELINQ**, **JOB**, **NINQ**, and **REASON**, has decreased by more than half. It is also evident that all remaining variables now have no more than 7% missing values, except for **DEBTINC**, which continues to have a significantly higher percentage of missing data.

5.3.3 Diagnosing Missing Data Mechanisms

Before choosing a method to address the remaining missing data, it is important to understand the nature of the missingness itself. There are three main types of missing data:

1. **MCAR (Missing Completely at Random):** Missing values occur independently of any observed or unobserved variables.

2. **MAR (Missing at Random):** The likelihood of missing data depends on observed data but not on the unobserved data itself. In this scenario, the missingness can be explained by other variables in the dataset, allowing for more sophisticated approaches to account for it.

3. **MNAR (Missing Not at Random):** Missingness is related to the unobserved data. In other words, the reason for the missingness is tied to the actual missing values.

This classification is important because it helps in choosing the appropriate approaches to handle incomplete data and maintain the reliability and integrity of the analysis. Therefore, to classify the missing data accordingly, exploration and diagnostic tests will be performed.

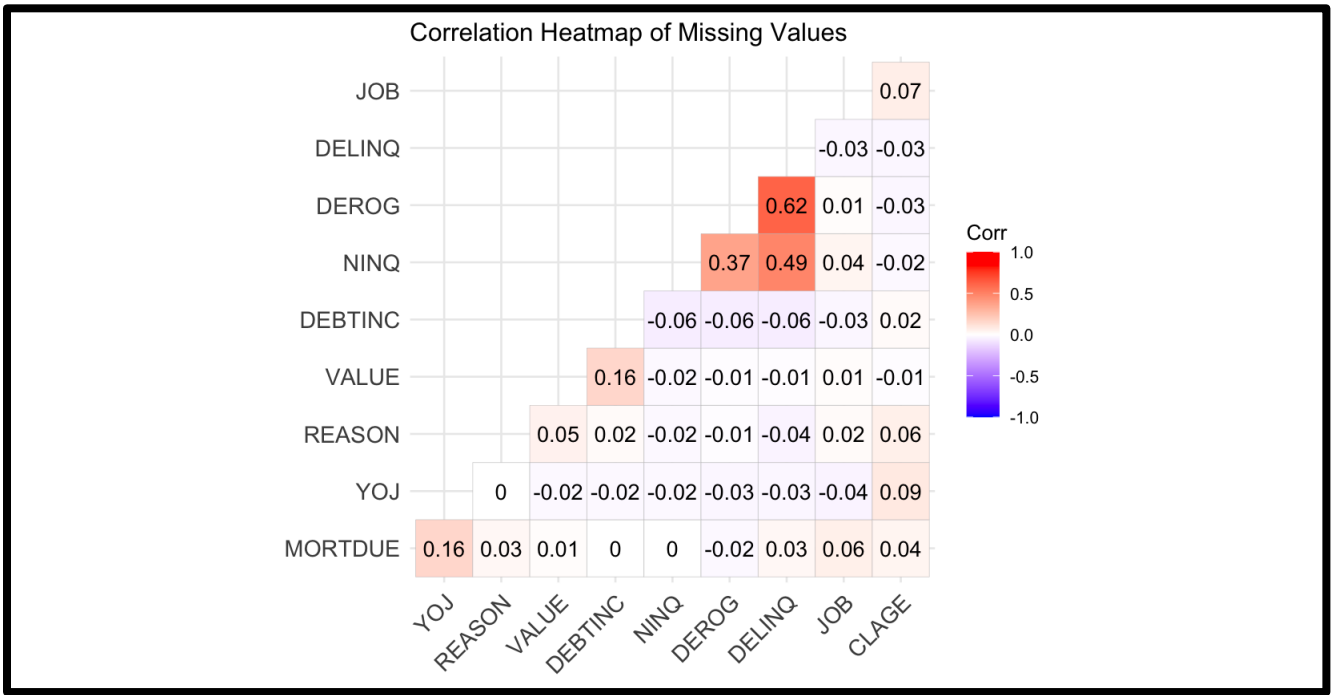

Figure 5.6: Correlation Heatmap of Missing Values

Figure 5.6 illustrates a correlation heatmap of the missing values in the dataset. The goal of this heatmap is to identify missing data patterns. It can reveal if missing values in one variable are correlated with missing values in another variable, suggesting potential dependencies.

The strongest correlations are:

- **DEROG and DELINQ (0.62):** It is possible to observe that there is a moderately high correlation for missing values between these variables. It is possible to observe that both are indicators of borrower's negative credit behavior.
- **DELINQ and NINQ (0.49):** It is possible to observe a moderate correlation for missing values between the number of deliquencies and the number of recent inquiries.
- **DEROG and NINQ (0.37):** Number of derogatory reports and number of inquires have a moderately weak correlation for missing values.

Overall, these correlations suggest that the variables DEROG, DELINQ, and NINQ tend to be missing together, potentially indicating common factors or processes affecting data collection for these variables.

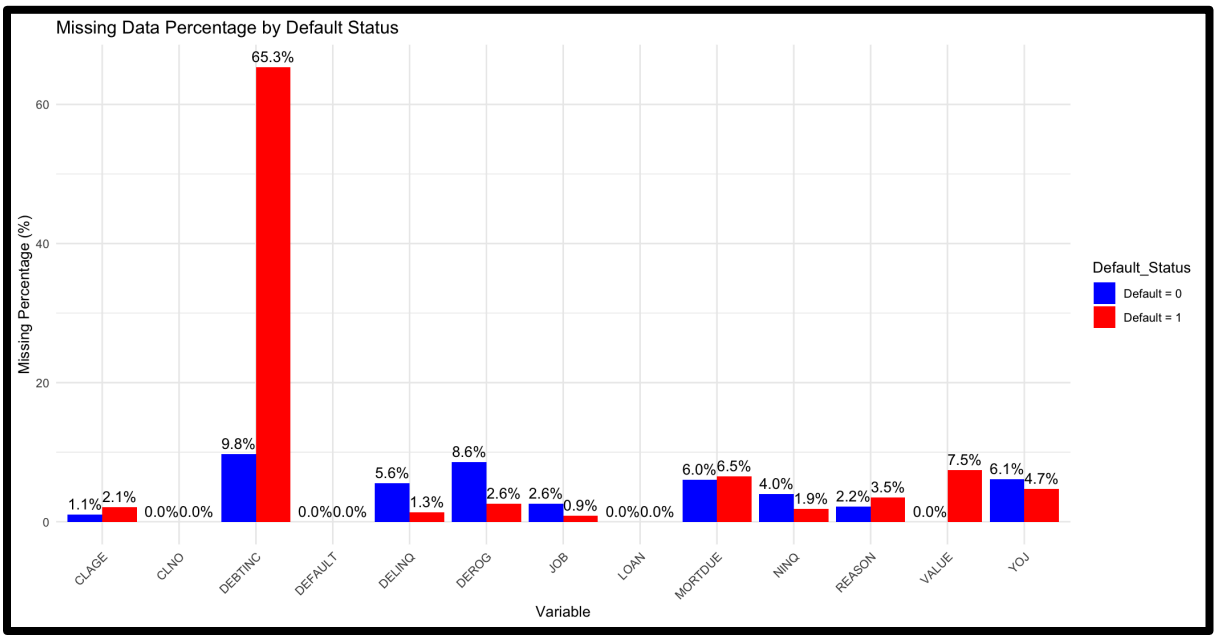

Figure 5.7: Percentage of Missing Values by Bad Status

Figure 5.7 shows the percentage of missing values for each variable in the dataset, grouped by whether the loan is classified as **BAD = 1**(bad loan) or **BAD = 0** (good loan). It is possible to observe that 65.3% of individuals who defaulted on their loans have missing values for DEBTINC, while 9.8% of the individuals who did not default have missing values for this variable.

MCAR TEST:

```
> print(mcar_result)
# A tibble: 1 × 4
  statistic    df p.value missing.patterns
      <dbl> <dbl>   <dbl>            <int>
1     6030.   684       0               65
```

Figure 5.8: Little's MCAR Test Results

To determine whether the missing data in the dataset is Missing Completely at Random (MCAR), Little's MCAR test was performed. This statistical test evaluates whether the pattern of missingness is entirely random, independent of both observed and unobserved data.

The test result shows a p-value of 0, which is less than the typical significance level (e.g., 0.05). This strongly suggests that the missing data is not completely random. Instead, the missingness is likely related to observed or unobserved variables in the dataset. This indicates that the data may fall under one of the other missingness mechanisms: Missing at Random (MAR) or Missing Not at Random (MNAR).

MAR TEST:

Logistic regression models were constructed to assess whether the missingness of variables in the dataset could be classified as Missing at Random (MAR). The goal was to determine if the missingness in one variable could be predicted by other observed predictors, thereby fitting the MAR assumption.

For each analysis, a logistic regression model was built with a binary target variable representing the missingness of a specific variable. This target variable was created as a new column where a value of 1 indicated that the variable had a missing value, and 0 indicated that it did not. Each logistic regression model included one predictor at a time, testing whether that predictor could explain the missingness of the target variable.

The p-values from these models were extracted and compared across all predictors to identify which variables were statistically significant in explaining the missingness. When significant predictors were identified, it indicated that the missingness of the target variable could be explained by other observed variables in the dataset, suggesting that the missingness follows the MAR pattern.

In summary, if significant relationships were found between other predictors and the missingness of the target variable, the missingness was considered MAR, as it was not completely random but could be explained by the values of other variables in the dataset.

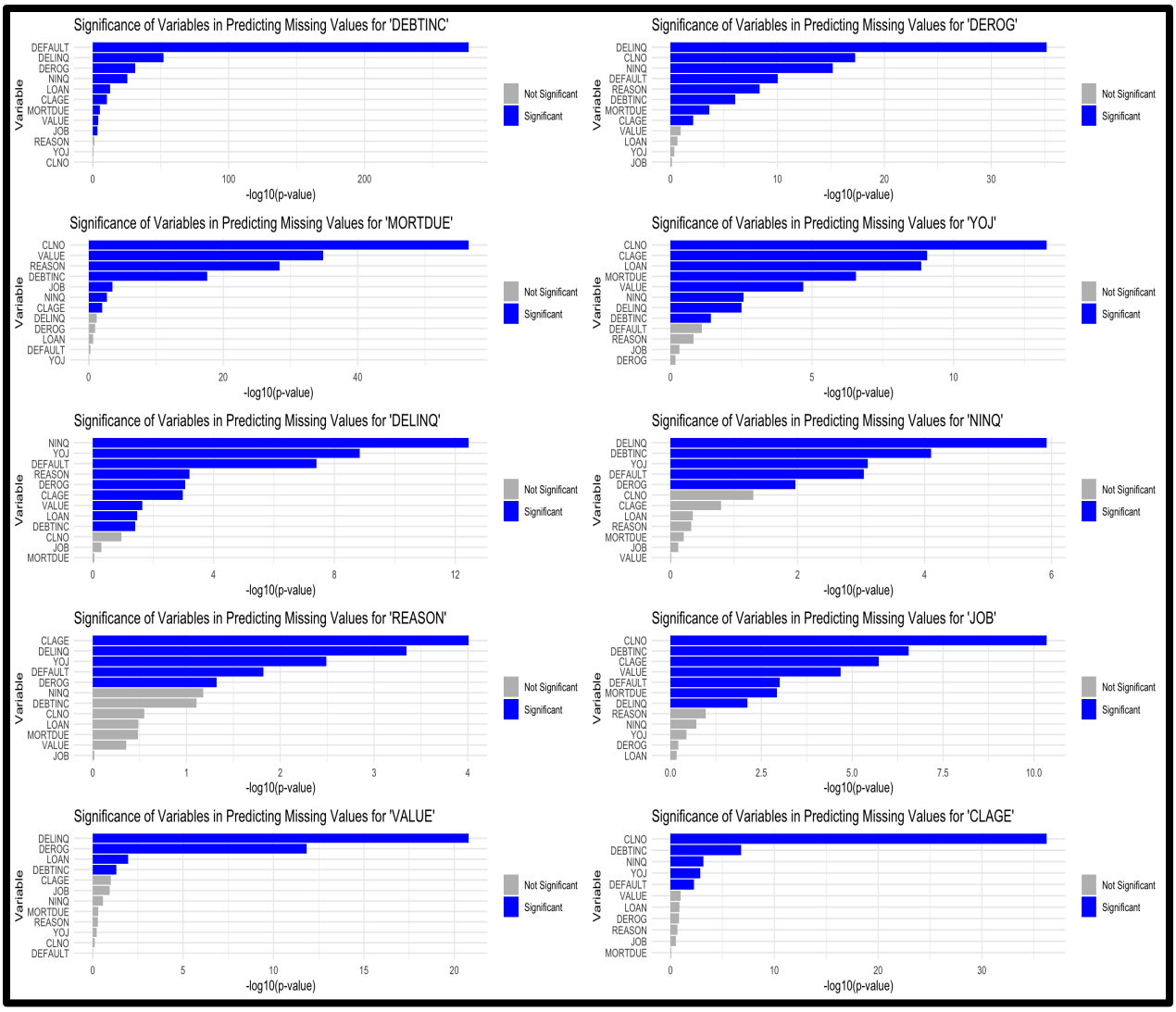

Figure 5.9: Plots of Significance of Variables in Predicting Missing Values

The plots in ***Figure 5.9*** display the p-values from the individual logistic regression models, with each bar representing the p-value of a model transformed using $-\log 10$ for better visualization. The longer the bar, the more significant the predictor is in explaining the missingness of the target variable. For example, in the first plot, values from the "DEFAULT" variable are the most significant in predicting missingness for "DEBTINC." In contrast, variables like "REASON," "YOJ," and "CLNO" are not significant.

In summary, these results suggest that the missingness in each variable can, to some extent, be explained by the values of other variables in the dataset. This supports the assumption that the missingness in this dataset follows the Missing At Random (MAR) mechanism.

The next steps involve discussing and determining the best strategy to handle the missing data based on these findings.

5.3.4 Methods for Handling MAR (Missing at Random)

Deletion Methods:

1. **Listwise Deletion:** This method involves deleting entire rows that contain missing data. This is a straightforward approach, but it can lead to the loss of a large amount of valuable information, especially if missing data is extensive.

2. **Pairwise Deletion:** Instead of deleting all rows with missing values, only the missing values are excluded from the analysis, maximizing data usage. The issue with this approach is that it can result in inconsistent sample sizes, as each analysis may include a different subset of rows depending on which variables are involved. This inconsistency can lead to instability in results and potential bias. Additionally, important patterns and relationships between variables may be lost, weakening the overall integrity of the model.

Imputation Methods:

1. **Simple Single Imputation:** This technique consists of filling the missing values once, with a single estimated value based on summary statistics like mean, mode, or median. It is straightforward and computationally efficient, but it tends to underestimate variability and can distort relationships between variables.

2. **Single Imputation (Predictive Methods):** This technique consists of filling the missing values once, with a single estimated value using predictive modeling techniques like KNN, Random Forest, or Regression. The goal is to predict missingness of a specific variable based on other variable values. They often preserve relationships between variables better than simple single imputation.

3. **Multiple Imputation:** This method handles missing data by filling in missing values multiple times to account for the uncertainty associated with missingness. Multiple imputation generates multiple plausible values for each missing data point. This helps in capturing variability and uncertainty, leading to more robust estimates.

Earlier, listwise deletion was applied to address rows with more than four missing values. However, due to the high proportion of missing data in this dataset, simple single imputation methods are insufficient. Therefore, this analysis will explore more sophisticated imputation techniques, including both single and multiple imputation approaches. The goal is to compare their performance and select the most effective method.

5.3.5 Performance Evaluation of Imputation Methods

To assess the performance of these methods, a complete-case dataset (i.e., all rows with missing values removed) will be used. This dataset will then be subjected to 20% artificially induced missing data, applied at random to the columns originally containing missing values. By comparing the imputed values with the known true values, we can evaluate the effectiveness of each method.

The imputation methods to be tested include Multiple Imputation by Chained Equations (MICE) using both Random Forest and Predictive Mean Matching (PMM), as well as single imputation methods such as Random Forest (MissForest) and K-Nearest Neighbors (KNN) for comparison.

Method: MICE Random Forest

MICE with Random Forest begins by performing an initial imputation using basic methods like the mean or median to temporarily fill in missing values. This ensures that no variable has missing data before applying Random Forest.

For each variable with missing values, Random Forest treats it as the target and uses other variables as predictors. A model is trained on the observed data, and then missing values are predicted based on non-missing values from other variables.

The process is iterative: after imputing one variable, MICE move to the next, repeating this cycle until the imputations stabilize. Random Forest builds multiple decision trees from data samples. Each tree predicts missing values, and the final imputed value is an aggregated prediction, usually the mean. This approach is powerful as it captures complex interactions and works with both categorical and continuous data.

Method: MICE Predictive Mean Matching (PMM)

The process begins with an initial imputation, where missing values are temporarily filled using simple methods like the mean or median. This allows the algorithm to start. Then, MICE iteratively works through each variable with missing values, applying PMM to impute them. PMM first fits a regression model for the variable using other variables as predictors. Instead of using the predicted values directly, PMM finds observed cases with the closest predicted values and uses the actual observed values from these "donors" to replace the missing data.

MICE repeats this process across all variables, refining the imputations in each iteration. The procedure continues until the imputations stabilize, meaning changes between iterations become minimal, indicating convergence or until it reaches a specified number of iterations.

The key advantage of MICE with PMM is that it preserves the dataset’s natural distribution by ensuring that imputed values are actual observed data points.

Method: Miss Forest

MissForest begins by applying simple imputation methods, like the mean for continuous variables or the mode for categorical ones, to temporarily fill missing values. Then, for each variable with missing data, a Random Forest model is built using the other variables as predictors. The missing values are replaced with predictions from the model, which are either averaged for continuous data or based on majority voting for categorical data.

MissForest operates iteratively, improving imputed values with each round. Newly imputed values are used in subsequent iterations until the imputed values stabilize, indicating convergence. This ensures that the imputations become more accurate over time.

Differently from MICE, this method results in a single imputed dataset after convergence. It does not produce multiple imputed datasets that capture variability in the imputation process.

Method: KNN

KNN imputation starts by identifying the K nearest neighbors for each missing value in the dataset. These neighbors are the data points that are closest to the one with the missing value, based on a distance metric like Euclidean distance. The distance is calculated using the non-missing values in the dataset, and the nearest data points are considered the most similar to the one with missing data.

Once the nearest neighbors are identified, the missing value is imputed based on their values. For continuous variables, the imputation is typically the mean or median of the K neighbors. For categorical variables, the missing value is replaced by the most frequent category (mode) among the neighbors. The process is repeated for each missing value, ensuring that all gaps in the dataset are filled.

The main advantage of KNN imputation is its simplicity and flexibility. It can accurately impute missing values without the need for complex modeling and it works for both continuous and categorical data, capturing local patterns effectively.

Performance Evaluation

The selected metric for evaluating the performance of the imputation methods is Root Mean Squared Error (RMSE), a widely used metric for assessing model accuracy. RMSE measures the average magnitude of errors between predicted (or imputed) values and actual values, providing insight into how closely a model's outputs align with real data. Importantly, RMSE expresses errors on the same scale as the original data, making it more interpretable. It is calculated by taking the square root of the mean of the squared differences between predicted and actual values. A lower RMSE indicates that predictions are closer to the actual values, signifying better model performance, whereas a higher RMSE reflects a larger discrepancy.

Because RMSE squares the differences before averaging, larger errors are penalized, making RMSE sensitive to large deviations. This helps ensure that imputation methods maintain data integrity by minimizing significant discrepancies from true values, preserving accuracy in the dataset.

MICE: PMM	
Variable	**Average RMSE**
DEBTINC	10.62
YOJ	10.32
MORTDUE	34,738.04
VALUE	40,358.41
DEROG	0.82
DELINQ	1.11
CLAGE	108.63
NINQ	2.04

Table 5.2: Average RMSE for Variables Using MICE with Predictive Mean Matching (PMM)

MICE: Random Forest	
Variable	**Average RMSE**
DEBTINC	8.92
YOJ	7.88
MORTDUE	28,638.90
VALUE	32,277.25
DEROG	0.71
DELINQ	0.93
CLAGE	82.58
NINQ	1.72

Table 5.3: Average RMSE for Variables Using MICE with Random Forest

MissForest	
Variable	**RMSE**
DEBTINC	5.58
YOJ	4.05
MORTDUE	12,919.87
VALUE	16,785.01
DEROG	0.51
DELINQ	0.65
CLAGE	44.10
NINQ	0.95

Table 5.4: RMSE for Variables Using MissForest Imputation

KNN	
Variable	**RMSE**
DEBTINC	6.85
YOJ	8.17
MORTDUE	40,485.32
VALUE	43,338.08
DEROG	0.64
DELINQ	0.80
CLAGE	76.23
NINQ	1.35

Table 5.5: RMSE for Variables Using K-Nearest Neighbors (KNN) Imputation

By comparing the performance of various imputation methods on continuous variables, it is evident that MissForest delivers the best results. For example, the average error for MORTDUE is 12,919.87 with MissForest, compared to 28,638.90 with Random Forest, 34,738.04 with PMM, and 40,485.32 with KNN.

Overall, PMM performs the worst, with the largest RMSE for most variables. Since MICE generates five different imputed datasets, the RMSE for each was averaged to assess its performance.

While MICE accounts for uncertainty better, given that the primary goal of this project is to predict loan defaults with high accuracy, MissForest, with its superior imputation accuracy, is the more suitable choice.

Final Imputation

Miss Forest imputation was used to impute the dataset with actual missing values.

```
> summary(imputed_loan)
 DEFAULT       LOAN          MORTDUE          VALUE          REASON         JOB           YOJ             DEROG            DELINQ            CLAGE
 0:4496   Min.   : 1100   Min.   :  2063   Min.   :  8000   DebtCon:3900   Mgr    : 748   Min.   : 0.000   Min.   : 0.0000   Min.   : 0.0000   Min.   :   0.0
 1:1125   1st Qu.:11300   1st Qu.: 44095   1st Qu.: 67000   HomeImp:1721   Office : 942   1st Qu.: 3.000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 115.8
          Median :16500   Median : 63005   Median : 90135                  Other  :2362   Median : 7.667   Median : 0.0000   Median : 0.0000   Median : 174.4
          Mean   :18846   Mean   : 72113   Mean   :103077                  ProfExe:1271   Mean   : 9.091   Mean   : 0.2412   Mean   : 0.4698   Mean   : 181.6
          3rd Qu.:23500   3rd Qu.: 90000   3rd Qu.:120724                  Sales  : 109   3rd Qu.:13.000   3rd Qu.: 0.0000   3rd Qu.: 1.0000   3rd Qu.: 232.5
          Max.   :89900   Max.   :399550   Max.   :855909                  Self   : 189   Max.   :41.000   Max.   :10.0000   Max.   :15.0000   Max.   :1168.2
      NINQ            CLNO          DEBTINC
 Min.   : 0.000   Min.   : 0.00   Min.   :  0.5245
 1st Qu.: 0.000   1st Qu.:15.00   1st Qu.: 30.2848
 Median : 1.000   Median :20.00   Median : 35.5498
 Mean   : 1.212   Mean   :21.45   Mean   : 34.6331
 3rd Qu.: 2.000   3rd Qu.:26.00   3rd Qu.: 39.3238
 Max.   :17.000   Max.   :71.00   Max.   :203.3122
```

Figure 5.10: Summary Statistics of Imputed Dataset

Figure 5.10 shows that the MissForest imputation method effectively filled the missing values in the loan dataset while preserving the data's overall structure and distribution. Key variables like MORTDUE, YOJ, DEROG, and DEBTINC, which had significant missing values previously, were successfully imputed.

By comparing this new summary with the summary in ***Figure 5.4***, it is possible to conclude that mean and median values remained consistent with the original dataset, ensuring that the data's integrity was maintained. For example, the DEBTINC variable, which previously had approximately 20% missing values, had a mean of 34.075. After imputation, the mean became 34.5859, demonstrating that the imputed values closely align with the original distribution.

Categorical variables such as REASON and JOB also saw successful imputation, with missing values filled and distributions remaining consistent. Overall, MissForest efficiently handled the missing data, producing a complete dataset that retains the original characteristics, making it suitable for further analysis or modeling.

```
$OOBerror
      NRMSE        PFC
0.12161457 0.04608395
```

Figure 5.11: OOB Error

The **OOB Error** output in MissForest refers to the performance of the imputation model based on **Out-of-Bag (OOB) samples** which provide an estimate of how well the model imputed the missing values. Since the actual values of the missing data are unknown, this is a great way to assess model performance. It serves as an additional evaluation metric alongside the earlier validation step, giving further confidence in the accuracy of the imputations.

It consists of two key metrics:

1. **NRMSE (Normalized Root Mean Squared Error)**: This value measures the imputation error for continuous variables. The NRMSE of 0.1216 means that the imputed values for continuous variables, on average, deviate by about 12.16% of the variance of the original data. A lower NRMSE indicates better imputation quality.

2. **PFC (Proportion of Falsely Classified)**: This metric applies to categorical variables and measures the proportion of misclassifications during the imputation process. The PFC of 0.04608 indicates that about 4.6% of the imputed categorical values were incorrectly classified, which suggests good accuracy in handling missing categorical data.

Together, these metrics indicate that the imputation performed well, with a low error for both continuous and categorical variables.

Section 6: Feature Engineering

The feature engineering section focuses on enhancing the quality and relevance of the data used for modeling. This process involves creating new variables based on existing ones to improve the predictive power of the models and uncover patterns in the data.

This step is critical to ensure that the machine learning algorithms can effectively learn from the data and provide accurate and meaningful predictions.

6.1 Equity:

Equity represents how much of the applicant's home is truly owned. It is calculated by subtracting the outstanding mortgage amount from the current market value of the property. When the property's value exceeds the mortgage owed, the applicant has positive equity.

$$\boldsymbol{Equity = VALUE - MORTDUE}$$

This measure is significant for several reasons. Higher equity generally indicates stronger financial stability for the applicant, as they have built more ownership in their property.

Additionally, in the event of a default, equity determines the amount that a bank with a junior lien (a loan secondary to the main mortgage) could recover. By calculating equity, it becomes possible to estimate how much might be recovered if the property needs to be sold to cover outstanding debts.

6.2 Unsecured Loan Amount

The **unsecured loan** represents the portion of a loan not covered by the collateral's equity. When the property's equity is lower than the loan amount, the unsecured loan amount becomes greater than zero. This variable quantifies the potential financial risk to the lender, highlighting the amount that could be lost if the applicant defaults and the collateral is insufficient to cover the loan.

$$\boldsymbol{Unsecured\ Loan = max(\ 0, LOAN - max(EQUITY, 0))}$$

6.3 Definitive Loss:

For applicants who have defaulted, the **definitive loss** represents the actual financial loss incurred by the lender. This metric is calculated when the borrower's property value or equity is insufficient to cover the loan amount. It quantifies the shortfall that directly impacts the lender due to the default.

$$\boldsymbol{Definitive\ Loss = \max(0, LOAN - \max(EQUITY, 0)) \cdot 1_{\{Default=1\}}}$$

6.4 Loan-to-Equity Ratio:

The **Loan-to-Equity (LTE) Ratio** compares the size of the loan to the applicant's equity. A high ratio means the loan is much larger than the property value, which can indicate higher financial risk for both the applicant and the lender.

$$\boldsymbol{LTE\ RATIO = \frac{LOAN}{EQUITY}}$$

6.5 YOJ-to-Loan Ratio:

This ratio compares how many years an applicant has been at their current job to the amount they are borrowing. Applicants with stable, long-term jobs may be seen as less risky compared to those with shorter employment histories, especially if they are requesting a large loan.

$$YOJ\ to\ LOAN\ RATIO = \frac{YOJ}{LOAN}$$

6.6 Loan-to-Value Ratio:

The **Loan-to-Value (LTV) Ratio** is a measure that compares the size of the loan to the value of the property. It is calculated by dividing the loan amount by the current market value of the property. This ratio is important for both lenders and applicants because it reflects the financial risk associated with the loan. A high LTV ratio means that the loan amount is a large percentage of the property value, which can indicate a higher risk for the lender.

$$LTV\ RATIO = \frac{LOAN}{VALUE}$$

6.7 Ownership Ratio

Ownership Ratio represents the proportion of the property that the borrower truly owns. It is calculated by dividing the applicant's equity by the current market value of the property. This ratio reflects the extent of financial ownership the borrower has in their property, providing insight into their overall financial stability.

$$OWNERSHIP\ RATIO = \frac{EQUITY}{VALUE}$$

This measure is significant because a higher ownership ratio suggests that the borrower has built substantial equity in their home, which reduces the risk of default. Lenders view borrowers with a higher ownership ratio as less risky because they have more financial stake in the property. On the other hand, a low ownership ratio, especially if it's negative (indicating negative equity), signals that the borrower may struggle to cover the loan if the property is sold.

By calculating the ownership ratio, lenders can assess how much of the property value is covered by the borrower's own funds (equity), as opposed to debt, offering a clearer picture of financial risk.

6.8 Credit Score:

To assess the creditworthiness of loan applicants, a custom credit score system was developed, inspired by the FICO score model. This system assigns points to several key factors that reflect the applicant's financial behavior and risk profile. The goal is to ensure transparency and provide a clear understanding of how each factor contributes to the overall score. The total score is then scaled to the traditional credit score range of 300 to 850, consistent with the FICO model.

Category	FICO Equivalent	Dataset Variables	Weight	Binning Logic
Payment History	35%	DELINQ (delinquencies)	**35%**	**0 delinquencies** = 50 points **1 delinquency** = 30 points **2+ delinquencies** = 10 points
Amounts Owed	30%	DEBTINC (debt-to-income ratio)	**30%**	**DEBTINC < 20%** = 50 points **20% ≤ DEBTINC ≤ 35%** = 30 points **DEBTINC > 35%** = 10 points
Length of Credit History	15%	CLAGE (credit age in months)	**15%**	**CLAGE > 120 months** = 50 points **60 ≤ CLAGE ≤ 120 months** = 30 points **CLAGE < 60 months** = 10 points
Ownership Ratio	-	OWNERSHIP (percentage of ownership in the house)	**10%**	**OWNERSHIP > 50%** = 50 points **50% ≤ OWNERSHIP ≤ 20%** = 30 points **EQUITY < 20%** = 10 points
New Credit	10%	NINQ (number of credit inquiries)	**10%**	**0 inquiries** = 50 points **1 inquiry** = 30 points **2+ inquiries** = 10 points

Table 6.1: Credit Scoring Model Variables and Weighting

The thresholds for each category were determined using a combination of data analysis, industry standards, and ensuring that the distribution of defaults aligns with real-world outcomes.

6.8.1 Payment History (DELINQ):

The **DELINQ** variable was selected over **DEROG** because it reflects missed payments rather than extreme events like bankruptcies or foreclosures, making it a more inclusive indicator of an applicant's financial behavior over time. The thresholds for delinquency were:

- **0 delinquencies = 50 points**: This represents the ideal scenario, indicating no missed payments and the lowest risk.
- **1 delinquency = 30 points**: This is considered manageable and reflects a less severe risk, as occasional missed payments can happen.
- **2 or more delinquencies = 10 points**: This signals a pattern of missed payments, representing higher risk.

These thresholds capture a realistic approach, recognizing that while one missed payment isn't catastrophic, repeated delinquencies increase risk.

6.8.2 Debt-to-Income Ratio (DEBTINC):

For the **Debt-to-Income (DTI)** ratio, the aim was to reflect how much debt an individual carries relative to their income, which is a critical measure of financial strain. While **industry standards** suggest that a **DTI of 35% or less** is favorable, applying these thresholds directly resulted in an unrealistically high proportion of **actual defaulters** (DEFAULT = 1) being classified in the **Fair** credit score category (around 40%), when the industry range in the real world is around 15-25%. Therefore, these thresholds were adjusted to:

- **0-20% DTI = 50 points**: Borrowers with a DTI of less than **20%** are considered to have strong financial health, as a smaller portion of their income is allocated to debt.
- **20-35% DTI = 30 points**: This range represents moderate financial strain, where debt takes up a significant portion of income, but it's still manageable.
- **35%+ DTI = 10 points**: Borrowers with a DTI higher than 35% face greater financial risk due to a large portion of their income being devoted to debt, leading to higher default rates.

By applying the adjusted thresholds, the proportion of actual defaulters in the Fair category dropped to **26%**, which aligns more closely with real-world data. This adjustment ensures the **default rates by credit score category** are more consistent with reality, providing a more accurate analysis of borrower behavior

6.8.3 Length of Credit History (CLAGE):

The **10-year mark** for credit history was chosen based on industry standards, where **10 years** is considered a long and established credit history. This is significant because negative items can remain on a credit report for up to 10 years, after which they are removed, reflecting improved financial behavior. The thresholds were:

- **10+ years = 50 points**: Indicates a long, reliable credit history.

- **5-10 years = 30 points**: Reflects a medium credit history with some risk.
- **Less than 5 years = 10 points**: Short credit history suggests higher risk.

6.8.4 Ownership Ratio (OWNERSHIP)

While **ownership ratio** (the proportion of equity relative to the property's value) is not traditionally part of FICO scores, it was included in this model to provide more comprehensive insights for lenders. This ratio offers a clearer picture of how much financial stake the borrower has in the property, directly reflecting their level of risk.

The thresholds for ownership ratio were determined based on borrower equity and their proportional stake in the property:

- **50%+ ownership ratio = 50 points**: Borrowers who own 50% or more of the property have significant equity, making them low risk. Their substantial ownership in the property reduces the likelihood of default.
- **20% to 50% ownership ratio = 30 points**: Borrowers with an ownership ratio between 20% and 50% still have a considerable financial stake but are at moderate risk. A **20% ownership threshold** aligns with common down payment requirements in mortgage lending.
- **Less than 20% ownership ratio = 10 points**: Borrowers with less than 20% ownership—including those with negative equity—pose the highest risk to lenders, as they have little to no financial stake in the property, increasing their chances of default.

6.8.5 New Credit (NINQ):

NINQ (Number of Inquiries) was used to represent **new credit** activity because frequent credit inquiries indicate a higher level of credit-seeking behavior, which can be a risk factor. The chosen thresholds were:

- **0 inquiries = 50 points**: Indicates no recent credit applications, reflecting lower risk.
- **1 inquiry = 30 points**: Shows moderate activity, but still acceptable.
- **2 or more inquiries = 10 points**: Suggests frequent applications for credit, signaling higher risk.

This approach ensures that the model captures the applicant's tendency to seek new credit, which is linked to potential over-leveraging.

After the points are calculated for each record, the formula used for scaling them in the FICO score range is:

$$\textbf{\textit{Scaled Score}} = 300 + \left(\frac{\textbf{\textit{Raw Score}}}{\textbf{\textit{Max Raw Score}}} \times 550\right)$$

By scaling the raw points to the 300-850 range, the credit scoring model produces a score that is comparable to FICO scores, making it easier for the bank to assess risk using a familiar scale.

6.9 Credit Score Category:

The **Credit Score Category** classifies borrowers based on their credit score, which reflects their creditworthiness and financial behavior. The classification follows the same ranges as the official FICO score categories, providing a standardized assessment of borrower risk.

- **Poor (300-579)**: This category represents the highest risk. Borrowers in this range typically have a history of missed payments, high debt-to-income ratios, short credit histories, and may frequently seek new credit. They are more likely to default on loans.

- **Fair (580-669)**: Borrowers in this range have moderate risk. They might have some delinquencies or higher debt levels, but their credit history shows they have been able to manage these issues to some extent. However, they still pose a moderate default risk.

- **Good (670-739)**: This category represents borrowers with good financial behavior. They typically have a manageable amount of debt, few or no delinquencies, and a longer credit history. They are less likely to default on loans and are generally considered low risk.

- **Very Good (740-799)**: Borrowers in this category have a strong credit profile with a long credit history, low debt, and few or no missed payments. They are highly reliable, and the chances of defaulting on a loan are minimal.

- **Excellent (800-850)**: The best credit category. Borrowers here have excellent financial habits, long-established credit histories, and extremely low debt levels. Their risk of default is almost negligible, making them ideal loan candidates.

Section 7: Exploratory Data Analysis

The Exploratory Data Analysis (EDA) phase is a critical component of any data science project, aimed at uncovering patterns, trends, and anomalies in the dataset. It employs statistical methods and visualizations to assess the dataset's structure, relationships between variables, and underlying distributions. EDA also plays a key role in evaluating data quality by identifying missing values, outliers, and inconsistencies that could impact subsequent analyses.

In this analysis, EDA focuses on variables relevant to loan defaults, such as loan amounts, mortgage dues, property values, debt-to-income ratios, credit line age, and years on the

job. Additional risk-related metrics, including equity, potential loss, and loan-to-equity ratios, are examined to assess the bank's exposure to financial risks. This phase sets the foundation for predictive modeling and decision-making by revealing critical insights through visualizations like histograms, boxplots, and correlation analyses.

7.1 Summary Statistics and Distributions

Summary statistics and visualizations provide a detailed view of the dataset's key variables. Metrics such as mean, median, and standard deviation capture central tendencies and variability, while histograms and boxplots illustrate the distribution and spread of critical variables. This analysis highlights potential outliers, skewness, and other patterns, contributing to a deeper understanding of the data and guiding further modeling efforts.

7.1.1 Original Variables

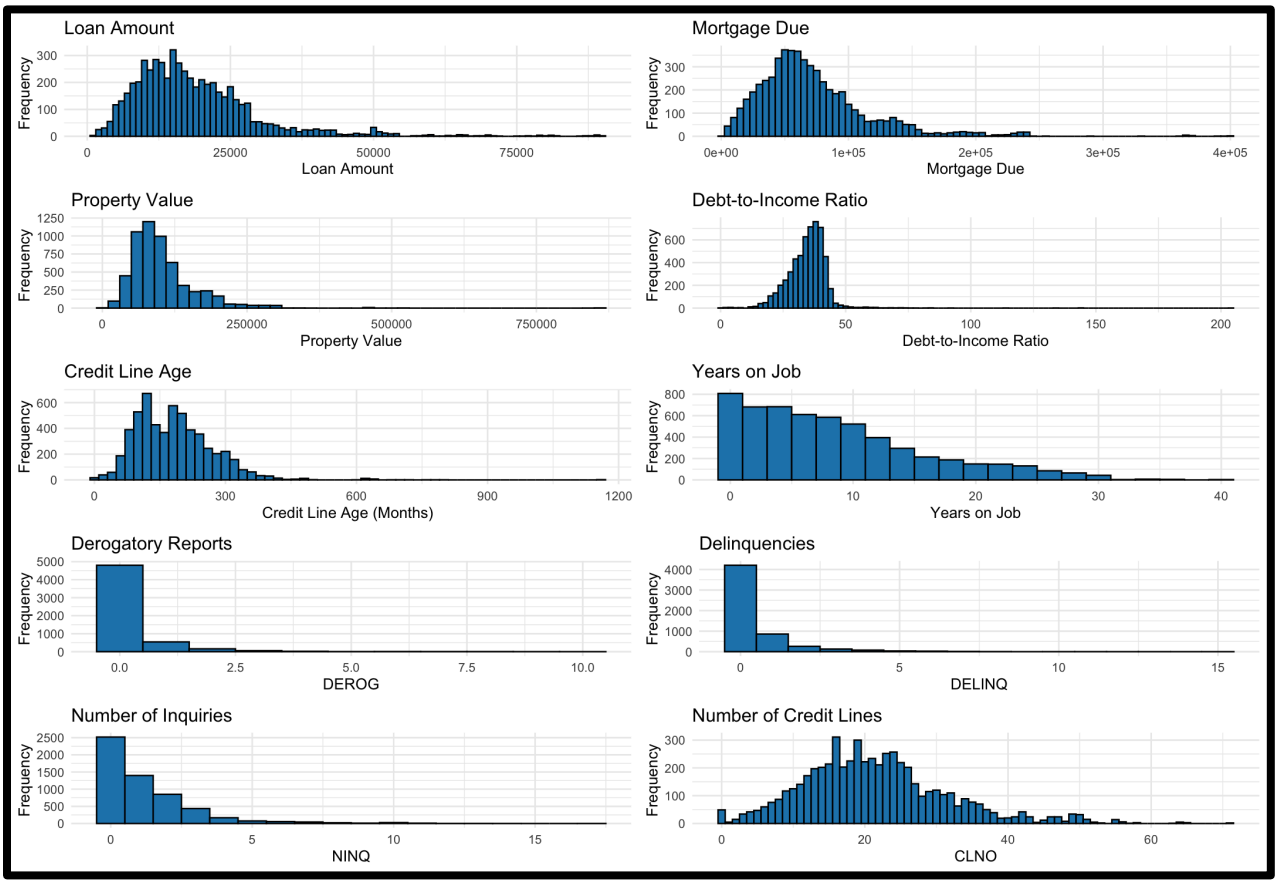

Figure 7.1: Histograms of Original Variables

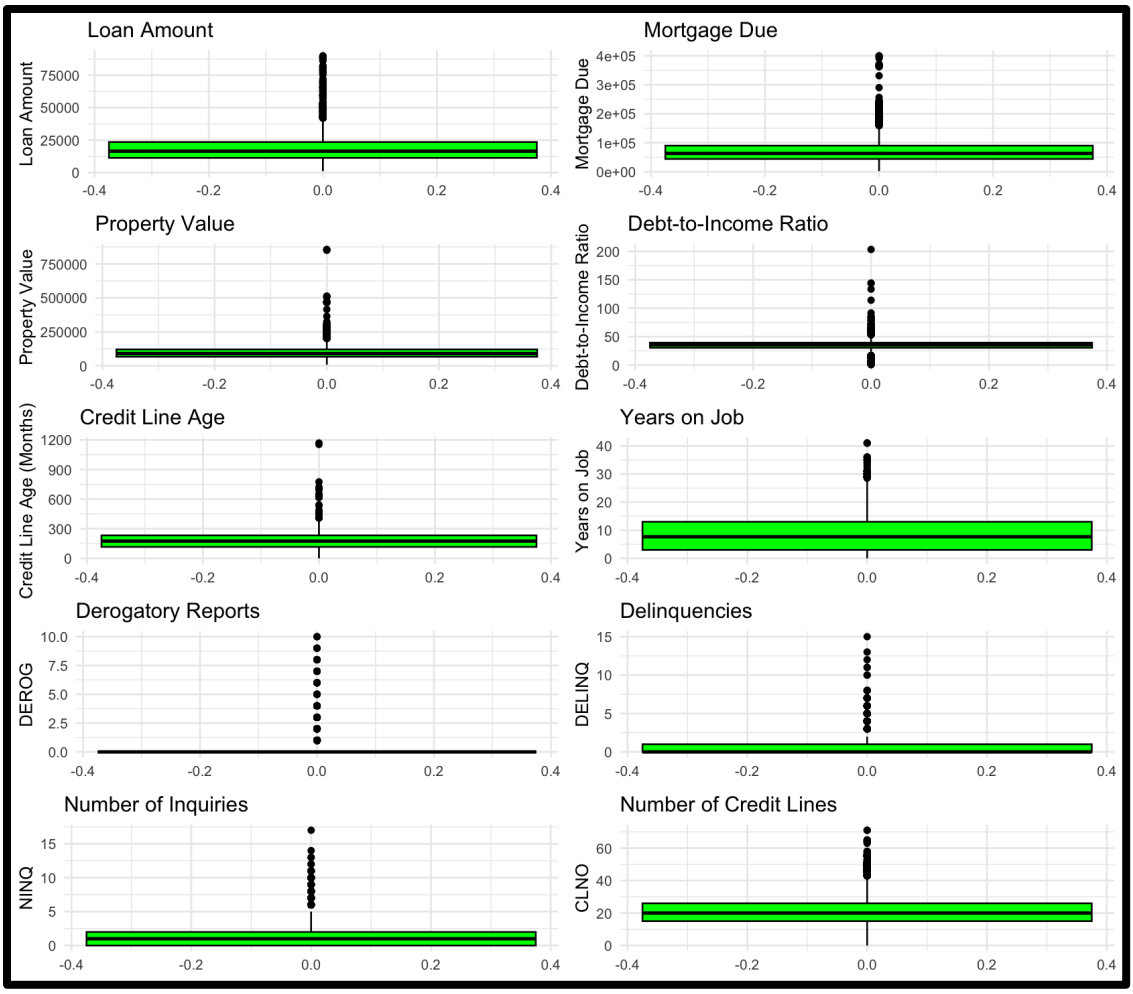

Figure 7.2: Boxplots of Original Variables

1. **LOAN (Loan Amount):**

The average loan amount is $18,846, with a relatively large spread of loan amounts. Most loans fall between $11,300 and $23,500, indicating that these are typically small to mid-sized loans, as it can be seen in both Figure 12 and Figure 13. It is also possible to observe the presence of outliers on the upper part of the histogram starting at approximately $40,000, suggesting that a small number of loans significantly exceed the average. These outliers represent larger, less common loans that deviate from the typical loan distribution.

2. **MORTDUE (Mortgage Due):**

Borrowers have mortgages with an average outstanding balance of $72,090. The median mortgage amount is lower than the mean, suggesting a right-skewed distribution (a few large mortgages). This distribution can be observed on Figure 12 and Figure 13, where the histogram shows higher frequencies between $25,000 and $75,000, and the box plot shows outliers ranging from approximately $150,000 to $399,550, suggesting

that the high mortgage balances seen in the outliers indicate that these borrowers have significant existing financial obligations. This may increase the overall risk to the lender if the borrowers already have large mortgages and are taking on additional loans.

3. **VALUE (Property Value):**

The average property value is just over $100,000. There are some higher-valued properties, but the majority of properties are valued between $68,000 and $120,000. It is possible to observe that some high-value properties skew the distribution.

4. **DEROG (Derogatory Reports):**

Most borrowers have no derogatory reports (bankruptcies, liens, etc.), as the median is 0. However, a small subset has as many as 10 derogatory marks, which could indicate higher risk for those borrowers.

5. **DELINQ (Delinquencies):**

Similar to DEROG, most borrowers do not have any delinquencies. However, some borrowers have up to 15 delinquencies, which indicates a history of missed payments for certain of borrowers.

6. **CLAGE (Credit Line Age):**

On average, borrowers have had their credit lines for about 15 years, which reflects a mature borrower profile. However, there are borrowers with very young credit lines as well as outliers with very old credit lines. While such extremely old credit lines seem unlikely, they require further investigation to determine if they reflect data anomalies or special cases.

7. **NINQ (Number of Inquiries):**

Most borrowers have had 1 or 2 recent inquiries, but a few have as many as 17 inquiries, which could indicate recent aggressive credit-seeking behavior (a potential red flag).

8. **CLNO (Number of Credit Lines):**

The average number of credit lines is around 21, with some borrowers having a very large number of lines (up to 71), which could indicate higher credit exposure.

9. **DEBTINC (Debt-to-Income Ratio):**

The average borrower has a debt-to-income ratio of around 35%. However, a small subset has extremely high debt-to-income ratios exceeding 100%, indicating they owe more than their annual income, a significant risk factor for default.

7.1.2 Derived Variables:

```
     EQUITY         UNSECURED_LOAN  DEFINITIVE_LOSS      LTE_RATIO          yoj_to_loan_ratio      LTV_RATIO          OWNERSHIP
 Min.   :-205445   Min.   :    0   Min.   :    0.0   Min.   :-177.5493   Min.   :0.0000000   Min.   :0.01518   Min.   :-6.1625
 1st Qu.:  17091   1st Qu.:    0   1st Qu.:    0.0   1st Qu.:   0.4538   1st Qu.:0.0001818   1st Qu.:0.12254   1st Qu.: 0.2046
 Median :  26268   Median :    0   Median :    0.0   Median :   0.6530   Median :0.0004310   Median :0.17031   Median : 0.2830
 Mean   :  31002   Mean   : 1268   Mean   :  242.7   Mean   :   0.6997   Mean   :0.0006429   Mean   :0.21490   Mean   : 0.2870
 3rd Qu.:  39285   3rd Qu.:    0   3rd Qu.:    0.0   3rd Qu.:   0.8478   3rd Qu.:0.0008462   3rd Qu.:0.24390   3rd Qu.: 0.3772
 Max.   : 641408   Max.   :56644   Max.   :50100.0   Max.   : 130.8947   Max.   :0.0095455   Max.   :2.32500   Max.   : 0.9795
  CREDIT_SCORE     SCORE_CATEGORY
 Min.   :432.0   Poor     : 726
 1st Qu.:630.0   Fair     :1695
 Median :674.0   Good     :1941
 Mean   :670.3   Very Good:1120
 3rd Qu.:718.0   Excellent:  90
 Max.   :850.0
```

Figure 7.3: Summary Statistics of Derived Variables

1. EQUITY

The average equity is $30,964, with borrowers having negative equity up to -$205,445, indicating some owe significantly more than their property's worth. This negative equity raises default risks and potential for foreclosure.

2. UNSECURED LOAN

The majority of loans are fully secured, with a median unsecured amount of $0. However, some loans are unsecured, reaching amounts as high as $56,644.

3. DEFINITIVE LOSS

Most loans result in no definitive loss (median = $0), though losses can go as high as $50,100 in extreme cases, reflecting substantial financial damage for the lender.

4. LTE RATIO

The LTE ratio, on average, is about 70%, indicating most loans are about 70% of the borrower’s equity. However, the wide range highlights cases of high-risk loans, with some exceeding borrower equity significantly.

5. YOJ_TO_LOAN_RATIO

This ratio remains quite small for most borrowers, indicating that the length of time in a job is not significantly large compared to loan size. It suggests that job tenure may not be a strong factor in securing larger loans for most borrowers.

6. LTV RATIO (Loan-to-Value Ratio)

The average loan-to-value ratio is about 21.6%, meaning that, on average, loans are a relatively small percentage of property values. However, a maximum LTV ratio of **2.325** indicates a few loans that significantly exceed property values, which is a considerable risk.

7. OWNERSHIP

On average, borrowers own about 28.76% of their properties, with negative ownership values indicating that some borrowers owe more than the total value of their property, which increases the risk of default and makes the loan more precarious.

8. CREDIT SCORE

With a mean credit score of 670.8, most borrowers fall in the **Fair to Good** range. This average suggests moderate credit risk for the lender, though the lowest scores (432) indicate significantly higher risk.

9. SCORE CATEGORY

Most borrowers fall in the **Fair** and **Good** credit score categories, with **Fair** having the largest portion of the dataset (1,695 observations). This indicates a moderate level of credit risk. A small number of borrowers (97) are in the **Excellent** category, representing low credit risk, while the **Poor** category (726 observations) highlights the most risky borrowers who could default on their loans.

7.2 Visual Exploration of Relationships

This section examines relationships between key variables using visualizations such as scatterplots and bar plots. These tools highlight interactions between borrower characteristics, loan attributes, and default risk, uncovering patterns and trends critical for feature selection and predictive modeling.

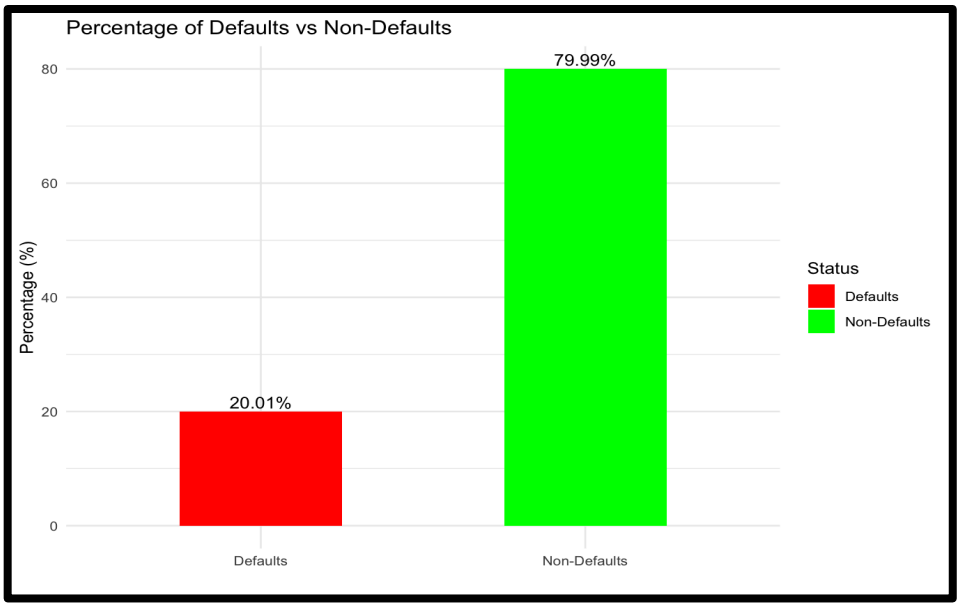

Figure 7.4: Percentages of Defaults vs Non-Defaults on Number of Loans

Figure 7.4 shows the percentage of defaults (79.99%) and the percentage of non-defaults (20.01%). This visualization illustrates the business problem the company is facing, where default rates are really high.

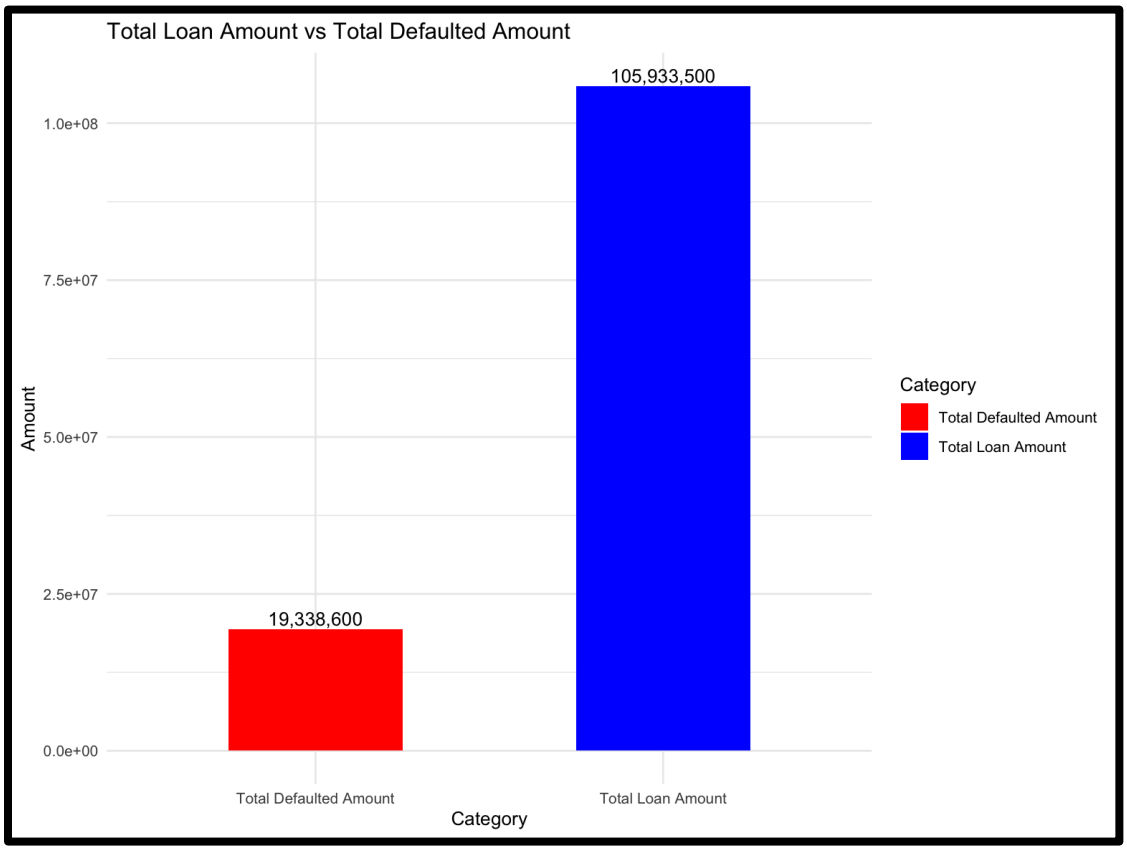

Figure 7.5: Percentages of Total Loan Amount vs Total Defaulted Amount

Figure 7.5 shows that the total dollar amount of loans the bank provided over the period covered by the data is $105,993,500. Of these, a total of $19,338,600 were defaulted loans.

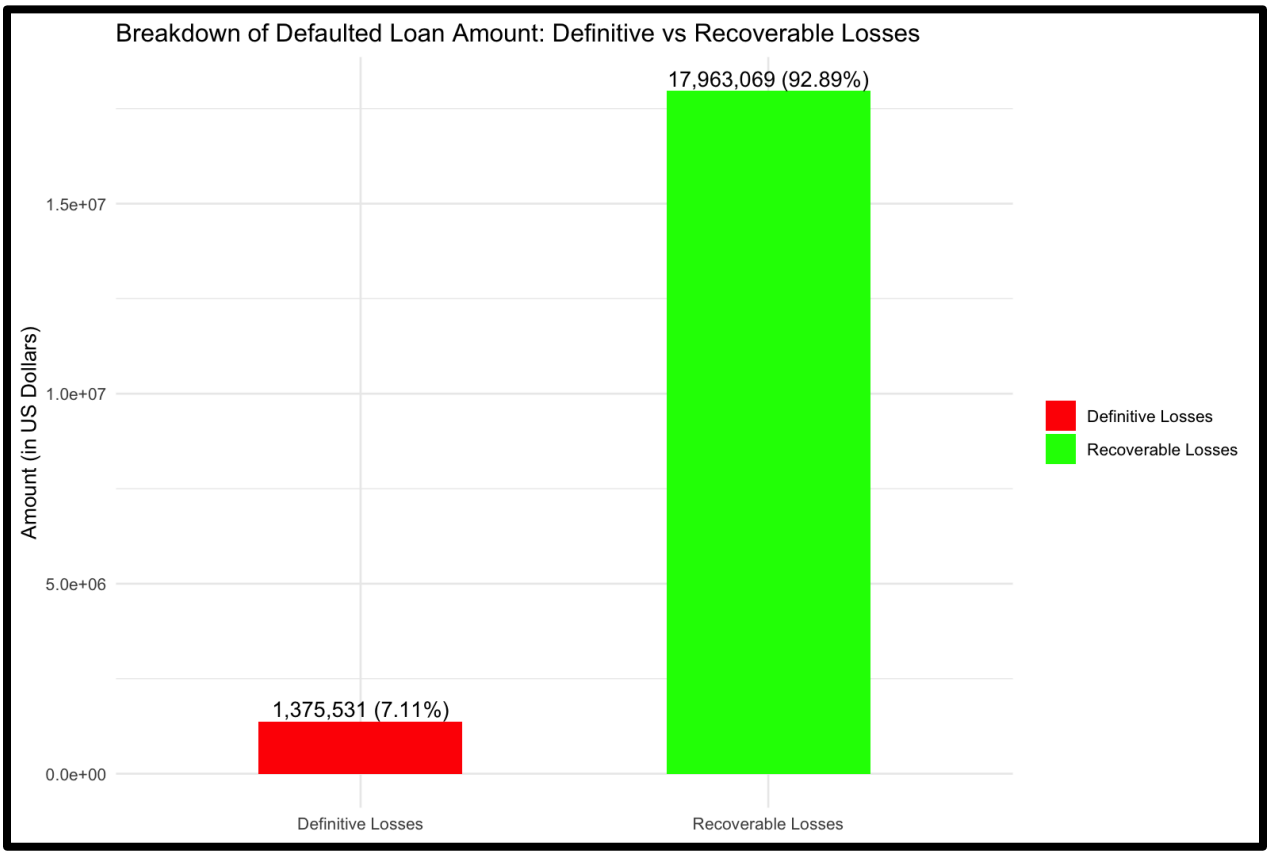

Figure 7.6: Percentages of Total Loan Amount vs Total Defaulted Amount

Figure 7.6 breaks down the total amount of defaulted loans into two categories:

- **Definitive losses** are those where the loan amount, either partially or fully, could not be recovered through collateral because the loan exceeded the borrower's available equity.

- **Recoverable losses** refer to the loans where the borrower defaulted but had sufficient equity to cover the outstanding amount.

The data shows that definitive losses amount to $1,375,531, while recoverable losses total $17,963,069. This means that the vast majority of defaulted loans (92.89%) could potentially be recovered, while only 7.11% were definitive losses.

From a broader perspective, this represents a relatively small impact, as the bank has issued $105,993,500 in total loans, and only 1.29% of that total has resulted in definitive losses.

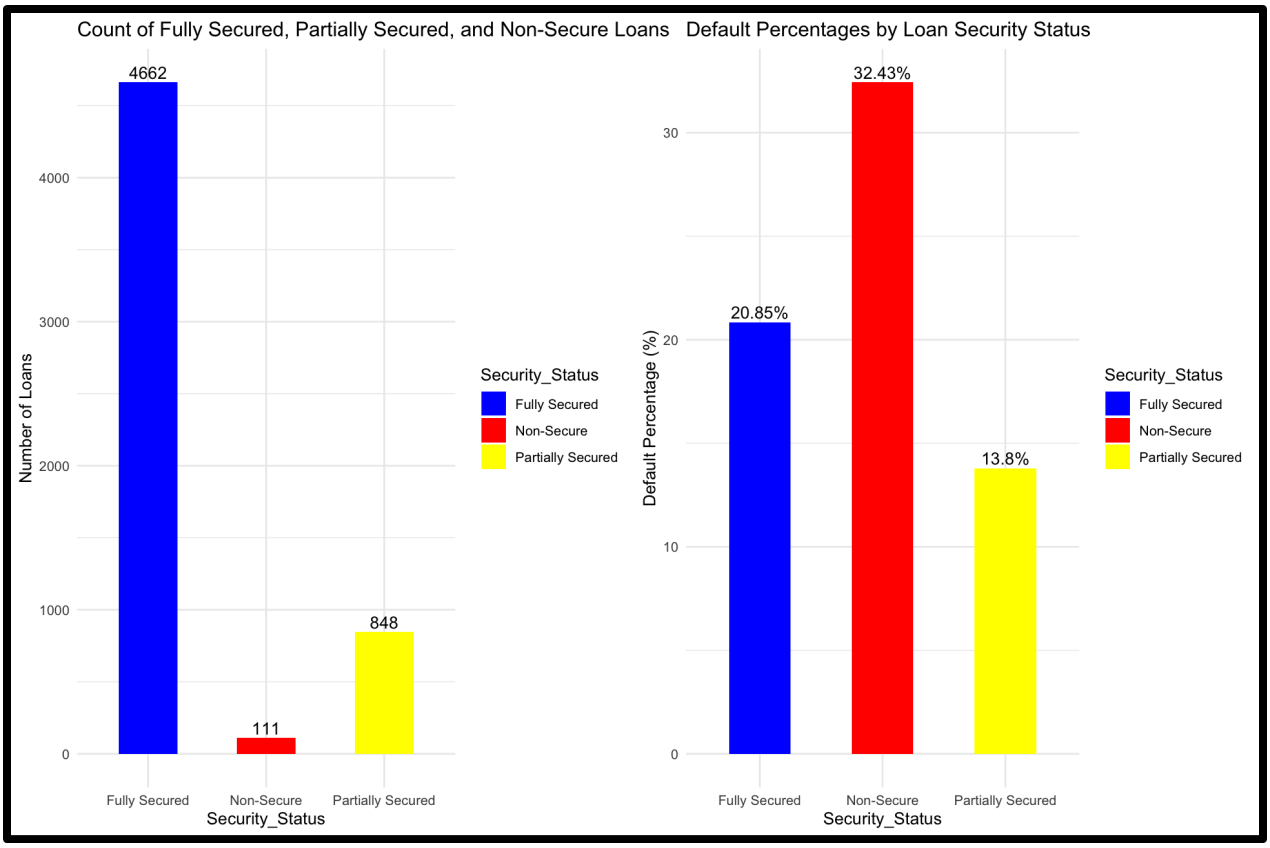

Figure 7.7: Loan Count and Default Percentages by Loan Security Status

Figure 7.7 displays the distribution of loans categorized as fully secured, partially secured, and unsecured, as well as the default rates for each.

- **Fully secured loans** are those where the borrower's equity is equal to or greater than the loan amount, meaning the loan is fully backed by collateral.

- **Partially secured loans** are those where the borrower's equity is positive but insufficient to cover the full loan amount in case of default, offering partial collateral.

- **Unsecured loans** are those where the borrower's equity is zero or negative, meaning there is no collateral to back the loan in the event of default, making these loans riskier.

The figure shows that the majority of loans provided are fully secured (82.93%), while partially secured loans make up around 15%, and unsecured loans account for less than 2%. This is a positive indicator, as it suggests the bank is largely protected against defaults, assuming economic conditions remain constant.

It is also notable that unsecured loans have a default rate of 32.43%, which makes them even riskier. This rate is approximately 12% higher than the already high average default rate of 20%, while the other loan categories have default rates below the average.

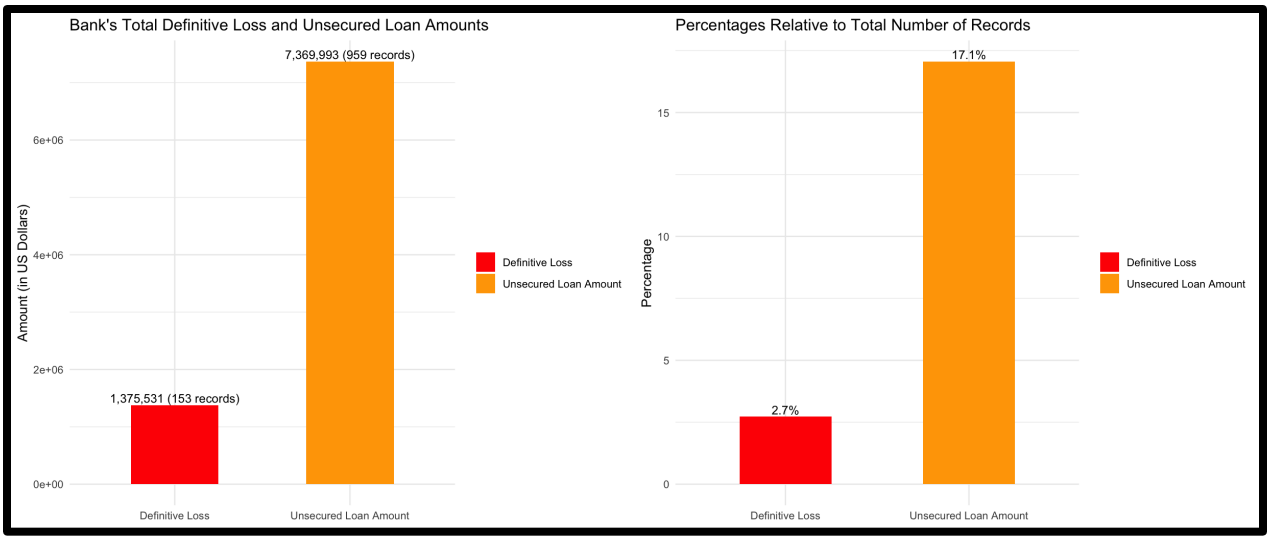

Figure 7.8: Comparison of Bank's Total Definitive Loss and Unsecured Loan Amounts: Absolute Values and Relative Percentages

Figure 7.8 compares the bank's total definitive loss to the dollar amount of loans that are uncovered in the event of default. The data shows that $7,369,993 is uncovered, representing the worst-case scenario, since this would be the total unrecoverable loss if all borrowers were to default. This amount constitutes 17.1% of the total loans the bank has issued.

In contrast, the definitive loss (actual losses incurred where recovery was not possible) stands at $1,375,531, which accounts for 2.7% of the total loans provided.

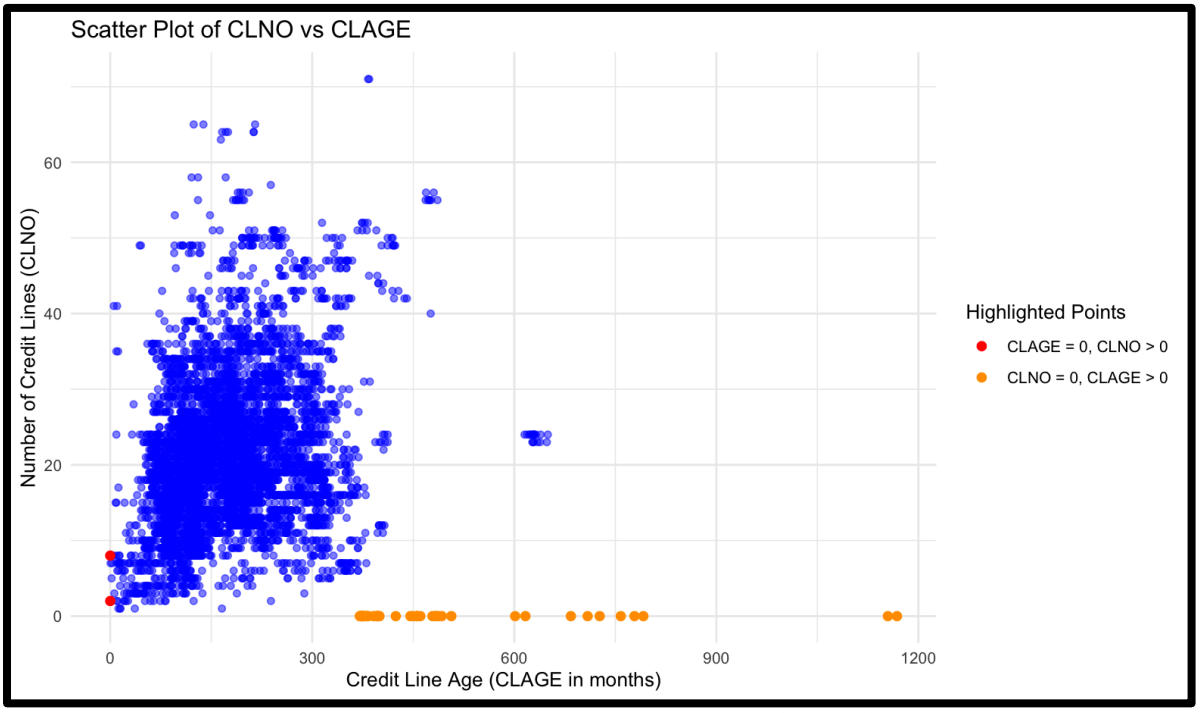

Figure 7.9: Scatter Plot of Number of Credit Lines (CLNO) vs Credit Line Age (CLAGE) with Highlighted Data Anomalies

By observing ***Figure 7.9***, the number of credit lines is plotted against the age of the oldest credit line. While this general relationship isn't particularly significant, what stands out are two records highlighted in red, where the number of credit lines is greater than zero (2 and 8), yet the age of the oldest credit line is 0. This inconsistency is unusual. One possible explanation is that these accounts were opened very recently. However, the fact that one of these accounts holds 8 credit lines raises suspicion. Additionally, this same account has 12 credit inquiries and has defaulted on the loan, which could indicate fraud.

Opening multiple credit lines within a short time, combined with numerous credit inquiries, suggests that the borrower may have been attempting to rapidly access credit beyond their ability to repay. The loan default further reinforces this possibility. Such behavior raises the concern of fraudulent activity, where the accounts may have been opened to exploit the available credit before defaulting.

Highlighted in orange, there are 49 individuals with no credit lines (CLNO = 0), yet the recorded age of their nonexistent lines ranges from over 25 years to as much as 97 years. Notably, 24 of them defaulted (49%), and coincidentally, all the 24 defaulters have the oldest credit lines.

These older credit line records may not accurately reflect current economic realities or lending standards, potentially making them less relevant for today's financial models. Factors such as inflation, fluctuating interest rates, and financial crises have heavily influenced lending practices over time.

Based on these observations, two approaches can be considered. The first is to interpret these as credit lines that have been inactivated but whose age is still being recorded, leading to these results, or as potential data inaccuracies. The second approach is to remove all records where the credit line age is greater than zero when the number of credit lines is zero. Given that these constitute only 49 records, to ensure data integrity, these entries will be deleted.

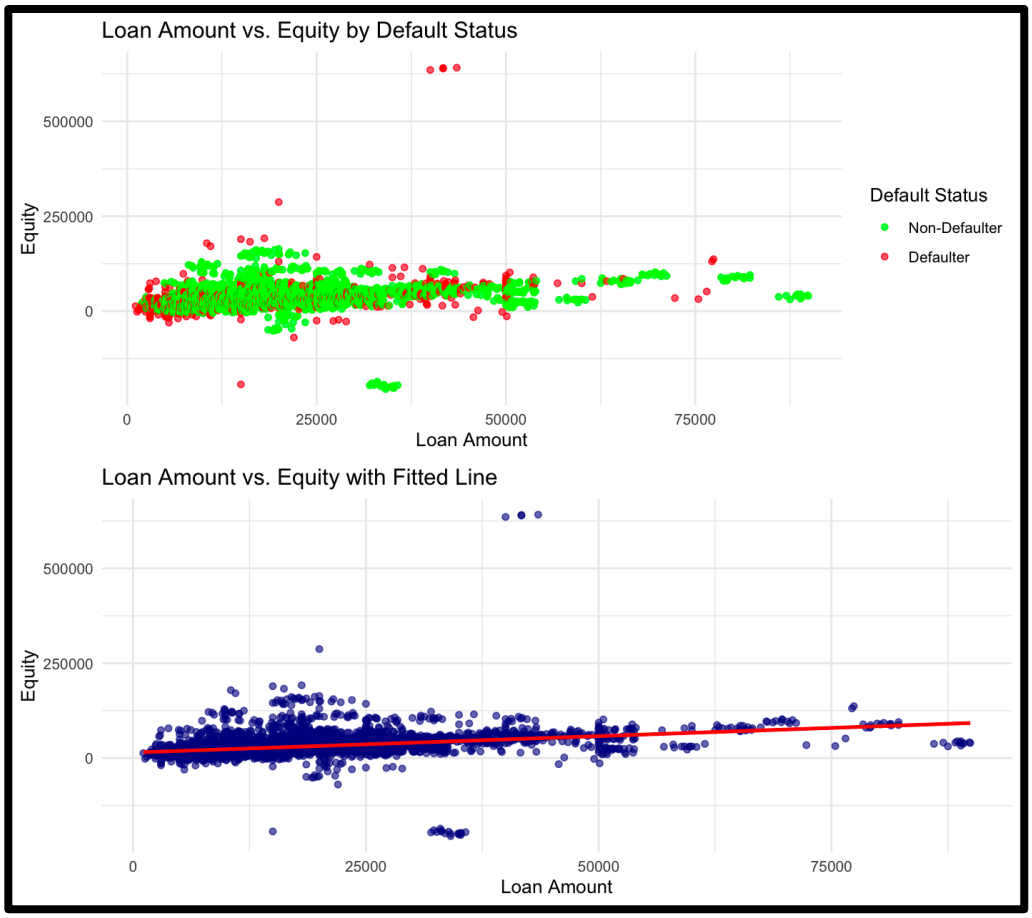

Figure 7.10: Analysis of Loan Amount vs Equity: Default Status and Linear Trend

In ***Figure 7.10*** it is evident that both defaulters and non-defaulters are distributed similarly across different equity levels, indicating that equity alone is not a strong predictor of default. Additionally, there are several applicants with high equity who still defaulted, suggesting that factors beyond equity, such as income stability or debt levels, play a significant role in determining default risk.

Among applicants with negative equity, 32.43% are defaulters, while 67.56 % are non-defaulters. This indicates that being "underwater" (owing more than the property's worth) does not automatically lead to default, as the majority of borrowers in this situation continue to make payments.

The second plot reveals a slight positive trend between loan amounts and equity, meaning that, on average, larger loans are associated with higher equity. However, the relationship is weak, indicating that loan amount and equity are not closely linked in a meaningful way for predicting default risk.

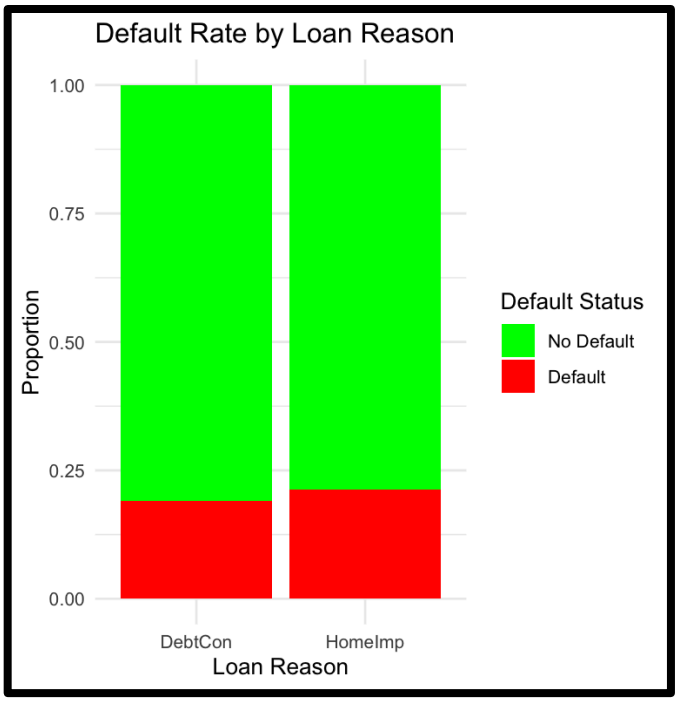

Figure 7.11: Proportion of Default Rate by Loan Reason

Figure 7.11 shows that default rates are similar for loans taken out for debt consolidation and home improvement. Approximately 20% of loans default in both categories, and around 80% of the loans do not default. This suggests that the type of loan may not be an effective predictor of default.

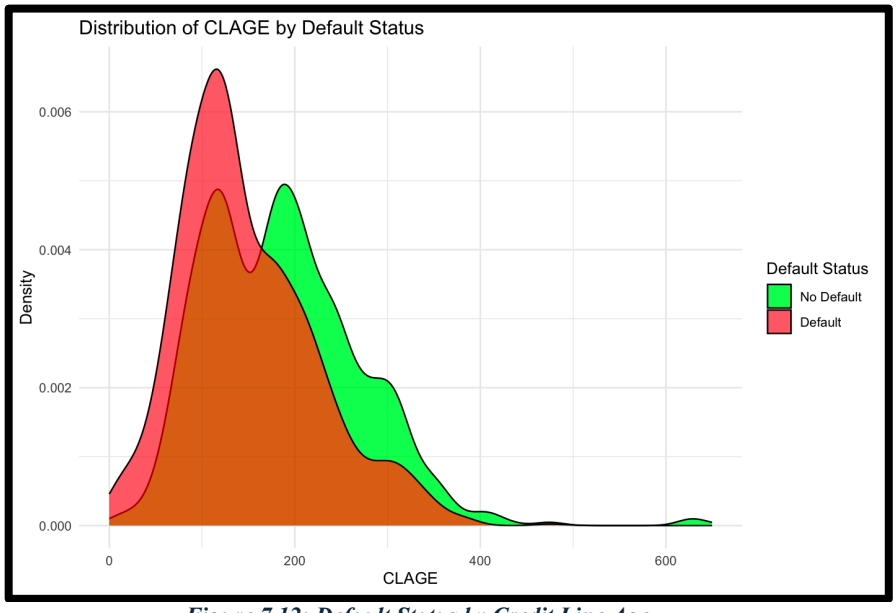

Figure 7.12: Default Status by Credit Line Age

Figure 7.12 show that loans with lower CLAGE values (0 to ~ 180 months) tend to have higher default rates, as shown by the red density curve being higher than the green curve. This suggests that borrowers with newer credit histories are at a greater risk of defaulting.

In contrast, as CLAGE values increase beyond 180 months, the likelihood of default decreases. The green curve becomes more prominent, indicating that borrowers with more established credit histories are less likely to default.

However, there is overlap between 0 to ~ 400 months, suggesting that while lower CLAGE helps to identify default risk, it is not the only factor influencing default likelihood.

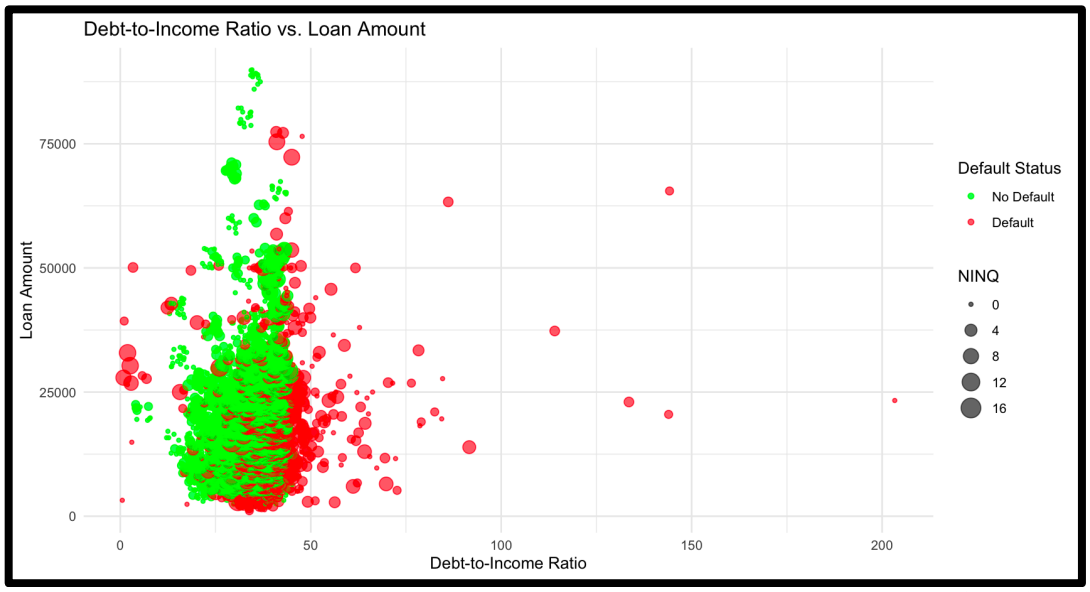

Figure 7.13: Default Status, Debt-to-Income Ratio vs Loan Amount, and Credit Inquiries (NINQ)

The scatter plot in ***Figure 7.13*** shows the relationship between Debt-to-Income Ratio, Loan Amount, and Default Status, with dot size indicating the number of inquiries (NINQ). Most non-defaults occur at lower Debt-to-Income Ratios and smaller loan amounts. As the Debt-to-Income Ratio rises, particularly above 50 and for loans exceeding $25,000, the number of defaults increases. Larger dots, representing higher NINQ, tend to correspond with more defaults, suggesting a link between multiple inquiries and a greater likelihood of default.

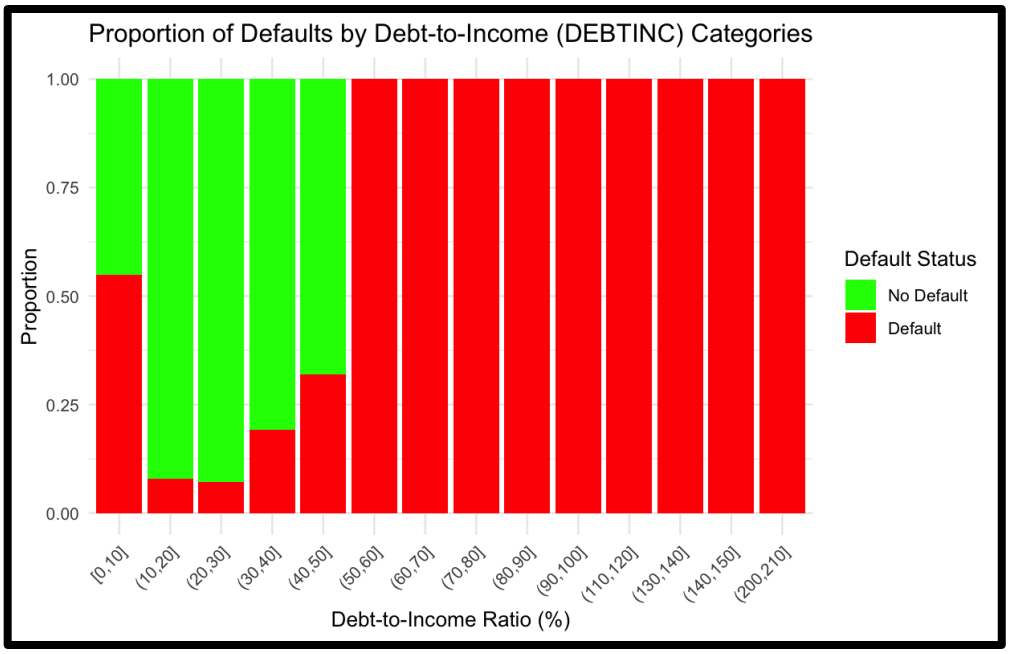

Figure 7.14: Proportions of Default and Non-default Across DEBTINC

Figure 7.14 reveals that borrowers with a Debt-to-Income Ratio (DEBTINC) above 50% face a 100% default rate, indicating an extremely high risk of financial instability in this category. Between 0 and 10%, default rates are unexpectedly high, exceeding 50%, suggesting that even borrowers with very low DEBTINC ratios may struggle with repayment, potentially due to other factors. The lowest default rates are observed in the 10–30% range, where borrowers demonstrate greater financial stability and a higher likelihood of fulfilling their loan obligations. This highlights a sweet spot for safer lending, while borrowers outside this range, particularly those above 50%, represent significant risk.

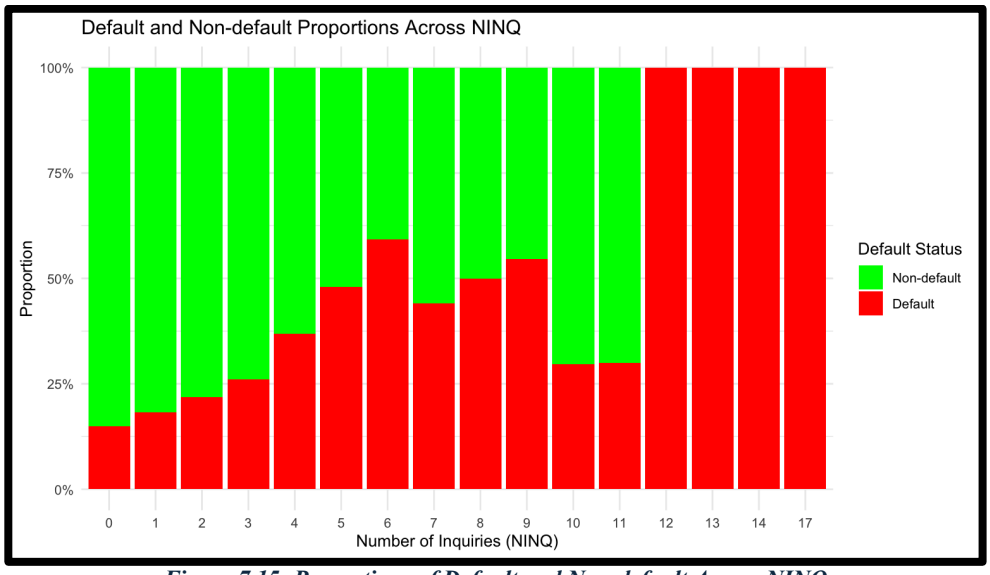

Figure 7.15: Proportions of Default and Non-default Across NINQ

The plot in ***Figure 7.15*** provides a detailed view of the relationship between the number of inquiries (NINQ) and defaults. With exception of inquires between 7 and 11, there is a clear

linear trend, with higher NINQ correlating with a higher default rate. Notably, inquiries above 12 are associated with a 100% default rate, highlighting the strong connection between multiple inquiries and increased financial risk.

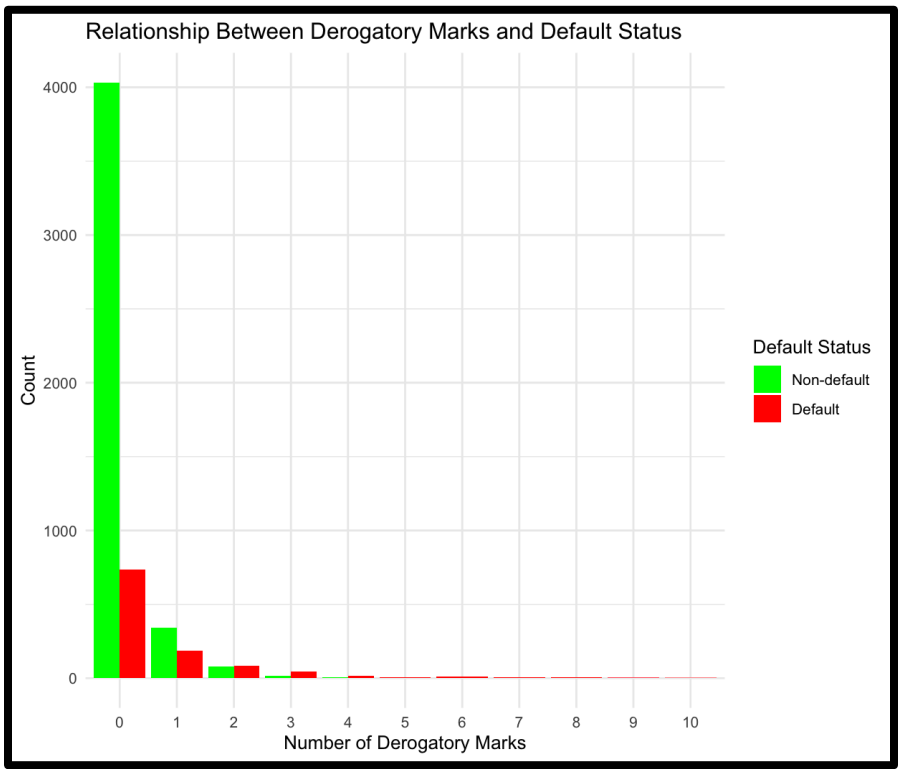

Figure 7.16: Proportions of Default Status Across Derogatory Marks

In ***Figure 7.16***, it is evident that as the number of derogatory reports increases, so does the likelihood of default. Beginning at 2 derogatory reports, the default rate surpasses the non-default rate. Once individuals accumulate 5 or more derogatory reports, all have defaulted. This suggests that Derogatory marks are a strong determinant of default.

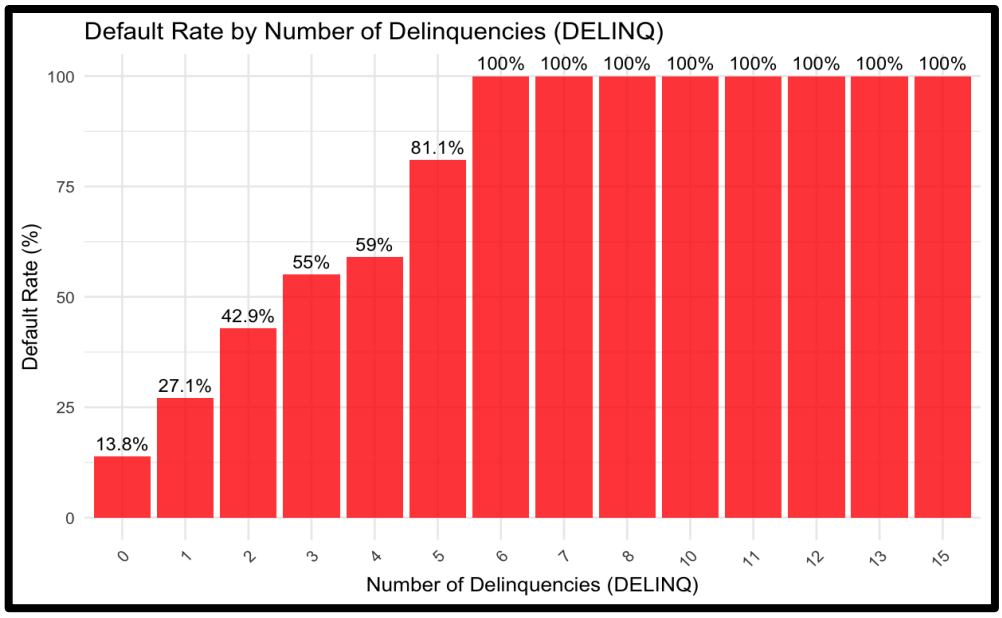

Figure 7.17: Default Rate Across Years of Job (YOJ) Ranges

Figure 7.17 illustrates the relationship between the number of delinquencies (DELINQ) and the default rate. Borrowers with no delinquencies have a default rate of only 13.8%, indicating that a clean credit record is a strong indicator of financial reliability. As the number of delinquencies increases, the default rate rises sharply. Borrowers with one delinquency exhibit a default rate of 27.1%, which nearly doubles to 42.9% for those with two delinquencies. Beyond three delinquencies, the default rate exceeds 50%, reaching 81.1% for borrowers with five delinquencies.

Once the number of delinquencies surpasses five, the default rate reaches 100%, indicating that borrowers with six or more delinquencies are virtually certain to default. This pattern strongly highlights the significance of delinquency history as a key predictor of default risk. Borrowers with even a single delinquency present heightened risk, while those with multiple delinquencies represent a progressively higher likelihood of default.

These findings suggest that delinquency history should be a critical factor in loan approval decisions. Borrowers with six or more delinquencies should be considered extremely high-risk, while those with fewer delinquencies may still require additional evaluation depending on their other financial indicators.

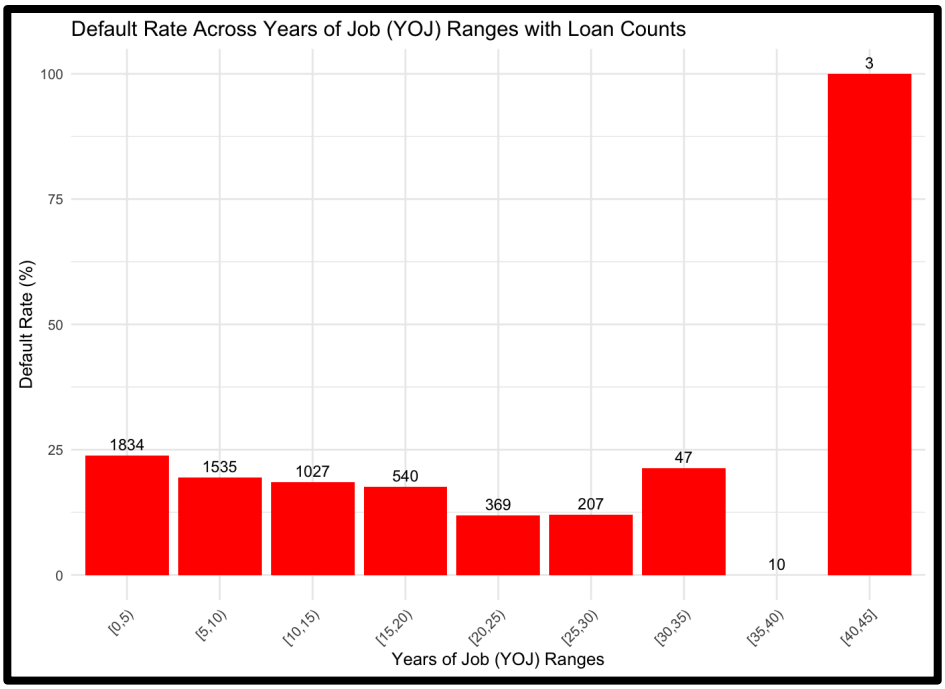

Figure 7.18: Default Rate Across Years of Job (YOJ) Ranges

In ***Figure 7.18***, the analysis of default rates across "Years of Job" (YOJ) reveals key insights. Borrowers with job tenures of 0-20 years show stable default rates between 20% and 25%, with a slight decrease in default for older groups. Borrowers with 20-30 years of tenure experience a greater drop in default rates, suggesting greater financial stability. In contrast, ranges of 40-45 years show sharply increased default rates, but these findings are based on very small loan samples, making the results less reliable. Overall, borrowers with less than 5 years of tenure pose the highest default risk, given their large loan count and relatively high default rate.

Additionally, it's important to note that the numbers displayed on top of each bar represent the total number of loans (records) within each YOJ range, not just the number of defaults. This provides context to better understand how many borrowers fall into each category and the reliability of the default rates for smaller groups.

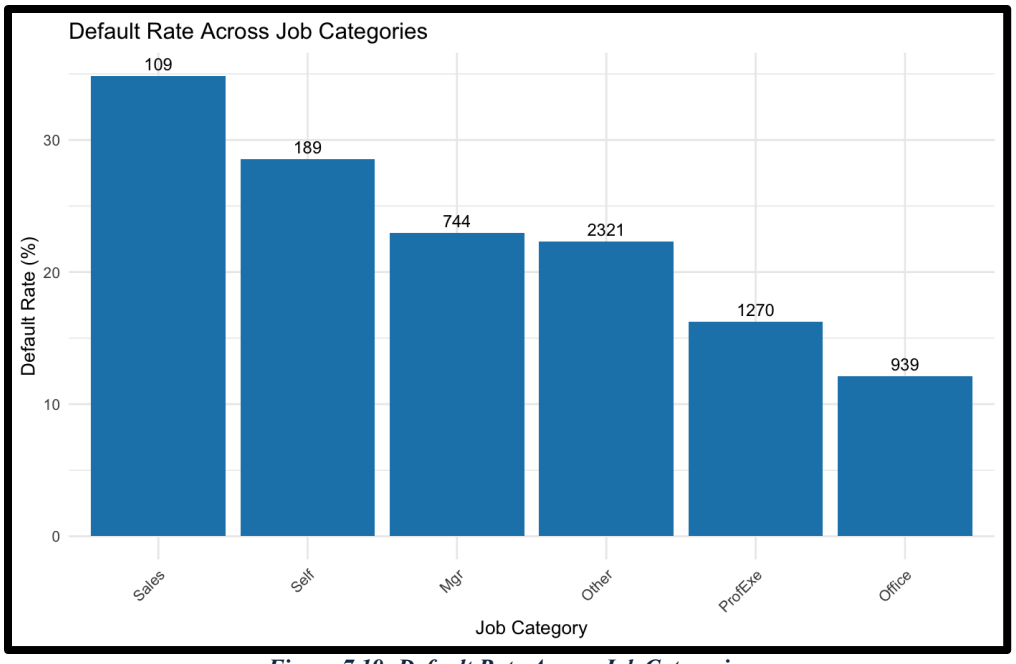

Figure 7.19: Default Rate Across Job Categories

The graph in ***Figure 7.19*** highlights key patterns in default rates across job categories. Sales workers have the highest default rate, over 30%, despite having fewer loans. Self-employed borrowers follow closely with nearly 30% defaulting, indicating financial instability in these groups.
Managers and the other job category have default rates around 25%, with "other" contributing most to overall defaults due to its larger loan volume. On the lower end, professionals/executives and office workers show the lowest default rates, both under 20%, suggesting more financial stability.

Overall, sales and self-employed borrowers pose the highest default risk, while professionals and office workers tend to be lower risk.

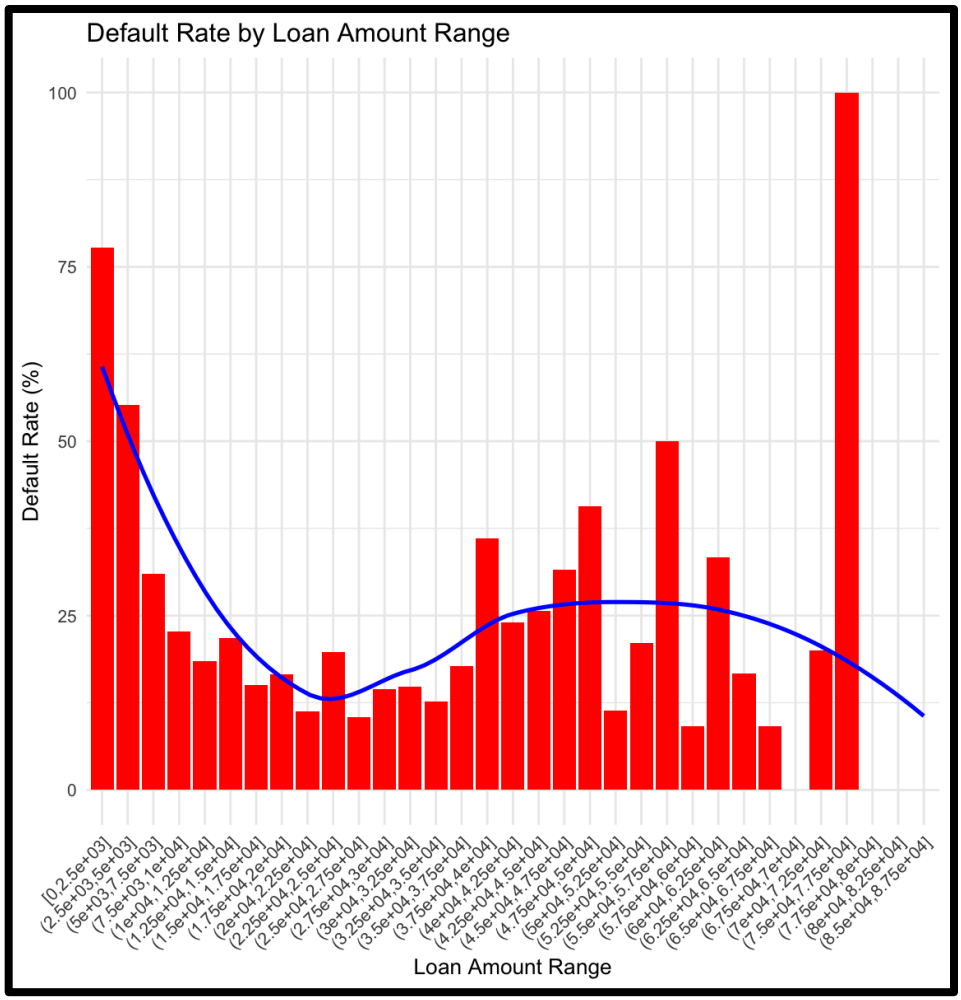

Figure 7.20: Default Rate by Loan Amount Range

The plot in ***Figure 7.20*** illustrates the relationship between loan amount ranges and default rates. It shows a high default rate for smaller loans, particularly those under $10,000, indicating a higher risk among borrowers in this category. As loan amounts increase, the default rate declines significantly for loans between $10,000 and $35,000, suggesting that borrowers within this range are more financially stable and less likely to default.

However, after the $35,000 threshold, the default rate begins to rise again, with a sharp increase for loans exceeding $75,000. This suggests that larger loan amounts are associated with increased risk, possibly due to over-leveraging or other financial challenges faced by borrowers. These trends highlight potential areas for adjusting lending strategies. Loans in the $10,000 to $35,000 range appear to carry the lowest default risk, making them attractive for approval.

Conversely, loans at the lower and higher ends may require more stringent risk assessments or adjusted terms to mitigate the higher likelihood of default.

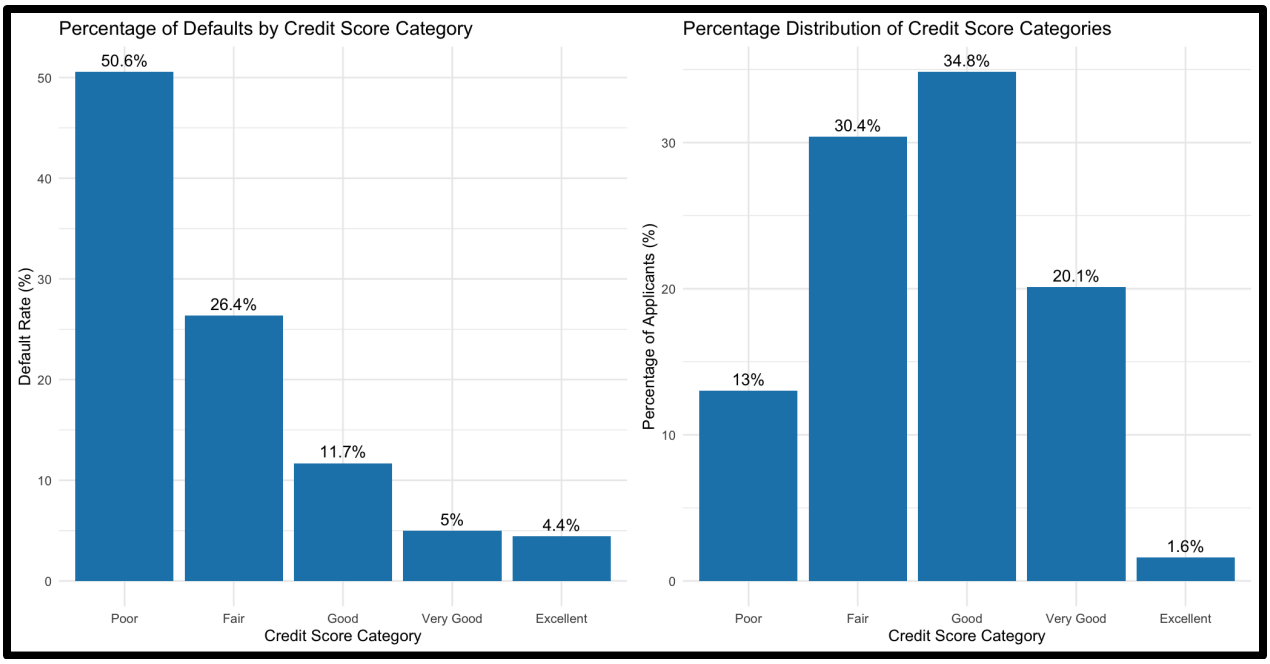

Figure 7.21: Default Rates and Applicant Distribution by Credit Score Categories

Figure 7.21 compares the default rates and the distribution of applicants across different credit score categories.

The chart on the left shows that borrowers in the **Poor** credit category (300-579) have the highest default rate, with over **50%** defaulting on their loans. As credit scores increase, the default rate decreases significantly. In the **Fair** category (580-669), about **26%** default, while in the **Good** category (670-739), only **11.7%** default. The **Very Good** and **Excellent** categories have the lowest default rates, at **5%** and **4.4%**, respectively.

The chart on the right highlights the distribution of applicants. The majority of applicants are in the **Fair** and **Good** credit score categories, accounting for **30.4%** and **34.8%** of the total, respectively. Despite the high default rate in the **Poor** category, only **13%** of applicants fall into this group. On the other end of the spectrum, just **1.6%** of applicants belong to the **Excellent** category, where the likelihood of default is the lowest.

It can be concluded that the credit score effectively captures borrower risk, as higher credit scores are consistently linked to lower default rates. Additionally, most applicants fall within the mid-range credit score categories, reinforcing the score's relevance in assessing creditworthiness

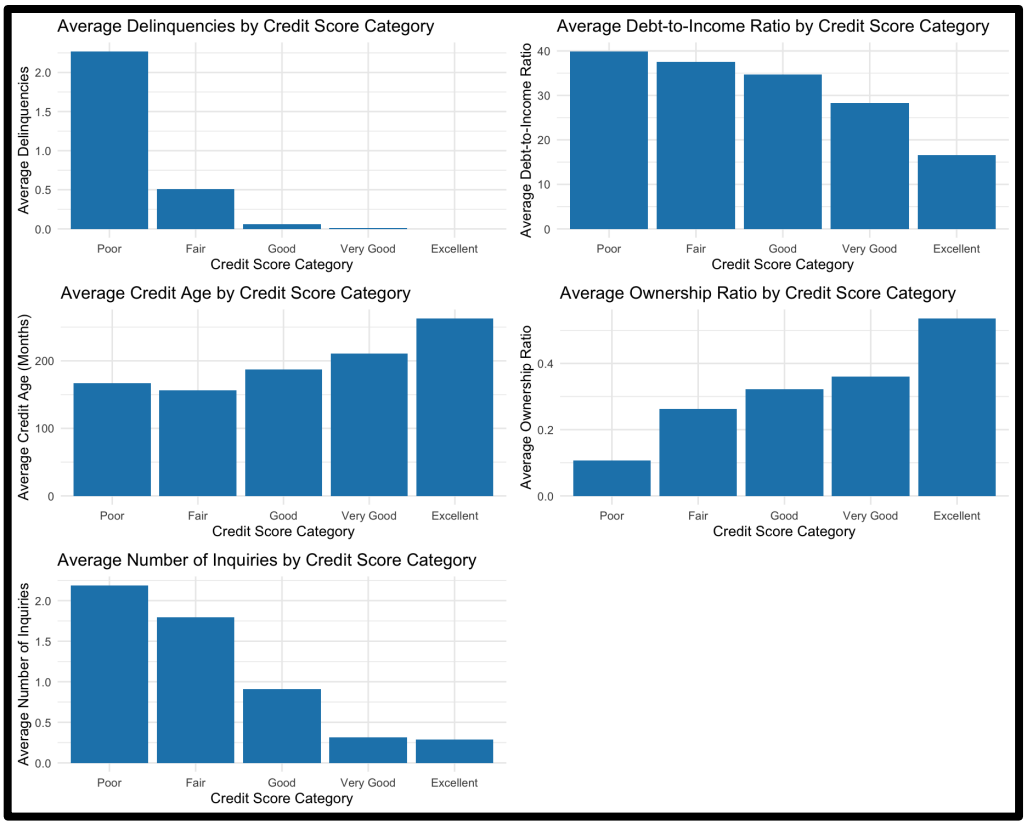

Figure 7.22: Key Credit Metrics by Credit Score Category

Figure 7.22 shows the distribution of Credit Score Categories across different credit metrics:

- **Delinquencies** significantly decrease as credit scores improve. Individuals with poor credit have the highest number of delinquencies, while those with excellent credit scores have almost none.

- **Debt-to-income ratio** is highest for those in the poor and fair credit categories, but it declines steadily as credit scores increase, indicating better financial management and lower debt burdens relative to income among higher credit score groups.

- **Credit age** shows a direct correlation with credit score. Those with excellent credit have the longest credit histories, while individuals with fair and poor credit tend to have shorter histories.

- **Ownership ratio** rises as credit scores improve. Individuals with excellent credit scores have significantly higher ownership ratios compared to those in the lower credit score categories.

- **Credit inquiries** are most frequent among individuals with poor credit, and the frequency declines with higher credit scores, suggesting that those with better credit are less reliant on frequent credit applications.

These patterns reflect expected financial behaviors, where higher credit scores are associated with fewer delinquencies, more responsible debt management, longer credit histories, higher ownership, and fewer credit inquiries.

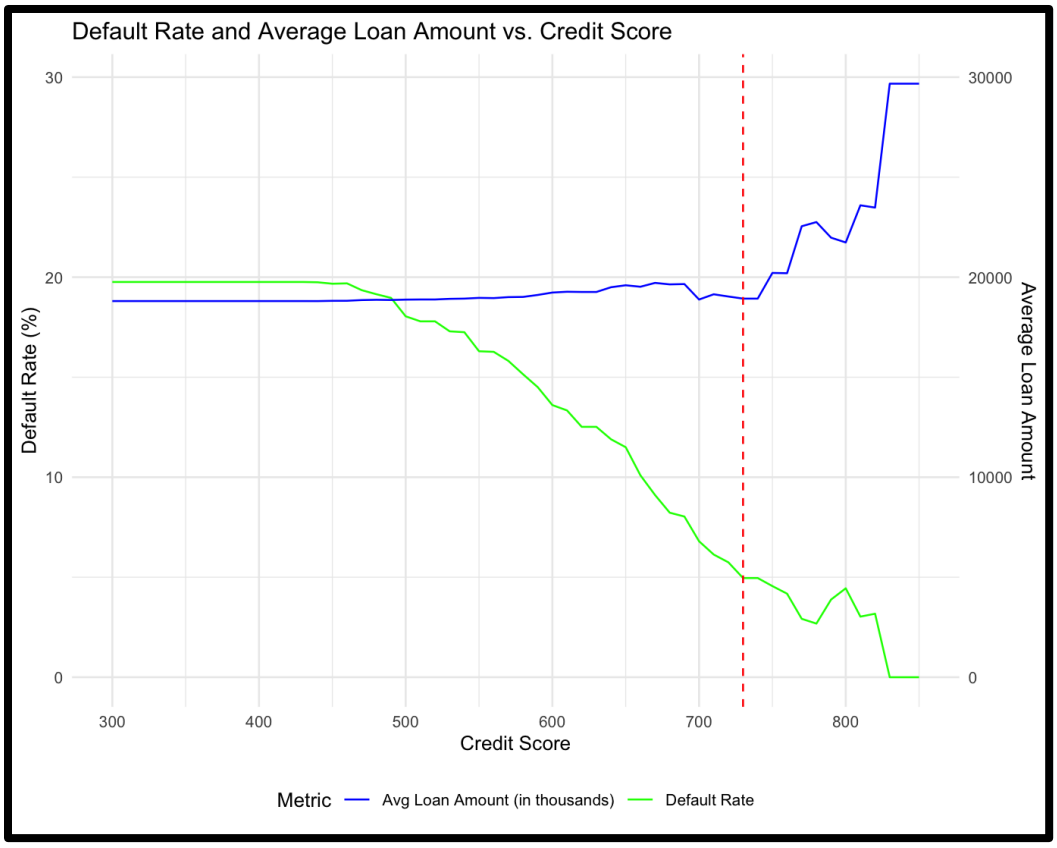

Figure 7.23: Default Rate and Average Loan Amount vs. Credit Score

Figure 7.23 highlights the relationship between credit scores, default rates, and average loan amounts. As credit scores increase, default rates drop significantly, particularly beyond the 730 threshold (marked by the red dashed line), where default rates fall below 5%. Borrowers with credit scores below 500 face a much higher default risk, while restricting loans to scores above 600 reduces the default rate to just under 15%, compared to 20% when including all credit scores.

For borrowers with credit scores below 700, average loan amounts remain stable at around $20,000. However, beyond the 730 threshold, loan amounts rise, exceeding $30,000 for scores above 800. This trend indicates that higher credit scores inspire lender confidence, resulting in larger approved loan amounts.

The 730 threshold serves as a critical point where lending risk is minimized, and borrowers are deemed low-risk, making this score an optimal benchmark for issuing larger loans with confidence.

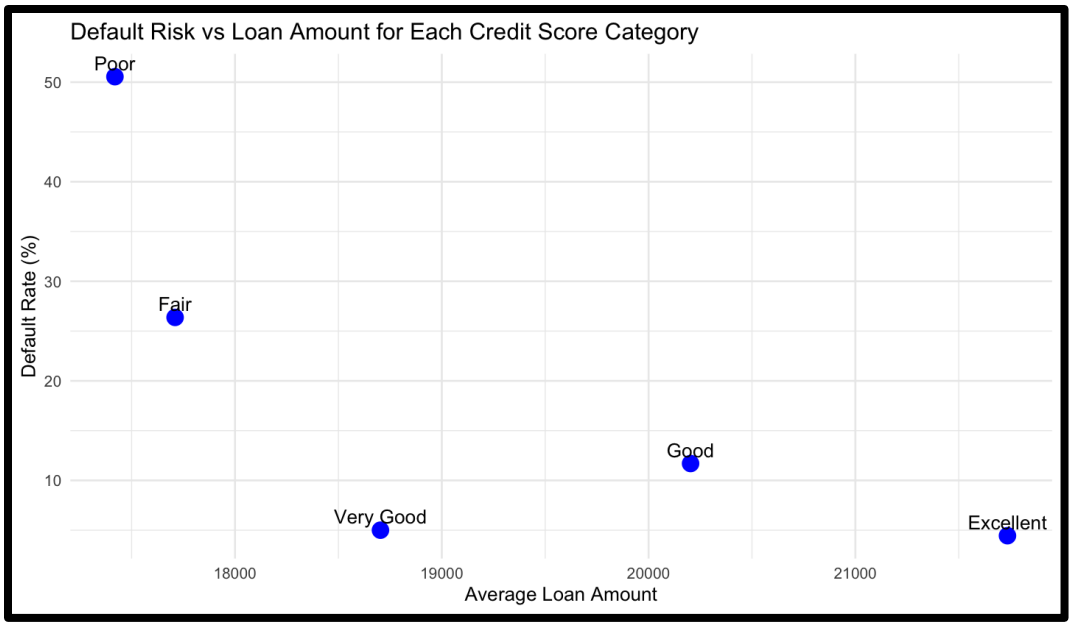

Figure 7.24: Default Risk vs Average Loan Amount for Each Credit Score Category

Figure 7.24 shows the relationship between creditworthiness categories, their average default rates, and average loan amounts.

Poor Score Category:

- This category has the highest default rate (around 50%) and the lowest average loan amount (below 18,000).
- Customers in this category pose significant risk, the bank should consider tightening loan approval criteria and charging higher interest rates to compensate for the elevated default risk.

Fair Score Category:

- Borrowers in this category have a default rate of approximately 30%, which is significantly lower than the "Poor" category but still represents moderate risk.
- These borrowers receive slightly higher average loan amounts than "Poor" borrowers, but additional measures like stricter loan terms, higher interest rates, or requiring higher collateral may help manage this risk.

Good Score Category:

- Borrowers in this category have a relatively low default rate (around 12%) and an average loan amount of approximately 20,000. Default risk is considerably lower

than the “Poor” and "Fair" categories but still slightly elevated compared to "Very Good" and "Excellent" groups.

- While these borrowers are generally lower risk, the bank should monitor this category closely and may consider implementing slightly higher interest rates , collateral or limited loan amounts to mitigate risk.

Very Good Score Category:

- This category exhibits a lower default rate (around 5%) and an average loan amount close to 19,000.

- Borrowers in this group are low risk and obtain lower loan amounts than "Good" borrowers. Offering competitive loan terms to attract these customers could further reduce overall risk in the bank's portfolio.

Excellent Score Category:

- This group has the lowest default rate (around 4.4%) and the highest average loan amount (over 21,000).

- Customers in this category are the most creditworthy, and the bank should prioritize offering them larger loan amounts with highly favorable terms to retain and attract more borrowers from this group.

The information in this plot supports implementing a risk-based pricing strategy. The bank should impose stricter eligibility criteria, increase interest rates, and require higher collateral for higher-risk borrowers ("Poor" and "Fair") to mitigate default risk and financial loss. Conversely, the bank can target lower-risk borrowers ("Good," "Very Good," and "Excellent") with larger loans and more favorable terms, as these groups offer a better balance of risk and return.

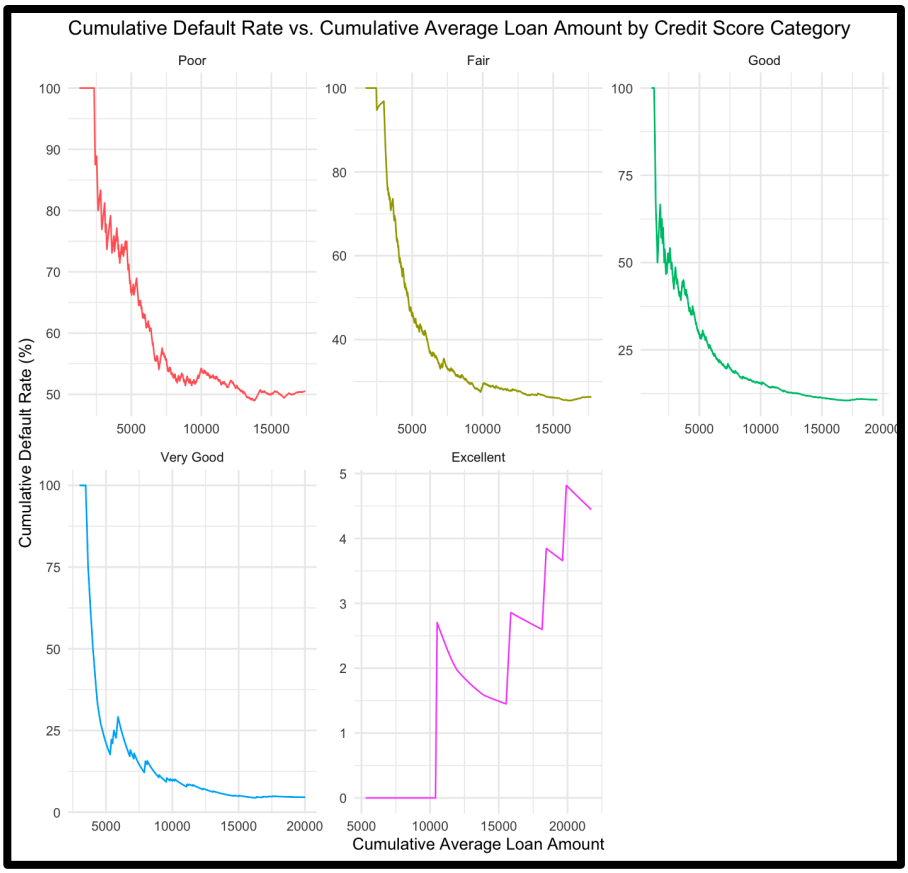

Figure 7.25: Cumulative Default Rate vs. Cumulative Average Loan Amount by Credit Score Category

The plot ***(Figure 7.25)*** shows the cumulative default rate versus cumulative average loan amount for each credit score category. Each panel represents a different credit score group: Poor, Fair, Good, Very Good, and Excellent. The y-axis shows the cumulative default rate (as a percentage), while the x-axis shows the cumulative average loan amount.

Poor Credit Score (Red):

- The default rate starts very high, near 100%, and gradually decreases as the cumulative average loan amount increases.
- For loans under $5,000, the default rate drops sharply but still remains high (above 50%) even as the cumulative loan amount increases toward $15,000.
- This suggests that borrowers with poor credit scores are much more likely to default, especially on smaller loans.

Fair Credit Score (Yellow):

- The pattern is similar to the "Poor" category but with lower starting default rates.
- The default rate drops from 100% to around 30% as the cumulative loan amount increases beyond $5,000. Beyond $10,000, the default rate stabilizes below 30%.
- Borrowers in this category are also risky, but there is a clear reduction in risk as the average loan amount increases.

Good Credit Score (Green):

- Borrowers with "Good" credit scores start with a default rate of 100%, which quickly drops below 25% by the time loan amounts reach $5,000.
- The default rate continues to decline steadily as the cumulative loan amount increases, approaching zero for higher loan amounts.
- This group shows a significant reduction in risk with higher loan values, suggesting that these borrowers are generally lower risk.

Very Good Credit Score (Blue):

- Default rates in this group start around 75% for very small loans, but quickly drop below 10% as the average loan amount exceeds $5,000.
- By $10,000, the default rate approaches zero, indicating that borrowers with "Very Good" credit scores are far less likely to default, even on smaller loans.
- The overall risk is minimal across all loan ranges.

Excellent Credit Score (Purple):

- In this group, the default rate starts very low, under 5%, even for the smallest loans.
- The trend is less smooth compared to other categories, but the default rate remains extremely low, hovering between 1% and 5%.
- Borrowers with excellent credit scores are extremely unlikely to default, regardless of loan amount.

General Insights:

- **Risk Reduction with Higher Loans**: For lower credit score categories (Poor, Fair, and Good), there is a clear pattern where the default rate decreases as the cumulative average loan amount increases. This suggests that offering larger loans to these groups may reduce risk, possibly because only more financially stable individuals receive larger loans.

- **Low Risk for High Credit Scores**: For borrowers in the "Very Good" and "Excellent" categories, the risk of default is already low and remains low regardless of the loan amount. This reinforces that credit score is a strong predictor of default risk.

- **Thresholds for Lending**: The cumulative default rate curves suggest that risk is highly concentrated in smaller loans for the lower credit score groups. As a result, lending strategies could focus on reducing risk by limiting small loan amounts for individuals with low credit scores.

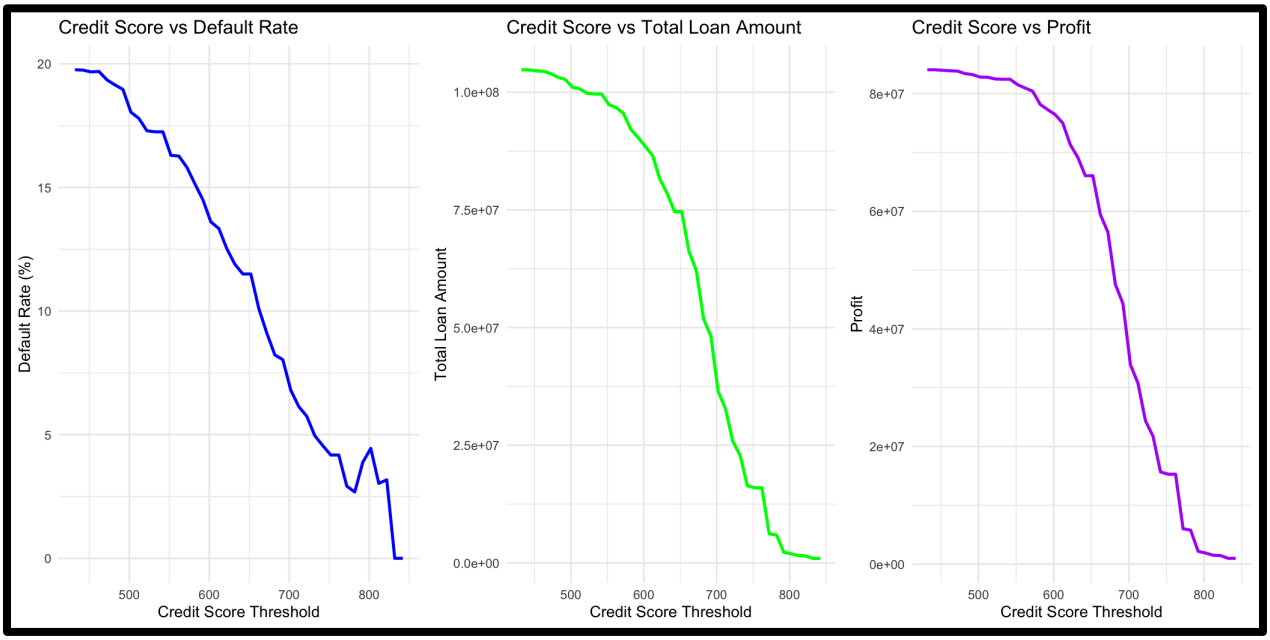

Figure 7.26: Impact of Credit Score Thresholds on Default Rate, Total Loan Amount, and Profit

The set of three plots in ***Figure 7.26*** illustrates the relationship between credit score thresholds and key metrics: default rate, total loan amount issued, and profit.

In the first plot, **Credit Score vs Default Rate**, there is a clear downward trend. As the credit score threshold increases, the default rate decreases significantly. For scores below 600, the default rate remains relatively high. However, as the credit score surpasses 732, the default rate drops to around 5%, indicating a strong relationship between higher credit scores and lower default risk.

The second plot, **Credit Score vs Total Loan Amount**, shows a steady decline in the total loan amount issued as the credit score threshold increases. This suggests that a higher credit score threshold excludes more borrowers, leading to fewer loans being issued. At lower credit scores, a large volume of loans is approved, but as the threshold tightens (particularly beyond 700), the number of loans significantly decreases.

In the third plot, **Credit Score vs Profit**, there is a similar pattern to the total loan amount. Profit decreases as the credit score threshold increases. Profit was calculated as the total loan amount issued minus the total amount lost to defaults, with no interest rates involved. Essentially, profit here represents the net amount the bank would retain after accounting for defaults. At lower credit scores (below 600), the profit remains high, but as the threshold increases profit sharply declines, reflecting fewer loans being issued, despite the lower default risk.

These plots highlight a trade-off: as credit score thresholds increase, both default rates and profit decrease. While higher credit scores reduce default risk, they also result in fewer loans and lower overall profit.

Section 8: Predictor Analysis and Relevancy

The goal of this phase of the project is to analyze and identify the key factors that affect the target variables. This requires closely reviewing the available data and determining which variables are most important for the model's accuracy. By selecting the most relevant predictors, we can create a strategy that is effective and meets the expectations of stakeholders, laying the groundwork for the next steps in the project.

Category	Variable Names
Target Variable	DEFAULT
Available Predictors	• MORTDUE, VALUE, REASON, JOB, YOJ, DEROG, DELINQ, CLAGE, NINQ, CLNO, DEBTINC, EQUITY, POTENTIAL_LOSS, ACTUAL_LOSS, LTE_RATIO, yoj_to_loan_ratio, LTV_RATIO, OWNERSHIP, and CREDIT_SCORE.

Table 8.1: Target Variable and Available Predictors

8.1 Correlation Matrix:

A correlation matrix is a table that displays the correlation coefficients between multiple variables. Each cell in the matrix shows the correlation between two variables, helping to identify relationships or patterns among them.

In a heatmap representation of the correlation matrix, colors are used to represent the strength of correlations. Dark red shows strong positive correlations, dark blue shows strong negative correlations, and white or light colors indicate weak or no correlation.

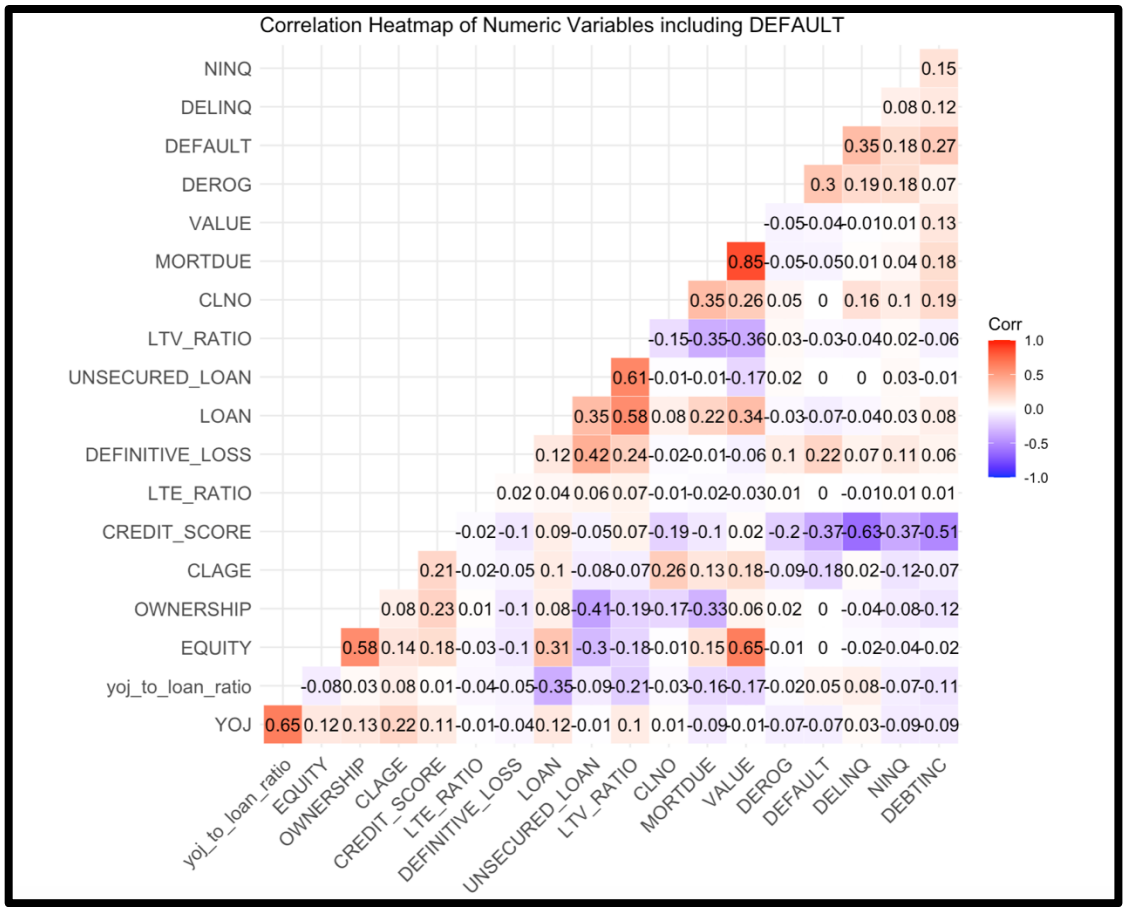

Figure 8.1: Correlation Heatmap

8.1.1 Key Correlation with Default:

- **DELINQ (0.35):**

  This moderate positive correlation suggests that past delinquencies play an important role in predicting default. The more delinquencies an applicant has, the more likely they are to default on a loan, aligning with common credit assessment practices.

- **DEROG (0.30)**:

  This indicates a moderate impact of derogatory marks on an applicant's credit history (like past bankruptcies) on the likelihood of default. The presence of derogatory remarks increases the risk of default.

- **DEBTINC (0.27)**:

  This suggests that as the debt-to-income ratio increases, the likelihood of default

also increases, though it's a weaker relationship compared to DELINQ and DEROG. Applicants with higher debt relative to income are more likely to default.

8.1.2 Less Significant Correlations with Default:

- **LOAN (-0.07)**:

  The weak negative correlation indicates that loan amount is not a strong predictor of default. This suggests that whether a loan is large or small doesn't significantly affect the likelihood of default. In your project, this supports the idea that focusing on reducing loan amounts for potential defaulters may not meaningfully reduce default risk.

- **MORTDUE (-0.05)** and **VALUE (-0.04)**:

  These variables have very weak negative correlations with default, indicating that the remaining mortgage balance and property value aren't key drivers of default either. While they might be important for other purposes (like determining equity or loan size), they are not strong indicators of whether a client will default.

- **YOJ (-0.07)**:

  This weak correlation shows that the number of years an applicant has been employed is not a major factor in predicting default risk in your dataset. Longer job tenure may not significantly reduce the likelihood of default.

8.1.3 Strong Correlations Among Predictors:

- **MORTDUE and VALUE (0.85)**:

  This strong positive correlation makes sense because the amount of mortgage due is usually closely tied to the value of the property. Properties with higher values tend to have larger mortgages, but neither of these variables strongly predicts default.

- **LOAN and LTV_RATIO (0.58)**:

  This moderate correlation shows that loan amount is significantly related to the loan-to-value ratio, which is expected. Higher loan amounts relative to property value increase the LTV ratio.
- **EQUITY and VALUE (0.65)**:

  Equity is strongly correlated with the value of the property, which makes sense since equity is a function of the property's value minus what is owed. However, EQUITY itself shows a very weak relationship to DEFAULT.

8.1.4 Credit Score and Default:

- **CREDIT_SCORE and DEFAULT (-0.37)**:

  A moderate negative correlation indicates that a higher credit score reduces the likelihood of default. This aligns well with the common understanding that credit scores serve as an important metric of financial health and stability.

8.1.5 Conclusion

The strongest correlations with default are observed for DELINQ (delinquencies), DEROG (derogatory marks), and DEBTINC (debt-to-income ratio), suggesting these factors may be important predictors of default. These variables reflect the applicant's financial behavior and ability to manage debt. In contrast, variables such as loan amount, mortgage due, and property value have weaker correlations with default, indicating they may be less significant individually.

The model for loan approval should likely prioritize factors related to financial behavior (e.g., delinquencies, derogatory marks, and debt-to-income ratio) over financial assets (e.g., loan amount and property value). This aligns with the idea that an applicant's past behavior with debt and current debt load relative to income are often more indicative of default risk than the size of the loan or property value. However, as these conclusions are based solely on correlations, further analysis using multivariate models is needed to validate their predictive power and to account for potential interactions and non-linear relationships.

Section 9: Data Partitioning

Data partitioning plays a vital role in the development of predictive models. Correctly dividing the data ensures that the model learns effectively and can generalize well to new, unseen data. In this phase of the project, the dataset will be split into three subsets: training, validation, and testing. Each subset serves a distinct purpose and is critical for building and deploying a reliable model.

9.1 Training Set

The training set forms the core of model development. It is the portion of data used to teach the model by identifying patterns and relationships between the predictors and the target variable. By adjusting its parameters through exposure to various examples, the model enhances its ability to make accurate predictions.

Purpose: To train the model by learning from the data provided.

9.2 Validation Set

The validation set is key to evaluating the model's performance after the training process. This data subset, which the model has not previously encountered, is used to compare different models and fine-tune their hyperparameters. Assessing the model on the validation set helps in selecting the best model for the task at hand. Additionally, the validation set helps prevent overfitting, where a model performs exceptionally well on training data but struggles to generalize to new data.

Purpose: To evaluate the model's performance and adjust hyperparameters.

9.3 Testing Set

Once the models have been trained and validated, the testing set is used for the final evaluation. This dataset is reserved to provide an unbiased assessment of how well the chosen model generalizes to unseen data. This step is crucial to confirm the model's effectiveness and reliability before its deployment in real-world applications.

Purpose: To provide an unbiased estimate of the final model's performance on unseen data.

9.4 Random Sampling

Random sampling techniques are applied to ensure that each subset reflects the overall dataset, minimizing biases and ensuring diversity in the data across the training, validation, and testing sets. In cases of imbalanced datasets, stratified sampling may be used to maintain consistent target class distributions across subsets.

Importance: Random sampling reduces bias and ensures the subsets are representative of the whole dataset, contributing to more reliable model evaluation.

9.5 Partition Summary

For this project, the dataset was partitioned into the following proportions:

- **Training Set:** 70% (3900 records)
- **Validation Set:** 20% (1114 records)
- **Testing Set:** 10% (558 records)

By following this approach to data partitioning, it is possible to build models that are robust, generalizable, and capable of making accurate predictions in real-world settings. For this dataset, the data was divided into 70% training, 20% validation and 10% testing.

Section 10: Feature Selection

The feature selection phase is an essential step in the modeling process, aimed at identifying the most relevant features that contribute to the model's predictive power. By selecting the most informative features and removing redundant or irrelevant ones, the objective is to improve model performance, enhance interpretability, and reduce the risk of overfitting.

During this phase, the dataset will be analyzed for potential **multicollinearity**, and the importance of each feature will be assessed using techniques such as **Variance Inflation Factor (VIF)** and feature importance algorithms like **Boruta**. The goal is to retain features that provide the most value to the model while discarding those that offer little or no benefit, ensuring a more efficient and accurate predictive model.

10.1 Exclusion of Outcome-Dependent Variable

During the modeling process, **Definitive Loss**, which was created for data exploration, will be removed as they do not contribute meaningfully to predictive performance. **Definitive Loss** is non-zero only for cases where default has occurred and remains zero for all non-defaulted cases. Therefore, this variable is highly dependent on the target variable, **DEFAULT**, as it represents outcomes of whether the borrower has defaulted or not.

Because these variables are directly linked to the outcome (default), they offer little independent predictive value. Instead, they represent the consequences of the default, not factors leading to it.

10.2 Addressing Redundancy and Multicollinearity

Highly correlated variables, such as **value** and **mortgage**, also require attention. Since **equity** and **ownership** are derived variables that capture the relevant information from a lender's perspective (specifically, whether there is sufficient value to recover in case of default and how much of the property an applicant truly owns) retaining these variables and removing **value** and **mortgage** is preferred to reduce redundancy and improve model clarity.

Additionally, **credit score,** although a useful indicator of creditworthiness, was excluded from the model. This decision was based on the fact that credit score in this dataset is derived from existing variables like delinquencies, debt-to-income ratio, and number of inquiries. Since these factors are already accounted for, including credit score would introduce redundancy without offering much additional predictive value. By focusing on the underlying financial behaviors, the model remains more streamlined and avoids unnecessary duplication of information.

To ensure the remaining variables contribute effectively to the model without introducing multicollinearity, a Variance Inflation Factor (VIF) analysis was conducted. This step helps to

identify any potential multicollinearity issues among the predictors, which can obscure the true relationships between variables and the target. By examining the VIF values, it is possible to determine whether any variables should be further adjusted or removed to improve the model's robustness and interpretability.

VIF quantifies how much the variance of a regression coefficient is inflated due to the correlation with other predictors. A VIF of 1 means no multicollinearity, while higher values indicate a growing degree of multicollinearity.

- **VIF < 5**: Generally considered acceptable. The multicollinearity is low, and the predictors are likely independent enough to keep in the model.
- **VIF between 5 and 10**: Moderately concerning. This range suggests that multicollinearity is present and may start to affect the stability and interpretability of the model.
- **VIF > 10**: Typically considered problematic. At this level, multicollinearity is high, meaning that the predictor is highly correlated with other predictors. This can make model coefficients unstable and reduce the model's reliability.

Variable	GVIF	Df	GVIF^(1/(2*Df))
LOAN	2.840918	1	1.685502
REASON	1.183926	1	1.088083
JOB	1.309384	5	1.027322
YOJ	2.737901	1	1.654660
DEROG	1.047420	1	1.023435
DELINQ	1.127500	1	1.061838
CLAGE	1.204032	1	1.097284
NINQ	1.087074	1	1.042629
CLNO	1.364823	1	1.168256
DEBTINC	1.179188	1	1.085904
EQUITY	1.595230	1	1.263024
UNSECURED_LOAN	1.843140	1	1.357623
LTE_RATIO	1.007491	1	1.003738
yoj_to_loan_ratio	3.137355	1	1.771258
LTV_RATIO	2.669152	1	1.633754
OWNERSHIP	1.576177	1	1.255459

Table 10.1: Variance Inflation Factor Results

As the table illustrates, the VIF values indicate that multicollinearity is not a significant issue in the model. No variables exceed the threshold of 5, which means the predictors can be retained without causing significant distortion in the model's predictive power or interpretation.

10.3 Boruta Feature Selection

The Boruta feature selection process identifies the most relevant predictors for the model by comparing the importance of each feature against randomized shadow features. It is particularly effective in identifying meaningful variables while avoiding the risk of discarding predictors that contribute subtly to the target variable.

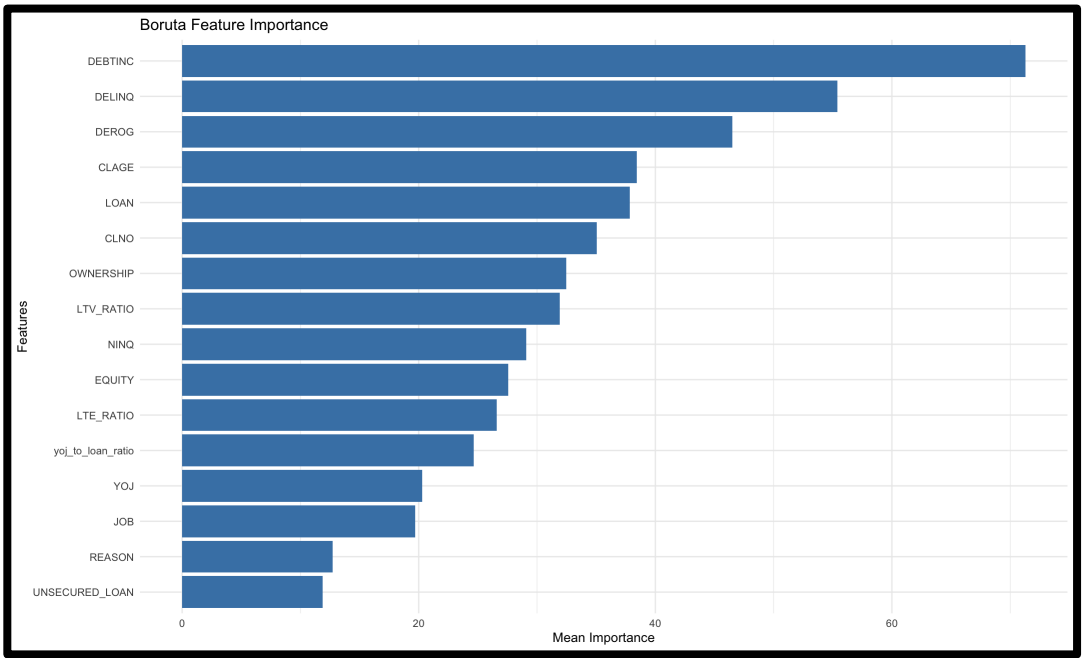

Figure 10.1: Boruta Feature Importance

The features ranked highest in importance, such as **DEBTINC**, **DELINQ**, **DEROG** and **CLAGE**, emerged as the most influential in predicting the target variable. While features like **JOB**, **YOJ**, **REASON**, and **UNSECURED_LOAN** were found to be less impactful, they still hold significance in the model. Notably, the Boruta algorithm identified all features as important, indicating that each contributes meaningfully to predicting the outcome.

Section 11: Model Selection

In the model selection phase, various classification models will be discussed to classify those who are likely to default on their loans. The models will be selected based on their ability to balance interpretability, flexibility, and predictive accuracy for both tasks.

10.1 Logistic Regression

Logistic regression is the most straightforward classification model. It predicts the probability of default based on borrower characteristics (features). The output is a probability between 0 and 1, which can be used to classify whether someone is likely to default (default if probability > 0.5, for example).

- **Pros:** Highly interpretable, easy to implement, and explains the contribution of each feature to the risk of default.
- **Cons:** Assumes linear relationships between features and the log odds of the default, which may not always be the case.

10.2 Decision Tree Classifier

Decision trees split the data into branches based on feature values. At each branch, the tree makes decisions that eventually classify an applicant as a default or non-default.

- **Pros:** Easy to visualize and interpret, handles non-linear relationships, works well with both categorical and continuous data.
- **Cons:** Can overfit the training data, leading to less generalizable predictions, but this can be mitigated with pruning or setting a maximum depth.

10.3 K-Nearest Neighbors (KNN)

KNN classifies applicants based on the majority class of their nearest "neighbors" in the feature space. For example, if most of the nearest neighbors to an applicant defaulted, the model will classify that applicant as likely to default.

- **Pros:** Simple and intuitive, no assumptions about the data distribution, handles non-linear relationships.
- **Cons:** Computationally expensive on large datasets, sensitive to the choice of k and distance metric, and less interpretable.

10.4 Random Forest

Random forest is an ensemble method that builds multiple decision trees and averages their predictions to improve accuracy and reduce overfitting. Each tree is trained on a different subset of the data, making it more robust than a single decision tree.

- **Pros**: Powerful model with high accuracy, handles both linear and non-linear relationships, reduces overfitting by averaging multiple trees.

- **Cons**: Not as interpretable as other models like logistic regression or decision trees. However, feature importance and SHAP (SHapley Additive exPlanations) can be used to understand the model's decisions by showing how much each feature contributes to the predictions.

Section 11: Model Fitting

In this section, the previously discussed algorithms will be trained on the dataset to capture patterns that help classify loan applicants as either defaulters or non-defaulters.

11.1 Logistic Regression

The logistic regression model was selected as the initial approach to predict the likelihood of a loan applicant defaulting. This method assumes a linear relationship between the predictor variables and the log-odds of default (the natural logarithm of the odds). The model estimates the log-odds as a linear function of the predictor variables, which is then transformed into a probability using the logistic function. This ensures that predicted probabilities are constrained between 0 and 1.

```
> summary(stepwise_model)

Call:
glm(formula = DEFAULT ~ LOAN + JOB + YOJ + DEROG + DELINQ + CLAGE +
    NINQ + CLNO + DEBTINC + EQUITY + yoj_to_loan_ratio + LTV_RATIO +
    OWNERSHIP, family = binomial, data = train_set)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.398e+00  3.982e-01 -13.556  < 2e-16 ***
LOAN              -2.412e-05  7.928e-06  -3.042  0.00235 **
JOBOffice         -7.672e-01  1.905e-01  -4.027 5.66e-05 ***
JOBOther           6.401e-02  1.492e-01   0.429  0.66786
JOBProfExe         9.085e-02  1.744e-01   0.521  0.60234
JOBSales           3.068e-01  3.618e-01   0.848  0.39644
JOBSelf            4.407e-01  2.939e-01   1.499  0.13380
YOJ               -6.620e-02  1.175e-02  -5.633 1.77e-08 ***
DEROG              6.688e-01  7.154e-02   9.348  < 2e-16 ***
DELINQ             6.723e-01  4.799e-02  14.010  < 2e-16 ***
CLAGE             -6.751e-03  7.369e-04  -9.161  < 2e-16 ***
NINQ               1.166e-01  2.556e-02   4.560 5.11e-06 ***
CLNO              -1.050e-02  5.610e-03  -1.873  0.06112 .
DEBTINC            1.224e-01  9.038e-03  13.547  < 2e-16 ***
EQUITY             5.836e-06  1.812e-06   3.221  0.00128 **
yoj_to_loan_ratio  7.295e+02  1.166e+02   6.255 3.98e-10 ***
LTV_RATIO          1.776e+00  4.397e-01   4.039 5.36e-05 ***
OWNERSHIP          5.672e-01  1.729e-01   3.281  0.00103 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3878.1  on 3899  degrees of freedom
Residual deviance: 2735.6  on 3882  degrees of freedom
AIC: 2771.6

Number of Fisher Scoring iterations: 6
```

Figure 11.1: Summary of Stepwise Logistic Regression Model

11.1.1 Model Formula

The logistic regression model predicts the binary target variable, "DEFAULT," using a range of predictors. These include continuous variables (e.g., LOAN, YOJ, CLAGE, etc.) and categorical variables (e.g., JOB, OWNERSHIP). For categorical variables, one category serves as the reference, and the intercept accounts for its effect. For instance, "Mgr" may be the baseline for JOB, meaning other categories like "Office" or "Sales" are compared against it.

Each coefficient represents the change in the log-odds of default for a one-unit increase in that predictor. A positive coefficient indicates that an increase in the variable raises the likelihood of default, while a negative coefficient suggests the opposite. The magnitude of a coefficient reflects the strength of its influence on default risk.

11.1.2 Standard Error

The standard error quantifies the precision of each coefficient's estimate. Smaller standard errors suggest more reliable estimates. If a predictor has a high standard error relative to its coefficient, it indicates greater uncertainty in the estimate for that variable.

11.1.3 P-Value

P-values determine the statistical significance of each predictor. A smaller p-value (generally < 0.05) implies that the predictor significantly affects the likelihood of default. Predictors with p-values less than 0.001 are considered highly significant and strongly influence default probability.

11.1.4 Significant Predictors

The most significant predictors of loan default, based on their p-values, include:

1. **JOB (Office)**: Coefficient = -0.767, $p < 0.001$. Applicants with office jobs are less likely to default compared to the reference category.

2. **YOJ (Years on Job)**: Coefficient = -0.066, $p < 0.001$. Longer tenure at a job reduces the likelihood of default, suggesting greater stability.

3. **DEROG (Derogatory Marks)**: Coefficient = 0.669, $p < 0.001$. More derogatory marks significantly increase default risk.

4. **DELINQ (Delinquencies)**: Coefficient = 0.672, $p < 0.001$. A higher number of delinquencies raises default probability.

5. **CLAGE (Credit Line Age)**: Coefficient = -0.0068, $p < 0.001$. Older credit lines lower default risk, signaling responsible credit behavior.

6. **NINQ (Number of Inquiries)**: Coefficient = 0.117, $p < 0.001$. More credit inquiries increase the likelihood of default.

7. **DEBTINC (Debt-to-Income Ratio)**: Coefficient = 0.122, $p < 0.001$. A higher debt-to-income ratio raises default probability, indicating financial strain.

8. **EQUITY**: Coefficient = 5.836e-06, $p = 0.001$. While small, increased equity slightly raises default probability, potentially due to over-leveraging.

9. **yoj_to_loan_ratio**: Coefficient = 729.5, $p < 0.001$. A higher ratio of years on the job to loan amount increases default risk.

10. **LTV_RATIO (Loan-to-Value Ratio)**: Coefficient = 1.776, $p < 0.001$. Higher loan-to-value ratios raise default risk, reflecting greater financial vulnerability.

11. **OWNERSHIP**: Coefficient = 0.567, $p = 0.001$. Ownership of assets increases default likelihood, potentially due to liquidity concerns.

11.1.5 Deviance

Deviance measures the model's goodness of fit. The null deviance (3878.1) reflects the model without predictors, while the residual deviance (2735.6) represents the model with predictors. The significant reduction in deviance indicates that the predictors improve the model's fit.

11.1.6 AIC (Akaike Information Criterion)

The AIC (2771.6) evaluates the model's quality by balancing goodness of fit and complexity. Lower AIC values indicate better model performance when comparing multiple models.

11.1.7 Summary

In summary, the logistic regression model identifies key predictors of loan default, including job type, years on the job, derogatory marks, delinquencies, debt-to-income ratio, loan-to-value ratio, and ownership. These variables provide meaningful insights into default risk. The model demonstrates good fit through its reduced deviance and relatively low AIC value, confirming its utility in predicting default behavior.

11.2 Decision Tree Classifier

The second model was built using the decision tree algorithm. This algorithm provides a hierarchical structure that clearly illustrates how various predictors, and their values lead to a specific outcome—whether an applicant is a defaulter (DEFAULT = 1) or not (DEFAULT = 0). As discussed earlier, decision trees make classifications by recursively splitting the data based on the feature values that result in the greatest reduction in impurity. This process helps effectively

distinguish between defaulters and non-defaulters by creating data subsets that are as homogenous as possible in relation to the target variable.

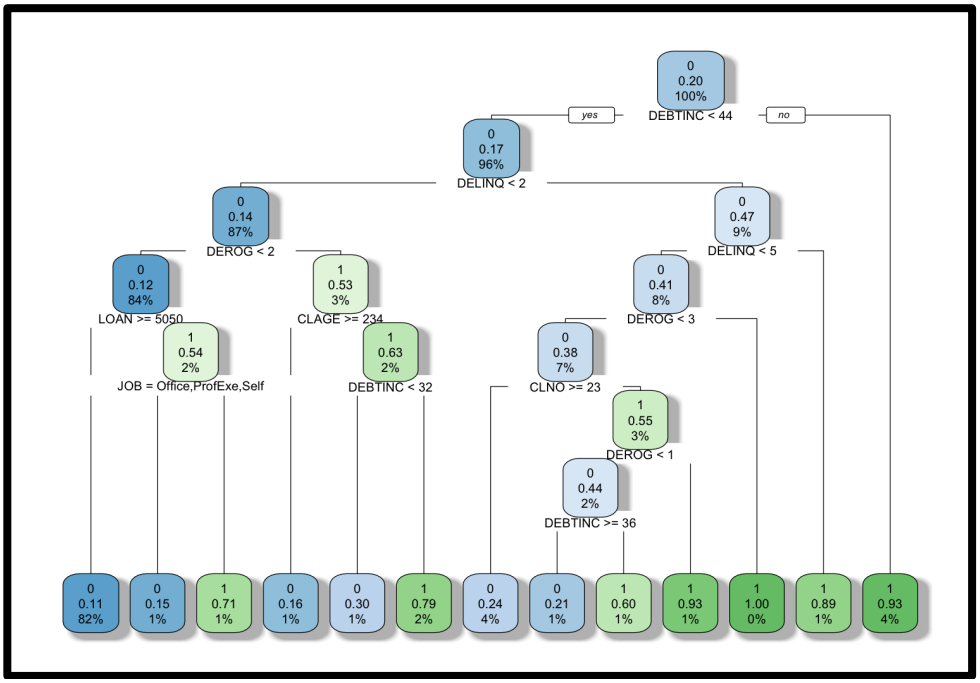

Figure 11.2: Decision Tree Model

11.2.1 Classification Tree Rules:

Rule 1: DEBTINC < 44, DELINQ < 2, DEROG < 2, LOAN ≥ 5050

- **Probability of default**: 11%
- **Records in this node**: 82%

Borrowers who have a debt-to-income ratio (DEBTINC) below 44, fewer than 2 delinquencies (DELINQ), fewer than 2 derogatory marks (DEROG), and loan amounts greater than or equal to $5050 are at low risk of default (11%). The majority of records (82%) fall into this category, indicating that these conditions are associated with lower risk.

Rule 2: DEBTINC < 44, DELINQ < 2, DEROG < 2, LOAN < 5050, JOB = Office, ProfExe, Self

- **Probability of default**: 15%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) below 44, fewer than 2 delinquencies (DELINQ), fewer than 2 derogatory marks (DEROG), loan amounts under $5050, and jobs categorized as Office, ProfExe, or Self have a 15% probability of default. This rule applies to only 1% of the records.

Rule 3: DEBTINC < 44, DELINQ < 2, DEROG < 2, LOAN < 5050, JOB ≠ Office, ProfExe, Self

- **Probability of default**: 71%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) below 44, fewer than 2 delinquencies (DELINQ), fewer than 2 derogatory marks (DEROG), loan amounts under $5050, and jobs not categorized as Office, ProfExe, or Self have a 71% probability of default. However, this rule applies to only 1% of the records, representing a small subset of borrowers.

Rule 4: DEBTINC < 44, DELINQ < 2, DEROG ≥ 2, CLAGE ≥ 234

- **Probability of default**: 16%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) below 44, fewer than 2 delinquencies (DELINQ), 2 or more derogatory marks (DEROG), and a credit line age (CLAGE) of 234 months or more have a 16% probability of default. This rule applies to only 1% of the records.

Rule 5: DEBTINC < 32, DELINQ < 2, DEROG ≥ 2, CLAGE < 234

- **Probability of default**: 30%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) below 32, fewer than 2 delinquencies (DELINQ), 2 or more derogatory marks (DEROG), and a credit line age (CLAGE) of less than 234 months have a 30% probability of default. This rule applies to only 1% of the records, identifying a small subset of borrowers.

Rule 6: 44 > DEBTINC ≥ 32, DELINQ < 2, DEROG ≥ 2, CLAGE < 234

- **Probability of default**: 79%
- **Records in this node**: 2%

Borrowers with a debt-to-income ratio (DEBTINC) below 44 but greater than or equal to 32, fewer than 2 delinquencies (DELINQ), 2 or more derogatory marks (DEROG), and a credit line age (CLAGE) of less than 234 months have a 79% probability of default. This rule applies to 2% of the records.

Rule 7: DEBTINC < 44, 2 ≤ DELINQ < 5, DEROG <3, CLNO ≥ 23

- **Probability of default**: 24%
- **Records in this node**: 4%

Borrowers with a debt-to-income ratio (DEBTINC) below 44, between 2 and 5 delinquencies (2 ≤ DELINQ < 5), fewer than 3 derogatory marks (DEROG), and at least 23 credit lines (CLNO ≥ 23) have a 24% probability of default. This rule applies to 4% of the records.

Rule 8: 36 ≤ DEBTINC < 44, 2 ≤ DELINQ < 5, DEROG < 1, CLNO < 23

- **Probability of default**: 21%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) between 36 and 44 (inclusive of 36 but less than 44), between 2 and 5 delinquencies (2 ≤ DELINQ < 5), no derogatory mark (DEROG < 1), and fewer than 23 credit lines (CLNO < 23) have a 21% probability of default. This rule applies to only 1% of the records.

Rule 9: DEBTINC < 36, 2 ≤ DELINQ < 5, DEROG < 1, CLNO < 23

- **Probability of default**: 60%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) below 36, between 2 and 5 delinquencies (2 ≤ DELINQ < 5), no derogatory mark (DEROG < 1), and fewer than 23 credit lines (CLNO < 23) have a 60% probability of default. This rule applies to only 1% of the records

Rule 10: DEBTINC < 44, 2 ≤ DELINQ < 5, 1≤ DEROG < 3, CLNO < 23

- **Probability of default**: 93%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) below 44, between 2 and 5 delinquencies (2 ≤ DELINQ < 5), between 1 and 3 derogatory marks (1 ≤ DEROG < 3), and fewer than 23 credit lines (CLNO < 23) have a 93% probability of default. This rule applies to only 1% of the records.

Rule 11: DEBTINC < 44, 2 ≤ DELINQ < 5, DEROG ≥ 3

- **Probability of default**: 100%
- **Records in this node**: 0.47%

Borrowers with a debt-to-income ratio (DEBTINC) below 44, between 2 and 5 delinquencies (2 ≤ DELINQ < 5), and 3 or more derogatory marks (DEROG ≥ 3) have a 100% probability of default. This rule applies to 18 (0.47%) records

Rule 12: DEBTINC < 44, DELINQ ≥ 5,

- **Probability of default**: 89%
- **Records in this node**: 1%

Borrowers with a debt-to-income ratio (DEBTINC) below 44 and 5 or more delinquencies (DELINQ ≥ 5) have an 89% probability of default. This rule applies to 1% of the records.

Rule 13: DEBTINC ≥ 44

- **Probability of default**: 93%
- **Records in this node**: 4%

Borrowers with a debt-to-income ratio (DEBTINC) of 44% or higher have a 93% probability of default. This rule applies to 4% of the records.

11.2.2 Conclusion:

The model categorizes borrowers into groups based on their financial behavior and credit history. Key factors such as debt-to-income ratio, number of delinquencies, derogatory marks, loan size, and credit line age play a critical role in determining default risk. Borrowers with lower debt-to-income ratios, fewer delinquencies, and older credit lines are associated with lower default probabilities, whereas those with higher debt ratios, more delinquencies, or shorter credit histories face higher risks. The model highlights critical decision rules that can guide lenders in assessing default risk more effectively.

11.3 K-Nearest Neighbors (KNN)

The third model created was based on the K-Nearest Neighbors (KNN) algorithm, where the goal is to identify the **k** records in the training set that are most similar to the new records being classified. Similarity is determined using Euclidean distance, which measures the straight-line distance between data points across all features. Once the nearest neighbors are identified, the classification of new records is based on the majority class among these neighbors.
The data was standardized instead of normalized, due to the skewed distributions and the presence of many outliers. Standardization was used to ensure that features with extreme values or skewness did not dominate the distance calculations, allowing for more balanced and reliable predictions.

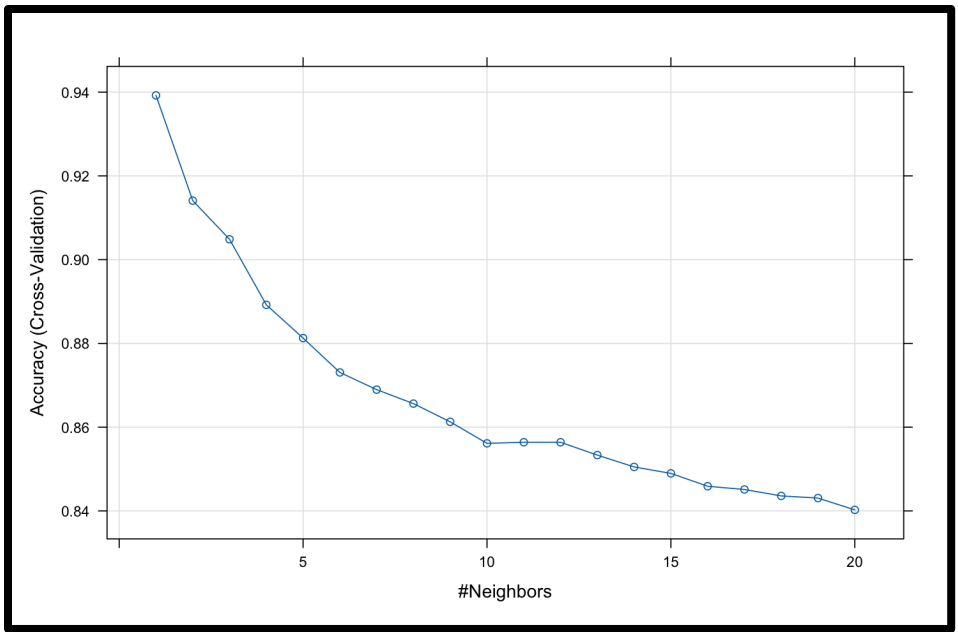

Figure 11.3: Accuracy Variation with Different K-Neighbors

For this model, k=5 was chosen to strike a balance between reducing the influence of noise and maintaining a local focus on the nearest neighbors. The plot indicates that while the accuracy is highest at k=1 (~94%), smaller k values are more sensitive to individual data points, which can lead to overfitting and instability in predictions. At k=5, the cross-validation accuracy remains high (~88%), but the model becomes less influenced by outliers or noise in the data. This choice ensures more stable and robust predictions while still capturing important local patterns effectively.

11.4 Random Forest

The last model was developed using the Random Forest algorithm. This approach works by creating multiple decision trees, each built from randomly chosen subsets of features and data samples. The randomness in both feature and sample selection allows the model to capture various patterns and interactions within the data. As a result, it minimizes the risk of overfitting and improves the model's generalizability, providing more reliable predictions compared to using a single decision tree. This ensemble method ensures a stronger and more stable performance by averaging the outcomes from multiple trees.

```
> print(rf_model)

Call:
 randomForest(formula = DEFAULT ~ . - EQUITY, data = train_set,      ntree = 500, importance = TRUE, classwt = c(`0` = 0.3, `1` = 0.7))
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 9.67%
Confusion matrix:
     0   1 class.error
0 3121   8 0.002556727
1  369 402 0.478599222
```

Figure 11.4: Random Forest Model Summary

The summary of the model in ***Figure 11.4*** shows that the target variable is "DEFAULT," and 500 classification trees were created, with 3 randomly selected variables considered at each split. In this instance, **class weights were adjusted**, assigning a weight of **0.3** to non-defaulters (class 0) and **0.7** to defaulters (class 1)**.** This adjustment was made to account for the imbalance between the two classes, giving more importance to defaulters in the classification process.

The Out-Of-Bag (OOB) estimate of the error rate is **9.67%**, indicating that the model performs well overall. The confusion matrix shows that the model misclassified **0.25%** of the non-defaulters as defaulters, while **47.85%** of the defaulters were misclassified as non-defaulters.

This suggests that while the model performs very well in predicting non-defaulters, its error rate is higher for defaulters, even with the adjusted class weights. This highlights the ongoing challenge in accurately identifying defaulters, though the class weighting helps improve the model's focus on this minority class.

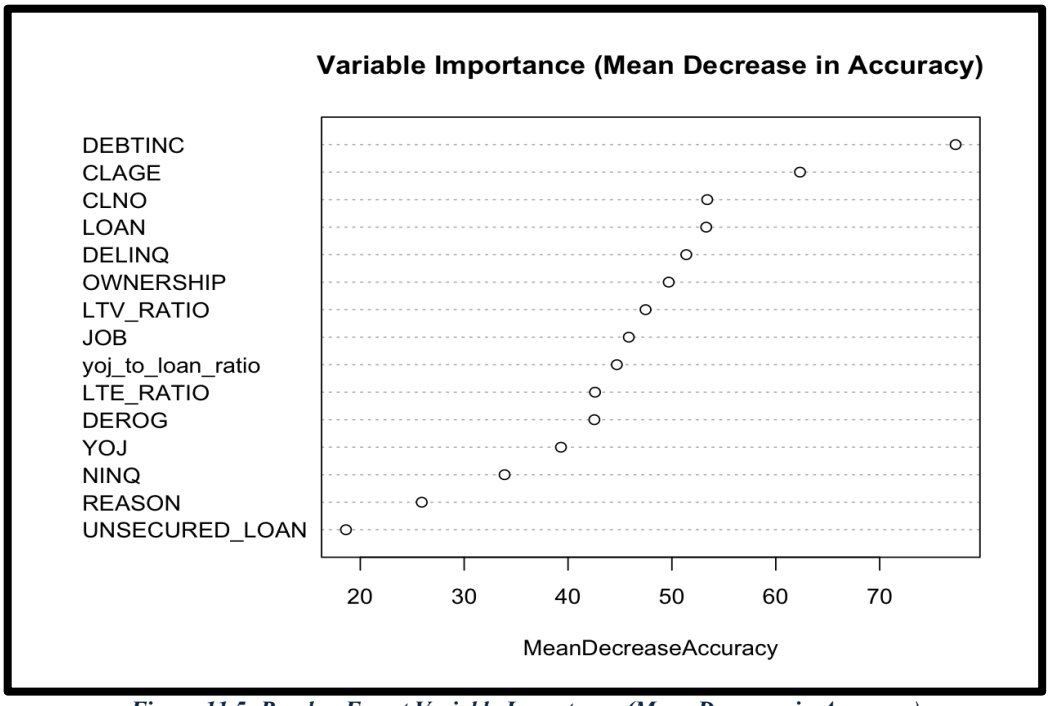

Figure 11.5: RandomForest Variable Importance (Mean Decrease in Accuracy)

This plot **(Figure 11.5)** displays the importance of the predictors used in the model based on the **Mean Decrease in Accuracy**, which evaluates how much the model's accuracy decreases when a specific predictor is randomly shuffled. This disruption breaks the relationship between the predictor and the target variable, providing insight into the predictor's contribution to the model. Simply put, higher values indicate that the variable plays a significant role in improving the accuracy of the model.

It is evident that **DEBTINC** (debt-to-income ratio), **CLAGE** (credit age), and **CLNO** are among the most important predictors, meaning they greatly influence the model's ability to predict loan defaults. On the other hand, features like **REASON** (reason for the loan), **NINQ** (number of inquiries), and **UNSECURED_LOAN** contribute less to the overall accuracy of the model.

In summary, identifying which variables most impact model accuracy not only helps improve performance but also provides valuable transparency. By highlighting the key predictors, the model aligns with the business goal of developing a high-performing yet transparent loan approval system, ensuring decision-makers understand which factors influence default predictions the most.

Section 12: Performance Evaluation

With all models trained, the next step is to assess their performance and determine which one offers the best results. This will be accomplished by applying the models to the validation set, which includes unseen data with known outcomes. This process allows for comparing the predicted and actual classifications. The validation set consists of **1,124** observations.

The evaluation will focus on key metrics such as **accuracy**, **sensitivity**, and **specificity**. These metrics will be obtained from the **confusion matrix**, which provides a summary of the model's predictions versus the true outcomes. The positive class in this analysis is **1**, which represents defaulters. Based on these metrics, the model that best predicts loan defaults will be selected. Based on these comparisons, the model that best predicts loan defaults will be selected.

1. **Accuracy**: This measures the overall correctness of the model, showing the proportion of correctly classified instances (both defaulters and non-defaulters) out of the total number of instances. It's useful for giving an overall sense of how well the model performs but can be misleading in imbalanced datasets, where one class dominates.

2. **Sensitivity (Recall for defaulters)**: This focuses on how well the model correctly identifies actual defaulters. It's particularly important in this case, as correctly identifying defaulters helps prevent financial loss. High sensitivity ensures that most defaulters are detected.

3. **Specificity (Recall for non-defaulters)**: This measures the model's ability to correctly classify non-defaulters. In loan default prediction, specificity ensures that reliable clients

are not mistakenly classified as defaulters, helping to maintain client trust and avoid unnecessary loan denials.

12.1 Confusion Matrices

12.1.1 Logistic Regression

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 712  59
         1 178 165

               Accuracy : 0.7873
                 95% CI : (0.762, 0.811)
    No Information Rate : 0.7989
    P-Value [Acc > NIR] : 0.8435

                  Kappa : 0.4476

 Mcnemar's Test P-Value : 1.789e-14

            Sensitivity : 0.7366
            Specificity : 0.8000
         Pos Pred Value : 0.4810
         Neg Pred Value : 0.9235
             Prevalence : 0.2011
         Detection Rate : 0.1481
   Detection Prevalence : 0.3079
      Balanced Accuracy : 0.7683

       'Positive' Class : 1
```

Figure 12.1: Confusion Matrix of Logistic Regression Model

The confusion matrix indicates an accuracy of 78.73%, showing that the model is effective in predicting loan defaults, performing better than the baseline. The threshold used for classification was set at 0.2, as it provided the best balance between sensitivity and specificity.

The model demonstrates a specificity of 80%, meaning it is good at correctly identifying non-defaulters. However, the sensitivity is 73.66%, indicating that while it performs moderately well in identifying defaulters, there is room for improvement in this area. The precision, or positive predictive value, for defaulters is relatively low at 48.10%, meaning that nearly half of the instances predicted as defaulters were incorrectly classified.

The decision threshold of 0.2 was specifically chosen to optimize the balance between detecting defaulters and maintaining overall accuracy. Although this threshold improves sensitivity, the trade-off is a lower precision specificity.

12.1.2 Classification Tree

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 824 100
         1  66 124

               Accuracy : 0.851
                 95% CI : (0.8287, 0.8714)
    No Information Rate : 0.7989
    P-Value [Acc > NIR] : 4.343e-06

                  Kappa : 0.5083

 Mcnemar's Test P-Value : 0.01043

            Sensitivity : 0.5536
            Specificity : 0.9258
         Pos Pred Value : 0.6526
         Neg Pred Value : 0.8918
             Prevalence : 0.2011
         Detection Rate : 0.1113
   Detection Prevalence : 0.1706
      Balanced Accuracy : 0.7397

       'Positive' Class : 1
```

Figure 12.2: Confusion Matrix of Classification Tree Model

The model was evaluated using a threshold of 0.2. The confusion matrix shows an accuracy of 85.1%, which is higher than the baseline, indicating that the model provides better predictions than a naive approach. The specificity is 92.58%, demonstrating that the model is very capable of correctly identifying non-defaulters. However, the sensitivity is only 55.36%, revealing that the model is not as effective at identifying defaulters. selected.

Despite the high specificity, the low sensitivity highlights the model's struggle to detect defaulters accurately, which is crucial in loan default prediction. The positive predictive value is 65.26%, meaning that when the model predicts a default, it is correct in most cases. However, it misses many actual defaulters, leading to an overall balanced accuracy of 73.97%.

12.1.3 KNN

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 889 133
         1   1  91

               Accuracy : 0.8797
                 95% CI : (0.8592, 0.8982)
    No Information Rate : 0.7989
    P-Value [Acc > NIR] : 6.333e-13

                  Kappa : 0.5197

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.40625
            Specificity : 0.99888
         Pos Pred Value : 0.98913
         Neg Pred Value : 0.86986
             Prevalence : 0.20108
         Detection Rate : 0.08169
   Detection Prevalence : 0.08259
      Balanced Accuracy : 0.70256

       'Positive' Class : 1
```

Figure 12.3: Confusion Matrix of KNN Model

The confusion matrix for the KNN model shows an accuracy of 87.97%, which is significantly higher than the baseline, indicating that the model provides strong predictions overall. The specificity is 99.88%, meaning the model is excellent at identifying non-defaulters, as it correctly classifies almost all non-defaulters in the dataset. However, the sensitivity is 40.63%, meaning the model struggles to identify defaulters, correctly detecting less than half of them.

The positive predictive value (precision) is 98.91%, suggesting that when the model predicts a default, it is almost always correct. However, the negative predictive value is 86.98%. The model achieves a balanced accuracy of 70.25%, which reflects the trade-off between sensitivity and specificity.

While the model shows strong specificity and precision, the relatively low sensitivity indicates the model is not great at detecting defaulters.

12.1.4 Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 810  16
         1  80 208

               Accuracy : 0.9138
                 95% CI : (0.8958, 0.9296)
    No Information Rate : 0.7989
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7577

 Mcnemar's Test P-Value : 1.277e-10

            Sensitivity : 0.9286
            Specificity : 0.9101
         Pos Pred Value : 0.7222
         Neg Pred Value : 0.9806
             Prevalence : 0.2011
         Detection Rate : 0.1867
   Detection Prevalence : 0.2585
      Balanced Accuracy : 0.9193

       'Positive' Class : 1
```

Figure 12.4: Confusion Matrix of Random Forest Model

The model was evaluated using a threshold of 0.2. The confusion matrix for the Random Forest model shows an accuracy of 91.38%, which is considerably higher than the baseline, indicating that the model performs well overall in predicting both defaulters and non-defaulters. The specificity is 91.01%, meaning the model is very effective at correctly identifying non-defaulters, while the sensitivity is 92.86%, showing that it also does an excellent job of identifying defaulters. A threshold of 0.2 was selected.

The positive predictive value (precision) is 72.22%, meaning that when the model predicts a default, it is correct 72.22% of the time. The negative predictive value is 98.06%, indicating that the model is highly reliable in identifying non-defaulters. The balanced accuracy of 91.93% reflects the model's strong performance in both sensitivity and specificity.

Overall, the Random Forest model provides a good balance between detecting defaulters and avoiding false positives. The high sensitivity and specificity make it the best option for predicting loan defaults.

12.2 ROC Curves

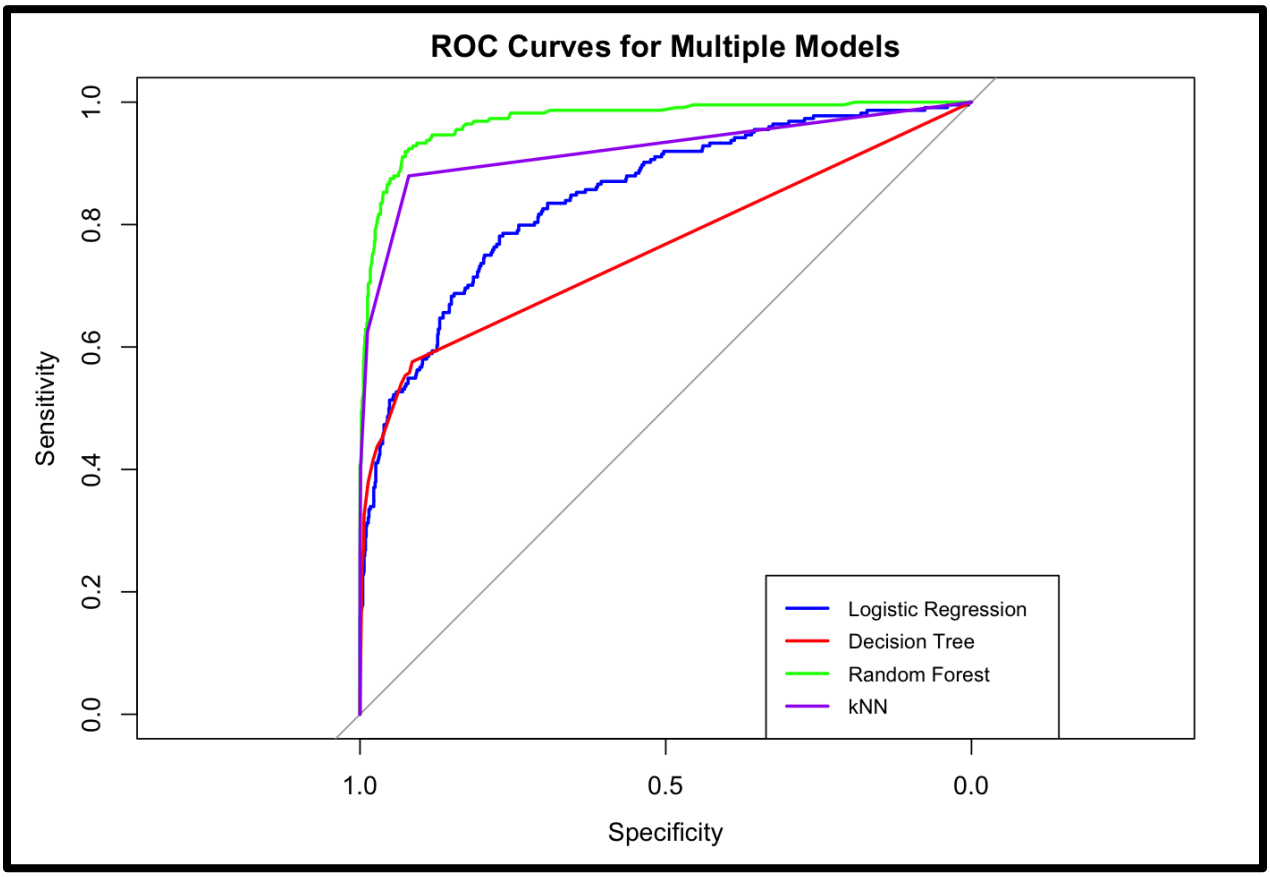

Figure 12.5: ROC Curves

The ROC (Receiver Operating Characteristic) curve visualizes the balance between sensitivity and specificity at various threshold levels, highlighting how shifts in one metric influence the other. It allows us to observe how well the model maintains specificity as sensitivity increases, showing the trade-offs between the two metrics.

This curve is particularly useful for comparing multiple models. Models that achieve high sensitivity and specificity with fewer compromises are considered more effective. The overall performance of each model is often summarized using the AUC (Area Under the Curve). AUC values range from 0 to 1, with a score of 1 indicating a perfect model. The closer the AUC is to 1, the better the model's capacity to distinguish between classes.

In the ROC plot, a diagonal line represents a random classifier with an AUC of 0.5, indicating no real predictive power. Models with AUC values higher than 0.5 demonstrate better-than-random performance.

By comparing the ROC curves for the models, Random Forest, with an AUC of 0.97, performs exceptionally well, displaying a strong ability to separate defaulters from non-defaulters. Logistic Regression follows with an AUC of 0.82, showing good performance, while

the Decision Tree, with an AUC of 0.71, performs less effectively but still offers moderate predictive power.

12.3 Performance Comparison

Method	Accuracy	Sensitivity	Specificity	Precision	AUC
Logistic Regression (0.2 Threshold)	78.73%	73.66%	80.00%	48.10%	0.8460
Decision Tree (0.2 Threshold)	85.10%	55.36%	92.58%	65.26%	0.7595
KNN	87.97%	40.63%	99.88%	98.91%	0.9216
Random Forest (0.2 Threshold)	91.38%	92.86%	91.01%	72.22%	0.9711

Table 12.1: Performance Metrics for Classification Models

The performance comparison of the models highlights that Random Forest outperforms the other methods due to its strong balance across key metrics like accuracy, sensitivity, precision, and AUC. With a sensitivity of 92.86%, Random Forest is particularly effective at correctly identifying defaulters—a critical factor in loan default prediction, where failing to capture high-risk clients can lead to substantial financial losses. While logistic regression demonstrates decent sensitivity at 73.66%, and models like KNN and decision trees perform worse in this area, they are less reliable in detecting defaulters.

In terms of precision, Random Forest also stands out with a precision of 72.22%, meaning that when the model predicts a default, it is correct most of the time. Although KNN has an impressive precision of 98.91%, its sensitivity is notably low at 40.63%, making it less dependable for identifying a sufficient number of defaulters. Random Forest, on the other hand, balances both precision and sensitivity more effectively, ensuring that its predictions are not only accurate but also actionable.

The Random Forest model's AUC of 0.9711 further underscores its superiority, as it nearly perfectly distinguishes between defaulters and non-defaulters. This high AUC confirms that Random Forest provides exceptional overall performance in separating the two classes, outperforming both the decision tree (AUC 0.7595) and logistic regression (AUC 0.8460).

Despite the challenges associated with the interpretability of Random Forest, its high sensitivity, precision, and AUC make it the most suitable model for this project. These qualities ensure that the model not only captures a large number of defaulters but also provides reliable and actionable predictions, making it the optimal choice for managing risk in loan default prediction.

Section 13: Enhancing Random Forest Transparency and Interpretability

In this phase, the focus shifts to improving the interpretability of the Random Forest model, which, despite its strong predictive performance, may be seen as not very transparent due to its complexity. Understanding how the model arrives at its decisions is crucial for gaining insights into the factors driving loan defaults and for ensuring transparency in decision-making. To achieve this, **SHAP (Shapley Additive Explanations)** will be used.

13.1 Introduction to SHAP for Interpretability

SHAP values will help explain the contribution of each feature to individual predictions, offering a clear view of how different variables influence the model's output. Meanwhile, feature importance plots will provide a broader understanding of which factors most significantly impact the model's overall predictions.

SHAP is based on the concept of **Shapley values** from cooperative game theory, which assigns a value to each feature by considering its contribution to the model's prediction. It does this by considering all possible combinations of features and their marginal contributions.

For each loan applicant, SHAP calculates a **SHAP value** for each feature, showing how much that feature is pushing the prediction towards a default (class 1) or away from it (class 0). For example, a high SHAP value for **DEBTINC** (debt-to-income ratio) might indicate that this feature is strongly contributing to a high probability of default for a particular applicant.

13.2 Explaining SHAP Values with Examples

	LOAN	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	EQUITY	UNSECURED_LOAN	LTE_RATIO	yoj_to_loan_ratio	LTV_RATIO	OWNERSHIP	rf_probabilities	shap_prob
1578	0.02196	0.00472	0.04412	0.03168	-0.00862	-0.01252	0.12096	-0.00264	0.00094	0.42110	0	6.480000e-03	0.01840	0.01288	0.01508	0.00666	0.846	0.878892308
700	0.00532	0.00046	0.01990	0.01116	-0.00768	-0.01526	-0.06652	0.08004	0.01722	0.00820	0	-2.780000e-03	-0.00556	-0.00556	0.02534	-0.00158	0.266	0.260392308
339	0.04002	0.00276	-0.01706	0.02374	-0.01986	0.07894	0.13410	0.01084	-0.01446	0.06374	0	8.200000e-03	0.00984	0.05438	-0.00894	-0.00952	0.524	0.554412308
2672	0.02548	-0.00034	-0.03486	0.02168	-0.00584	-0.01296	-0.01762	-0.00406	-0.00292	-0.00568	0	-1.522000e-02	-0.02450	0.00018	-0.00760	-0.00966	0.048	0.103772308
2370	-0.00156	-0.00024	0.00066	0.00284	-0.02912	-0.01710	-0.11330	0.04132	-0.00860	0.00238	0	-1.060000e-03	-0.00918	-0.00636	-0.02170	-0.00348	0.014	0.033192308
530	-0.03422	-0.00362	0.00752	0.00706	0.07288	0.01390	-0.06014	-0.01078	-0.00856	-0.01202	0	-4.640000e-03	-0.03134	-0.01222	0.00644	0.01490	0.082	0.142852308
1039	0.00778	0.00376	-0.00410	0.00184	-0.02372	0.10166	0.00708	-0.02102	-0.02022	-0.10638	0	6.000000e-05	0.00404	-0.01902	-0.02968	-0.04138	0.064	0.058392308
1391	0.03164	0.00036	0.01476	0.02524	-0.04520	-0.02718	0.04354	0.00016	-0.00580	0.05376	0	4.420000e-03	0.00998	-0.00114	0.00870	-0.00498	0.300	0.305952308
4867	-0.00644	-0.00072	0.01764	-0.00738	-0.01076	-0.03034	-0.03428	-0.00802	-0.00178	-0.07502	0	2.000000e-04	0.00354	0.00622	-0.00436	-0.00452	0.030	0.041672308
301	0.01100	0.00198	0.01764	-0.01400	0.20760	0.21048	-0.01940	0.05404	-0.00950	0.05928	0	2.880000e-03	0.01340	-0.04348	0.00302	0.01232	0.662	0.704952308
911	0.02570	0.00908	-0.04234	0.01562	-0.01574	0.21584	-0.05896	-0.01028	0.01268	-0.00424	0	2.320000e-03	0.00830	0.01946	-0.02360	-0.02602	0.298	0.325512308
628	-0.00640	0.00198	0.00920	-0.00364	-0.00954	0.29132	0.03882	-0.01586	-0.00012	0.05334	0	3.480000e-03	0.01402	-0.01564	0.02290	0.01302	0.536	0.594572308
3868	-0.03104	0.00030	-0.01856	-0.01070	-0.01344	-0.02586	-0.04924	-0.01280	-0.00116	0.02074	0	5.800000e-04	0.00816	0.00204	0.01148	0.00822	0.032	0.086412308

Figure 13.1: Sample of SHAP Value Contributions and Model Probabilities for Loan Default Predictions

In this sample ***(Figure 13.1)***, the SHAP values demonstrate how each feature influences the Random Forest model's prediction for individual loan applicants. Each row represents a

specific applicant, and the SHAP values show whether a feature pushes the prediction toward a higher or lower probability of default.

A key aspect of SHAP is the base value, which is the average default probability across the entire training set. In this case, the base default probability is 0.1977, meaning that, on average, an applicant in the training set has a 19.77% chance of defaulting. The SHAP values for each feature adjust this baseline (either increasing or decreasing it) based on the applicant's specific characteristics.

Example 1 (Row 1578 - Default):

- **Base Probability**: 19.77% (0.1977)
- **Model's Predicted Probability (rf_probabilities)**: 84.60%
- **SHAP-calculated Probability (shap_prob)**: 87.89%

For row 1578, the model starts with the base probability of 19.89%. Features such as DEBTINC (SHAP value: +0.4211) and CLAGE (SHAP value: +0.1209), combined with other features, push the SHAP-calculated probability higher to 87.89%. The Random Forest model's predicted probability is slightly lower at 84.60%, indicating a small discrepancy. Despite the difference, both probabilities strongly suggest a high likelihood of default.

Example 2 (Row 2370 - No Default):

- **Base Probability**: 19.77% (0.1977)
- **Model's Predicted Probability (rf_probabilities)**: 1.4%
- **SHAP-calculated Probability (shap_prob)**: 3.31%

For row 2370, the model again starts with the same base probability of 19.89%. Features like CLAGE (SHAP value: -0.1133) and DEROG (SHAP value: -0.0291), combined with other features, lower the SHAP-calculated probability to 3.31%. The model's predicted probability is even lower at 1.40%, again showing a slight discrepancy. Both probabilities indicate a very low likelihood of default.

13.3 SHAP vs. Model Predictions

The discrepancy between the Random Forest model’s predicted probability and the SHAP-calculated probability occurs because SHAP values explain each feature’s contribution independently, while the model captures complex interactions between features. These non-linear interactions lead to slight differences in predicted probabilities, but this reflects the model's complexity rather than a flaw in SHAP.

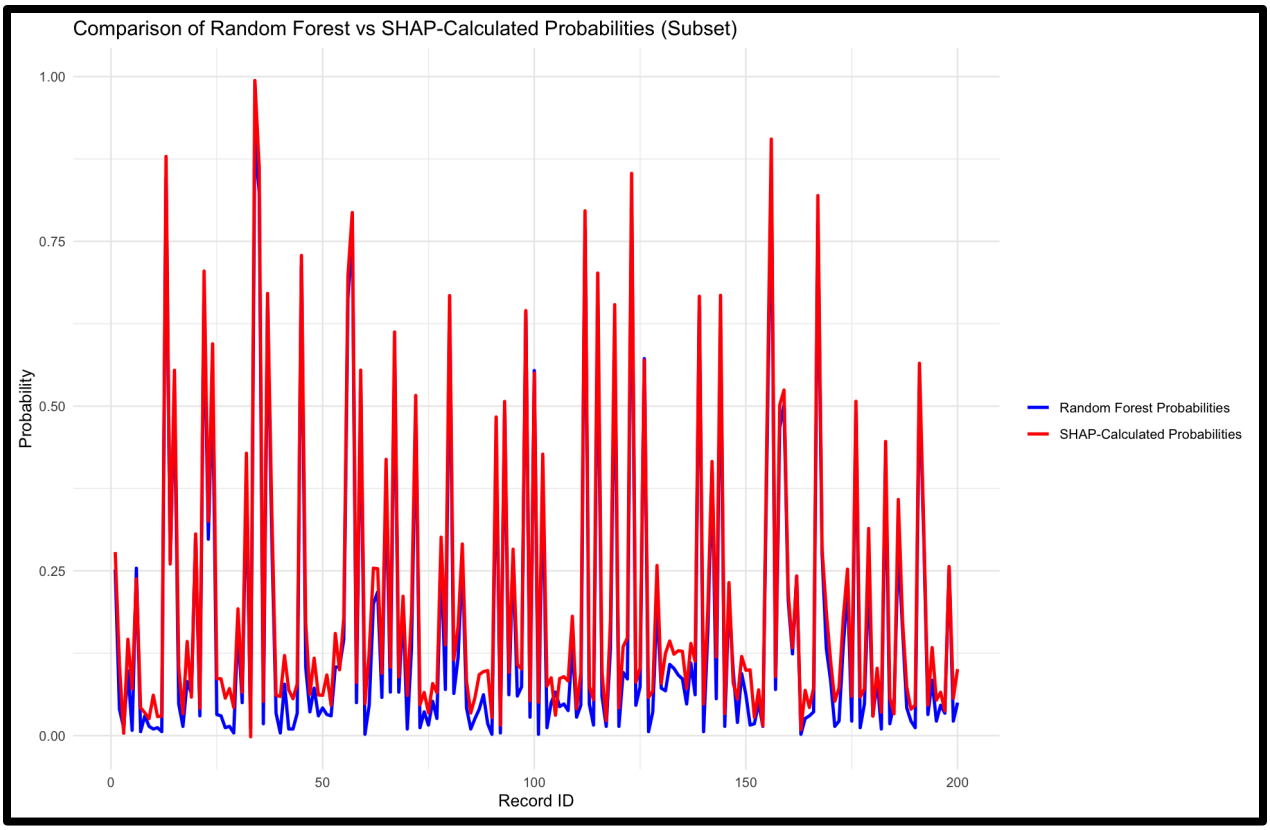

Figure 13.2: Comparison of SHAP-Calculated Probabilities and Random Forest Model Predictions for a Subset of Records

This plot compares the Random Forest model's predicted probabilities (blue) with the SHAP-calculated probabilities (red) for 200 records. In many cases, the two probabilities align closely, especially for lower probability predictions. There are some discrepancies at higher probabilities, where SHAP-calculated values show sharp spikes. However, with a Mean Absolute Error (MAE) of only 3.23%, these differences are generally small, suggesting that SHAP provides an accurate representation of the model's predictions while still capturing how features contribute to the outcome.

Section 14: Test Phase

After developing and assessing several models using different algorithms, the project now moves into the test phase. Based on prior evaluations, the Random Forest model has demonstrated itself to be the most effective and well-suited for deployment. Before moving forward, the model will undergo a final evaluation to ensure its robustness and practical applicability. This phase involves applying the model to a new, unseen dataset consisting of 562 observations to confirm its reliability and performance in a real-world setting.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 417   9
         1  35  97

               Accuracy : 0.9211
                 95% CI : (0.8956, 0.9421)
    No Information Rate : 0.81
    P-Value [Acc > NIR] : 1.251e-13

                  Kappa : 0.7658

 Mcnemar's Test P-Value : 0.000164

            Sensitivity : 0.9151
            Specificity : 0.9226
         Pos Pred Value : 0.7348
         Neg Pred Value : 0.9789
             Prevalence : 0.1900
         Detection Rate : 0.1738
   Detection Prevalence : 0.2366
      Balanced Accuracy : 0.9188

       'Positive' Class : 1
```

Figure 14.1: Confusion Matrix of Random Forest Model on Test Set

The final evaluation of the Random Forest model on the test set shows that its performance is consistent with the earlier results. With an accuracy of 92.11%, high sensitivity at 91.51%, and a balanced ability to correctly identify both defaulters and non-defaulters, the model is well-suited for real-world use. These results confirm that the model is reliable and ready for implementation.

Section 15: Data Driven Improvements

The predictive model has played a crucial role in refining the bank's decision-making process by providing clear insights into loan default probabilities. With data-driven risk assessments, the model has allowed for more informed and proactive decisions, reducing uncertainty and improving the approval process. By identifying high-risk applicants more accurately, the bank can now manage credit risk more effectively, optimizing profitability while maintaining transparency.

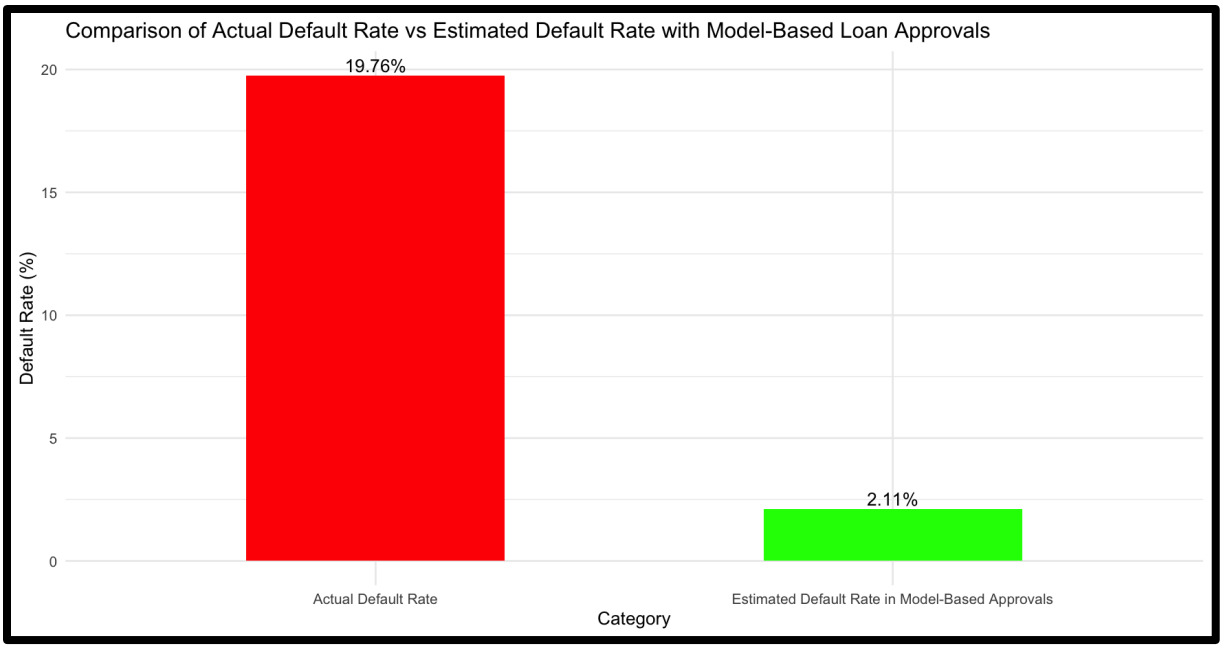

Figure 15.1: Comparison of Actual Default Rate vs Estimated Default Rate with Model-Based Loan Approvals.

Figure 15.1 compares the actual historical default rate of 19.76% to the default rate of 2.11% when using the model for loan approvals, as calculated from the test set. This demonstrates an approximate 89% reduction in default rates, underscoring the model's effectiveness in identifying high-risk applicants and minimizing defaults.

These results confirm that by relying on the model's predictions, the bank can significantly lower financial risk, reduce losses, and improve profitability. The model enables smarter loan approval decisions, directing resources to applicants with a higher likelihood of repayment and enhancing the overall stability of the bank's credit portfolio.

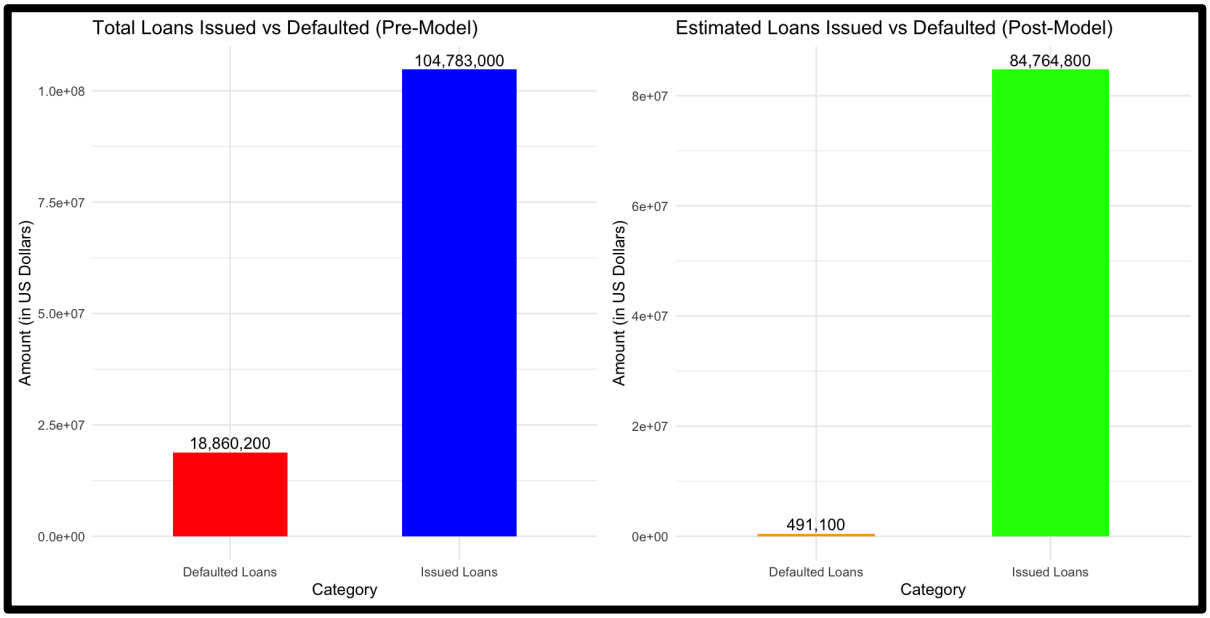

Figure 15.2: Comparison of Loans Issued vs Defaulted: Pre-Model and Post-Model Implementation

The two plots in ***Figure 15.2*** compare loan issuance and defaults before and after implementing the model-based loan approval strategy. In the pre-model scenario, $104,783,000 in loans were issued, with $18,860,200 defaulted. This reflects the historical approach, where high-risk applicants were not effectively filtered out, resulting in significant financial losses.

In the post-model scenario, the estimated loan issuance drops to $84,764,800, with only $491,100 expected to default based on the model's predictions. To obtain these values, the model was applied to the entire dataset for comparison purposes, simulating how the bank's loan issuance and defaults would look if the model had been used historically. This approach provides an estimate of the model's effectiveness in reducing defaults.

By using the model, the bank achieves an approximately 97% reduction in defaulted loan amounts, significantly lowering exposure to financial losses. While the total loan issuance decreases by about 19%, the model ensures that approved loans are far less likely to default, leading to improved portfolio quality and financial stability. These results highlight the model's potential to enhance decision-making and mitigate risk in a real-world scenario.

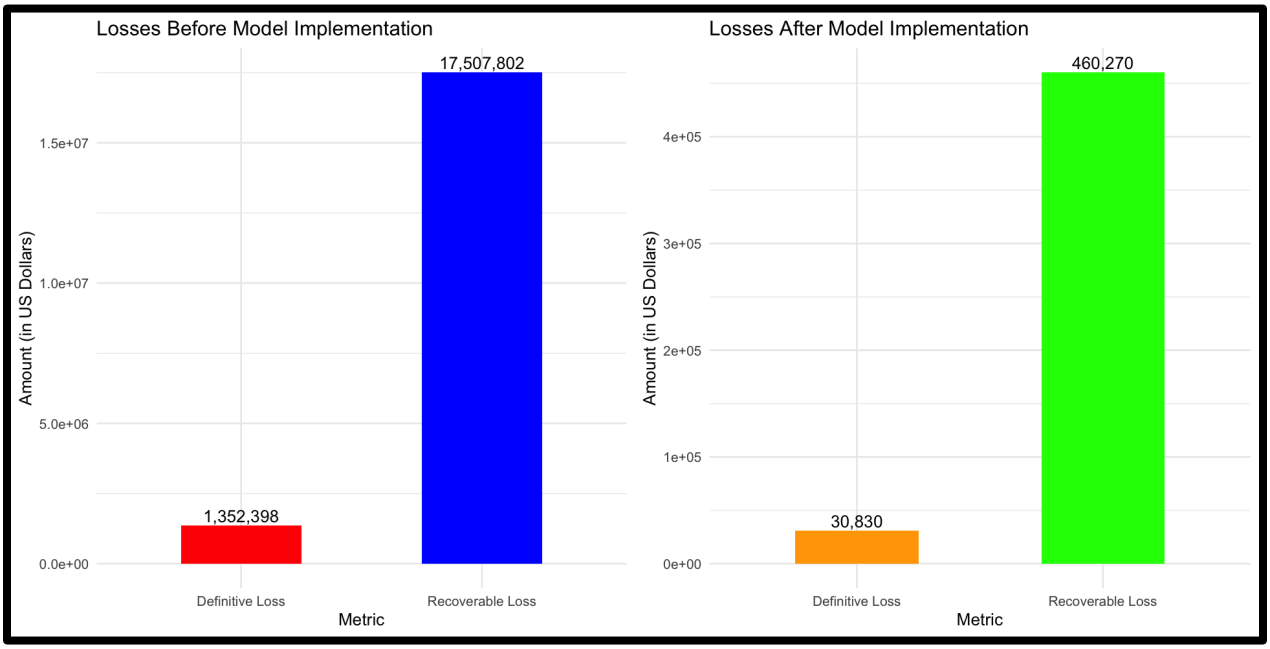

Figure 15.3: Comparison of Definitive vs Recoverable Losses: Pre-Model and Post-Model Implementation

The plots show a significant reduction in both definitive and recoverable losses after implementing the model-based loan approval strategy. In the pre-model scenario, definitive losses were $1,352,398, and recoverable losses totaled $17,507,802. The large recoverable loss indicates a substantial portion of defaulted loans could potentially be recovered, but this process is both uncertain and resource intensive.

Post-model, definitive losses are estimated to drop to $30,830 (a reduction of $1,321,568 or approximately 98%), while recoverable losses fall to $460,270 (a reduction of $17,047,532 or approximately 97%). These values were obtained by applying the model to the entire dataset for comparison purposes, simulating its impact historically.

This substantial reduction underscores the model's effectiveness in mitigating financial risk. By minimizing both definitive and recoverable losses, the model improves overall portfolio quality and ensures a more efficient allocation of resources.

Conclusion

This project successfully demonstrated the potential of predictive analytics and machine learning to optimize bank loan approvals, significantly reducing default rates and financial risks. By implementing a Random Forest model with high accuracy (92.11%) and sensitivity (91.51%), the estimated default rate was reduced from 20% to 2.11%, showcasing the model's ability to identify high-risk applicants effectively.

The use of SHAP methodology enhanced the model's transparency, ensuring compliance with ECOA regulations and enabling clear communication of decisions to stakeholders. Additionally, the development of a FICO-inspired scoring system provided actionable insights, aiding in risk segmentation and improving credit decision-making.

In conclusion, the model has greatly enhanced the bank's ability to manage risk while maintaining transparency and client trust. The insights from this model provide a robust foundation for future decision-making, balancing profitability with responsible lending practices. This solution not only addresses the immediate challenge of high default rates but also equips the bank with a powerful tool for sustainable long-term growth.

Financially, the project demonstrated a 97% reduction in defaulted loan amounts and a 98% decrease in unrecoverable losses. These results underscore the transformative impact of data-driven lending practices, enabling the bank to make informed, transparent, and equitable credit decisions while fostering long-term portfolio stability and growth.

Recommendations

Transparent Client Education:

- Focus on educating clients about how their credit score affects loan approvals. Provide resources to help them improve their financial health for better loan opportunities.

Integrate the Predictive Model into Lending Processes:

- Incorporate the model into loan approval workflows to filter high-risk applicants and reduce default rates and ensure data-driven decisions.

- Combine the model with SHAP methodology to make decisions more transparent, ensuring compliance with ECOA regulations and providing clear explanations for approvals or denials.

Risk-Based Pricing and Targeted Pre-Approvals:

- Implement targeted pre-approvals based on credit scores, focusing on clients with scores above 730 who are less likely to default. Offer competitive rates to attract these low-risk borrowers.

Careful Approach to Loans Below $10,000:

- Since loans below $10,000 are linked to higher default risks, implement stricter approval criteria or higher interest rates for these smaller loans to manage exposure.

Focus on High-Value Clients:

- Prioritize applicants with scores over 730, as they are reliable borrowers who typically request larger loans.

- Offer exclusive benefits, such as competitive rates, higher loan limits, or loyalty rewards, to attract and retain these clients.