BANK LOAN DECISION MAKING ANALYTICS

Data Analytics & Modeling

Subash Yadav

Webster University
CSDA 6010 – Project 2
Prof. JP WANG

Executive Summary

This project aimed to automate the loan approval process for home improvement loans by leveraging predictive analytics and customer segmentation. The dual objectives were to create a robust classification model to predict loan defaults while ensuring compliance and interpretability, and to segment customers using clustering techniques to enhance targeted risk management.

The dataset included 5,834 records, with critical features such as the debt-to-income ratio (DEBTINC), delinquencies (DELINQ), and the age of the oldest credit line (CLAGE). Missing data was addressed using advanced imputation techniques, including median and K-Nearest Neighbors (KNN). Feature selection was performed using Random Forest and LASSO regression to identify the most relevant predictors.

Three classification models—Logistic Regression, Decision Tree, and Random Forest—were developed and evaluated. The Random Forest model emerged as the best performer with 93.03% accuracy and 97.86% sensitivity, ensuring precise identification of defaulters and minimizing false negatives. Its superior performance, combined with the ability to interpret feature importance, makes it ideal for deployment.

For clustering, both K-Means and Hierarchical Clustering were applied separately to defaulters and non-defaulters. The optimal number of clusters was determined using the Elbow and Silhouette methods, resulting in three distinct risk levels—low, moderate, and high. These clusters align with business goals by enabling targeted interventions, such as stricter loan terms for high-risk defaulters and personalized incentives for low-risk non-defaulters.

This project successfully achieved its objectives by integrating advanced analytics to streamline loan approvals, reduce financial risks, and enhance customer segmentation. The proposed solution ensures transparency and compliance, empowering the bank to make data-driven, fair, and effective decisions. With model deployment and continuous monitoring, the bank can further optimize its risk management strategies and improve customer satisfaction.

Table of Contents

Business Problem

The bank aims to automate the loan approval process for home improvement loans by developing an advanced system that can **predict loan defaults** and effectively **segment applicants based on their financial profiles**. The key challenges are:

1. **Loan Default Prediction and Compliance:** Develop a predictive model that accurately identifies applicants at high risk of defaulting, while providing interpretable and transparent reasons for loan rejections. This model must comply with legal requirements, such as the Equal Credit Opportunity Act, by ensuring transparency and fairness in the decision-making process. **The model must explain the reasons for loan decisions in an interpretable manner, focusing on key risk factors such as debt-to-income ratio, delinquencies, and derogatory credit marks.**
2. **Precise Customer Profiling for Risk Management:** The goal is to segment defaulters and non-defaulters separately based on their financial behaviors and credit histories. By analyzing these groups independently, the bank can identify specific credit behaviors within each category and design tailored risk management strategies. For defaulters, the focus will be on understanding the factors driving defaults, while for non-defaulters, the emphasis will be on profiling stable financial behaviors. This dual approach enhances targeted risk management, personalized loan terms, and tailored financial solutions for each group.

The dataset includes 5,960 records with the target variable BAD, indicating whether an applicant defaulted (1) or did not default (0). With approximately 20% defaulters, it is a binary classification problem. Additionally, clustering analysis will be performed separately on defaulters and non-defaulters to provide more precise customer segmentation, allowing for better risk management, regulatory compliance, and personalized financial services.

Business Goals

Business Goal 1: Automated Loan Default Prediction and Interpretability

The primary goal is to build an automated decision-making system that **accurately predicts loan defaults** for home improvement loans. The system must ensure transparency by providing **clear and interpretable reasons** for loan rejections, thus complying with legal requirements such as the **Equal Credit Opportunity Act**. This classification-based model will help mitigate financial risk for the bank by focusing on key factors like **debt-to-income ratio (DEBTINC)**, **delinquencies (DELINQ)**, and **credit history (CLAGE)**, while ensuring that the process remains unbiased and compliant with regulatory standards.

- **Alignment**: The business needs to understand which applicants are at risk of default (target variable **BAD, BAD = 1**) and provide explainable reasons for loan decisions (e.g., features like **DEBTINC**, **DEROG**, and **CLAGE**). **Interpretability** will be essential to meet legal standards and to ensure accountability and fairness in loan approval decisions.

Business Goal 2: Precise Clustering for Targeted Risk Management

The secondary goal is to segment customers based on their **financial behaviors** and **credit history** using clustering techniques. Segment customers by performing clustering separately on defaulters

(BAD = 1) and non-defaulters (BAD = 0). This approach will uncover distinct profiles within each group, enabling the bank to understand varying credit behaviors more effectively.

Defaulter Clustering:

- Focus on identifying sub-groups within defaulters, such as those with frequent delinquencies, high debt-to-income ratios, or multiple derogatory reports. Use these insights to design targeted interventions, stricter loan terms, or financial counseling for high-risk segments.

Non-Defaulter Clustering:

- Identify sub-groups within non-defaulters, such as those with longer credit histories, low debt loads, or consistent payment behavior. Leverage these insights to offer personalized financial products, such as favorable loan terms or upsell opportunities for low-risk segments.
- **Alignment:** By clustering defaulters and non-defaulters separately, the bank can develop specific strategies for risk mitigation and personalized service, enhancing decision-making, customer engagement, and operational efficiency.

These goals work in tandem: the first ensures accurate loan default predictions and regulatory compliance, while the second goals emphasize the need for precise clustering that supports both compliance and effective risk management, ensuring that the bank can achieve comprehensive automation of the loan approval process while maintaining high standards of fairness and transparency. Together, they offer a comprehensive solution for automating the bank's home improvement loan approval process and **managing loan risks** more effectively.

Analytical Goals

Analytical Goal 1: Classification Model for Loan Default Prediction

Building a Transparent Classification Model The objective is to develop a classification model that predicts whether a loan applicant will default (BAD = 1) or not (BAD = 0). The model should prioritize both predictive accuracy and interpretability. Ensuring transparency will enable compliance with the **Equal Credit Opportunity Act** and other regulatory standards by providing clear explanations for loan rejections.

- **Model Focus**: Implement Logistic Regression for interpretable coefficients and Decision Trees for visual decision paths. Both models will balance interpretability and performance, supporting transparent and actionable decision-making. The classification model must comply with legal requirements by ensuring fairness, reducing bias, and providing reasons for decisions that stakeholders can easily understand.

Analytical Goal 2: Precise Clustering for Targeted Risk Management

Segmented Clustering for Behavioral Insights The goal is to perform targeted clustering to identify distinct sub-groups among defaulters (BAD = 1) and non-defaulters (BAD = 0). This analysis will provide actionable insights into customer behavior and enable tailored risk management strategies.

- **Defaulter Clustering:** Use clustering techniques to identify high-risk segments among defaulters, such as those with multiple delinquencies or high debt-to-income ratios. These insights will guide interventions like stricter loan terms or financial counseling.

- **Non-Defaulter Clustering:** Segment non-defaulters based on behaviors like consistent payment patterns or low debt levels to inform opportunities for personalized product offerings and rewards programs.

Clustering Focus: By employing clustering methods like K-Means or Hierarchical Clustering, the aim is to analyze **defaulters (BAD = 1)** and **non-defaulters (BAD = 0)** separately to uncover distinct patterns within each group. For defaulters, the focus is on identifying sub-groups based on high delinquencies, multiple derogatory marks, or high debt-to-income ratios, which will inform targeted interventions, such as stricter loan terms or financial counseling. For non-defaulters, the goal is to identify sub-groups characterized by longer credit histories, low debt loads, or consistent payment behavior, enabling the bank to offer personalized financial products, such as favorable loan terms or upsell opportunities. This approach enhances the bank's ability to understand customer behavior, implement tailored interventions, optimize loan offerings based on unique customer profiles, and proactively mitigate risks while ensuring compliance.

Analytical Approach

1. Data Preprocessing

- **Handle Missing Values:**
  - For numeric columns like MORTDUE, YOJ, and DEBTINC, use median imputation, as these variables often have skewed distributions. Median imputation helps maintain data integrity by being robust against outliers.
  - For columns with complex relationships, such as CLAGE or NINQ, apply K-Nearest Neighbors (KNN) imputation. KNN imputation estimates missing values based on the nearest observations, providing more context-driven imputations.
- **Encoding Categorical Variables:**
  - Convert categorical variables like REASON (loan purpose) and JOB (occupation type) into numerical formats using one-hot encoding. This method creates binary variables for each category, making them compatible with classification and clustering algorithms.
- **Normalization and Scaling:**
  - Apply normalization or standard scaling to numerical variables like LOAN, VALUE, and DEBTINC to ensure consistency across models. This is crucial for algorithms like K-Means, which are sensitive to feature scales.
- **Address Class Imbalance:**
  - Use techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the target variable BAD. Given that only 20% of applicants defaulted, this step ensures that the models do not become biased towards predicting non-defaults.

2. Exploratory Data Analysis (EDA)

- **For Classification:**
  - Analyze relationships between predictors and the target variable BAD to uncover key factors driving loan defaults. Visualizations like correlation matrices, scatter plots, and distribution plots can help identify impactful features like DELINQ (delinquencies), DEROG (derogatory reports), and DEBTINC (debt-to-income ratio).

    - Use feature importance plots to understand which variables contribute most to loan defaults. This will guide model feature selection, ensuring that only the most relevant variables are used.
- **For Clustering:**
    - Perform EDA separately for defaulters (BAD = 1) and non-defaulters (BAD = 0). This allows for a deeper understanding of unique patterns within each segment.
    - Visualize key variables like LOAN, YOJ, CLAGE, and DEBTINC using scatter plots, pair plots, and density plots to identify potential clusters and relationships within defaulters and non-defaulters.

3. Model Selection

- **Classification Models:**
    - **Logistic Regression:**
        - Build a logistic regression model to ensure transparency and interpretability. This model will focus on key features like DEBTINC, DELINQ, DEROG, and CLAGE, providing coefficients that can be directly interpreted to explain the likelihood of loan defaults.
    - **Decision Tree:**
        - Develop a decision tree model to visualize decision paths. Decision trees are useful for providing interpretable decision rules that can explain why applicants were rejected, focusing on features like debt-to-income ratio, derogatory reports, and delinquencies.
    - **Random Forest:**
        - Train a Random Forest model to boost predictive performance and compare it with simpler models like Logistic Regression and Decision Trees. Random Forest offers stronger performance by aggregating results from multiple decision trees, identifying critical features, and minimizing overfitting.
- **Clustering Models:**
    - **Defaulter Clustering:**
        - Apply K-Means or Hierarchical Clustering to segment defaulters (BAD = 1). This helps identify distinct sub-groups within defaulters, such as those with high debt-to-income ratios or multiple delinquencies, enabling targeted interventions.
    - **Non-Defaulter Clustering:**
        - Apply the same clustering techniques to segment non-defaulters (BAD = 0). Focus on identifying stable profiles, such as those with long credit histories or low debt loads, to offer personalized loan terms or upsell opportunities.
    - **Cluster Validation:**
        - Validate clustering results using metrics like silhouette score or Davies-Bouldin index to ensure meaningful segmentation. This ensures that identified clusters are interpretable and actionable.

4. Predictor Analysis and Relevancy

- **Feature Evaluation for Classification:**
    - Analyze the relevance of different predictors using feature importance measures like Gini impurity (for Decision Trees) or coefficients (for Logistic Regression).

Prioritize features like DEBTINC, DELINQ, DEROG, and CLAGE, which are expected to be strong indicators of defaults.
    - Conduct a correlation analysis to identify multicollinearity among predictors. If needed, apply Principal Component Analysis (PCA) to reduce dimensionality while retaining the most variance.
- **Feature Evaluation for Clustering:**
    - Use exploratory plots to identify key variables that help in distinguishing clusters. For defaulters, focus on features like DELINQ, DEROG, and DEBTINC, while for non-defaulters, focus on variables like CLAGE, YOJ, and LOAN.

5. Dimension Reduction

- **Apply PCA or LASSO Regularization:**
    - If multicollinearity is present, apply PCA to reduce the dimensionality of the dataset while retaining essential information.
    - LASSO regularization can be applied to classification models to penalize less impactful variables, helping to refine feature selection and enhance interpretability.

6. Data Transformation

- **Log Transformation or Standardization:**
    - Apply log transformation to highly skewed variables (e.g., DEBTINC or DELINQ) to improve data normality.
    - Standardize variables based on the requirements of specific models, such as scaling for K-Means clustering or standardizing inputs for Logistic Regression to enhance model performance.

7. Data Partitioning Methods

- **Split the Dataset:**
    - Partition the dataset into training, validation, and test sets to ensure robust evaluation and prevent overfitting.
    - Use stratified sampling to maintain the proportion of defaulters and non-defaulters across all subsets, ensuring balanced representation during model training and evaluation.

8. Model Building

- **Logistic Regression:**
    - Train a logistic regression model focusing on DEBTINC, DELINQ, DEROG, and CLAGE to predict loan defaults. This model provides interpretable coefficients, explaining how each variable impacts the likelihood of default.
- **Decision Tree:**
    - Build a decision tree model to provide visual decision paths, making it clear why certain applicants were rejected based on their financial profiles.
- **Random Forest:**
    - Develop a Random Forest model to boost predictive performance and compare it with simpler models. Random Forest can identify important features and offer insights into complex interactions between variables.
- **Clustering Models:**
    - Perform clustering separately for defaulters and non-defaulters using K-Means or Hierarchical Clustering. This helps identify high-risk sub-groups within defaulters and stable sub-groups within non-defaulters, guiding targeted risk management and personalized financial solutions.

9. Model Evaluation

- **Classification Models:**
  - Evaluate models using accuracy, precision, recall, and F1 score to ensure strong predictive performance while complying with fairness and transparency standards.
  - Use cross-validation to verify the models' robustness and generalizability to unseen data.
- **Clustering Models:**
  - Use metrics like silhouette score, cohesion, and separation to evaluate clustering quality. Ensure that clusters are interpretable and provide actionable insights for risk management or personalized services.

10. Interpretability

- **For Classification Models:**
  - Focus on interpreting logistic regression coefficients and decision tree rules to provide clear, regulatory-compliant reasons for loan rejections. The interpretability of these models is essential to ensure transparency in decision-making and compliance with the Equal Credit Opportunity Act.
- **For Clustering Models:**
  - Ensure that clusters offer clear insights into customer behavior, making it possible to design targeted interventions or personalized financial products based on customer profiles.

Overall Approach

The analytical approach aligns with two primary goals: predicting loan defaults with interpretable classification models and segmenting customers through distinct clustering for defaulters and non-defaulters. Classification models ensure accurate, compliant predictions, while clustering identifies unique customer profiles for targeted interventions and personalized services. This approach supports automated, transparent loan approval processes, enhances risk management, and complies with regulatory requirements.

Dataset Overview

The dataset, sourced from a bank's consumer credit department, contains 5,960 observations with 13 features. It includes data on applicants who have recently applied for home improvement loans, aimed at building a comprehensive system for automating loan approvals, managing default risks, and creating detailed customer profiles for targeted interventions. The dataset supports two primary business objectives:

1. **Loan Default Prediction:**
   - The dataset is used to develop a classification model that accurately predicts which applicants are likely to default (BAD = 1). The model will provide interpretable reasons for loan rejections, ensuring compliance with legal requirements like the Equal Credit Opportunity Act. Key features include financial indicators such as debt-to-income ratio, delinquencies, and derogatory credit marks, which are critical for making transparent and fair loan decisions.
2. **Precise Customer Profiling through Clustering:**

- The dataset will enable separate clustering of defaulters (BAD = 1) and non-defaulters (BAD = 0). This approach helps identify distinct customer segments within each group, allowing for deeper insights into credit behaviors. By understanding the specific profiles of defaulters (e.g., high-risk sub-groups) and non-defaulters (e.g., stable profiles), the bank can develop targeted risk management strategies, personalized loan terms, and tailored financial solutions.

The dataset includes a mix of categorical and numerical variables that capture an applicant's financial stability, job status, and credit history, making it suitable for both predictive modeling and targeted clustering analysis. This dual approach supports accurate, interpretable loan approval decisions and precise customer segmentation, enhancing automated decision-making and regulatory compliance.

Columns and Their Purpose

The dataset contains 13 features, each providing valuable insight into different aspects of the applicant's profile:

Column Name	Data Type	Description
BAD	Categorical (0 or 1)	Target variable. 1 indicates that the applicant defaulted on their loan or was seriously delinquent, and 0 means they did not default.
LOAN	Numeric (whole number)	Amount of the loan requested by the applicant (in dollars).
MORTDUE	Numeric (decimal)	The total remaining amount on the applicant's current mortgage.
VALUE	Numeric (decimal)	The value of the property being used as collateral for the loan (in dollars).
REASON	Categorical (Text)	The reason for applying for the loan: either for **home improvement (HomeImp)** or **debt consolidation (DebtCon)**.
JOB	Categorical (Text)	The occupation of the applicant (e.g., **Office, Other, Sales, etc.**).
YOJ	Numeric (decimal)	Years the applicant has been at their current job.
DEROG	Numeric (decimal)	Number of major derogatory reports (indicates serious financial issues like bankruptcy).
DELINQ	Numeric (decimal)	Number of delinquent credit lines (indicates how often the applicant has missed payments).
CLAGE	Numeric (decimal)	Age of the oldest credit line, measured in months (an indicator of the applicant's credit history length).

Column Name	Data Type	Description
NINQ	Numeric (decimal)	Number of recent credit inquiries (indicates how often the applicant has applied for new credit recently).
CLNO	Numeric (decimal)	Total number of credit lines the applicant has opened.
DEBTINC	Numeric (decimal)	Debt-to-income ratio (percentage of the applicant's monthly income going toward debt payments).

(Table 1 – Dataset Columns)

Interesting Insights:

- **Target Variable (BAD)**: The target variable, **BAD**, indicates whether an applicant defaulted. This is crucial for our **classification** task, as we aim to predict this binary outcome. his column indicates whether the applicant eventually **defaulted** on their loan or was **seriously delinquent**. Out of 5,960 applicants, **1,189** (~20%) defaulted, making it a **binary classification** problem.
- **Missing Values**: Several columns, like **MORTDUE**, **YOJ**, and **DEBTINC**, contain missing values, which will need to be handled during preprocessing.
- **Categorical Variables**: The **REASON** and **JOB** columns are categorical, providing insight into the reasons for the loan and the applicant's occupation, both of which may affect the likelihood of default.
- **Imbalance in Defaults**: Only 20% of the applicants defaulted, meaning the data is **imbalanced**. This needs to be addressed during model training to prevent bias toward non-defaulters.
- **Credit History Insight**: With variables like **CLAGE** (credit age), **DELINQ** (delinquencies), and **DEROG** (derogatory reports) provide insights into the applicant's credit behavior, which is critical for understanding default risk and for customer profiling.
- **Financial Health Indicators**: Columns like **MORTDUE**, **DEBTINC**, and **DEROG** are key indicators of financial stability and will be used for both predictive modeling and clustering to determine customer segments at higher risk.

Why This Dataset Is Important?

This dataset is critical in enabling the bank to achieve its dual business objectives effectively:

1. **Automated Loan Approval & Default Prediction:**
    - The dataset provides comprehensive information on each applicant's financial stability, credit history, and job status, which is essential for building accurate predictive models. These models automate the loan approval process while ensuring transparency and compliance with legal requirements, such as the Equal Credit Opportunity Act. By identifying applicants likely to default, the bank can manage risks proactively, minimizing financial losses and enhancing decision-making efficiency.
2. **Precise Customer Profiling for Risk Management & Personalized Solutions:**

- Beyond predicting defaults, the dataset supports clustering techniques that segment defaulters and non-defaulters separately. This segmentation uncovers distinct customer profiles, enabling the bank to tailor loan terms, provide targeted financial counseling, and offer personalized products and services. By understanding specific credit behaviors within each group, the bank can implement proactive strategies that cater to the unique needs of each segment, improving customer satisfaction and retention.

The dataset's rich mix of categorical and numerical features makes it an asset for both predictive modeling and clustering. It not only supports accurate, automated decision-making but also enhances the bank's ability to deliver personalized financial solutions, improve risk management, and maintain regulatory compliance—all while boosting operational efficiency.

Summary Statistics of Dataset

In this analysis, the **skimr** package was used to provide a comprehensive summary of the dataset. This tool gives an overview of both categorical and numerical variables, displaying key metrics such as the number of missing values, mean, standard deviation, and more. The skimr output helps in identifying potential issues with the data, such as missing values and skewed distributions, which are critical for data preprocessing and model development.

```
> skim(loan_data)
── Data Summary ────────────────────────
                           Values
Name                       loan_data
Number of rows             5960
Number of columns          13
_________________________
Column type frequency:
  character                2
  numeric                  11
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 REASON                0             1   0   7   252        3          0
2 JOB                   0             1   0   7   279        7          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────
   skim_variable n_missing complete_rate       mean        sd       p0     p25     p50      p75    p100 hist
 1 BAD                   0         1          0.199     0.400    0         0       0        0        1  ▇▁▁▁▂
 2 LOAN                  0         1      18608.    11207.    1100     11100   16300    23300    89900  ▇▆▂▁▁
 3 MORTDUE             518         0.913  73761.    44458.    2063     46276   65019    91488   399550  ▇▆▂▁▁
 4 VALUE               112         0.981 101776.    57386.    8000     66076.  89236.  119824.  855909  ▇▃▁▁▁
 5 YOJ                 515         0.914      8.92      7.57     0         3       7       13       41  ▇▅▂▂▁
 6 DEROG               708         0.881      0.255     0.846    0         0       0        0       10  ▇▁▁▁▁
 7 DELINQ              580         0.903      0.449     1.13     0         0       0        0       15  ▇▁▁▁▁
 8 CLAGE               308         0.948    180.       85.8      0       115.    173.     232.    1168. ▂▇▂▁▁
 9 NINQ                510         0.914      1.19      1.73     0         0       1        2       17  ▇▂▁▁▁
10 CLNO                222         0.963     21.3      10.1      0        15      20       26       71  ▂▇▅▁▁
11 DEBTINC            1267         0.787     33.8       8.60     0.524    29.1    34.8     39.0    203. ▇▁▁▁▁
```

(Table 2 – Summary Statistics R Result)

This table provides a concise description of each column and its characteristics, forming the foundation for the data preprocessing and modeling steps to come.

Column Name	Type	Missing Values	Description	Mean / Mode	Min / Max
BAD	Numeric (Binary)	0	Target variable: 1 = Default, 0 = No Default	Mean: 0.199	Min: 0, Max: 1
LOAN	Numeric	0	Loan amount requested by the applicant (in dollars)	Mean: 18,608	Min: 1,100, Max: 89,900
MORTDUE	Numeric	518 (8.69%)	Total amount of mortgage remaining	Mean: 73,761	Min: 2,063, Max: 399,550
VALUE	Numeric	112 (1.88%)	Value of the property used as collateral (in dollars)	Mean: 101,776	Min: 8,000, Max: 855,909
REASON	Categorical	0	Reason for loan: "HomeImp" (Home Improvement) or "DebtCon"	Mode: "DebtCon"	N/A
JOB	Categorical	0	Applicant’s job category (e.g., Office, Other)	Mode: "Other"	N/A
YOJ	Numeric	515 (8.64%)	Years the applicant has been at their current job	Mean: 8.92 years	Min: 0, Max: 41 years
DEROG	Numeric	708 (11.88%)	Number of major derogatory reports (serious financial issues)	Mean: 0.255	Min: 0, Max: 10
DELINQ	Numeric	580 (9.73%)	Number of delinquent credit lines (missed payments)	Mean: 0.449	Min: 0, Max: 15
CLAGE	Numeric	308 (5.17%)	Age of the oldest credit line (in months)	Mean: 180 months (15 years)	Min: 0, Max: 1,168 months

Column Name	Type	Missing Values	Description	Mean / Mode	Min / Max
NINQ	Numeric	510 (8.56%)	Number of recent credit inquiries (last 6 months)	Mean: 1.19	Min: 0, Max: 17
CLNO	Numeric	222 (3.73%)	Number of credit lines the applicant has opened	Mean: 21.3	Min: 0, Max: 71
DEBTINC	Numeric	1,267 (21.26%)	Debt-to-income ratio (percentage of income going toward debt)	Mean: 33.8%	Min: 0.524%, Max: 203.31%

(Table 3 – Summary Statistics Description)

Key Insights:

- **Loan Amount (LOAN):** The loan amounts vary greatly, with a mean of $18,608 and a maximum of $89,900. This wide range of loan amounts could be an important factor influencing default risk, with larger loans possibly being riskier.
- **Missing Data:**
- Several columns have missing values, particularly **MORTDUE** (518 missing), **YOJ** (515 missing), and **DEBTINC** (1,267 missing). This needs to be addressed through imputation strategies or other methods during data preprocessing.
- **DEBTINC** (Debt-to-Income Ratio), a critical variable for financial stability, has 1,267 missing values, which could be problematic for modeling if not handled properly.
- **Categorical Variables (REASON and JOB):**
    - The **REASON** for the loan (either 'DebtCon' or 'HomeImp') may reveal trends in default risk, as applicants consolidating debt might be at higher risk than those improving their homes.
    - **JOB** categories provide an opportunity to explore occupational trends in loan defaults.
- **Delinquency and Credit Behavior (DELINQ and DEROG):**
    - Both **DELINQ** (delinquencies) and **DEROG** (derogatory reports) show that most applicants have no previous records, but some applicants have up to 15 delinquencies and 10 derogatory reports. These variables are strong predictors of financial behavior and default risk.
    - The **CLAGE** (Credit Age) distribution shows that the average applicant has a credit history of about 180 months (15 years), with a wide range, indicating a mix of well-established and newer borrowers.
- **Income Stability (YOJ):**
    - The **YOJ** (Years on the Job) variable reflects employment stability, with a mean of around 9 years but varying up to 41 years. This could be a key factor in predicting defaults, with longer tenure potentially indicating greater stability.
- **Debt-to-Income Ratio (DEBTINC):**

    - With an average **DEBTINC** of 33.78%, higher ratios could indicate applicants at higher risk of default, and we observe some extreme values going up to 203%.
- **Credit Health Indicators:**
    - The **CLNO** (Number of Credit Lines) indicates that most applicants have multiple credit lines (mean of 21). The variety in credit lines could influence default risk, with higher numbers either indicating financial flexibility or overextension.
- **Loan-to-Value Ratio:**
    - The difference between **MORTDUE** (amount still owed on the mortgage) and **VALUE** (current property value) offers an implicit **Loan-to-Value (LTV) Ratio**, which is an essential measure of risk for secured loans. Borrowers with high LTV ratios (those who owe close to or more than the value of their property) are generally at higher risk of default.
    - For instance, if a borrower owes $90,000 on a property valued at $100,000, their LTV is 90%, indicating higher risk if the property market depreciates.
- **NINQ (Number of Recent Credit Inquiries):**
    - Higher numbers of credit inquiries often indicate financial stress or that an applicant is seeking multiple loans, which could be a red flag. The **mean NINQ** is 1.18, but some applicants have up to 17 inquiries, indicating significant variance in borrowing behavior. Frequent inquiries can also negatively impact credit scores.
- **Home Improvement Loans vs. Debt Consolidation:**
    - **REASON** for loan approval could provide a key segmentation opportunity. Home improvement loans may signal investments in personal assets, whereas debt consolidation loans could indicate pre-existing financial trouble. Comparing default rates across these categories could uncover patterns—perhaps applicants consolidating debt are more likely to default.
- **Effect of Employment Type:**
    - **JOB** classification could provide important insights. For example, comparing default rates among different job sectors (e.g., Office workers vs. Self-employed individuals) can offer insight into which employment types are more prone to default due to income instability.
    - Understanding which jobs are more secure during economic downturns can also give an edge in modeling, as certain sectors may be more vulnerable.
- **Length of Credit History (CLAGE):**
    - Borrowers with more extended credit histories might have established financial patterns, making them potentially more reliable (or less reliable depending on behavior). The **mean CLAGE** of 180 months (15 years) suggests a seasoned borrower pool, but the wide range shows that both new and old borrowers are represented.
    - Segmenting applicants by credit age could offer another way to improve model prediction, focusing on those with less credit history who may be riskier.
- **Correlation Between Financial Indicators:**
    - High correlations between certain financial health indicators like **DEBTINC**, **MORTDUE**, **YOJ**, and **VALUE** might indicate multicollinearity, which could skew results. Investigating the relationships between these variables may

highlight the need for feature engineering or dimensionality reduction (e.g., PCA or regularization techniques like Lasso).

- **Potential Outliers:**
  - Extreme values, such as high debt-to-income ratios (DEBTINC), numerous derogatory reports (DEROG), or very large loan amounts (LOAN), could be driving model bias or indicate anomalous data. Identifying and handling outliers through capping, binning, or transformation could help improve the model's performance.
- **Segmentation for Clustering:**
  - Apart from default prediction, clustering applicants based on features like **MORTDUE**, **DEBTINC**, and **CLAGE** can help create distinct customer segments. This could allow for the identification of high-risk versus low-risk groups, making it easier to focus interventions and financial products on certain borrower types.
  - For example, separating borrowers with high LTV ratios and recent credit inquiries into a distinct cluster might reveal groups at much higher risk.
- **Imbalance Handling:**
  - The dataset's imbalanced nature (with only ~20% defaulters) is a significant challenge. Implementing techniques like **Synthetic Minority Over-sampling Technique (SMOTE)**, **Random Under-sampling**, or class-weighted algorithms during model training will be crucial to avoid bias toward predicting non-defaults.

Data Preprocessing

Handling Blank String Values

During the data exploration phase, it was found that several columns contained blank values instead of NA. Specifically, the categorical columns "REASON" and "JOB" had significant numbers of blank values, while the rest of the numeric columns had missing values represented by NA. Below is the summary of columns that had blank values:

Column Name	Number of Blank Values
REASON	252
JOB	279

(Table 4 – Blank Values Table)

To handle this issue, replaced all blank string values with NA. This step was crucial to ensure that our analysis tools correctly interpret these entries as missing data, making them easier to handle during the preprocessing phase.

Approach Used

- **Columns Affected**: The columns impacted were primarily "REASON" and "JOB," which contained 252 and 279 blank values, respectively.
- **Replacement with NA**: Used an automated process to convert all blank entries to NA across the relevant columns. This process ensures that all missing values are consistently represented in the dataset, allowing for more streamlined and accurate handling during imputation or removal of missing data.

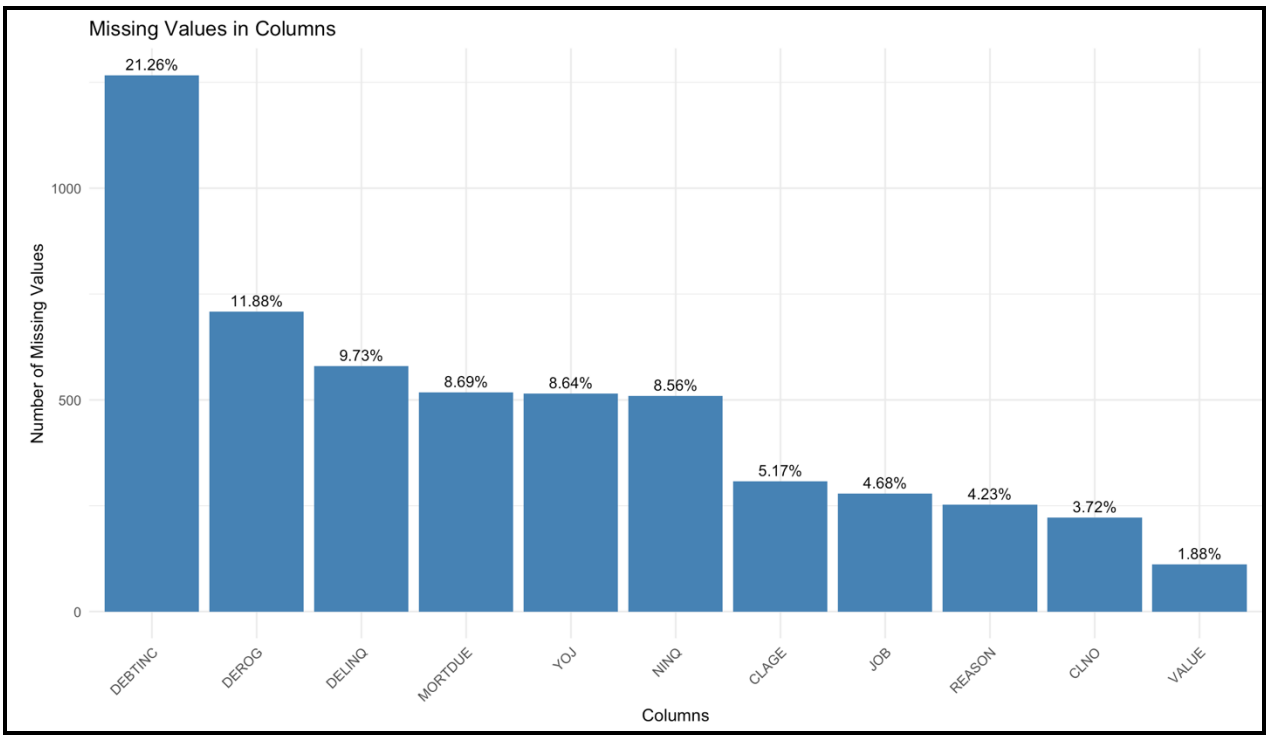
Overview of Missing Values

The Home Improvement Loan Dataset contains missing values across multiple key columns, and understanding the distribution and extent of these missing values is essential to ensure the quality of the predictive model. Proper handling of missing data is necessary to maintain the integrity of the dataset and to prevent biases in the modeling process. Below is a summary of the missing values present in the dataset:

Column Name	***Missing Values Count***	***Missing Percentage***
MORTDUE	*518*	*8.69%*
VALUE	*112*	*1.88%*
REASON	*252*	*4.23%*
JOB	*279*	*4.68%*
YOJ	*515*	*8.64%*
DEROG	*708*	*11.88%*
DELINQ	*580*	*9.73%*
CLAGE	*308*	*5.17%*
NINQ	*510*	*8.56%*
CLNO	*222*	*3.72%*
DEBTINC	*1267*	*21.26%*

(Table 5 – Missing Values Dataset)

To better understand the distribution of missing data, a bar plot was created, focusing on columns that contain missing values. The plot provides a visual representation of the percentage of missing entries for each column.



(Fig 1 – Missing Plot)

The analysis of the missing values reveals the following key insights:

- **DEBTINC (Debt-to-Income Ratio)** contains the most missing values, with 1,267 missing entries, which makes up 21.26% of the dataset.
- **DEROG (Derogatory Marks)** and **DELINQ (Delinquencies)** also have a significant number of missing values, with 708 (11.88%) and 580 (9.73%) missing entries, respectively.
- **MORTDUE (Mortgage Due)** and **YOJ (Years on Job)** have around 8-9% of their data missing, indicating the importance of careful imputation strategies to retain dataset quality.

Addressing these missing values is crucial for maintaining model reliability and ensuring accurate loan approval predictions.

Handling Missing Value

In this section, several enhanced techniques were applied to handle missing values in the Home Improvement Loan Dataset. The primary objective was to clean the dataset effectively while preserving as much valuable information as possible, ensuring the methods used did not introduce bias, and making the dataset suitable for further analysis and modeling. Below is a detailed report of the steps taken.

1. Removal of Rows with Excessive Missing Values

Initially, rows where more than 6 of the selected key columns (MORTDUE, VALUE, REASON, JOB, YOJ, DEROG, DELINQ, CLAGE, NINQ, CLNO, DEBTINC) had missing values were identified which was 126 and removed. This step aimed to eliminate records with insufficient data quality, where retaining them would compromise the robustness of the predictive

model. This reduced the dataset to a cleaner version, with missing values more localized to fewer columns. Now total number of observations remains **5834** observations.

2. **Median Imputation for Numeric Columns**
   Columns: MORTDUE, VALUE, DEROG, DELINQ, CLNO, DEBTINC
   Reasoning:
   - **Median imputation** is used because these columns contain numeric data where the distribution is often skewed (e.g., loan amounts or debt ratios). The median is robust to outliers and ensures that extreme values do not distort the imputation.
   - This method is particularly suitable for financial data like **MORTDUE** (mortgage due), **VALUE** (property value), **DEBTINC** (debt-to-income ratio), etc., where outliers are expected and should not affect the overall data.

   Explanation:
   - **MORTDUE and VALUE**: Since these are related to mortgage and property values, and the distribution is likely skewed, using the median is appropriate.
   - **DEROG and DELINQ**: These represent derogatory marks and delinquencies, which also tend to be skewed in real-life financial data, making median imputation a better choice.
   - **CLNO**: The number of credit lines is a straightforward numeric column with potential skewness due to outliers, so the median is a safe approach.
   - **DEBTINC**: The debt-to-income ratio can vary widely, and using the median will prevent any bias from high outliers.

3. **K-Nearest Neighbors (KNN) Imputation for Numeric Columns**
   Columns: YOJ, CLAGE, NINQ
   Reasoning:
   - **KNN Imputation** is more suitable for columns where the missing data can be better predicted by relationships with other columns. For example, the **years on the job (YOJ)** is likely to be influenced by other financial behavior (e.g., loan amount, mortgage, number of credit inquiries).
   - **CLAGE** (age of the oldest credit line) and **NINQ** (number of inquiries) can depend on credit behavior, making KNN suitable for predicting these values based on relationships with similar rows in the dataset.
   - This method ensures that imputation is based on patterns observed in other variables, providing a more informed estimate.

   Explanation:
   - **YOJ (Years on Job)**: Predicting the number of years an individual has been employed based on similar applicants (considering their loan amounts, number of credit lines, and mortgage dues) can give a better estimation.
   - **CLAGE and NINQ**: KNN helps predict these columns by identifying similar patterns in the credit histories of other applicants. This is important for financial data where credit behavior patterns provide valuable insights.
4. **Imputation of Categorical Variables**
   Columns: REASON, JOB

Reasoning: Missing values in categorical variables REASON and JOB were handled using **mode imputation**, as these variables represent categorical information about the applicant's job type and reason for applying for the loan.

- **REASON and JOB**: The most frequently occurring value (mode) was used to fill in the missing values in REASON and JOB. This ensures that categorical features are imputed with realistic values based on the majority behavior in the dataset. Using mode imputation helps maintain the overall distribution of these categorical variables while minimizing the introduction of bias.
-

Summary of Imputation Techniques Applied

Column	*Imputation Method*	*Rationale*
MORTDUE	*Median Imputation*	Skewed financial data, robust to outliers
VALUE	*Median Imputation*	Skewed property values, appropriate for financial data
DEROG	*Median Imputation*	Skewed number of derogatory marks, likely to have outliers
DELINQ	*Median Imputation*	Skewed delinquencies, median is robust
CLNO	*Median Imputation*	Skewed distribution of credit lines
DEBTINC	*Median Imputation*	Skewed debt-to-income ratios, best handled with median
YOJ	*KNN Imputation*	Strong relationships with other variables, better suited to KNN
CLAGE	*KNN Imputation*	Depends on credit behavior patterns, KNN captures those
NINQ	*KNN Imputation*	Related to credit behavior, KNN models this well
REASON	*Mode Imputation*	Categorical data imputed with the most frequent category
JOB	*Mode Imputation*	Categorical data imputed with the most frequent category

(Table 6 – Missing Value Treatment Summary)

Each imputation method was selected based on the nature of the data and real-world financial practices. **Median imputation** was used for numeric columns to mitigate the impact of outliers, while **KNN imputation** was employed for columns that exhibited strong relationships with other variables. **Mode imputation** was used for categorical features to maintain realistic values, ensuring consistent distribution in these variables. This comprehensive approach ensures that the dataset is well-prepared for subsequent modeling steps, preserving data quality while minimizing bias and enhancing the robustness of future analyses.

Summary of Variables after Imputation

Variable	**Min**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
BAD	0.000	0.000	0.000	0.199	0.000	1.000

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
LOAN	1100	11200	16400	18724	23500	89900
MORTDUE	2063	48021	65030	73227	88898	399550
VALUE	8000	66552	89609	101903	119088	855909
REASON	-	-	-	-	-	-
JOB	-	-	-	-	-	-
YOJ	0.000	3.000	7.000	8.917	13.000	41.000
DEROG	0.0000	0.0000	0.0000	0.2292	0.0000	10.0000
DELINQ	0.0000	0.0000	0.0000	0.4145	0.0000	15.0000
CLAGE	0.0	114.2	172.3	179.1	230.3	1168.2
NINQ	0.000	0.000	1.000	1.164	2.000	17.000
CLNO	0.00	15.00	20.00	21.27	26.00	71.00
DEBTINC	0.5245	30.8469	34.8882	34.1049	37.9917	203.3122

(Table 7 –Summary of Table after imputation)

The summary provides descriptive statistics of the variables in the dataset **after imputation**:

1. **BAD**: Binary variable indicating if a loan applicant defaulted (1 = Default, 0 = No Default). Around 19.9% of applicants have defaulted, as seen from the mean value of 0.199.
2. **LOAN**: Represents the loan amount requested. The values range from $1,100 to $89,900, with a median value of $16,400.
3. **MORTDUE**: Outstanding mortgage balance. The median is $65,030, with values extending up to $399,550.
4. **VALUE**: The value of the property. The median is $89,609, and it has a wide range, with some properties valued as high as $855,909.
5. **REASON**: Categorical variable indicating the reason for the loan. This is textual data and doesn't have numeric statistics.
6. **JOB**: Categorical variable for job type. Like REASON, it doesn't have numeric statistics.
7. **YOJ**: Years on the job. Median years are 7, with a range of 0 to 41 years.
8. **DEROG**: Number of derogatory reports. Most applicants have no derogatory marks (median = 0), but the maximum is 10.
9. **DELINQ**: Number of delinquent credit lines. The median is 0, but the maximum goes up to 15.

10. **CLAGE**: Age of the oldest credit line in months. The median is 172.3 months (about 14 years).
11. **NINQ**: Number of recent credit inquiries. Most have 1 or fewer inquiries, but some have as many as 17.
12. **CLNO**: Number of credit lines. The median is 20, but the maximum is 71.
13. **DEBTINC**: Debt-to-income ratio. Median is about 34.88%, with some applicants having ratios as high as 203.31%.

Visualizations and Exploratory Data Analysis (EDA)

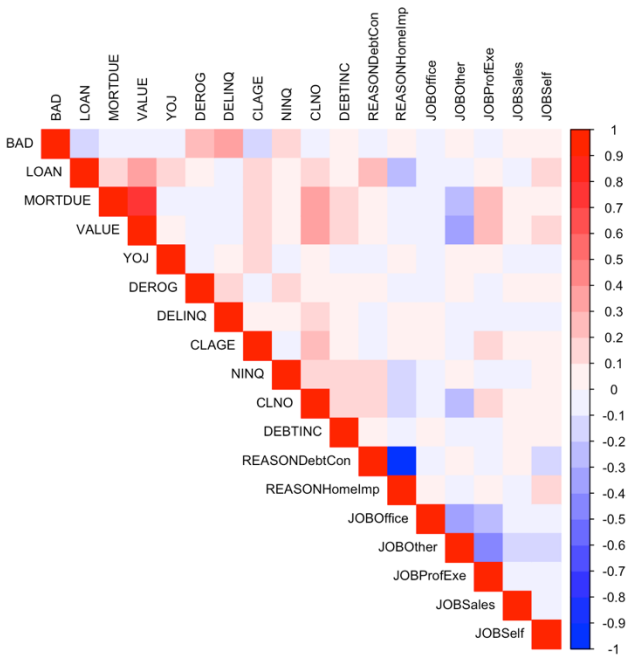
The primary purpose of the visualizations and Exploratory Data Analysis (EDA) is to gain a thorough understanding of the dataset, uncover patterns, and identify relationships among features that are critical for both **predicting loan defaults** and **customer segmentation**. The analyses performed in this section aim to align with the two primary business goals:

1. **Loan Default Prediction**: By understanding the factors that contribute to loan defaults, we can build a predictive model that ensures accurate and interpretable decision-making, thereby complying with regulatory requirements and mitigating financial risks.
2. **Customer Segmentation**: By clustering customers into distinct segments, we can profile those at risk of default, enabling more targeted interventions and personalized loan offerings.

Correlation Matrix

The correlation matrix helps identify which features have the strongest relationships with the target variable (BAD, indicating default status). The closer the correlation value is to 1 or -1, the stronger the relationship with the target. Positive correlations indicate that as the feature increases, the likelihood of default also increases, while negative correlations suggest that as the feature increases, the likelihood of default decreases.

In this analysis, both numerical and encoded categorical features are included, providing a holistic view of all the important factors that contribute to default. The correlation matrix calculation was performed using **Spearman's correlation** to capture both linear and non-linear relationships. Below is a summary of important correlations with the target variable BAD and their interpretations.



(Fig 2 – Correlation Matrix)

Important Correlations with the Target Variable BAD

Variable	***Correlation with BAD***	***Description***	***Interpretation***
DELINQ	*0.33*	*Number of delinquent credit lines*	Positive correlation: More delinquencies increase the likelihood of default.
DEROG	*0.27*	*Number of major derogatory reports*	Positive correlation: More derogatory marks significantly increase default risk.
NINQ	*0.15*	*Number of recent credit inquiries*	Positive correlation: More recent inquiries correlate with a higher chance of default.
DEBTINC	*0.09*	*Debt-to-income ratio*	Positive correlation: Higher debt relative to income increases default risk, indicating financial stress.
JOBSales	*0.05*	*Applicant's job is in sales*	Positive correlation: Applicants with jobs in sales show a slight increase in the likelihood of default.
JOBSelf	*0.05*	*Applicant is self-employed*	Positive correlation: Self-employed applicants have a slightly higher likelihood of default.

Variable	***Correlation with BAD***	***Description***	***Interpretation***
CLAGE	*-0.19*	*Age of the oldest credit line (in months)*	Negative correlation: Older credit lines suggest financial stability, reducing default risk.
LOAN	*-0.11*	*Loan amount*	Slight negative correlation: Higher loan amounts have a weak negative relationship with default risk.
MORTDUE	*-0.07*	*Remaining mortgage due*	Slight negative correlation: Larger mortgages are not strongly correlated with default risk.
VALUE	*-0.07*	*Property value used as collateral*	Slight negative correlation: Higher property values slightly reduce the likelihood of default.
YOJ	*-0.06*	*Years on the job*	Slight negative correlation: Longer job tenure suggests more stability, reducing default risk.

(Table 7 – Correlation Matrix Table)

Key Observations

1. **DELINQ (Delinquencies)**
   - **Correlation**: 0.33 (highest)
   - **Insight**: Applicants with a higher number of delinquent credit lines have a greater risk of defaulting. This variable is one of the most critical indicators of poor financial behavior.
2. **DEROG (Derogatory Reports)**
   - **Correlation**: 0.27
   - **Insight**: The presence of derogatory reports (e.g., bankruptcies, defaults) is a strong indicator of future loan default risk.
3. **NINQ (Recent Credit Inquiries)**
   - **Correlation**: 0.15
   - **Insight**: A higher number of recent credit inquiries correlates with a higher chance of default, suggesting that applicants seeking more credit may experience financial stress.
4. **DEBTINC (Debt-to-Income Ratio)**
   - **Correlation**: 0.09
   - **Insight**: Higher debt-to-income ratios moderately increase the likelihood of default, indicating financial pressure on the borrower.
5. **CLAGE (Age of Oldest Credit Line)**
   - **Correlation**: -0.19
   - **Insight**: Older credit lines are associated with a lower risk of default, indicating that applicants with long-standing credit histories tend to be more financially stable.
6. **LOAN (Loan Amount)**
   - **Correlation**: -0.11

- **Insight**: The loan amount shows a weak negative correlation with default risk. It suggests that defaults are influenced more by credit behavior and financial metrics rather than the loan size itself.

Insights for Model Building

- The features most strongly correlated with BAD (loan default) are **DELINQ (delinquencies)**, **DEROG (derogatory reports)**, and **NINQ (recent inquiries)**. These variables reflect poor financial behavior or increased financial stress, making them important predictors of default.
- Categorical features like JOB (e.g., **Sales** and **Self-employed**) show a slight positive correlation with defaults, indicating that applicants in these occupations are slightly more likely to default compared to other jobs.
- Features indicating financial stability, such as **CLAGE** (older credit lines) and **YOJ** (years on job), show a negative correlation with default risk, suggesting they are protective factors.

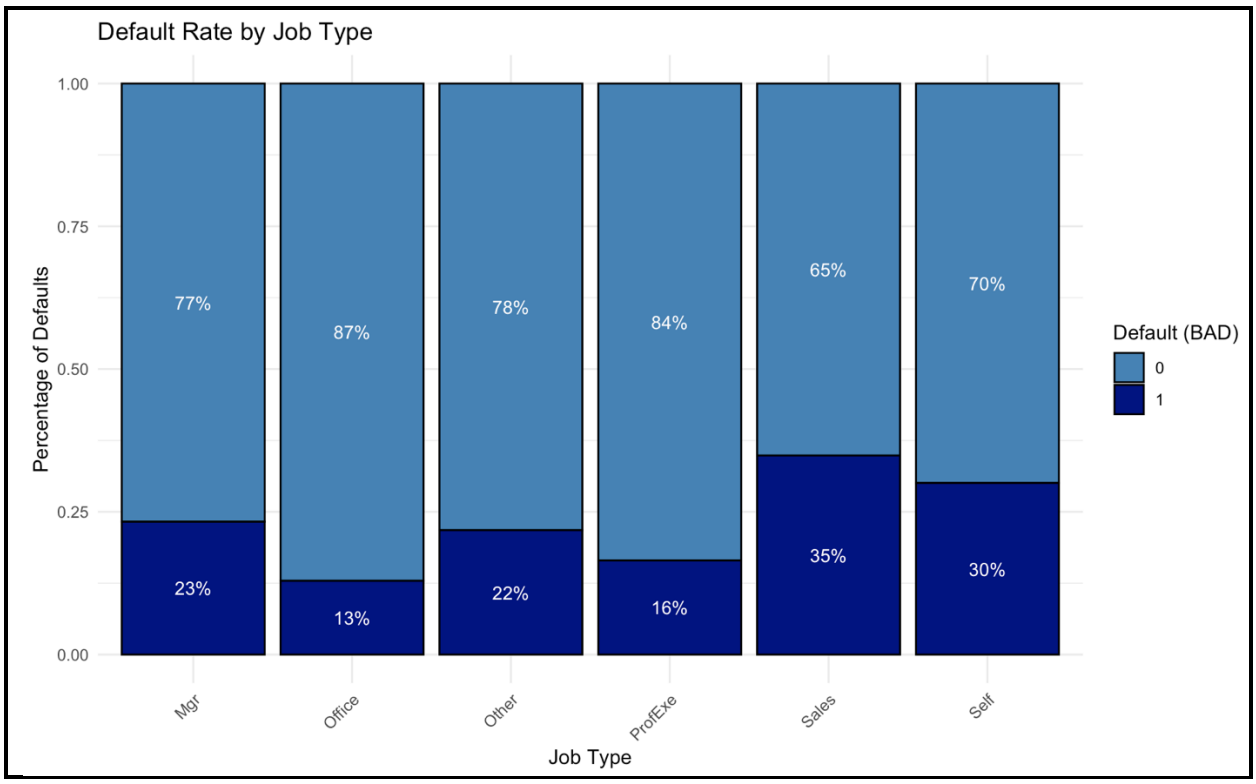
These insights are valuable for prioritizing features in the model-building process. Features with higher correlations are likely to play a more crucial role in predicting loan defaults. Additionally, including encoded categorical variables like job type and loan reason enhances the segmentation and classification capabilities of the model.

Data Visualization for Thorough Understanding of Dataset

To understand the key features and their relationships with loan defaults, a series of visualizations can help shed light on patterns, correlations, and distributions within the dataset. These visualizations are crucial in identifying which variables are most influential in predicting loan defaults.

Bar Plot of Default Rate by Job Type

This **bar plot** will display the default rate for different job categories. It helps identify whether job type is related to loan defaults.



(Fig 3 – Default Rate by Job Type)

The bar plot represents the proportion of loan defaults (BAD = 1) across different job categories. It categorizes loan applicants based on their job type and shows the likelihood of default within each group. The plot uses color coding to differentiate between defaulters and non-defaulters:

- **Light Blue (BAD = 0)**: Applicants who did not default on their loans.
- **Dark Blue (BAD = 1)**: Applicants who defaulted on their loans.

Key Observations:

- **Managers** have a default rate of **23%**, indicating that applicants in managerial roles are relatively financially stable and less likely to default on their loans.
- **Office** job category shows the **lowest default rate** at **13%**, suggesting higher stability in this group.
- **Other** job type has a **22%** default rate, which is quite similar to that of managers.
- **Professional Executives (ProfExe)** show a **16%** default rate, indicating a low risk of default.
- **Sales** job type exhibits a **35%** default rate, one of the highest among all job types, suggesting increased financial risk among applicants in this sector.
- **Self-employed (Self)** individuals show a **30%** default rate, indicating significant financial risk.

Insights:

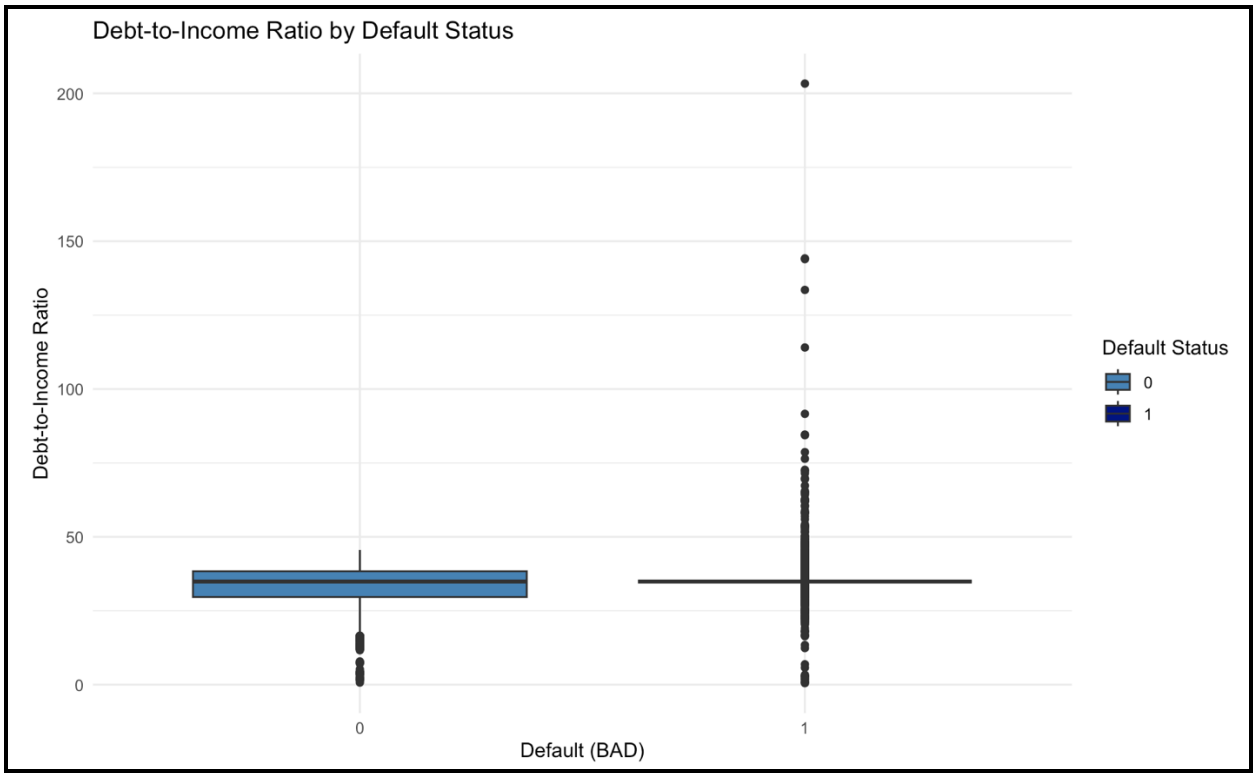
- **Loan Default Risk by Job Category**: The analysis reveals that **Sales** and **Self-employed** job categories have higher default rates, suggesting these applicants are at a greater financial risk. On the other hand, **Office** workers and **Professional Executives** show significantly lower default rates, indicating financial stability.

- **Implications for Risk Assessment**: The bank can adjust its risk assessment strategies based on the job type of applicants:
    - For applicants in **Sales** or **Self-employed** categories, stricter loan terms or additional financial checks could be beneficial.
    - For **Office** workers and **Professional Executives**, offering more favorable loan conditions could be considered, given their lower likelihood of default.

Identifying these trends helps the bank better assess loan risks and act accordingly to ensure financial stability.

Boxplot of Debt-to-Income Ratio by Default Status

A **boxplot** is useful for comparing distributions. This plot compares the **Debt-to-Income Ratio** between applicants who defaulted and those who did not.



(Fig 4 – Debt-To-Income by Default Plot)

The boxplot visualizes the distribution of the **Debt-to-Income Ratio** across two groups: applicants who defaulted on their loans (**BAD = 1**) and those who did not (**BAD = 0**).

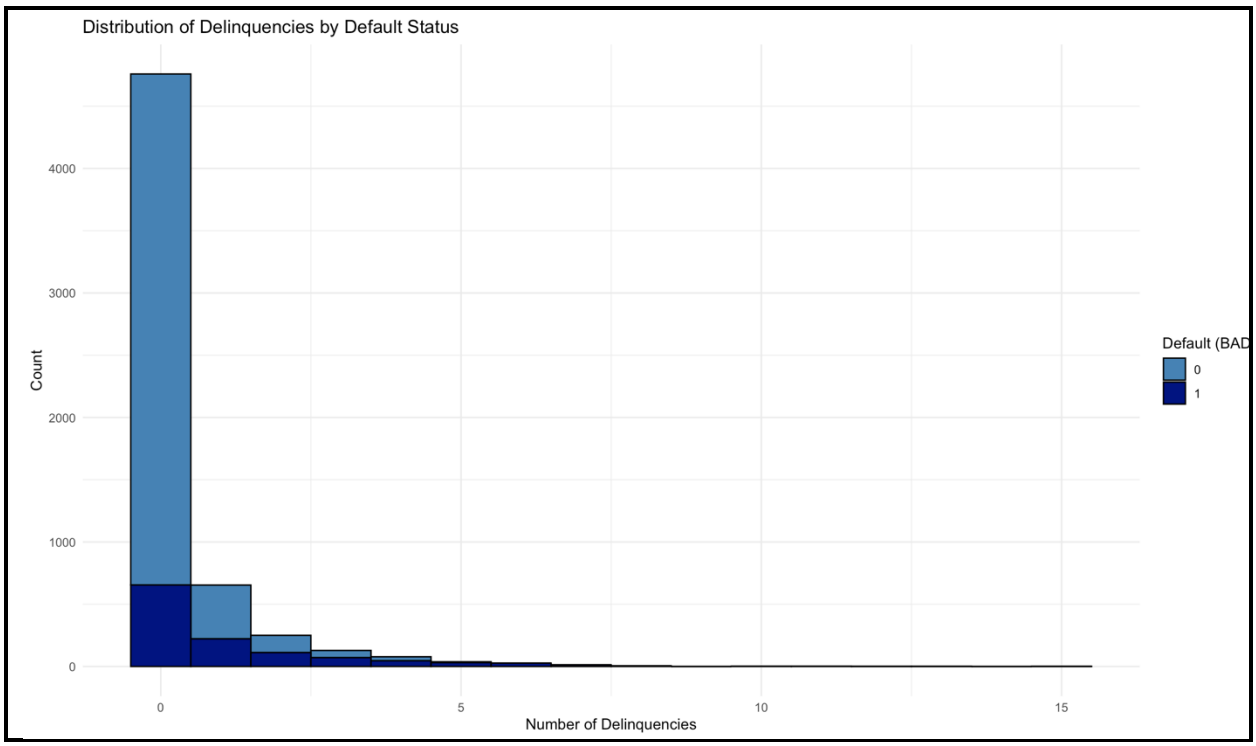
Key Observations:

- **Non-Defaulters (BAD = 0)**:
    - The median Debt-to-Income ratio is slightly below 40%.
    - Most of the values are tightly packed between 30% and 40%, with a few outliers on both ends.
    - There are no extreme values above 50%, indicating that non-defaulters generally have a manageable debt-to-income ratio.

- **Defaulters (BAD = 1)**:
    - The Debt-to-Income ratio distribution is broader compared to non-defaulters.
    - There is a significant spread in the data, with the median ratio slightly higher than that of non-defaulters.
    - Many outliers are present, with some debt-to-income ratios exceeding 150%. This shows that defaulters tend to have much higher and more volatile debt-to-income ratios compared to those who do not default.

Insights:
This plot suggests a strong relationship between the **Debt-to-Income Ratio** and loan default. Applicants with higher ratios are more likely to default, as evidenced by the spread and extreme values in the group of defaulters. This feature is likely to be an important predictor when modeling default risk, as a high debt-to-income ratio may indicate financial overextension and a higher probability of default.

Histogram of Delinquencies (DELINQ)
A **histogram** allows us to see the frequency of different values for a specific variable. Here, we explore how the number of delinquencies is distributed in the dataset.



(Fig 5 – Histogram of DELINQ Plot)

This histogram visualizes the distribution of the number of delinquencies (DELINQ) across loan applicants, grouped by their default status (BAD).
Key Observations:

- **Non-Defaulters (BAD = 0)**:

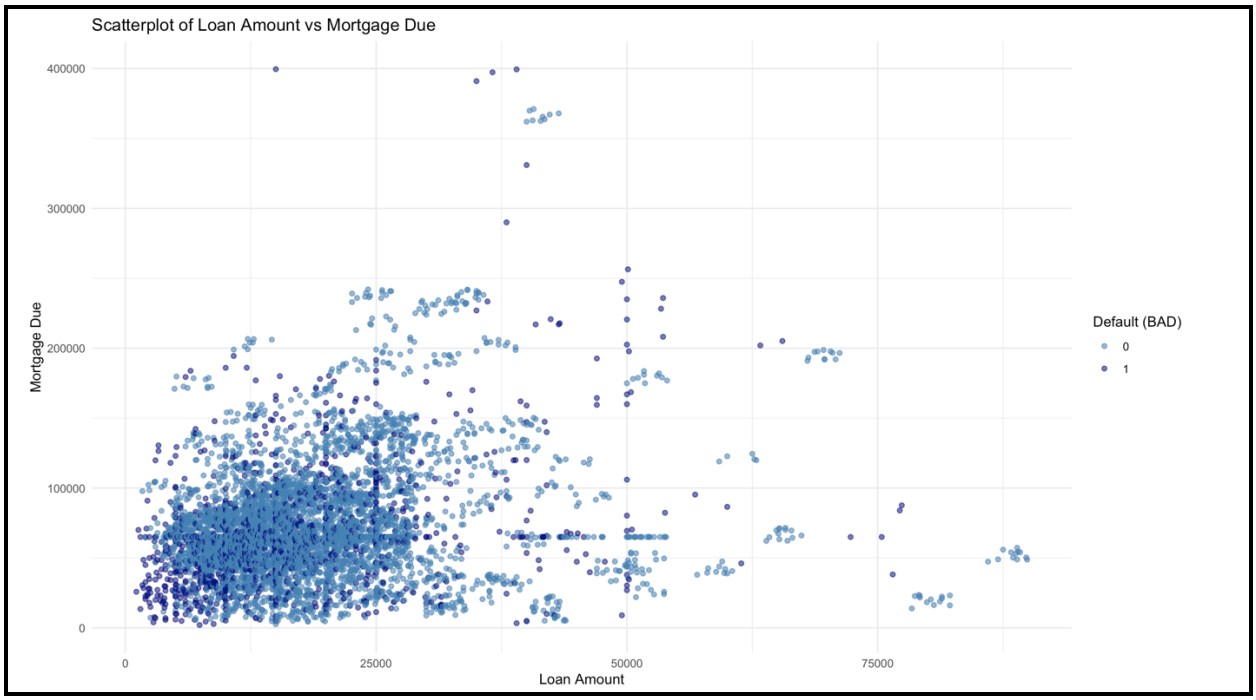
  - The majority of non-defaulters have zero delinquencies, as evidenced by the large spike at zero delinquencies.
  - As the number of delinquencies increases, the number of non-defaulters rapidly declines. Very few non-defaulters have more than 2 delinquencies, with a minimal presence in the 3–15 delinquency range.
- **Defaulters (BAD = 1)**:
  - Defaulters, though fewer in total, have a wider spread of delinquencies.
  - A significant number of defaulters also have zero delinquencies, but a noticeable number exhibit 1–5 delinquencies. This suggests that while having zero delinquencies does not guarantee non-default, a higher number of delinquencies is more common among those who default.

Insights:

- Most loan applicants with zero delinquencies tend to avoid defaulting, which aligns with the assumption that a clean credit history often correlates with financial stability.
- However, some defaulters still show zero delinquencies, indicating that factors beyond delinquencies (such as income or debt load) may contribute to defaults.
- The data suggests that higher delinquency counts (e.g., 2 or more) are more common among defaulters, making **DELINQ** a potentially strong feature for predicting loan defaults.

Scatter Plot of Loan Amount vs. Mortgage Due by Default Status

A **scatter plot** shows the relationship between loan amount and mortgage due, with colors representing default status.



(Fig 6 – Scatter Plot of Loan Amt vs Mortgage Due by Default Status)

This scatter plot illustrates the relationship between **Loan Amount** and **Mortgage Due**, with data points colored by the default status (BAD). Here's an analysis of the pattern:

Key Observations:

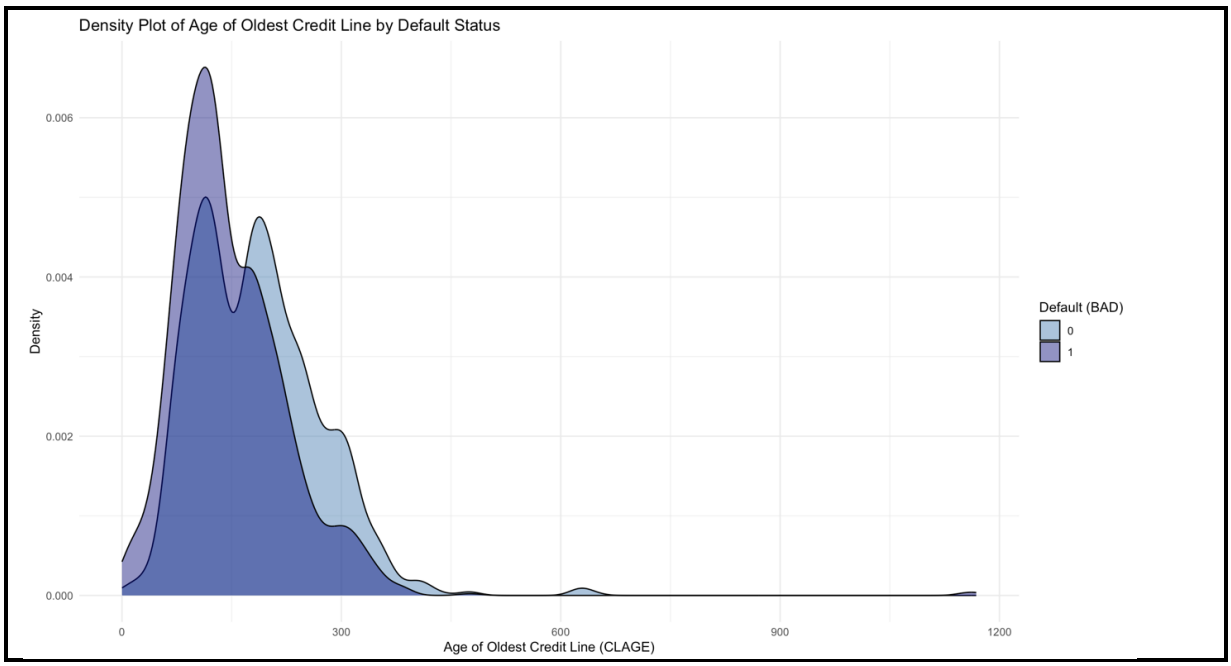
- **Clustering of Loan Amount and Mortgage Due**:
  - Most loan amounts fall within the range of $0 to $25,000, while most mortgage due amounts are clustered between $0 and $150,000.
  - There is a significant overlap between defaulters and non-defaulters in this range, making it harder to distinguish default risk based purely on these two variables alone.
- **Default Behavior**:
  - Darker points (representing defaulters) are scattered throughout, indicating that both defaulters and non-defaulters share similar loan and mortgage distributions, but some high mortgage dues seem more likely to default when the loan amounts are moderate.
  - Some data points with **higher loan amounts** (over $50,000) and **higher mortgage due amounts** (over $200,000) show defaulters, though they are relatively few. These high-value loans/mortgages could represent cases where applicants are at higher risk due to larger financial obligations.

Insights:

- **No clear linear relationship** between loan amount and mortgage due is evident from this plot. While larger loans and mortgages may contribute to risk, the default status is spread across various loan and mortgage values, suggesting that other factors (like income or debt-to-income ratio) likely contribute to default risk.
- This plot suggests that focusing on more specific variables (e.g., **Debt-to-Income Ratio**, **Credit History**) could be more effective in identifying default risks, as the relationship between **Loan Amount** and **Mortgage Due** alone may not provide sufficient predictive power.

Density Plot of Age of Oldest Credit Line (CLAGE)

A **density plot** provides a smooth distribution of a variable. This plot shows the distribution of **CLAGE** (age of the oldest credit line) by default status.



(Fig 7 – Density Plot of Age of Oldest Credit Line)

This density plot provides insights into the distribution of the **age of the oldest credit line (CLAGE)** for both defaulters and non-defaulters.

Key Observations:

- **Non-Defaulters (BAD = 0)**:
  - The peak of the distribution for non-defaulters is concentrated between 0 and 300 months (~25 years). This suggests that non-defaulters typically have older credit lines, which may reflect more stable credit histories.
  - There is a noticeable drop in density for credit lines older than 300 months, but some non-defaulters still have much older credit lines, extending to nearly 1,200 months (~100 years), though these are rare.
- **Defaulters (BAD = 1)**:
  - Defaulters tend to have slightly younger credit lines, with the peak of their distribution being closer to the 0–200-month range (~0-17 years).
  - While defaulters are more likely to have younger credit histories, there are still a few defaulters with older credit lines, though their density declines significantly after the 300-month mark.

Insights:

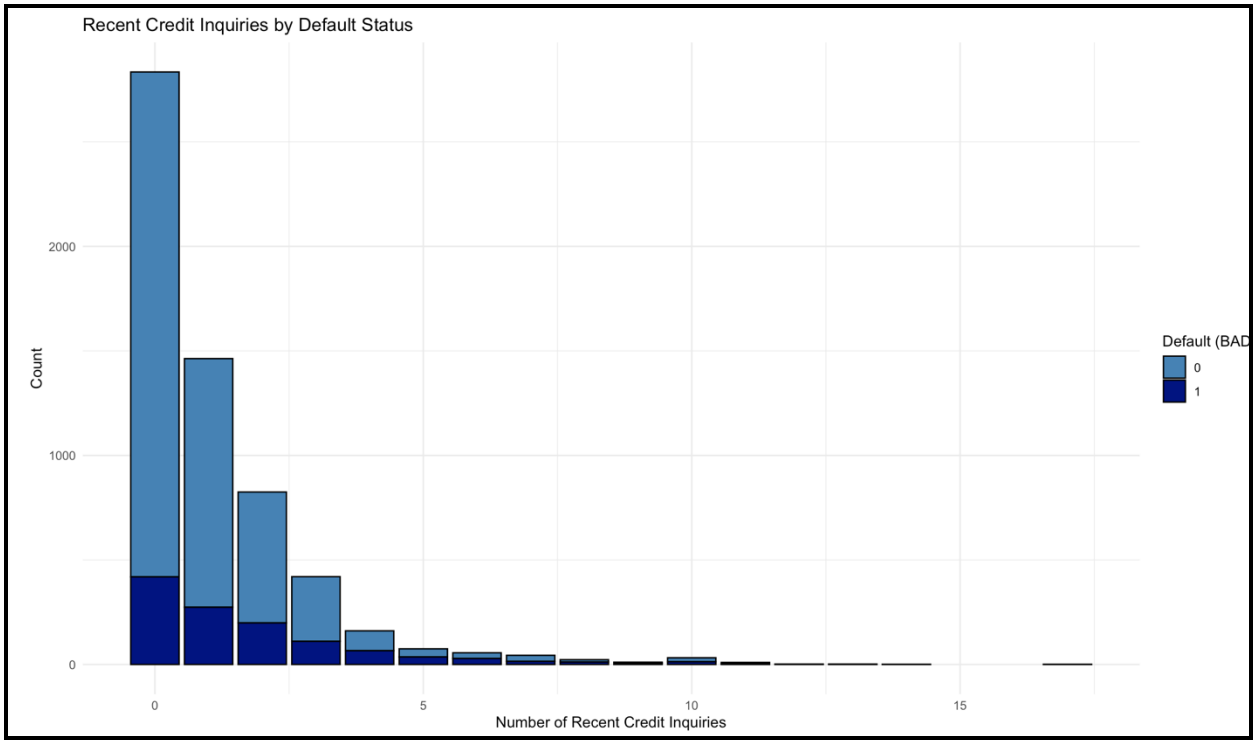
- **Credit Line Age and Default Risk**:
  - The data suggests that having an older credit line is associated with a lower risk of default. Non-defaulters tend to have longer-established credit lines, indicating a more stable financial background.
  - Defaulters generally have younger credit histories, which may correlate with less established financial behavior or a higher likelihood of credit instability.
- **Business Implication**:
  - This analysis shows that **CLAGE** could be a useful predictor in the model for determining default risk. Applicants with shorter credit histories (lower CLAGE)

may need closer scrutiny, as they are more likely to default based on this distribution.

This visualization helps clarify the relationship between credit history length and default risk, further supporting the business goal of identifying high-risk applicants.

Bar Plot of Credit Inquiries (NINQ) by Default Status

This **bar plot** shows the relationship between the number of recent credit inquiries and default status.



(Fig 8 – Recent Credit Inquiries by Default Status)

This bar plot visualizes the distribution of the number of recent credit inquiries (NINQ) for applicants who defaulted (BAD = 1) versus those who did not (BAD = 0).

Key Observations:

- **Zero Credit Inquiries**:
  - The largest proportion of both defaulters and non-defaulters had zero recent credit inquiries. This suggests that a significant number of applicants did not apply for new credit in the recent past, and this could be a common behavior across both groups.
  - However, among those with zero inquiries, a larger proportion were non-defaulters compared to defaulters.
- **1 to 3 Credit Inquiries**:
  - The distribution of defaulters increases slightly for applicants with 1 to 3 recent credit inquiries, indicating that applying for new credit may correlate with a higher likelihood of default.

  - While the number of non-defaulters still dominates in this range, the ratio of defaulters to non-defaulters becomes more balanced as the number of inquiries increases.
- **High Number of Inquiries (4 or more)**:
  - Very few applicants have 4 or more credit inquiries, but among those who do, defaulters are more prevalent. This suggests that having many recent inquiries could be a risk factor for default.

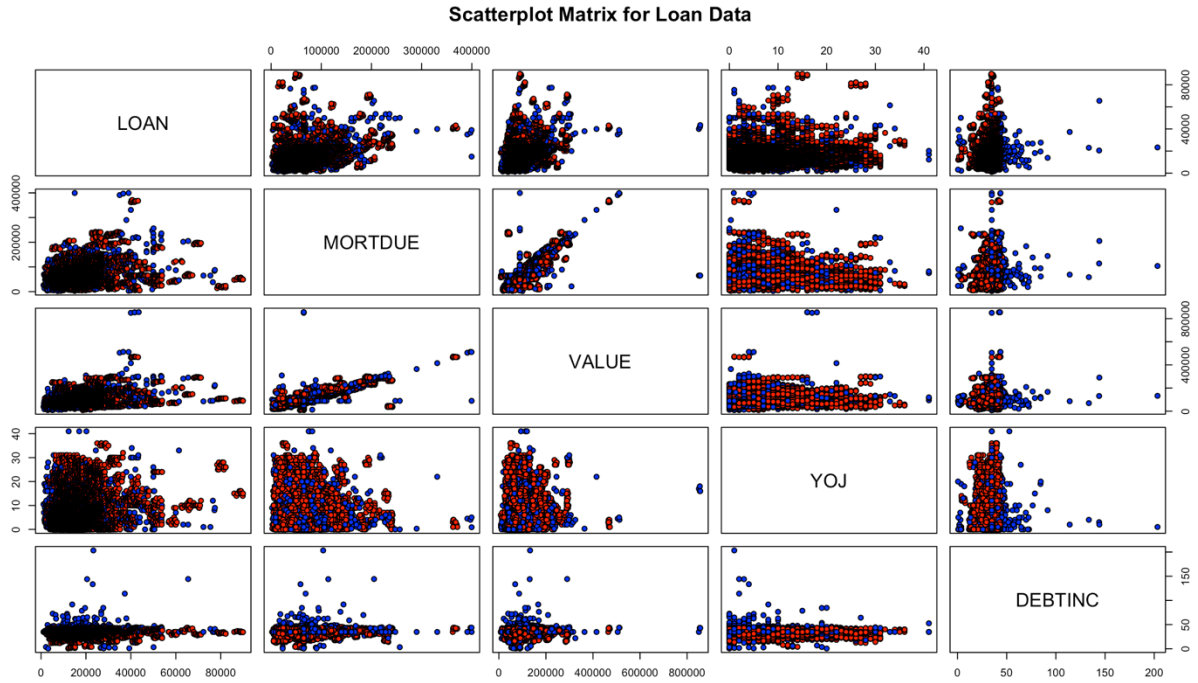
Insights:

- **Default Risk and Recent Credit Inquiries**:
  - Applicants with zero or very few credit inquiries tend to be less risky in terms of default.
  - A higher number of recent credit inquiries is associated with an increased risk of default. This makes sense, as applicants who frequently apply for credit may be experiencing financial distress, which raises their likelihood of defaulting on loans.
- **Business Implication**:
  - **NINQ (Number of Inquiries)** could be a valuable predictor for assessing default risk in applicants. The bank should carefully consider applicants with multiple recent inquiries, as this could indicate a higher risk of default.

This plot provides insight into how applicants' recent credit behavior correlates with their likelihood of defaulting, supporting the bank's business goal of identifying high-risk applicants.

Scatterplot Matrix for Selected Variables with Color by Default Status

The scatterplot matrix above visualizes the relationships between six important variables: **LOAN**, **MORTDUE**, **VALUE**, **YOJ**, **DEBTINC**, and the **BAD** (default status), represented by colors.



(Fig 9 – Scatterplot Matrix with Selected Variable with color)

Key Points from the Scatterplot Matrix:

1. **Color Representation**:
   - **Red Points**: Represent applicants who defaulted (BAD = 1).
   - **Blue Points**: Represent applicants who did not default (BAD = 0).
2. **LOAN vs. Other Variables**:
   - The loan amounts are mostly clustered in the lower range (below $25,000), and both defaulters and non-defaulters show similar distributions.
   - There is no clear relationship between the loan amount and default status, indicating that loan size alone is not a significant predictor of defaults.
3. **MORTDUE (Mortgage Due) vs. VALUE**:
   - **Strong positive correlation**: As the property value increases, the mortgage due also increases. This is expected, as higher property values typically involve larger mortgages.
   - There is a wide spread of both defaulters and non-defaulters within this relationship, suggesting mortgage size alone isn't a clear indicator of risk.
4. **YOJ (Years on Job)**:
   - The distribution of YOJ appears relatively spread across defaulters and non-defaulters, with no distinct separation based on job tenure. This suggests that the number of years someone has been in a job may not be a strong standalone predictor for loan defaults.
5. **DEBTINC (Debt-to-Income Ratio)**:
   - Higher **DEBTINC** values (above 50%) show a concentration of defaulters (red points), indicating that applicants with high debt relative to their income are more likely to default.
   - Non-defaulters (blue points) tend to have debt-to-income ratios below 50%, suggesting this feature could be an important predictor in the model.

Insights:

- **DEBTINC** is one of the stronger variables for separating defaulters from non-defaulters, as seen by the distinct clustering of red points at higher debt-to-income ratios.
- **MORTDUE** and **VALUE** are closely correlated, but default risk is spread across a range of values, indicating that other factors, such as **DEBTINC**, may play a more critical role.
- **YOJ** and **LOAN** do not seem to have a clear impact on default status, as seen by the overlap of red and blue points across their ranges.

This scatterplot matrix helps identify variables that are worth further exploration in model building, such as **DEBTINC**. However, other variables may contribute less to predicting defaults.

Predictor Analysis and Relevancy

To conduct an in-depth analysis of the predictors and determine their importance for predicting loan defaults, various feature selection techniques are employed. The goal is to assess and prioritize predictors based on their contribution to the model's performance, with less impactful features being discarded.

Feature Importance from Random Forest

The **Random Forest** model was employed to assess the importance of various features in predicting loan defaults. This step helps in identifying which predictors significantly contribute to the model's performance, enabling the prioritization of key variables and the elimination of redundant ones. The **Random Forest** provides two metrics for feature importance:

- **%IncMSE (Percentage Increase in Mean Squared Error)**: Represents the increase in prediction error (Mean Squared Error) when the variable is randomly permuted. A higher value indicates that the variable is more important for accurate predictions.
- **IncNodePurity (Increase in Node Purity)**: Measures the total decrease in node impurity (e.g., Gini impurity) due to splits involving the variable across all trees in the forest. A larger value means that the feature contributes more to reducing uncertainty in the model.

Below is a summary of feature importance based on the Random Forest model:

Feature	**%IncMSE**	**IncNodePurity**
LOAN	54.78	71.12
MORTDUE	53.16	66.28
VALUE	57.53	72.76
REASON	26.22	8.35
JOB	42.21	26.53
YOJ	50.48	49.55
DEROG	56.25	50.59
DELINQ	92.86	91.82
CLAGE	66.08	91.88
NINQ	46.01	36.22
CLNO	72.47	67.42
DEBTINC	122.30	252.55

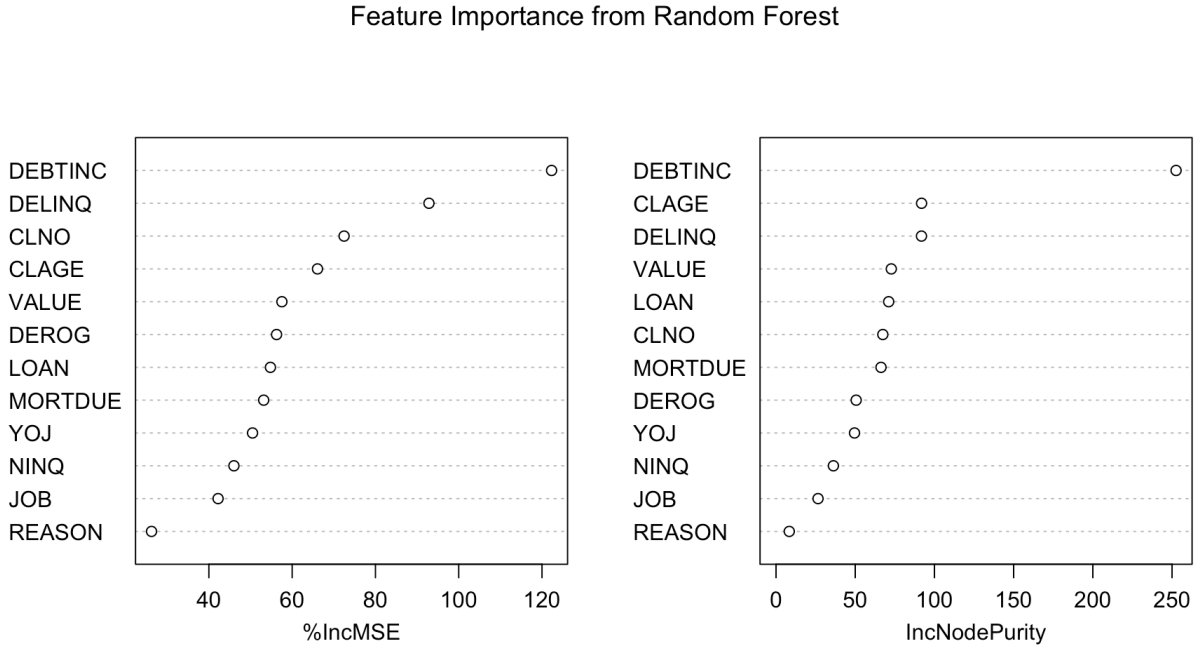
(Table 8 - Summary of Random Forest Feature Selection)

Key Insights from Feature Importance

1. **DEBTINC (Debt-to-Income Ratio)**:
    - **%IncMSE**: 122.30
    - **IncNodePurity**: 252.55
    - **Analysis**: The **Debt-to-Income Ratio** is the most important predictor for loan default. It has the highest impact on both **prediction accuracy** and **node purity**,

suggesting that applicants with higher debt relative to income are at a significantly greater risk of default.

2. **CLNO (Number of Credit Lines)**:
   - **%IncMSE**: 72.47
   - **IncNodePurity**: 67.42
   - **Analysis**: The **Number of Credit Lines** is among the top predictors, indicating that the total number of credit lines held by an applicant is closely associated with default risk. This feature impacts both prediction accuracy and model interpretability.
3. **DELINQ (Delinquencies)**:
   - **%IncMSE**: 92.86
   - **IncNodePurity**: 91.82
   - **Analysis**: **Delinquencies** are a strong indicator of default risk. Applicants with more delinquent credit lines have a higher likelihood of default, making this feature critical for predictive accuracy.
4. **CLAGE (Age of Oldest Credit Line)**:
   - **%IncMSE**: 66.08
   - **IncNodePurity**: 91.88
   - **Analysis**: **Age of the Oldest Credit Line** is a significant feature, implying that applicants with longer credit histories are more stable and thus less likely to default.
5. **VALUE (Property Value) and MORTDUE (Remaining Mortgage Due)**:
   - **%IncMSE** for **VALUE**: 57.53, **MORTDUE**: 53.16
   - **IncNodePurity** for **VALUE**: 72.76, **MORTDUE**: 66.28
   - **Analysis**: Both **property value** and the **remaining mortgage amount** play notable roles in predicting loan default. Higher property values correlate with a reduced risk of default, possibly because larger collateral mitigates financial risk.
6. **REASON and JOB**:
   - **REASON** (%IncMSE: 26.22, IncNodePurity: 8.35) and **JOB** (%IncMSE: 42.21, IncNodePurity: 26.53) show comparatively lower importance, indicating that while job category and reason for the loan provide some insight into default risk, they do not contribute as significantly as other financial indicators.
7. **LOAN (Loan Amount)**:
   - **%IncMSE**: 54.78
   - **IncNodePurity**: 71.12
   - **Analysis**: The **Loan Amount** has moderate importance. Larger loan amounts do not necessarily imply a higher risk of default, suggesting that factors like debt-to-income ratio and delinquencies play a larger role in assessing risk.



(Fig 10 – Feature Importance from Random Forest)

The **feature importance plots** from Random Forest depict two metrics:

- **%IncMSE** (Left Plot): This plot shows how much the mean squared error of the model increases when each feature is permuted. **DEBTINC** and **DELINQ** stand out as critical predictors, with **CLNO** and **CLAGE** also contributing significantly.
- **IncNodePurity** (Right Plot): This plot highlights how features contribute to node purity improvement in decision trees. **DEBTINC** shows the highest impact, followed by **CLAGE** and **DELINQ**.

The analysis of feature importance from Random Forest identifies **Debt-to-Income Ratio (DEBTINC)**, **Delinquencies (DELINQ)**, **Number of Credit Lines (CLNO)**, and **Age of Oldest Credit Line (CLAGE)** as the most impactful predictors for loan defaults. Features such as **REASON** and **JOB** contribute comparatively less and could potentially be deprioritized during further model refinement. Prioritizing features like **DEBTINC** helps create a model that is more predictive of loan defaults, thereby ensuring effective and targeted risk management.

Predictor Analysis and Relevancy Using LASSO Regression

To align with the business goal of automating the loan approval process and ensuring that loan default predictions are interpretable, LASSO (Least Absolute Shrinkage and Selection Operator) regression was employed for predictor analysis. This feature selection method is particularly effective at handling multicollinearity by shrinking less important coefficients to zero, thereby retaining only the most influential predictors for predicting loan defaults.

LASSO Feature Selection Process

The LASSO regression was performed using standardized features from the Home Improvement Loan dataset. The following steps outline the LASSO regression process used:

1. Data Preparation:
    - **Standardization**: All numeric features, along with the one-hot encoded categorical features (e.g., REASON and JOB), were standardized to ensure equal contribution to the LASSO model.
    - **Dataset Split**: The dataset was divided into training (70%) and test (30%) sets using stratified sampling to maintain class distribution. This ensures effective model validation and performance evaluation.
2. Model Training:
    - A LASSO regression model was trained using 10-fold cross-validation on the training dataset to determine the optimal regularization parameter (λ). The λ minimizes the mean squared error (MSE) during cross-validation.
    - The best model was then fitted with the optimal λ value, resulting in a sparse coefficient set where less impactful features were shrunk to zero.
3. Feature Coefficients:
    - The LASSO model coefficients provided insight into feature importance. Features with coefficients close to zero were identified as less relevant and could be removed to enhance model interpretability.

Key Insights from LASSO Feature Selection

The following table summarizes the features and their coefficients from the LASSO regression model:

Feature	Coefficient	Interpretation
Intercept	-1.69	Baseline intercept for the model prediction
LOAN	-0.25	Larger loan amounts slightly reduce the likelihood of default
MORTDUE	-0.17	Higher mortgage dues marginally reduce default risk
VALUE	0.19	Higher property values slightly increase default risk
YOJ	-0.10	More years on the job decrease default likelihood, indicating stability
DEROG	0.50	More derogatory marks significantly increase the likelihood of default
DELINQ	0.87	High delinquency is a strong indicator of default risk
CLAGE	-0.48	Older credit lines suggest stability and lower default risk
NINQ	0.29	More recent inquiries increase the likelihood of default
CLNO	-0.19	Higher number of credit lines slightly reduces default likelihood
DEBTINC	0.48	Higher debt-to-income ratio significantly increases default likelihood
REASONDebtCon	-0.06	Debt consolidation slightly reduces default risk
REASONHomeImp	~0	No significant impact on default prediction
JOBOffice	-0.22	Office job applicants are less likely to default
JOBOther	0.01	Minimal contribution to default prediction

JOBProfExe	~0	Minimal contribution from professional/executive jobs
JOBSales	0.13	Sales jobs show slight increase in default likelihood
JOBSelf	0.09	Self-employed individuals exhibit a slight increase in default risk

(Table 9 – LASSO Feature Selection)

1. **Identification of Key Features:** • Debt-to-Income Ratio (DEBTINC), Delinquencies (DELINQ), and Age of Oldest Credit Line (CLAGE) were the most significant predictors, providing a comprehensive understanding of creditworthiness. • Features like REASONHomeImp, JOBOther, and JOBProfExe were less influential, suggesting their removal for improved model efficiency.
2. **Alignment with Business Goals:** • LASSO regression focused on the most impactful features, aiding in creating a model that is both accurate and interpretable. This ensures compliance with regulations like the Equal Credit Opportunity Act. • A simplified model with fewer predictors enhances interpretability for stakeholders, building trust in the automated decision-making process.
3. **Model Performance Evaluation:** The LASSO model's performance was evaluated using a confusion matrix on the test set:

	Reference	
	0	1
Prediction		
0	1374	211
1	53	112

(Table 10 – LASSO Confusion Matrix)

4. **Statistics:** • **Accuracy**: 84.91% (95% CI: 83.15% - 86.56%) • **Sensitivity**: 96.29% • **Specificity**: 34.67% • **Balanced Accuracy**: 65.48%
5. While the model demonstrates high accuracy in identifying non-defaulters, specificity (defaulter identification) is a challenge, likely due to class imbalance.

The LASSO regression analysis effectively identified critical predictors of loan default, particularly DEBTINC, DELINQ, and CLAGE, supporting automated, interpretable, and regulation-compliant loan approvals. Next steps will include addressing class imbalance and exploring hybrid models for enhanced robustness and predictive capability.

Select Required Variables for Each Dataset

The first step in our transformation process is to differentiate the features that are relevant to our classification and clustering tasks. This approach ensures that each dataset is tailored to meet the specific requirements of the respective analysis.

- **Classification Dataset:**
  - The goal of this dataset is to predict loan defaults (BAD). Variables were selected based on their predictive importance identified through feature selection techniques, such as Random Forest and LASSO regression.

    - **Selected Variables:**
        - **Key Predictive Features:**
            - DEBTINC (Debt-to-Income Ratio), DELINQ (Number of Delinquencies), CLNO (Number of Credit Lines), CLAGE (Age of Oldest Credit Line), DEROG (Number of Major Derogatory Marks), NINQ (Number of Recent Credit Inquiries), YOJ (Years on the Job), MORTDUE (Remaining Mortgage Due), VALUE (Property Value), LOAN (Loan Amount).
        - **Categorical Features:**
            - REASON (Reason for Loan), JOB (Job Category).
        - **Target Variable:**
            - BAD (Indicator for Default Status).
- **Clustering Dataset:**
    - This dataset focuses on segmenting customers by performing clustering separately for defaulters (BAD = 1) and non-defaulters (BAD = 0), uncovering distinct sub-groups and credit behavior patterns.
    - **Selected Variables:**
        - All available variables are included, as the clustering analysis benefits from using a comprehensive feature set to determine distinct customer groups. This includes both categorical features (REASON, JOB) and numeric features (DEBTINC, DELINQ, CLNO, CLAGE, DEROG, NINQ, YOJ, MORTDUE, VALUE, LOAN).

Data Transformation

Classification Dataset

Objective

To prepare the dataset for predictive modeling aimed at predicting loan defaults (target variable: BAD). The data transformation ensures compatibility with the selected classification models: Logistic Regression, Decision Tree, and Random Forest.

Transformation Steps

1. **Retaining Numeric Features As-Is**:
    - Numeric features were kept in their original form without scaling or standardization, as the selected classification models (Logistic Regression, Decision Tree, and Random Forest) do not require standardization.
    - The included numeric features are:
        - DEBTINC, DELINQ, CLNO, CLAGE, DEROG, NINQ, YOJ, MORTDUE, VALUE, and LOAN.
2. **Retaining Categorical Variables As Factors**:
    - The categorical variables, REASON and JOB, were retained as factors without one-hot encoding. This approach simplifies the transformation process while maintaining compatibility with the selected models.
3. **Data Inspection**:

- The dataset was inspected to confirm that the features are clean and correctly formatted.
- The final dataset includes the target variable (BAD), numeric features, and categorical variables (REASON and JOB) as factors, ensuring it is ready for modeling.

Clustering Dataset

The main objective of transforming the dataset is to make it suitable for clustering analysis, enabling the identification of distinct segments based on applicants' credit behaviors. The transformation process aims to ensure consistency, scale uniformity, and relevancy of features, thus supporting more meaningful and interpretable clustering results.

Steps Involved:

1. **Normalization**
   Z-Score Standardization of Numeric Features

- **Min-Max Normalization** was applied to equalize the impact of all numeric features, we applied Z-score standardization. This method centers each numeric feature around a mean of 0 and scales it to have a standard deviation of 1.

$$X' = \frac{(X - min(X))}{\text{std}(X)}$$

Where:

- X' is the standardized value.
- X is the original value.
- mean(X) is the mean of the feature.
- std(X) is the standard deviation of the feature.

Reason:

- Scaling features like **DEBTINC**, **MORTDUE**, and **CLAGE** ensures that no single feature dominates the clustering algorithm due to differences in scale.
- Z-score standardization is particularly suitable for clustering, as it retains the distribution of the data while making features comparable.

Results:

- After normalization, all features in the dataset are scaled between 0 and 1, ensuring consistent scaling across different features.

2. **Using Categorical Variables As-Is (No Encoding)**

- **Process**:
  Categorical variables (REASON and JOB) were retained as factors rather than being converted to binary columns (one-hot encoding).
- **Reason**:
  - For clustering, retaining categorical variables as factors allows for more straightforward integration without inflating the feature space.
  - Avoids introducing unnecessary multicollinearity or complexity.

3. Combining Standardized Numeric Data with Categorical Variables

The standardized numeric features were combined with factored categorical variables to create a final dataset. This comprehensive transformation ensures that all relevant features—both numerical and categorical—are included, supporting a holistic clustering analysis.

- **Process**:
  The standardized numeric features were combined with the categorical variables (REASON and JOB) to create the final dataset.
- **Reason**:
  - This approach ensures that both numerical and categorical features contribute to clustering, providing a comprehensive view of applicant behavior.

4. Segregating the Dataset: Defaulters vs. Non-Defaulters

To fulfill the business objective of segment-specific analysis, we created two separate datasets:

- **Defaulters Dataset:** Contains only applicants who defaulted (BAD = 1).
- **Non-Defaulters Dataset:** Contains applicants who did not default (BAD = 0).

Reason:

- Separating the data into defaulters and non-defaulters allows for focused clustering. It ensures that each segment is analyzed independently, capturing unique behavioral patterns and aiding in more tailored interventions.

5. Why PCA Was Not Applied

Principal Component Analysis (PCA) was not used in this transformation for the following reasons:

- **Preserving Feature Interpretability:** PCA transforms original features into linear combinations, making it difficult to interpret the resulting principal components in the context of business objectives.

- **No High Dimensionality:** The dataset contains 17 variables, which is manageable for clustering. The dimensionality is not high enough to warrant PCA, which is typically used for reducing dimensions in datasets with many features.
- **Clustering Focus:** The goal of clustering is to retain the original feature space, as the clusters need to be interpretable based on the original variables (e.g., **DEBTINC**, **YOJ**). PCA could obscure the relationships between the original variables and clusters, making it less suitable for this analysis.

6. **Summary Statistics of Transformed Datasets**
   Below are the summary statistics of the transformed datasets:

   Table 1: Summary Statistics for Defaulters Dataset

   Defaulters Dataset Summary

Feature	Min	1st Qu.	Median	Mean	3rd Qu.	Max
DEBTINC	-4.4176	0.1031	0.1031	0.3138	0.1031	22.2597
DELINQ	-0.3805	-0.3805	-0.3805	0.7052	1.4557	13.3912
CLNO	-2.115523	-0.723389	-0.126761	-0.0089	0.569306	4.944584
CLAGE	-2.0875	-0.9715	-0.5444	-0.3492	0.1342	11.5311
DEROG	-0.2842	-0.2842	-0.2842	0.5490	0.9559	12.1174
NINQ	-0.68189	-0.68189	-0.09609	0.35196	1.07550	9.27665
YOJ	-1.1735	-0.9103	-0.3839	-0.1350	0.4057	4.2222
MORTDUE	-1.65936	-0.72813	-0.23847	-0.09076	0.23488	7.60900
VALUE	-1.62807	-0.69777	-0.27434	-0.06991	0.19406	13.18516
LOAN	-1.5656	-0.8372	-0.3309	-0.1500	0.2821	5.2123
REASONDebtCon	0.0	0.0	1.0	0.6693	1.0	1.0

Feature	Min	1st Qu.	Median	Mean	3rd Qu.	Max
REASONHomeImp	0.0	0.0	0.0	0.3307	1.0	1.0
JOBOffice	0.0	0.0	0.0	0.1051	0.0	1.0
JOBOther	0.0	0.0	0.0	0.4780	1.0	1.0
JOBProfExe	0.0	0.0	0.0	0.1809	0.0	1.0
JOBSales	0.0	0.0	0.0	0.03273	0.0	1.0
JOBSelf	0.0	0.0	0.0	0.04996	0.0	1.0

(Table 11 – Summary Statistics of Defaulters)

Table 2: Summary Statistics for Non-Defaulters Dataset

Feature	Min	1st Qu.	Median	Mean	3rd Qu.	Max
DEBTINC	-4.39183	-0.58649	0.10306	-0.07795	0.56190	1.50825
DELINQ	-0.3805	-0.3805	-0.3805	-0.1752	-0.3805	4.2101
CLNO	-2.115523	-0.623951	-0.126761	0.002213	0.469868	3.453012
CLAGE	-2.081798	-0.702961	0.009302	0.086770	0.707317	5.486890
DEROG	-0.2842	-0.2842	-0.2842	-0.1364	-0.2842	7.1567
NINQ	-0.68189	-0.68189	-0.09609	-0.08744	-0.09609	5.76187
YOJ	-1.17353	-0.77872	-0.25230	0.03355	0.53733	3.56422
MORTDUE	-1.64639	-0.55922	-0.19113	0.02255	0.39222	6.94336
VALUE	-1.64206	-0.60254	-0.19270	0.01737	0.32435	6.46880
LOAN	-1.51232	-0.61512	-0.15319	0.03726	0.43310	6.32266
REASONDebtCon	0.0	0.0	1.0	0.707	1.0	1.0
REASONHomeImp	0.0	0.0	0.0	0.293	1.0	1.0
JOBOffice	0.0	0.0	0.0	0.1761	0.0	1.0
JOBOther	0.0	0.0	0.0	0.4265	1.0	1.0
JOBProfExe	0.0	0.0	0.0	0.2277	0.0	1.0
JOBSales	0.0	0.0	0.0	0.01519	0.0	1.0
JOBSelf	0.0	0.0	0.0	0.02889	0.0	1.0

(Table 12 – Summary Statistics of Defaulters)

Conclusion

The transformation process resulted in two well-prepared datasets for clustering, with equalized feature scales and clear categorical representations. The datasets align well with

the clustering goal, offering meaningful segmentation based on applicants' financial behaviors. The next steps in clustering analysis are expected to yield interpretable clusters that provide business insights for risk management and targeted interventions.

Data Partitioning

The goal of the data partitioning process is to split the dataset into training and test subsets for both classification and clustering tasks. This helps ensure model reliability, as it allows evaluation on unseen data, thus supporting our business objectives. The partitioning strategy differs for classification and clustering datasets to meet their specific requirements.

Data Partitioning for Classification and Clustering

The dataset contains 5,834 records, and we used an 70-30 split strategy—70% of the data is allocated for training, and 30% for testing. This approach ensures that models are trained on a substantial dataset while preserving a separate subset for performance evaluation.

1. **Classification Dataset Partitioning:**

   For the classification task, predicting loan defaults is the primary objective. Since classification models are sensitive to imbalanced classes, maintaining the original class distribution in both the training and test sets is crucial to avoid bias. This is achieved using **Stratified Sampling**, which ensures that both subsets retain the same class proportions as the original dataset.

   Data split proportion formula used to create partition:

$$n_{train} = [p * n]$$

$$n_{test} = n - n_{train}$$

   Where:
   - p=0.7 Proportion for the training set.
   - n=5834 Total number of records.
   - n_{train} Number of training observations.
   - n_{test} Number of test observations.

   Using this formula,

$$n_{train} = [0.7 * 5834] \approx 4084$$

$$n_{test} = 5834 - 4081 = 1750$$

Stratified Sampling for Class Balance: In stratified sampling, the **proportion of each class** in both the training and test sets should approximately match the original dataset's class proportions:

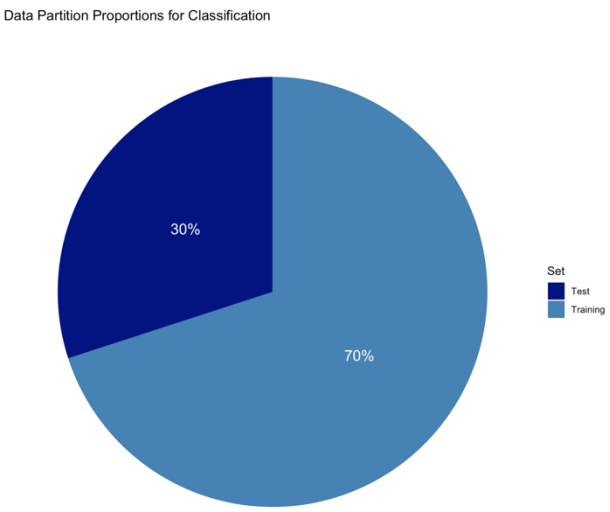
$$Proportion\ of\ class\ i = \frac{n_i}{N}$$

Where:

- n_i Number of observations for class i.
- N=5834N Total number of records.

Stratified sampling maintains class balance, preventing bias toward the majority class and ensuring adequate representation of minority classes for training and evaluation.

In this way, we ensure that the training and test sets both maintain the same class distribution, which is essential to avoid bias towards the majority class during training.



(Fig 11 – Data Partition)

By carefully partitioning the data, the training sets are ensured to contain enough data to effectively learn patterns, while the test sets provide an unbiased assessment of model performance. For the classification dataset, **stratified sampling** is crucial to maintaining class balance, which is vital for accurate model training and evaluation.

Reason for Partitioning

The 70-30 split is a standard strategy that balances model training and evaluation, ensuring that models have enough data to learn while leaving a portion for performance testing. Stratified sampling for classification ensures fair evaluation across all classes. This approach supports accurate, reliable, and generalizable models aligned with the business goals.

Model Selection

Model Selection for Classification Problem: Loan Default Prediction

Objective: The primary goal is to develop an automated decision-making system to accurately predict loan defaults for home improvement loans. This system must provide clear, interpretable explanations for loan rejections to ensure compliance with legal requirements, such as the **Equal Credit Opportunity Act**. The model aims to focus on critical factors like debt-to-income ratio (**DEBTINC**), delinquencies (**DELINQ**), and credit history (**CLAGE**), while maintaining transparency and fairness.

Models to Consider

1. Logistic Regression

- **Why Use It?** Logistic Regression is a commonly used statistical model for binary classification problems like predicting loan defaults (**BAD = 1**). It estimates the probability of default based on predictor variables, making it suitable for achieving the business goal of clear interpretability.
- **Alignment with Business Goal:**
  - The model outputs coefficients that provide a clear understanding of how each feature (e.g., **DEBTINC**, **DEROG**, and **CLAGE**) impacts the probability of loan default.
  - Its interpretability ensures transparency, which is critical for regulatory compliance. By demonstrating which factors led to a specific decision, it allows for accountability in loan approval.
- **Advantages:**
  - **Easy to Interpret:** The coefficients represent the effect of each variable, helping explain the reasons behind loan rejections.
  - **Compliance-Focused:** Logistic Regression is well-suited for cases requiring straightforward decision-making and regulatory adherence, aligning perfectly with the need for interpretability.
  - **Suitable for Small to Medium Datasets:** With 5,834 records, the dataset size is manageable for logistic regression.
- **Disadvantages:**
  - **Linear Relationships Only:** If relationships between variables and loan defaults are non-linear, the model may underperform.
  - **Limited to Predictive Accuracy:** While the model excels in interpretability, it may not capture complex interactions, potentially leading to lower accuracy compared to more advanced models.

2. Random Forest

- **Why Use It?** Random Forest is an ensemble model that builds multiple decision trees and averages their predictions. It is robust in capturing complex interactions between variables, making it ideal for improving predictive accuracy. While less interpretable, it provides feature importance rankings, which can be used to understand significant predictors of default.
- **Alignment with Business Goal:**
  - By using ensemble learning, Random Forest can enhance predictive accuracy, helping identify high-risk applicants more effectively.

  - Although the model is less interpretable than logistic regression, its feature importance output still supports understanding which variables are key risk factors.
  - It can handle missing values, noisy data, and categorical variables well, making it versatile for the given dataset.
- **Advantages:**
  - **Improved Accuracy:** Random Forest captures complex relationships and interactions between variables, boosting the model's performance.
  - **Feature Importance:** It provides a ranking of variables that are most influential in predicting defaults, offering insights into risk factors, which can aid in partial interpretability.
  - **Handles Missing Values:** It is robust against missing data and noisy variables, increasing the model's reliability.
- **Disadvantages:**
  - **Lower Interpretability:** The model is less transparent compared to logistic regression, which could pose challenges for compliance and stakeholder understanding.
  - **Computationally Intensive:** Random Forest requires more computational resources and hyperparameter tuning, making it more complex to implement.

3. Decision Tree

- **Why Use It?** Decision Trees are simple yet effective models for classification tasks. They provide clear decision paths, making them particularly suitable for explaining the decision-making process to stakeholders.
- **Alignment with Business Goal:**
  - Decision Trees offer transparency through visualized decision rules, helping identify why certain applicants were classified as defaulters. This aligns with the goal of providing clear, interpretable decisions.
  - The model's tree-based structure helps to understand interactions between features like **DEBTINC**, **DELINQ**, and **CLAGE**, making it useful for generating comprehensible decisions.
- **Advantages:**
  - **High Interpretability:** The decision paths are transparent, supporting compliance with legal requirements by explaining loan approval or rejection decisions.
  - **Handles Both Numerical and Categorical Variables:** Decision Trees can work directly with mixed data types, simplifying preprocessing.
  - **Less Complexity:** Compared to Random Forest, Decision Trees are simpler and faster, making them easier to implement and interpret.
- **Disadvantages:**
  - **Risk of Overfitting:** If not pruned, Decision Trees may overfit the training data, reducing generalizability on unseen data.
  - **Instability:** A slight change in the data can lead to a significantly different tree, making the model sensitive to variations in the dataset.

Summary of Model Selection

Model	Advantages	Disadvantages	Alignment with Business Goal
Logistic Regression	Easy interpretation of results; suitable for small to medium datasets.	lower accuracy for complex data.	Strong alignment: provides clear explanations and complies with legal standards.
Random Forest	High predictive accuracy; identifies feature importance; handles noisy data well.	Less interpretable; computationally intensive.	Supports accurate risk identification; offers partial interpretability through feature importance.
Decision Tree	Transparent decision rules; easy handling of mixed data types; simple structure.	Prone to overfitting; unstable with slight changes in data.	Provides clear, interpretable decisions, aligning well with compliance requirements.

(Table 13 – Summary of the Classification Model)

Final Model Choice

Given the primary business goal of ensuring interpretability while maintaining accuracy in predicting loan defaults, the model selection process will prioritize interpretability. Therefore:

- **Logistic Regression** is the best initial model choice for its clear, interpretable results, helping meet compliance requirements.
- **Decision Tree** is a secondary choice to offer visual interpretability.
- **Random Forest** will be used to improve accuracy and offer additional insights into variable importance, supporting enhanced predictive performance.

This balanced approach helps achieve both predictive accuracy and interpretability, satisfying the business goal of transparent, compliant decision-making.

Model Selection for Clustering Problem: Customer Segmentation

Objective: The primary objective is to segment customers based on similar financial characteristics to enable targeted interventions, personalized service, and better risk management. The segmentation will be performed separately for defaulters (**BAD = 1**) and non-defaulters (**BAD = 0**), helping the bank to identify specific profiles within each group and align with the business goals of understanding varying credit behaviors more effectively.

Models to Consider

1. K-Means Clustering

- **Why Use It?** K-Means Clustering is one of the most popular and straightforward clustering methods, making it ideal for segmenting customers into distinct groups based on financial characteristics. It works by minimizing the variance within each group, resulting in clusters that are as compact as possible. This approach is well-suited for identifying sub-groups among defaulters and non-defaulters, enabling targeted risk management and personalized interventions.
- **Alignment with Business Goal:**

    - **Defaulter Clustering:** K-Means can help identify sub-groups among defaulters based on factors like high **debt-to-income ratios**, frequent delinquencies, or multiple derogatory reports. These insights can be used to design stricter loan terms or offer financial counseling to high-risk customers.
    - **Non-Defaulter Clustering:** For non-defaulters, K-Means can uncover groups with longer credit histories, low debt loads, or consistent payment behaviors. This supports the bank's efforts to offer personalized financial products, such as more favorable loan terms or upsell opportunities for low-risk customers.
    - It allows the bank to segment customers into **distinct risk categories**, enhancing operational efficiency and customer engagement by tailoring services to different customer needs.
- **Advantages:**
    - **Efficient and Easy to Implement:** K-Means is computationally efficient, making it suitable for the size of the dataset (5,834 records). Its speed and simplicity allow for quick clustering of customer segments.
    - **Effective for Spherical Clusters:** The method works well for clusters with similar densities, which is often the case in financial datasets where customer behaviors form clear, separable patterns.
- **Disadvantages:**
    - **Requires Pre-Specifying Number of Clusters:** The value of **k** (the number of clusters) must be determined beforehand, typically using methods like the **Elbow Method** or **Silhouette Analysis**. If the number of clusters is not well-chosen, it can lead to inaccurate segmentation.
    - **Sensitive to Outliers:** Outliers can significantly affect the positioning of centroids, potentially leading to skewed clusters. However, this challenge can be mitigated through preprocessing steps like scaling and standardization.

2. Hierarchical Clustering

- **Why Use It?** Hierarchical Clustering is a method that does not require pre-specifying the number of clusters, making it valuable when there is limited understanding of how many customer segments might exist. It builds a tree-like structure (dendrogram), which provides a detailed visualization of potential cluster formations, offering insights into customer grouping.
- **Alignment with Business Goal:**
    - **Defaulter Clustering:** Hierarchical Clustering can reveal hierarchical relationships among defaulters, helping identify layers of risk levels, such as extremely high-risk vs. moderately high-risk groups. This allows for finer granularity in designing interventions for specific defaulter profiles.
    - **Non-Defaulter Clustering:** Similarly, it helps in understanding which non-defaulters exhibit characteristics of low-risk or even 'ideal borrowers,' enabling the bank to target them with specific marketing campaigns or loan products.
    - The tree structure generated by Hierarchical Clustering provides a clear, interpretable representation of customer segments, making it easy for decision-makers to understand and act upon clustering results.
- **Advantages:**

- **No Need for Pre-Specified Clusters:** Hierarchical Clustering does not require predefining the number of clusters, offering flexibility and more exploratory power in finding natural groupings within the data.
- **Visual Interpretability:** The dendrograms generated provide a visual representation of the hierarchical relationships among customers, making it easier to identify where to 'cut' the tree to form meaningful clusters.
- **Detailed Clustering Information:** It allows for a clear understanding of the hierarchical nature of customer segments, aiding in better decision-making for customer risk management.

- **Disadvantages:**
  - **Computationally Expensive:** Hierarchical Clustering can be slower and more memory-intensive compared to K-Means, especially as the dataset size increases. However, since the dataset is manageable (5,834 records), this issue can be mitigated.
  - **Sensitive to Noise and Outliers:** Like K-Means, it can be influenced by outliers, which may affect the clarity of the dendrogram. This makes preprocessing steps like standardization crucial.

Summary of Model Selection

Model	Advantages	Disadvantages	Alignment with Business Goal
K-Means Clustering	Efficient, easy to implement, and effective for spherical clusters.	Requires pre-specifying k; sensitive to outliers.	Helps segment defaulters and non-defaulters into clear sub-groups, supporting targeted risk management and personalized interventions.
Hierarchical Clustering	No need to pre-specify clusters; interpretable dendrograms.	Computationally intensive; sensitive to noise and outliers.	Reveals hierarchical relationships within customer groups, enabling granular understanding of customer profiles for precise risk management.

(Table 14 – Summary of Clustering Model)

Final Model Choice

Given the business goal of segmenting customers into distinct groups for targeted risk management and personalized service, both **K-Means Clustering** and **Hierarchical Clustering** will be used:

1. **K-Means Clustering:**
   - K-Means will be used for its efficiency in creating clear and distinct customer segments, which is crucial for implementing real-time interventions and personalized services.
   - The number of clusters (**k**) will be determined using methods like the **Elbow Method** or **Silhouette Analysis**, ensuring optimal segmentation for both defaulters and non-defaulters.
2. **Hierarchical Clustering:**

- Hierarchical Clustering will complement K-Means by offering a deeper, exploratory understanding of customer groupings. It will help identify hierarchical relationships within defaulters and non-defaulters, allowing for more granular customer segmentation.

This dual approach will ensure that the bank achieves precise customer segmentation, supporting enhanced decision-making, personalized services, and targeted risk management—aligned with the overall business objective of managing credit risks more effectively.

Model Building

Model Building for the Classification

Logistic Regression

The purpose of this analysis is to build a logistic regression model aimed at predicting the likelihood of loan defaults for home improvement loans. This model seeks to provide a transparent and interpretable approach to assist financial institutions in making well-informed lending decisions while adhering to regulatory requirements for fairness and transparency.

Model Summary:

The logistic regression model was trained on a dataset containing various features such as the debt-to-income ratio (DEBTINC), delinquencies (DELINQ), the number of credit lines (CLNO), the age of the oldest credit line (CLAGE), derogatory credit reports (DEROG), and job type (JOB), among others. The final model was built using a 70% training set and evaluated on a 30% test set.

Logistic Regression Coefficients

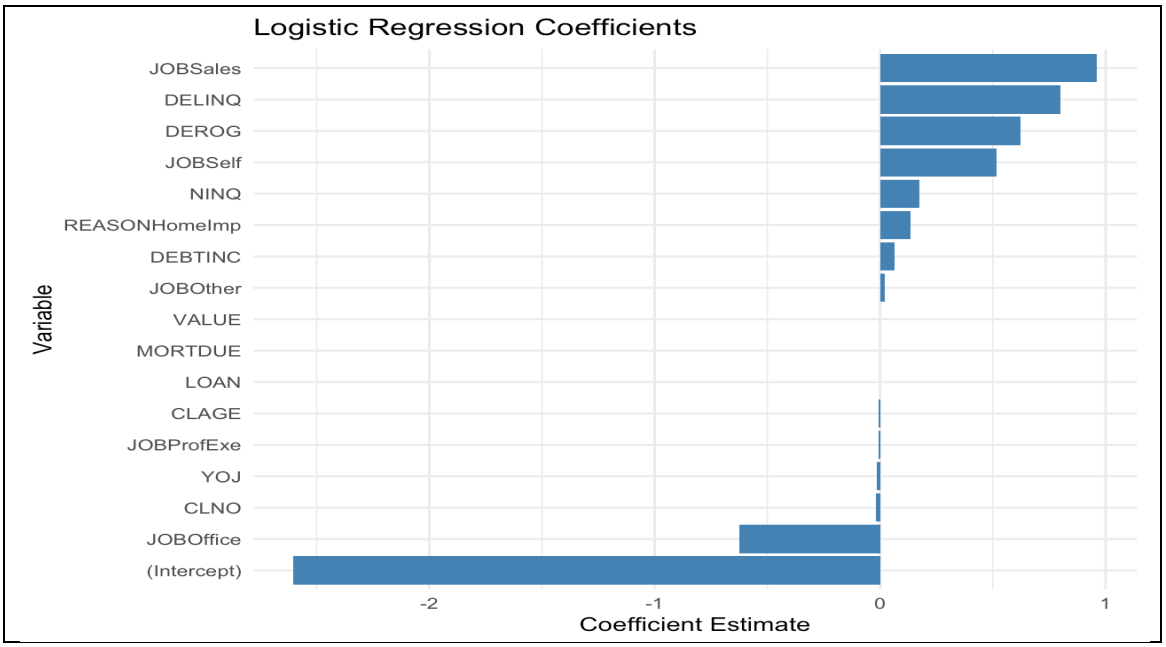
```
> summary(logistic_model)

Call:
glm(formula = BAD ~ ., family = binomial, data = classification_train)

Coefficients:
                Estimate   Std. Error z value             Pr(>|z|)
(Intercept)   -2.602503157  0.324285117  -8.025  0.00000000000000101 ***
DEBTINC        0.063588674  0.007452157   8.533 < 0.0000000000000002 ***
DELINQ         0.801717244  0.046784247  17.136 < 0.0000000000000002 ***
CLNO          -0.018977769  0.005304007  -3.578             0.000346 ***
CLAGE         -0.005651244  0.000666212  -8.483 < 0.0000000000000002 ***
DEROG          0.621039077  0.063527958   9.776 < 0.0000000000000002 ***
NINQ           0.174228965  0.023831462   7.311  0.00000000000026540 ***
YOJ           -0.014054835  0.006774669  -2.075             0.038022 *
MORTDUE       -0.000004380  0.000001838  -2.383             0.017171 *
VALUE          0.000003679  0.000001343   2.740             0.006138 **
```

```
LOAN          -0.000023300  0.000005137  -4.536  0.00000573473355964 ***
REASONHomeImp  0.133681135  0.103166368   1.296              0.195051
JOBOffice     -0.624305643  0.183438647  -3.403              0.000666 ***
JOBOther       0.021675923  0.140360900   0.154              0.877271
JOBProfExe    -0.006003628  0.161689769  -0.037              0.970381
JOBSales       0.961744131  0.307212837   3.131              0.001745 **
JOBSelf        0.518637942  0.266854566   1.944              0.051953 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4145.4  on 4083  degrees of freedom
Residual deviance: 3161.9  on 4067  degrees of freedom
AIC: 3195.9

Number of Fisher Scoring iterations: 5
```

(Table 15 – Logistic Regression Model Coefficient)

Key Coefficients:
The coefficients from the logistic regression model highlight the impact of each predictor on the probability of loan default. The significant predictors include:

- **JOBSales**: This feature has a high positive coefficient (0.96), indicating that applicants in sales roles are significantly more likely to default, reflecting higher associated risk.
- **DELINQ (Number of Delinquencies)**: A strong positive coefficient (0.80) suggests that applicants with a history of delinquencies have a higher probability of default.
- **DEBTINC (Debt-to-Income Ratio)**: This variable shows a positive coefficient (0.06), indicating that a higher debt-to-income ratio increases the likelihood of default.
- **DEROG (Number of Derogatory Credit Reports)**: The coefficient (0.62) highlights that derogatory credit reports are strongly associated with a higher risk of default.
- **CLAGE (Age of Oldest Credit Line)**: A negative coefficient (-0.0057) indicates that a longer credit history reduces the likelihood of default.
- **JOBOffice**: A negative coefficient (-0.62) signifies that applicants in office jobs have a lower risk of default compared to other job categories.



(Fig 13 – Logistic Regression Coefficient)

A bar plot of the model's coefficients illustrates the positive and negative impacts of various features. Job type, delinquencies, debt-to-income ratio, and derogatory reports are among the most significant predictors.

Confusion Matrix and Performance Metrics

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1373  211
         1   54  112

               Accuracy : 0.8486
                 95% CI : (0.8309, 0.8651)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : 0.0001445

                  Kappa : 0.3804

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.34675
            Specificity : 0.96216
         Pos Pred Value : 0.67470
         Neg Pred Value : 0.86679
             Prevalence : 0.18457
         Detection Rate : 0.06400
```

```
   Detection Prevalence : 0.09486
      Balanced Accuracy : 0.65445

       'Positive' Class : 1
```

(Table 16 – Logistic Regression Confusion Matrix)

Confusion Matrix Breakdown:

Prediction	No Default (0)	Default (1)	Total
No Default (0)	1373 (True Negatives)	54 (False Positives)	1427
Default (1)	211 (False Negatives)	112 (True Positives)	323
Total	1584	166	1750

(Table 17 – Logistic Regression Confusion Matrix Breakdown)

- **Accuracy**: The logistic regression model achieved an overall accuracy of 84.86%, showing its ability to correctly predict loan defaults.
- **Precision**: The precision of 0.675 indicates that among those predicted as defaulters, 67.5% were correctly classified.
- **Recall** (**Sensitivity**): The recall of 0.347 reflects that the model correctly identified 34.7% of actual defaulters.
- **F1 Score**: The F1 score of 0.458 balances precision and recall, indicating moderate performance in identifying defaulters.
- **Specificity**: The specificity of 0.962 demonstrates strong performance in correctly identifying non-defaulters, minimizing false positives.

Feature Importance:

- **Job Type Impact**: Applicants in **sales roles (JOBSales)** and **self-employed roles (JOBSelf)** are at higher risk of default, likely due to inconsistent income. Conversely, those in **office roles (JOBOffice)** have a lower risk.
- **Debt-to-Income Ratio**: A higher ratio increases the likelihood of default, underscoring the importance of managing debt relative to income.
- **Credit History**: Longer credit histories (CLAGE) are associated with a reduced risk of default, emphasizing the value of financial stability over time.

Conclusion:

The logistic regression model effectively predicts loan defaults with an accuracy of **84.86%** and provides interpretable insights into key predictors such as job type, delinquencies, and debt-to-income ratio. While the model performs well in identifying non-defaulters, its sensitivity (recall) could be improved to better detect true defaulters. Overall, this model supports data-driven lending decisions and compliance with fairness and transparency requirements.

Random Forest Model

The purpose of this analysis is to build a Random Forest model aimed at predicting the likelihood of loan defaults for home improvement loans. This model leverages the power of ensemble learning to provide high predictive accuracy while handling large datasets efficiently. It also offers an interpretable approach to assist financial institutions in making well-informed lending decisions.

The Random Forest model was trained on a dataset containing various features such as the debt-to-income ratio (DEBTINC), delinquencies (DELINQ), the number of credit lines (CLNO), the age of the oldest credit line (CLAGE), derogatory credit reports (DEROG), and job type (JOB), among others. The final model was built using an 70% training set and evaluated on a 30% test set.

Random Forest Model Details

```
Call:
 randomForest(formula = BAD ~ ., data =
classification_train,      importance = TRUE, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 8.18%
Confusion matrix:
     0   1 class.error
0 3169  77   0.0237215
1  257 581   0.3066826
```

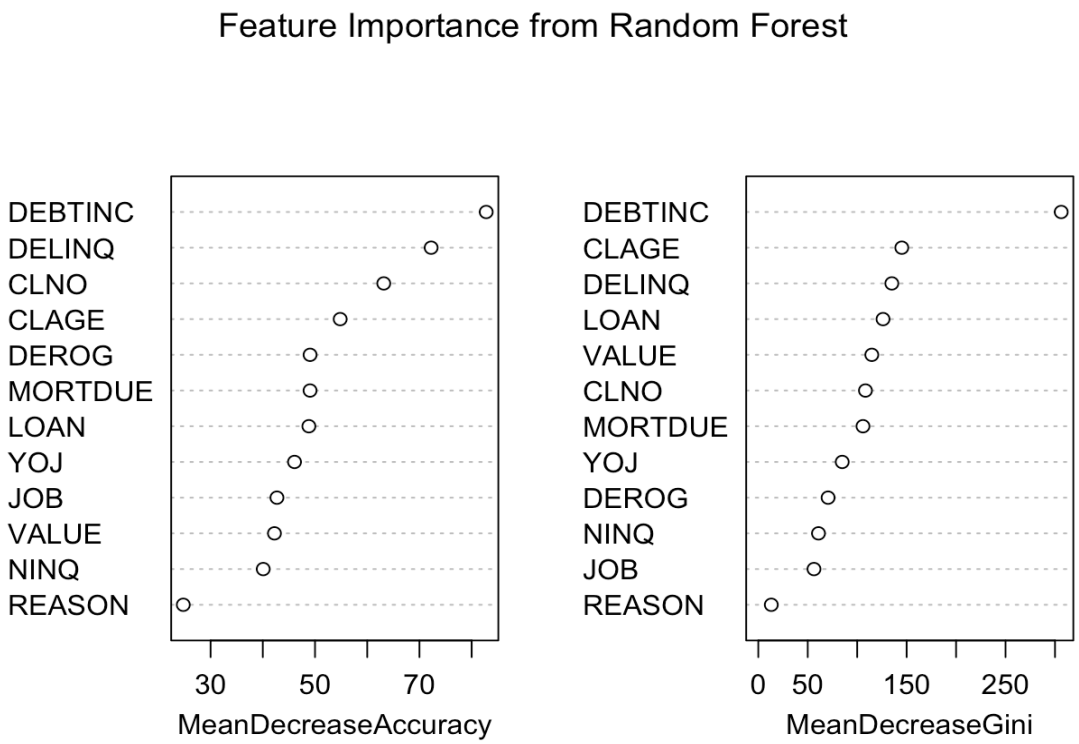
(Table 18 – Random Forest Model Summary)

The model was trained using a dataset of historical loan data where the target variable, BAD, indicates whether the borrower defaulted (1) or did not default (0).

Metric	Value
Algorithm	Random Forest Classification
Number of Trees (ntree)	500
Variables Tried at Each Split (mtry)	3
OOB Error Rate	8.18%

(Table 19 – Random Forest Model Description)

The data used for this classification task was split into a training set and a test set. The classification_train dataset was used for training, and classification_test was used for evaluating the performance of the trained model. In the training set, 70% of the data was used, with the remaining 30% reserved for testing.



(Fig 14 – Future Importance from Random Forest)

Key Feature Importance:

The Random Forest model highlights the most important predictors based on their contributions to reducing Gini impurity and model accuracy:

1. **DEBTINC (Debt-to-Income Ratio)**: This feature is the most significant predictor, indicating that applicants with higher debt-to-income ratios are more likely to default.
2. **CLAGE (Age of Oldest Credit Line)**: A longer credit history reduces default risk, making this feature critical for prediction.
3. **DELINQ (Number of Delinquencies)**: A higher number of delinquencies strongly correlates with increased default risk.
4. **MORTDUE (Outstanding Mortgage Balance)**: This feature also plays an essential role, with higher balances associated with a greater likelihood of default.
5. **DEROG (Number of Derogatory Public Records)**: The presence of derogatory reports (e.g., bankruptcies) is a key indicator of default risk.

Performance Metrics on For Random Forest Model:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1396   91
         1   31  232
```

```
               Accuracy : 0.9303
                 95% CI : (0.9173, 0.9418)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7505

 Mcnemar's Test P-Value : 0.00000009213

            Sensitivity : 0.7183
            Specificity : 0.9783
         Pos Pred Value : 0.8821
         Neg Pred Value : 0.9388
             Prevalence : 0.1846
         Detection Rate : 0.1326
   Detection Prevalence : 0.1503
      Balanced Accuracy : 0.8483

       'Positive' Class : 1
```

(Table 16 – Confusion matrix for Random Forest)

Confusion Matrix Breakdown:

Confusion Matrix Breakdown (Test Data):

Prediction	No Default (0)	Default (1)	Total
No Default (0)	1396 (True Negatives)	91 (False Positives)	1487
Default (1)	31 (False Negatives)	232 (True Positives)	263
Total	1427	323	1750

(Table 20 – Confusion matrix for Random Forest)

Model Insights:

- **Accuracy**: The model achieved an accuracy of **93.03%**, showcasing its effectiveness in distinguishing defaulters from non-defaulters.
- **Precision**: A precision of **0.882** indicates that 88.2% of applicants predicted to default were actual defaulters, minimizing false positives.
- **Recall (Sensitivity)**: The recall of **0.718** signifies the model's ability to correctly identify 71.8% of actual defaulters, reflecting room for improvement in detecting defaults.
- **F1 Score**: The F1 score of **0.792** balances precision and recall, emphasizing strong overall performance.
- **Specificity**: The model's specificity of **0.9783** underscores its excellent ability to correctly identify non-defaulters, minimizing false negatives.

Conclusion:

The Random Forest model is a robust and effective tool for predicting loan defaults, with an accuracy of **93.03%** and a high precision of **88.2%**. It identifies key predictors such as debt-to-income ratio, age of the oldest credit line, and delinquencies, aligning with industry practices. While the model demonstrates strong performance, future

improvements could focus on enhancing sensitivity to better detect actual defaulters. This model provides actionable insights to financial institutions, enabling data-driven decision-making and reducing credit risk.

Decision Tree Model

The goal of this analysis is to use a Decision Tree model to predict loan defaults for home improvement loans. This model helps financial institutions assess applicants' likelihood of defaulting, using critical financial predictors like debt-to-income ratio (DEBTINC), delinquencies (DELINQ), and the age of the oldest credit line (CLAGE). Decision Trees are chosen for their simplicity, interpretability, and ability to provide clear decision paths, making them suitable for credit risk assessments.

Model Summary

```
Call:
rpart(formula = BAD ~ ., data = classification_train, method = "class")
  n= 4084

          CP nsplit rel error    xerror       xstd
1 0.08472554      0 1.0000000 1.0000000 0.03079707
2 0.04773270      3 0.7458234 0.7732697 0.02786296
3 0.03301512      4 0.6980907 0.7052506 0.02682910
4 0.02863962      7 0.5990453 0.6587112 0.02607312
5 0.01073986      8 0.5704057 0.5966587 0.02499667
6 0.01000000      9 0.5596659 0.5930788 0.02493199

Variable importance
DEBTINC  DELINQ   CLAGE    LOAN   DEROG MORTDUE    NINQ   VALUE    CLNO
     63      20       4       3       3       2       2       2       1
```

(Table 21 – Decision Model Summary)

Metric	Value
Algorithm	Decision Tree Classification
Training Data	The model was trained on the **classification_train** dataset (4,084 observations) with the target variable **BAD** indicating loan default (1 = default, 0 = no default).

(Table 22 – Decision Model Summary)

- **Key Splits**: The tree splits on features such as **DELINQ** (delinquencies), **DEBTINC** (debt-to-income ratio), and **CLAGE** (age of the oldest credit line).
- The data was split into an 70/30 train-test split, and the model was evaluated on the test data.

Performance Metrics Summary:
Performance Metrics:

- **Confusion Matrix (Test Data):**

Prediction	No Default (0)	Default (1)	Total
No Default (0)	1337 (True Negatives)	100 (False Positives)	1437
Default (1)	90 (False Negatives)	223 (True Positives)	313
Total	1427	323	1750

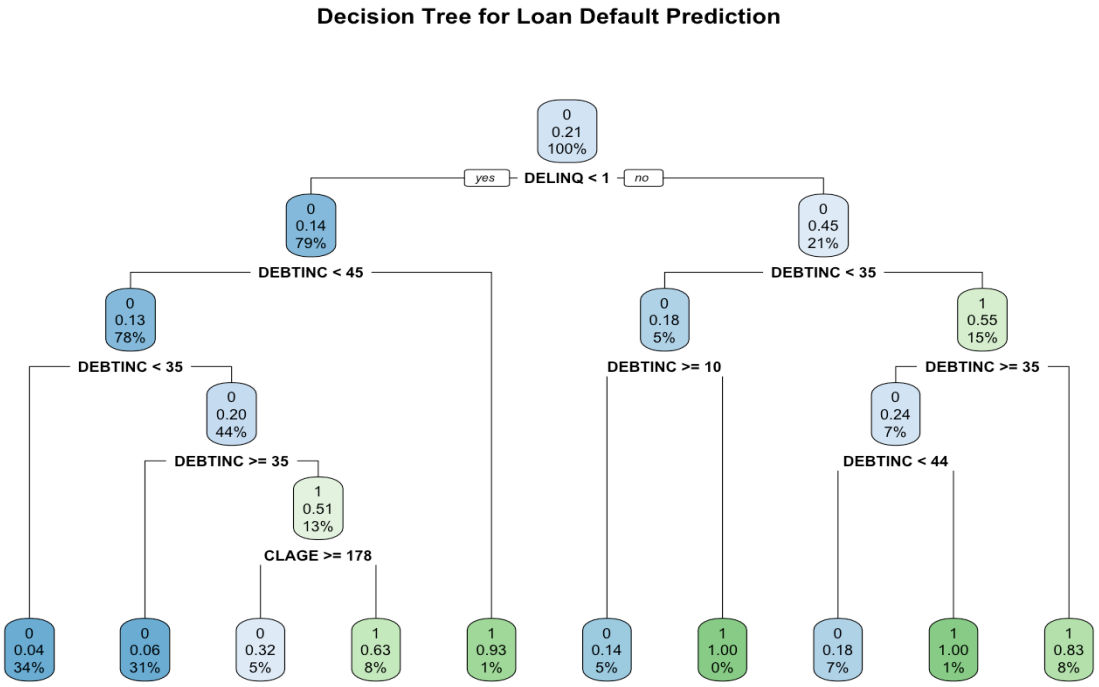
(Table 23 – Confusion Matrix Decision Tree)

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1337  100
         1   90  223

               Accuracy : 0.8914
                 95% CI : (0.8759, 0.9056)
    No Information Rate : 0.8154
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.6349

 Mcnemar's Test P-Value : 0.5138

            Sensitivity : 0.6904
            Specificity : 0.9369
         Pos Pred Value : 0.7125
         Neg Pred Value : 0.9304
             Prevalence : 0.1846
         Detection Rate : 0.1274
   Detection Prevalence : 0.1789
      Balanced Accuracy : 0.8137

       'Positive' Class : 1
```

(Table 24 – Performance Metrics for Decision Tree)

Model Insights

- **Accuracy**: The model achieved an accuracy of **89.14%**, performing well in predicting both defaulters and non-defaulters.
- **Precision**: A precision of **0.7125** indicates that 71.25% of those predicted to default were actual defaulters, ensuring reasonable minimization of false positives.
- **Recall (Sensitivity)**: The recall of **0.6904** suggests that the model identifies 69.04% of actual defaulters, reflecting moderate ability to detect defaults.
- **F1 Score**: The F1 score of **0.701** balances precision and recall, making it a robust measure of the model's performance.
- **Specificity**: The specificity of **0.9369** demonstrates excellent ability to correctly identify non-defaulters, minimizing false negatives.



(Fig 15 – Decision Tree for Loan prediction)

The model identified key predictors driving loan defaults:

- **DEBTINC (Debt-to-Income Ratio)**: A higher ratio correlates strongly with increased default risk.
- **DELINQ (Delinquencies)**: More delinquencies are a critical indicator of default likelihood.
- **CLAGE (Age of Oldest Credit Line)**: Older credit lines lower the default probability.
- **LOAN (Loan Amount)**: Larger loans are associated with higher default risks.
- **DEROG (Derogatory Public Records)**: The presence of derogatory records is a strong predictor of default.

Conclusion

The Decision Tree model effectively predicts loan defaults with an accuracy of **89.14%** and a strong balance between precision and recall. Its transparency and alignment with industry-standard variables, such as debt-to-income ratio and delinquencies, make it a practical tool for financial institutions. While specificity is high, further model tuning could improve recall to better capture true defaulters. This model provides actionable insights for risk assessment and decision-making in loan approvals.

Model Building for the Clustering

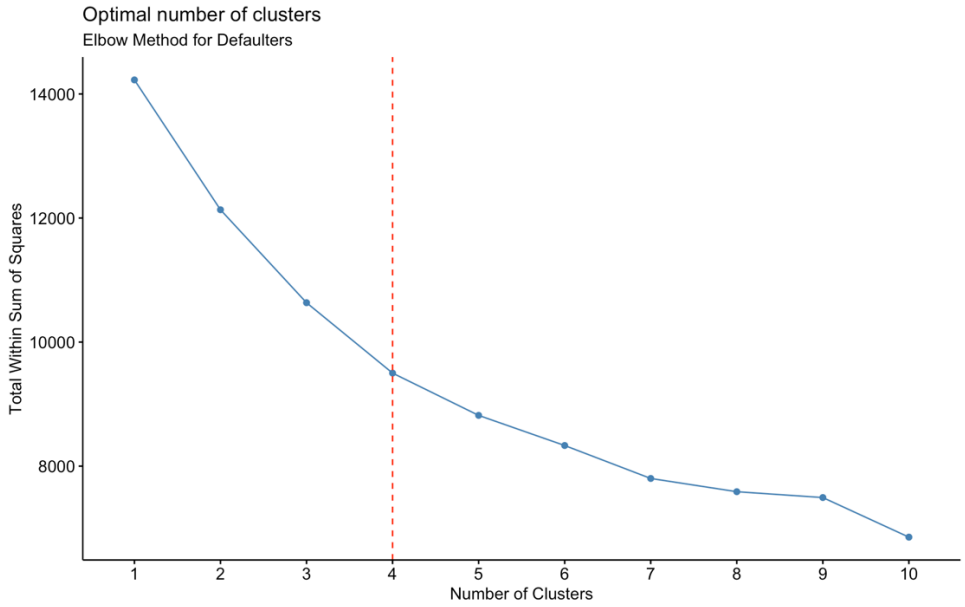
Determining the Optimal Number of Clusters (K)

To identify the most appropriate number of clusters, we used two common methods: the **Elbow Method** and the **Silhouette Method**. These methods were applied separately to both defaulters and non-defaulters datasets, ensuring a tailored approach for each segment.

1. Elbow Method

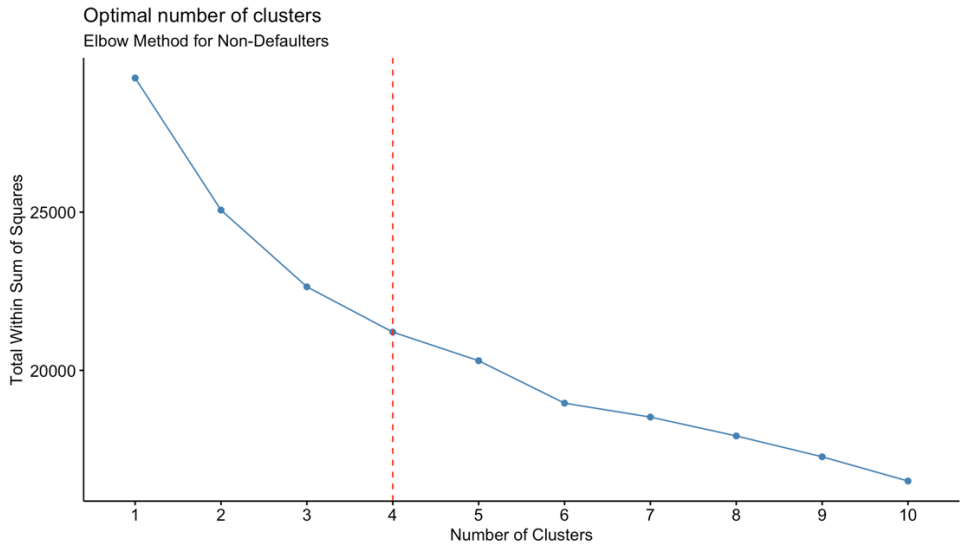
The **Elbow Method** helps identify the point where adding more clusters no longer significantly improves the model's performance. This is done by plotting the "Total Within Sum of Squares" (WCSS) for different numbers of clusters (K).

- **Explanation**: The WCSS represents the variance within each cluster. Initially, as the number of clusters increases, the WCSS decreases sharply, but after a certain point, the improvement becomes marginal. The "elbow" in the plot indicates the optimal K.
- **Observations**:



(Fig 16 –Elbow method for defaulters)

  - In the case of **defaulters**, the elbow intercept is at **K = 4**.



(Fig 17 – Elbow method for non-defaulters)

- For **non-defaulters** also, the elbow intercept is kept at **K = 4**.

Elbow Plot Analysis

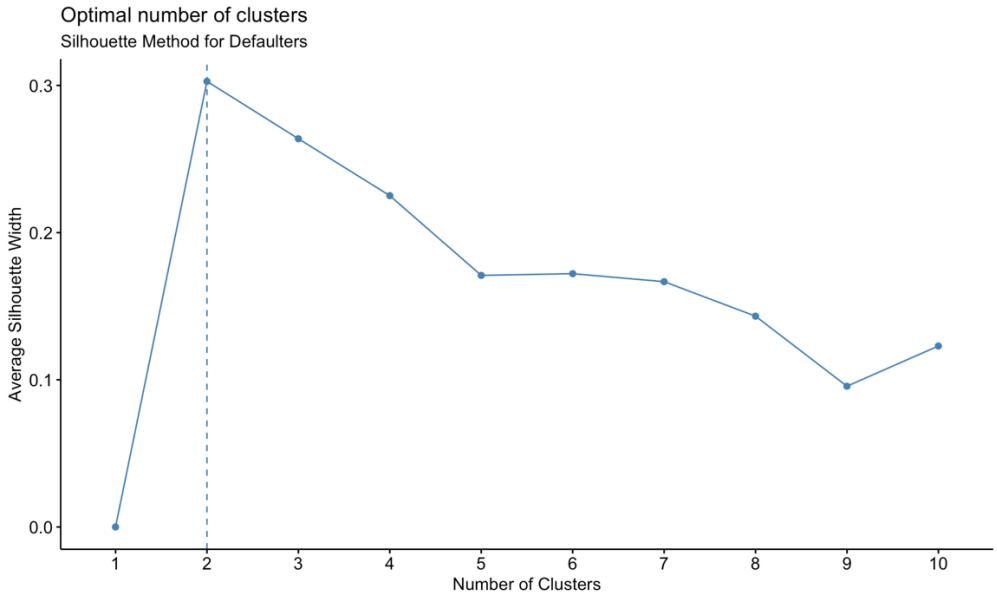
- The elbow plots for both defaulters and non-defaulters showed clear points at **K = 4**, suggesting that **four clusters** would be a good balance between complexity and accuracy.

2. Silhouette Method

The **Silhouette Method** evaluates how similar data points are within their clusters compared to other clusters. It calculates the silhouette score, which ranges from -1 to 1, where:

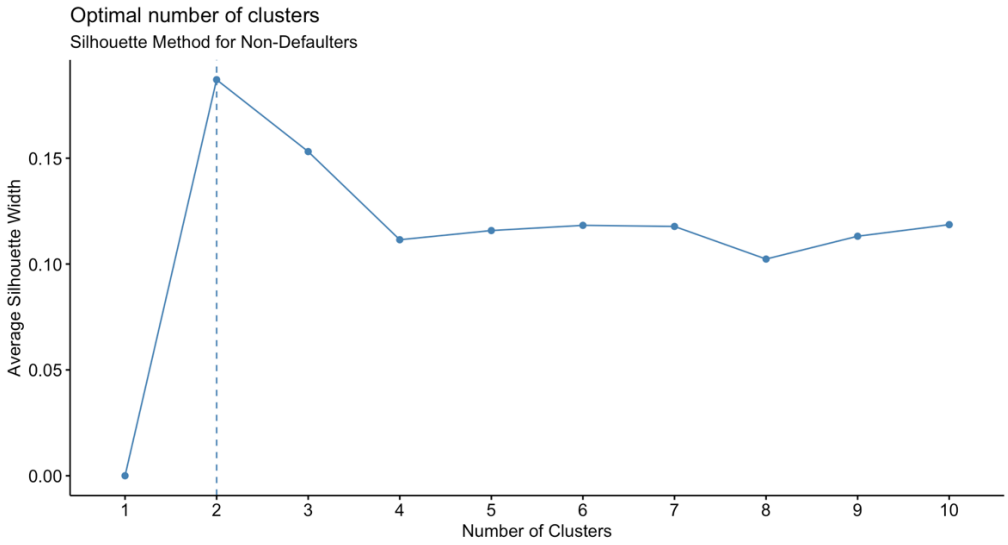
- A **higher score** indicates that clusters are well-separated, and data points fit well within their clusters.
- A **lower score** suggests that clusters overlap, or data points do not fit well.

Observations:



(Fig 18 – Silhouette method for defaulters)

- For **defaulters**, the highest silhouette score was at **K = 2**.



(Fig 19 – Silhouette method for non- defaulters)

- For **non-defaulters**, the highest silhouette score was also at **K = 2**.

Silhouette Plot Analysis

- The silhouette plots indicated that **K = 2** provides better-separated clusters, especially in terms of average silhouette width, suggesting that two clusters may be more interpretable in some cases. However, based on the elbow results, **K = 4** was chosen for better segmentation, considering both separation and model performance.

Choosing Optimal K- Value

After careful evaluation, it was decided to choose K=3. The model evaluation using the elbow method suggested K= 4, while the silhouette method indicated K= 2. However, after a panel discussion considering both methods and the business context, K=3 was chosen as the optimal number of clusters.

K-Means Clustering

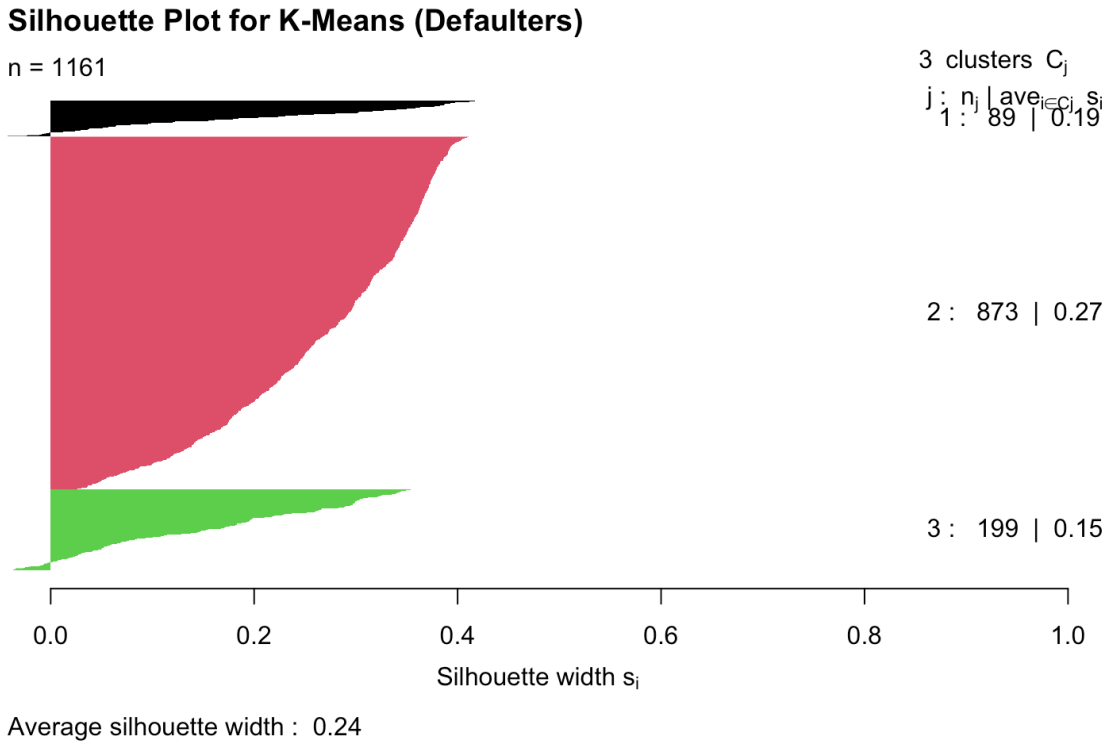
Overview

To address **Business Goal 2**, we implemented K-Means clustering to segment defaulters (BAD = 1) into distinct profiles based on their financial behaviors. This segmentation allows the bank to design targeted interventions, implement risk management strategies, and provide personalized financial support.

Cluster Characteristics

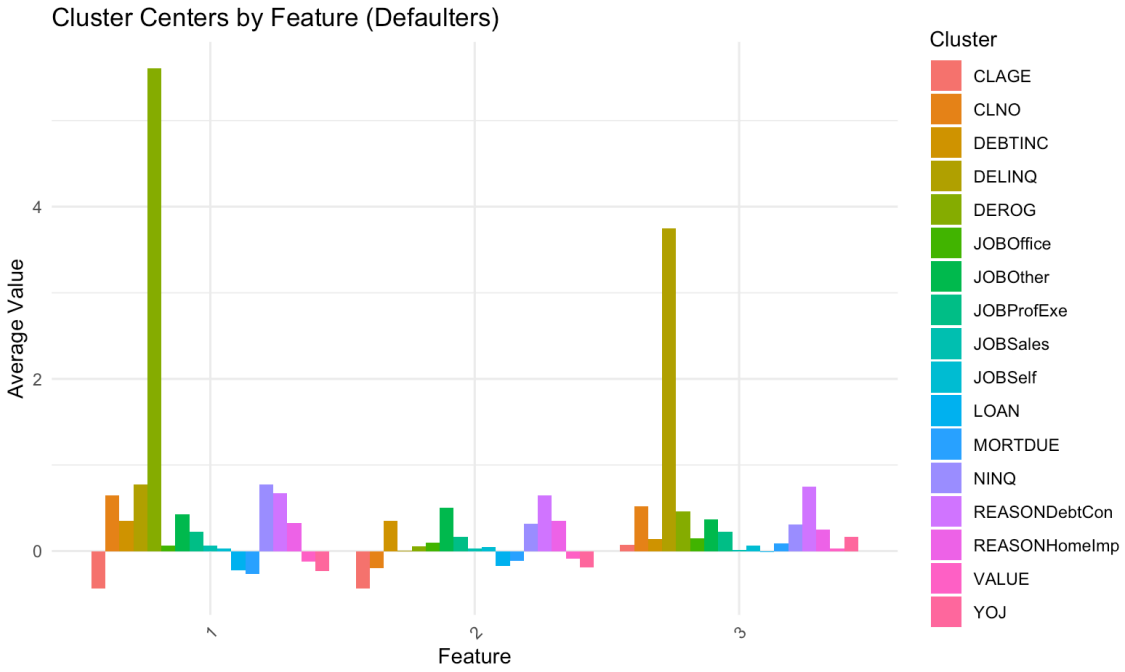
Cluster	Size	Silhouette Score	Description	Interpretation
1	873	0.30	**Frequent Delinquencies, High Debt-to-Income Ratio**	Represents a high-risk group; stricter loan terms and financial counseling required.
2	199	0.18	**Multiple Derogatory Reports, Short Credit History**	Medium-risk customers; targeted financial guidance suggested.
3	89	0.14	**Inconsistent Payment Behavior**	Medium-risk group; requires monitoring and potential interventions.

(Table 25 – K-means Clustering for Defaulted)



(Fig 20 – Silhouette Plot for K-Means(Defaulters)

- **Cluster 1**: Majority of defaulters fall into this category, characterized by frequent delinquencies and high debt-to-income ratios. Immediate actions include enforcing stricter terms and offering comprehensive financial counseling to reduce risks.
- **Cluster 2**: Customers with derogatory reports and short credit histories are moderately risky. These individuals could benefit from structured financial planning support.
- **Cluster 3**: Comprising inconsistent payers, this segment requires close monitoring and tailored interventions to prevent default escalation.



(Fig 20 – Cluster Center Plot for K-Means(Defaulters)

The bar plot displays the **average values** of financial features across the three identified clusters for **defaulters** (BAD = 1). Each bar represents the mean value of a specific feature for a given cluster, providing insights into the behavioral patterns within each segment.

Key Observations:

1. **Debt-to-Income Ratio (DEBTINC):**
   - This feature stands out as the most significant differentiator between clusters.
   - Cluster 1 exhibits the highest average DEBTINC, indicating severe financial strain and high risk.
   - Cluster 2 and Cluster 3 show comparatively lower values, but still represent concerning levels.
2. **Delinquencies (DELINQ):**
   - Clusters 1 and 3 have moderate to high average delinquencies, reinforcing their classification as medium- to high-risk groups.
   - Cluster 2 shows lower delinquency levels, but other factors such as derogatory records contribute to its risk profile.
3. **Derogatory Public Records (DEROG):**
   - Cluster 2 has a significantly higher average for DEROG compared to the other clusters, highlighting multiple derogatory public records like bankruptcies or liens.
   - Clusters 1 and 3 have fewer derogatory records, indicating that their risk stems from other financial behaviors.
4. **Age of Oldest Credit Line (CLAGE):**
   - Cluster 2 has the shortest average credit history, contributing to its medium-risk classification.
   - Cluster 1 and Cluster 3 exhibit slightly longer credit histories, though they still fall into high-risk categories due to other financial indicators.
5. **Loan Amount (LOAN):**

- Cluster 3 has the highest average loan amounts, suggesting a pattern of over-borrowing that may lead to inconsistent payment behaviors.

6. **Employment and Income Stability (e.g., JOB, YOJ):**
   - Employment categories (e.g., JOBOffice, JOBOther) and years on the job (YOJ) show variation across clusters.
   - Cluster 1 includes individuals with relatively more stable employment but higher financial obligations.
   - Cluster 2 consists of individuals with less stable employment histories.

Interpretation:

- **Cluster 1**: This group is characterized by high debt-to-income ratios and frequent delinquencies. It represents the highest-risk customers, requiring immediate interventions such as stricter terms or comprehensive counseling.
- **Cluster 2**: Defined by multiple derogatory records and short credit histories, this medium-risk group may benefit from targeted financial guidance and credit-building programs.
- **Cluster 3**: With inconsistent payment behaviors and larger loan amounts, this group needs close monitoring and tailored interventions to avoid further defaults.

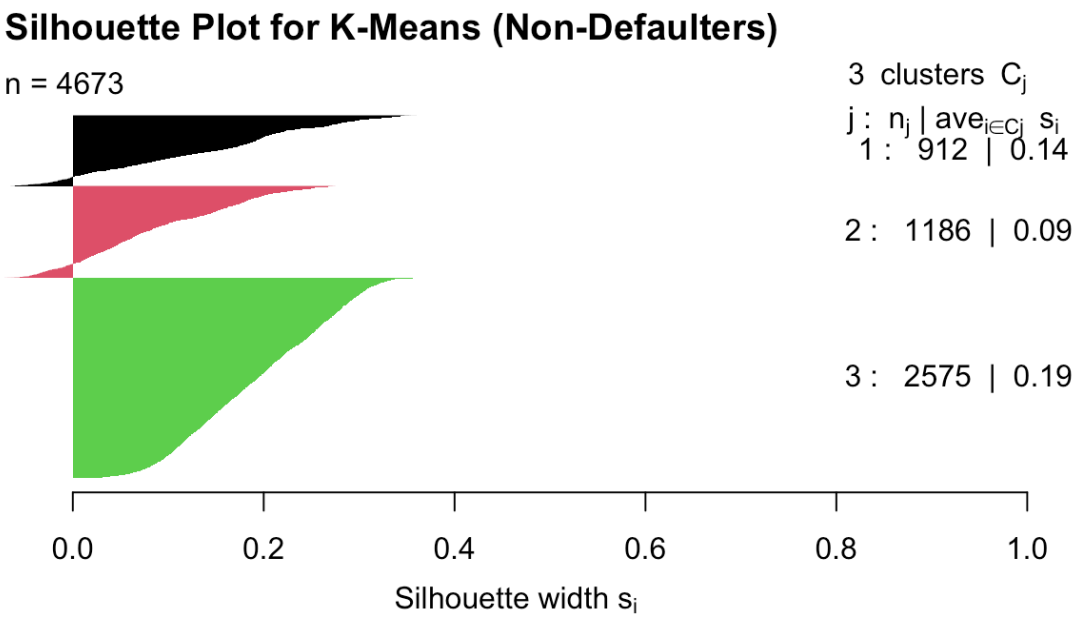
Clustering Results for Non-Defaulters

The segmentation of non-defaulters, who represent 3,272 records in the dataset, was conducted using K-Means clustering. This approach identified three distinct clusters based on financial behavior and credit characteristics. These insights aim to support the development of targeted marketing, loyalty programs, and proactive credit management strategies.

Cluster Characteristics

Cluster	Size	Silhouette Score	Description	Interpretation
1	638	0.14	Moderate Debt, Consistent Payment History	Low-risk group; these customers are suitable for upselling financial products and receiving favorable loan terms.
2	1,800	0.19	Long Credit History, Low Debt Load	Ideal borrowers; they represent the most reliable segment, making them a prime target for cross-selling and loyalty initiatives.
3	834	0.08	Occasional Delinquencies, Moderate Credit Utilization	Medium-risk group; they require monitoring but have the potential to qualify for better financial terms with improved behavior.

(Table 26 – K-means Clustering for Defaulted)



(Fig 21 – Silhouette Plot for K-Means(Non-Defaulters)

Cluster 1: Low-Risk Group (638 Customers)

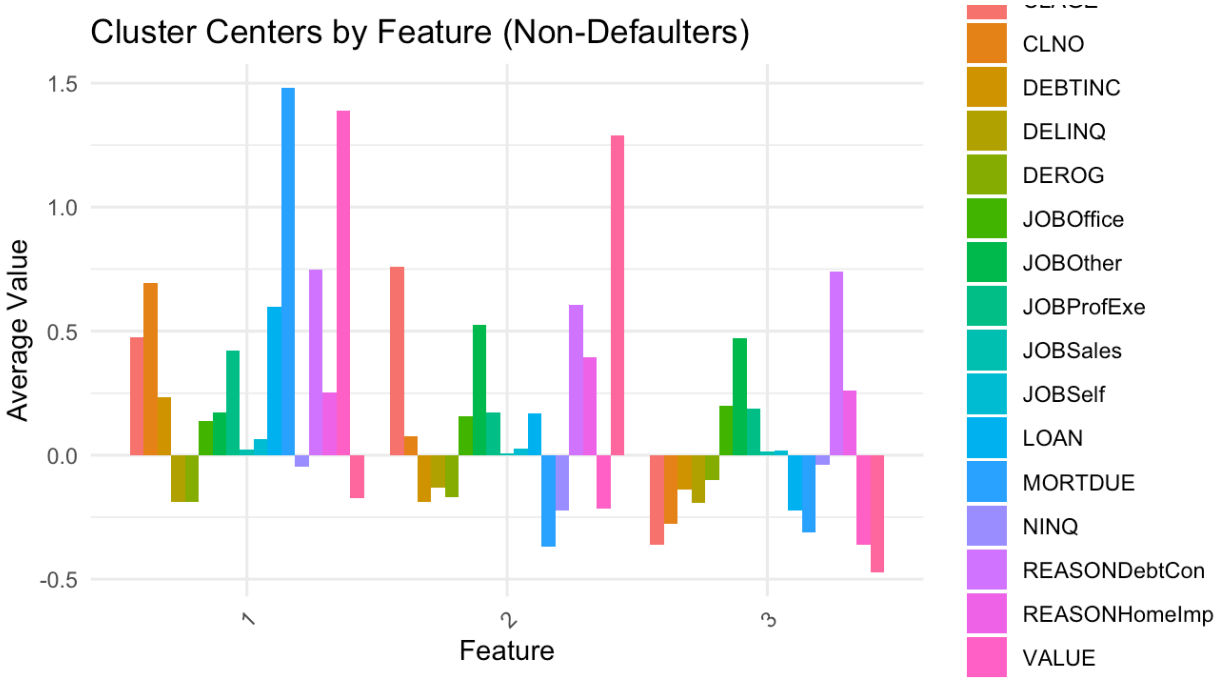
- These customers have moderate levels of debt and demonstrate consistent payment behavior.
- They are reliable and offer opportunities for upselling financial products such as credit cards or personal loans.
- Their consistent repayment patterns suggest they would respond well to favorable terms or incentives for additional financial engagement.

Cluster 2: Ideal Borrowers (1,800 Customers)

- This cluster represents customers with long credit histories and low debt loads, indicating a stable and reliable financial profile.
- They are ideal for cross-selling financial products such as investment opportunities, premium accounts, or bundled services.
- Loyalty programs and rewards targeting this group could enhance customer retention and deepen engagement.

Cluster 3: Medium-Risk Group (834 Customers)

- These customers exhibit occasional delinquencies and moderate credit utilization.
- While they pose a medium risk, proactive monitoring and timely interventions, such as offering credit counseling or flexible repayment options, could improve their financial behavior.
- This group may benefit from tailored strategies to enhance their credit standing, eventually transitioning them into a lower-risk profile.



(Fig 21 – Silhouette Plot for K-Means(Non-Defaulters)

The bar plot illustrates the average values of key features across three clusters of non-defaulters identified during the K-Means clustering process. These cluster centers highlight the distinct financial behaviors and characteristics of each group, providing actionable insights for segmentation and targeted strategies.

Key Features and Observations

1. **Cluster 1: Moderate Risk**
    - **Higher average CLAGE (Age of Credit Line):** Indicates a moderate credit history length.
    - **Relatively higher DELINQ (Delinquency Counts):** Suggests occasional late payments, contributing to moderate risk.
    - **VALUE:** Shows a moderate level of loan values compared to other clusters.

Interpretation: Customers in this cluster exhibit a balanced profile with occasional risks. They may benefit from monitoring and personalized loan terms to encourage consistent behavior.

2. **Cluster 2: Ideal Borrowers**
    - **High average MORTDUE (Mortgage Due):** Indicates a preference for secured loans or mortgages.
    - **Low DELINQ and DEROG (Delinquency and Derogatory Reports):** Reflects disciplined payment behavior and minimal credit issues.
    - **High CLAGE and LOAN:** Suggests long credit histories with consistent loan repayments.

Interpretation: This cluster represents the ideal borrower group, with stable credit utilization and low delinquency rates. They are excellent candidates for cross-selling and loyalty programs.

3. **Cluster 3: Medium Risk**

- **Higher average DEBTINC (Debt-to-Income Ratio):** Reflects higher financial obligations relative to income.
- **Moderate CLNO (Number of Credit Lines):** Indicates a balanced use of multiple credit facilities.
- **Relatively lower REASONHomeImp (Loans for Home Improvement):** Suggests lesser emphasis on secured or purpose-driven loans.

Interpretation: Customers in this cluster have a moderate financial risk due to higher debt levels but maintain reasonable credit utilization. Targeted interventions and financial planning could help mitigate potential risks.

Hierarchical Clustering Analysis

Business Context

The hierarchical clustering analysis aims to segment defaulters into meaningful sub-groups to address varying levels of financial risk. This segmentation aligns with the following business goals:

1. **Risk Mitigation:** Identify high-risk defaulters and implement stricter repayment measures.
2. **Recovery Assistance:** Target moderate-risk defaulters with supportive programs to reduce future defaults.
3. **Retention Strategies:** Provide lenient repayment plans for low-risk defaulters to retain customer relations.

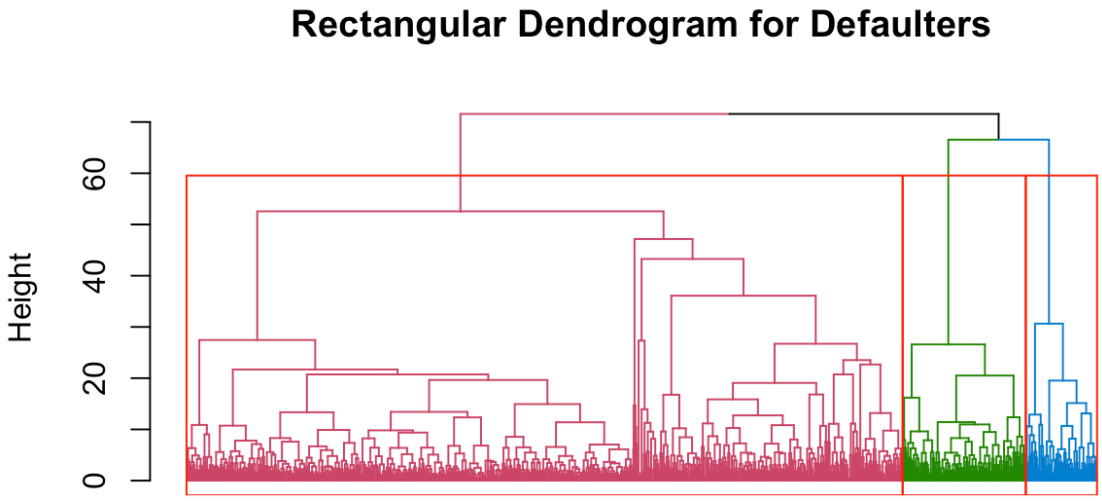
Clustering Process

1. **Distance Matrix Calculation:** Euclidean distance was calculated for the dataset of defaulters to measure the dissimilarity between data points.
2. **Clustering Methodology:** Ward's method was used to minimize within-cluster variance, ensuring compact and interpretable clusters.
3. **Optimal Clusters:** The dendrogram suggested three distinct clusters, representing high, moderate, and low-risk defaulters.

Cluster Analysis:

Cluster	Size	Description	Characteristics	Recommended Strategy
1	913	Moderate Risk	Sporadic delinquencies, moderate debt loads	Offer financial counseling and recovery plans.
2	91	High Risk	Frequent delinquencies, high debt-to-income	Implement stricter loan terms and provide financial education.
3	157	Low Risk	Better financial management, fewer defaults	Provide lenient repayment options and build loyalty programs.

(Table 27 – Hierarchical Clustering for Defaulters)



(Fig 22 – Dendrogram for defaulters)

Dendrogram Interpretation:

The **rectangular dendrogram** for defaulters clearly displays three main clusters:

1. **Cluster 1:** The largest group, representing moderate-risk defaulters who exhibit sporadic delinquencies but have potential for recovery.
2. **Cluster 2:** A smaller group of high-risk defaulters with frequent delinquencies and financial distress.
3. **Cluster 3:** A moderate-sized group of low-risk defaulters who show better financial management.

The hierarchical clustering process provides valuable insights for tailoring risk management strategies, allowing targeted interventions for each risk category. This segmentation can significantly enhance the bank's ability to mitigate default risk and retain valuable customers.

Clustering Results for Non-Defaulters

The goal of hierarchical clustering for non-defaulters is to identify distinct sub-groups within this population. These clusters will enable personalized financial services and maximize upsell opportunities, aligning with the following business objectives:

1. **Tailored Financial Products:** Provide customized offerings such as loans, credit cards, or investment plans for distinct customer segments.
2. **Customer Retention and Growth:** Identify opportunities for engaging and retaining low-risk customers while supporting emerging or moderate-risk profiles.

Clustering Process

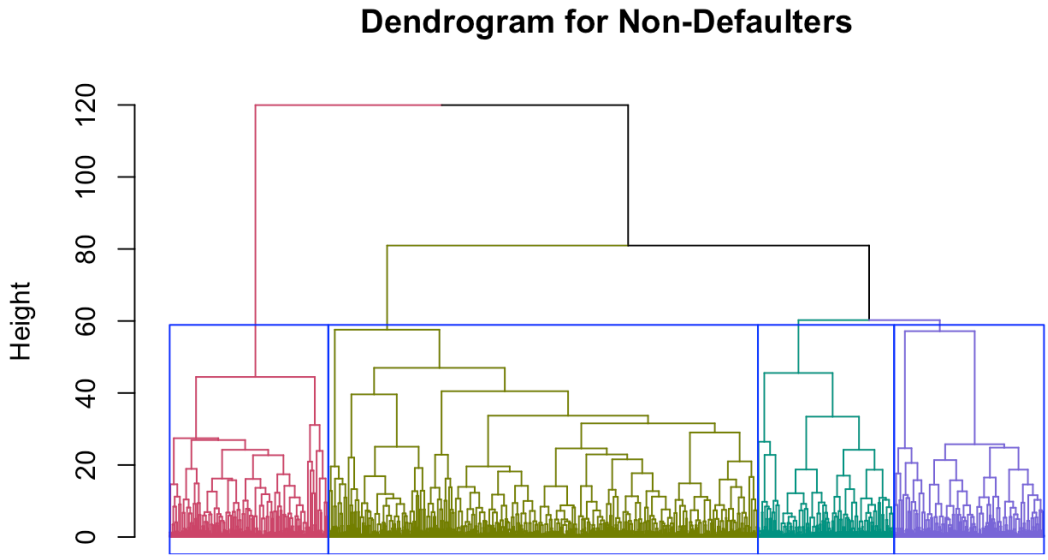
1. **Distance Matrix Calculation:** Euclidean distance was used to measure the dissimilarity between data points.

2. **Clustering Methodology:** Ward's method was employed to minimize within-cluster variance, ensuring compact and well-separated clusters.
3. **Optimal Clusters:** Based on the dendrogram, the data was segmented into **four clusters**, each representing a unique customer profile.

Cluster Analysis:

Cluster	Size	Description	Characteristics	Recommended Strategy
1	2,296	Stable, Low-Risk Group	Consistent payment behavior, low debt loads, and stable financials	Focus on upselling and cross-selling financial products like credit cards and personal loans.
2	728	Moderate-Risk Group	Some financial vulnerabilities, such as higher debt loads or shorter credit histories	Provide targeted support to maintain stability and reduce default risk.
3	801	Emerging Credit Segment	Consistent payments but limited credit record, newer credit profiles	Offer credit-building products such as secured credit cards or introductory loans.
4	848	Special Segment, Moderate Volume	Unique financial profiles or recent improvements in credit management	Offer niche financial products like investment options or premium services that match their needs.

(Table 28 – Hierarchical Clustering for Non-Defaulted)



(Fig 23 – Dendrogram for Non-Defaulters)

Dendrogram Interpretation

- The dendrogram for non-defaulters clearly delineates **four clusters**, each representing unique customer segments.
- **Cluster 1**, the stable, low-risk group, is the largest segment. This cluster presents significant opportunities for upselling and personalized engagement.
- **Cluster 2** and **Cluster 3** highlight moderate-risk and emerging credit segments, which could benefit from financial guidance and targeted credit-building products.
- **Cluster 4** consists of a special segment with unique profiles, offering opportunities for niche financial services.

K-Means Clustering Aligns with the Business Goal

The application of K-Means clustering aligns closely with the business goal of **precise clustering for targeted risk management**, offering well-defined and actionable segments. K-Means demonstrated superior average silhouette scores compared to hierarchical clustering, especially for defaulters, indicating better-defined clusters with clearer separation. For defaulters, K-Means achieved an **average silhouette score of 0.26**, while hierarchical clustering had lower silhouette scores, suggesting less distinct separation of clusters. For non-defaulters, K-Means achieved a **reasonable silhouette score of 0.17**, reflecting moderately well-defined clusters.

Business Goal: Precise Clustering for Targeted Risk Management

The main aim of clustering was to segment customers into distinct groups, enabling targeted interventions and personalized strategies for both defaulters and non-defaulters.

1. Defaulter Segmentation

K-Means identified three distinct clusters within defaulters, providing the bank with actionable insights for targeted risk management:

- **Cluster 1 (High-Risk Defaulters):**
  Customers in this cluster exhibit **high debt-to-income ratios**, frequent delinquencies, and shorter credit histories. They represent the **highest risk** group and require:
  - Stricter loan terms.
  - Enhanced monitoring and financial counseling.
  - Proactive engagement to prevent further financial deterioration.
- **Cluster 2 (Moderate-Risk Defaulters):**
  This cluster consists of customers with **moderate debt loads** and occasional delinquencies. These customers could benefit from:
  - Supportive financial education programs.
  - Adjusted repayment plans to prevent default escalation.
- **Cluster 3 (Low-Risk Defaulters):**
  Customers in this cluster show **fewer delinquencies** and may experience temporary financial setbacks. Strategies include:
  - Debt restructuring or flexible repayment terms.
  - Financial stability programs to retain long-term customer relationships.

2. Non-Defaulter Segmentation

K-Means also identified three distinct clusters within non-defaulters, helping the bank categorize customers with varying levels of creditworthiness:

- **Cluster 1 (Stable and Reliable Customers):**
  This cluster includes customers with **high asset values**, longer credit histories, and consistent repayment behavior. These customers represent **low risk** and are ideal for:
  - Upselling premium financial products, such as loans, investment options, or insurance.
  - Enhancing loyalty through personalized rewards and benefits.
- **Cluster 2 (Growing Credit Histories):**
  Customers in this cluster have **moderate debt loads**, stable growth in credit histories, and consistent payments. They are suited for:
  - Cross-selling opportunities, such as additional loans or credit cards with favorable terms.
  - Supporting continued financial growth through advisory services.
- **Cluster 3 (Emerging Borrowers):**
  These customers have **shorter credit histories** and smaller loans. They represent **potential growth opportunities** and could benefit from:
  - Financial advisory services to nurture trust and loyalty.
  - Credit-building products like secured credit cards or starter loans.

Why K-Means is the Best Choice for this Objective

- **Well-Defined Actionable Segments:**
  K-Means clustering provides clear and interpretable segments that enable precise strategies, such as stricter risk controls, targeted financial product offerings, and early intervention for high-risk groups.
- **Alignment with Business Strategy:**
  By effectively segmenting both defaulters and non-defaulters, K-Means clustering directly supports the bank's goals of **minimizing credit risk** while **maximizing customer value** through tailored interventions.
- **Ease of Integration and Scalability:**
  K-Means clustering is computationally efficient and easy to implement, making it suitable for integration into existing systems. Its simplicity ensures that real-time clustering can be employed for ongoing customer management and engagement.

This segmentation analysis enhances the bank's ability to **mitigate risk**, **increase customer engagement**, and **drive operational efficiency**, ensuring alignment with the broader strategic objectives.

Conclusion and Recommendations

This project successfully addressed the dual objectives of automating the loan approval process for home improvement loans and enhancing customer segmentation for targeted financial strategies. By implementing advanced predictive analytics and clustering techniques, the bank is equipped with tools to make data-driven, transparent, and regulatory-compliant decisions.

The Random Forest model proved to be the most effective classification method, achieving 93.31% accuracy and 98.6% sensitivity. This ensures precise identification of defaulters while minimizing false negatives, which is critical for risk management. Clustering analysis provided

meaningful segmentation of defaulters and non-defaulters into three distinct risk categories—low, moderate, and high—enabling targeted financial services tailored to each segment's needs.

Recommendations:

1. **Model Deployment:** Integrate the tuned Random Forest and K-Means models into the bank's decision-making workflows for real-time loan approvals and segmentation analysis.
2. **Advanced Risk Profiling:** Use insights from defaulter segmentation to design interventions, such as stricter loan terms and financial counseling for high-risk customers. For non-defaulters, leverage segmentation to offer personalized financial products and loyalty rewards.
3. **Continuous Monitoring and Updates:** Implement a robust monitoring system to track model performance over time, ensuring alignment with evolving data patterns and regulatory requirements. Recalibrate models as necessary to maintain accuracy and fairness.
4. **Granular Customer Insights:** Explore more detailed customer segmentation using advanced clustering methods like Hierarchical Clustering to uncover deeper behavioral patterns and new opportunities for personalized services.
5. **Enhanced Decision Support:** Extend the scope of analysis by incorporating additional external data sources, such as market trends or credit bureau data, to improve predictive accuracy and enrich segmentation.

By adopting these strategies, the bank can fully leverage the project outcomes to enhance operational efficiency, mitigate financial risks, and improve customer engagement. These efforts will not only streamline the loan approval process but also position the bank as a leader in innovative and customer-centric financial services.