Contents

1. [INTRODUCTION 4](#_bookmark0)
   1. [Problem Description 4](#_bookmark1)
   2. [Objectives 5](#_bookmark2)
   3. [Data Overview 6](#_bookmark3)
2. [Literature Review 6](#_bookmark4)
3. [Methodology 8](#_bookmark5)
   1. [Data Collection 8](#_bookmark6)
   2. [Data Exploration 8](#_bookmark7)
   3. [Data Preprocessing 13](#_bookmark8)
      1. [Dropping the Loan\_ID Column 13](#_bookmark9)
      2. [Handling Missing Values 13](#_bookmark10)
      3. [Feature Engineering 20](#_bookmark11)
   4. [Exploratory Data Analysis 21](#_bookmark12)
   5. [Outliers 25](#_bookmark13)
4. [ENCODING 27](#_bookmark14)
   1. [Random Forest 33](#_bookmark15)
   2. [XGBOOST 37](#_bookmark16)
   3. [MODEL EVALUATION 40](#_bookmark17)
   4. [Validation Results Explanation 41](#_bookmark18)
   5. [BUSINESS INSIGHTS 42](#_bookmark19)
   6. [Limitations 43](#_bookmark20)
   7. [Future Work 43](#_bookmark21)
   8. [Conclusion 44](#_bookmark22)
5. [REFERENCES 45](#_bookmark23)
   * 1. [Additional Works 47](#_bookmark24)

# INTRODUCTION

Housing Finance plays a significant role in housing delivery. One of the most expensive assets that most families own over their lifetime is a house and its purchase needs external financial aid (Doling et al. (2013). The primary source for the financial aid for Housing is banks and mortgages (Kavaarpuo, Mintah and Donkor-Hyiaman (2023). Globally, the housing finance market size was valued at around 4500 billion in the year 2021 and is expected to grow seven times more by the year 2031. (Allied Market Research, 2022).

The rising need for money to purchase housing among individuals and businesses has boosted the growth of the housing finance market. Due to the demand of the loans, Dream Housing Finance Company, a player in the home loan sector and attending to urban, semi-urban and rural regions, seeks to automate the loan eligibility process to promote efficiency.

Through my project, I utilized machine learning models such as Random Forest and XGBoost methods to predict loan eligibility based on the customer data. This will provide a quick and simple approach to pick deserving loan applicants. It will save time for the banks’ employees and enhance operational efficiency within the Housing Finance Company. (Marc Escapa, 2023).

### Problem Description

Dream Housing Finance Company provides home loans. Their process involves customer applications for house loans, followed by the company validating eligibility manually. This traditional process involves significant paperwork, leading to delays and frustration among potential clients. Other challenges may include time wastage when loan officers do an analysis to check a customer’s creditworthiness. With the extensive paperwork involved, various departments manage the verification process which increases data entry errors. (MonJa: Intelligent Document Automation for Lenders, 2023b).

These delays lead to late approvals and increase inaccuracies. The company aims address these challenges by automating the eligibility process. Banks benefit more

from loans and need an easier process to determine who is genuine and who will return the loan (Yamparala et al., 2022). When looking for genuine clients, the volume of loan applicants keeps growing rapidly, with an average increase of 1.5% per year between 2015 and 2022 (Cheriyan et al., 2024). This creates a cumbersome workload for banks, often accompanied by higher costs and increased operational risks.

Before banking automation, banks used to be so crowded, and employees in the branches were always busy handling the day-to-day affairs of the banks. Even branch managers could not get enough time with their big clients (Najafbagy et al., 2010).

In today’s world, it is important for banks to have a culture shift and adopt technology to optimize the existing processes and ensure efficiency (Wingard, 2024). In 2021, one of the largest banks increased its technology budget by more than 10% and moved 70% of its applicants to cloud and made plans to increase the number by 30% yearly and this was projected to save the bank about 2 billion dollars in the long run and shift the balance of its operations from 60% on maintaining systems to almost 50% being spent on innovations .(Baig, Sohoni and Lhuer (2024).

By automating systems, live updates and tracking capabilities allow various stakeholders like underwriters and managers to monitor the applications and identify outstanding tasks. This ensures timely progress and helps users in prioritizing their responsibilities effectively (Vanner, 2024). By approving loans automatically, the loan providers can approve more loans, which will lead to increased revenue within the banks (Vanner, 2024).

Automating the loan approval system will streamline operations, reduce costs, and enhance the customer experience in Dream Housing Finance Company.

### Objectives

The PRIMARY objective of this study is to develop a predictive model for loan approval that forms a critical component of enabling the loan eligibility automation process.

To achieve this, I will also evaluate and compare the performance of the machine learning models to identify the most reliable approaches for predicting loan eligibility.

My last objective is to lay the groundwork for the automation of the loan eligibility process by creating a reliable prediction system that will be perfectly integrated with real-time application.

### Data Overview

The Dream House financing company datasets were obtained from Kaggle. The purpose of these datasets is to assess and predict loan eligibility based on the customer information such as their credit history, financial status, marital status, employment status and more. By doing the analysis, the datasets aim to find patterns and criteria impacting loan approval. This will streamline and enhance decision making for the House Financing company.

Two datasets were provided, one is the training dataset and the other is a testing dataset. The training dataset contains 614 entries and 13 columns, and the test datasets has 367 entries and 12 columns. (with exception of the target variable).

The columns in the datasets are as follows:

|  |  |
| --- | --- |
| **Key Name** | **Description** |
| **Loan\_ID** | Unique Loan ID of applicant |
| **Gender** | Male/Female |
| **Married** | Marital Status (Y/N) |
| **Dependents** | Number of Dependents |
| **Education** | Graduate / Undergraduate |
| **Self\_Employed** | Is applicant self-employed(Y/N) |
| **ApplicantIncome** | Applicant Income |
| **CoapplicantIncome** | Coapplicant’s Income |
| **LoanAmount** | Amount of loan in Thousands |
| **Loan\_Amount\_Term** | Term of the loan in Months |
| **Credit\_History** | Prior credit guidelines met |
| **Property\_Area** | Urban/ Semi-Urban / Rural |
| **Loan\_Status** | Loan Approval (Y/N) |

# LITERATURE REVIEW

**Loan Eligibility and Approval Process in Traditional Banking**

For the longest time, the loan approval process was being verified manually and included a lot of paperwork and human intervention. Banks heavily relied on face-to- face interaction with the applicants and this led to a lot of delays (Najafbagy et al. (2010). Despite offering an applicant an opportunity to have a discussion with the bank about any challenges or questions they may have about the loans, it would waste so much time for both the lender and the client. (MonJa: Intelligent Document Automation for Lenders, 2023c) .

Yearly, the volumes of loan applications continue to grow making it hard for manual handling of application cumbersome and in the long-run, unsustainable. Cheriyan et al. (2024). This meant that a more efficient approach was desperately needed to meet the growing demand for loans in banks. (Yamparala, R., et al. (2022).

**Automating Loan Eligibility**

By automating bank services, efficiency, and reduction in cost of labour and errors is enabled Vanner (2024). By use of online application systems to gather customer data, banks can easily decide if the applicant qualifies for a loan. Streamlining benefits both banks and customers which makes the whole process smoother and faster. In the recent years, banks are now using cloud-based technologies to optimize their systems. Wingard (2024). The transition to digital platforms automatically streamlines day-to-day procedures enhancing loan application.

Banks are anticipating the millions of dollars that they will save just from the digital automation and the better resource management (Baig, Sohoni, and Lhuer, 2024). Furthermore, this automation will improve customer experience by providing faster loan processing periods (Marc Escapa, 2023). The speed is crucial for customer satisfaction where it matters to clients in competitive markets.

**Performance of Machine Learning in Loan Eligibility**

Machine learning has gained attention worldwide with its capacity to improve loan eligibility . it creates models that can be trained to give output of accurate predictions and how the models will respond to new data using the known data that it was trained with Gogula and Chattu (2024). Models such as XGBoost are more effective when it comes to data that involves binary outcomes and it has an ability to handle both large and small datasets Wang(2024).

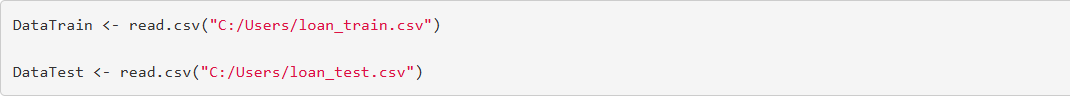
Random Forest is a power model because of its capacity to manage complex data using multiple decision trees Ellis (2023). Both models can predict loan eligibility and providing resourceful insights. In this study, machine learning is useful as it will aid in the process of validating a customer, knowing who has the least chances of defaulting and a dependable one.

In conclusion, transitioning from traditional manual ways to automated marks a significant growth in the banking industry in matters related to efficiency and effectiveness. Continuous integration of technologies and new systems will ensure sustainable and customer-friendly loan eligibility process.

# METHODOLOGY

### Data Collection

The datasets in this study were obtained from Kaggle. The training dataset (loan\_train.csv) has historical loan application data and the testing data (loan\_test.csv) used to predict the model’s predictions. Both datasets were loaded into R as shown in figure 1 using the (“read.csv.) function and all libraries needed for the data preparation loaded.



### Data Exploration

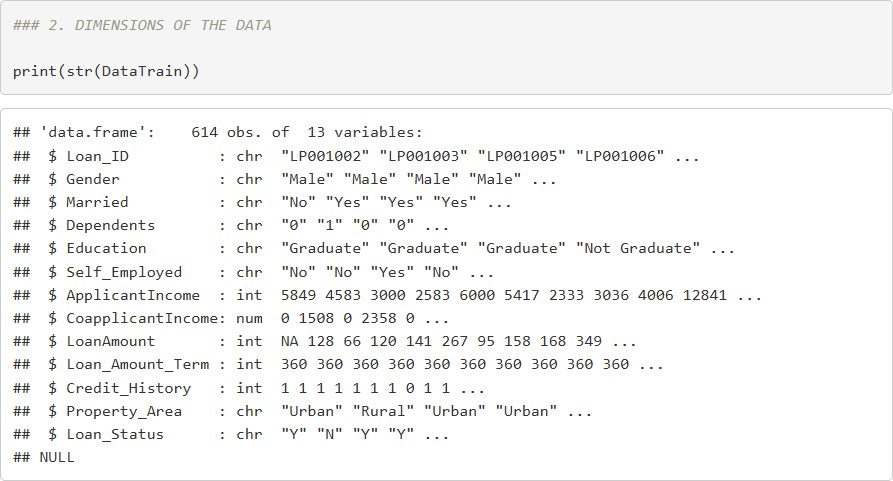
*Loan Report Figure 1*

**Structure and Summary of the Datasets**

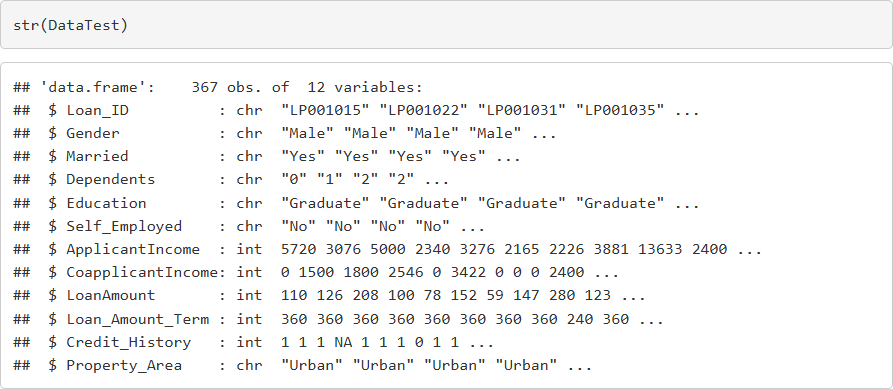
The first step in data exploration is understanding the dimensions of our datasets. The visualizations represent the various columns in our datasets and help understand the distribution of the data.

The code ‘str ()’ function provides a human readable summary of our datasets. Particularly, it is useful in understanding the types of data in the data frames. Through the ‘str ()’ function, there are both categorical and numerical variables in our datasets. There are missing values that will need to be addressed in data preprocessing. The

datasets provide a comprehensive overview of the loan applicants which can be used for further analysis to make predictions of loan approval.

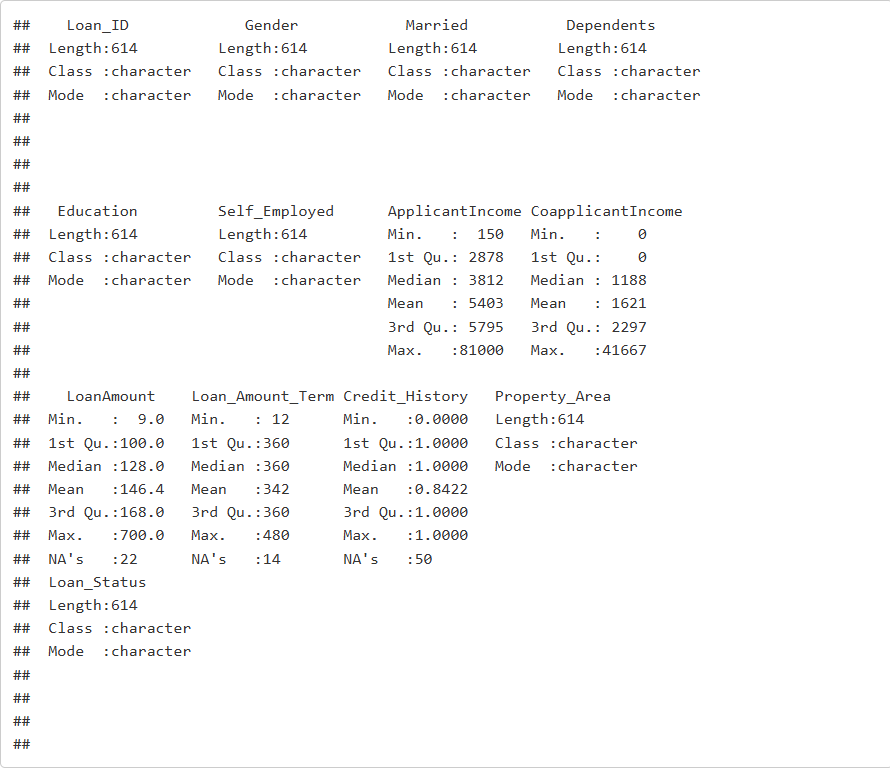


*Loan Report Figure 2*

**

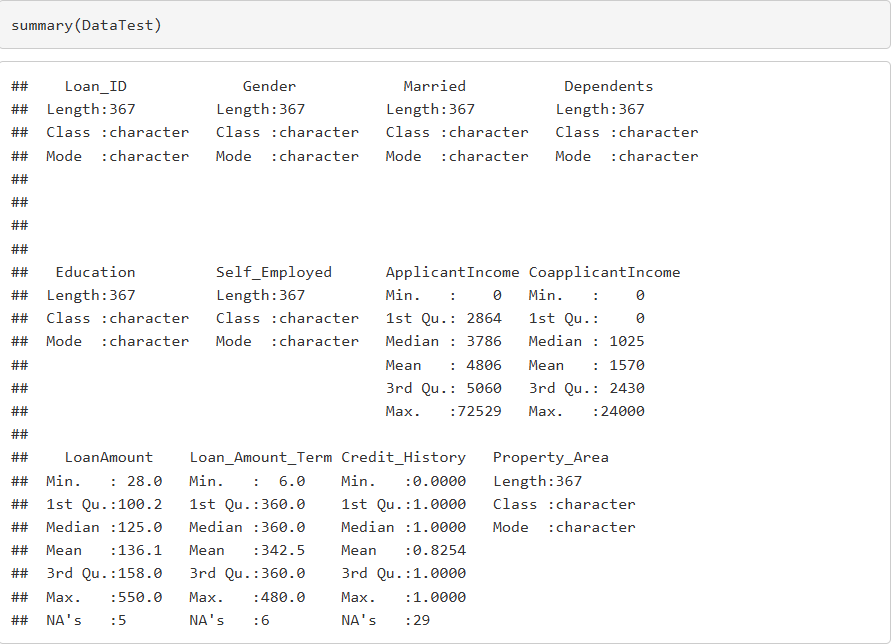
*Loan Report Figure 3*

The DataTrain has 614 observations and 13 variables. This observation shows that there are missing values in LoanAmount, Loan\_Amount\_Term, and Credit\_History. The Income and Loan Amount distributions suggest a wide range of financial backgrounds among the Applicants, which could influence loan approval decisions.



*Loan Report Figure 4*

In the Data Test, there are 367 observations and 12 variables. There are also missing values that will need attention and handling in the data preprocessing.



*Loan Report Figure 5*

**Feature exploration**

Understanding the characteristics of the categorical and handling the noncategorical features is crucial for effective prediction and use of models. Through R, I developed a function to show the variables within both datasets. The understanding of the data is also important to make informed decisions during data preprocessing, exploratory data analysis and model development in the later stages of analysis.

The function developed uses the ‘explore\_object\_type’ to show the number of times a unique value or text appears in each column. Figure 6 shows the variables in DataTrain and figure 7 shows the variables in the DataTest.



*Loan Report Figure 6*

**

*Loan Report Figure 7*

### Data Preprocessing

#### Dropping the Loan\_ID Column

The Loan\_ID column is a unique identifier that is assigned to every individual who applies for a loan. It holds no meaningful relation with our target variable or any other independent variable.

Since our models rely on patterns, this column would not contribute to improve the predictions. This is a standard practice as it has no contextual significance for understanding our problem.

By setting the Loan\_ID column to NULL, it was effectively removed from the datasets ensuring only relevant columns in the datasets remained. To verify if the columns were successfully removed, I checked the column names of the datasets as shown in figure

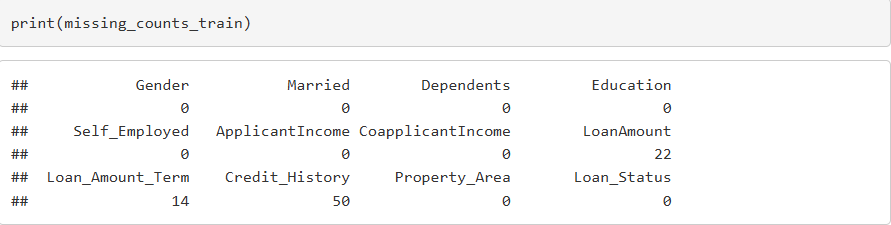
8. Here are the results that confirm that the Loan\_ID was removed successfully from both datasets.



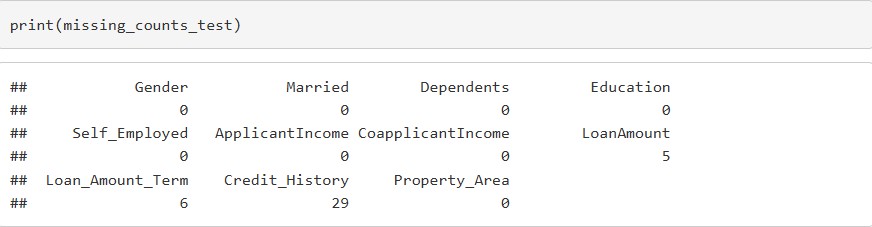
*Loan Report Figure 8*

#### Handling Missing Values

It is essential to identify and handle any missing values in the datasets. Most of missing data is attributed to human error of omission of some answers by respondents or machine error. They can impact the results of your analysis. They reduce the amount of usable data, and this compromises the insights that could be retrieved from the data. Figure 9 and figure 10 show the number of missing values in the datasets.



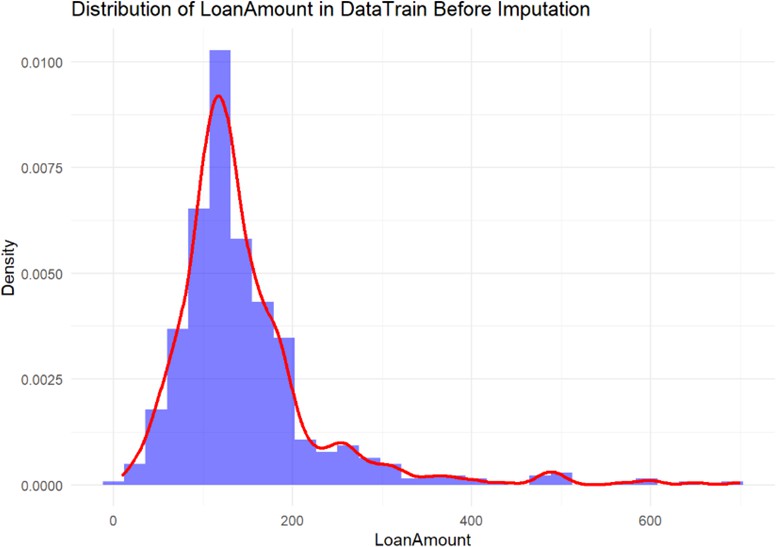
*Loan Report Figure 9*

**

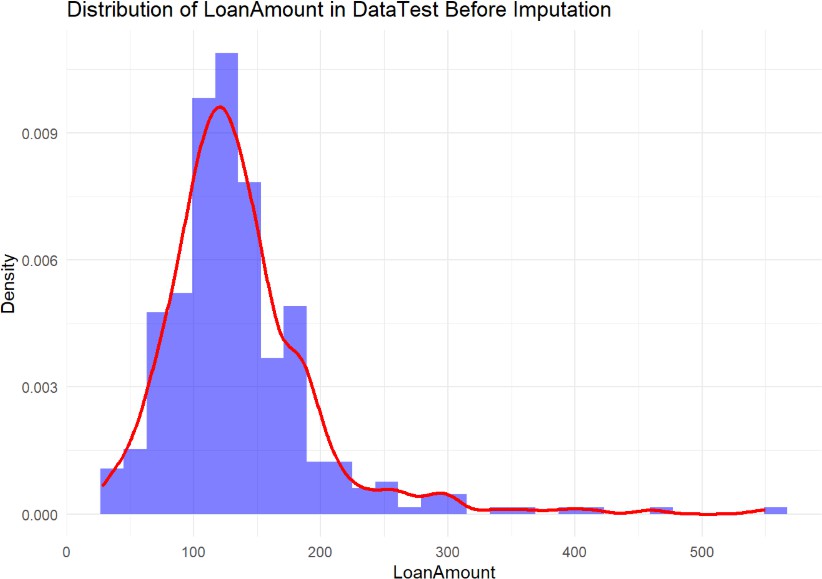
*Loan Report Figure 10*

To handle the missing values in LoanAmount, it is important to visualize both datasets and see the distribution of the data. Through this, it is easy to identify which method would be used to handle the missing value.

In the analysis of the datasets, I created dataframe that summarizes both missing and non-missing values then reshaped them for plotting purposes. The histograms were then generated to show the data distribution in both DataTrain and DataTest.



*Loan Report Figure 11*

**

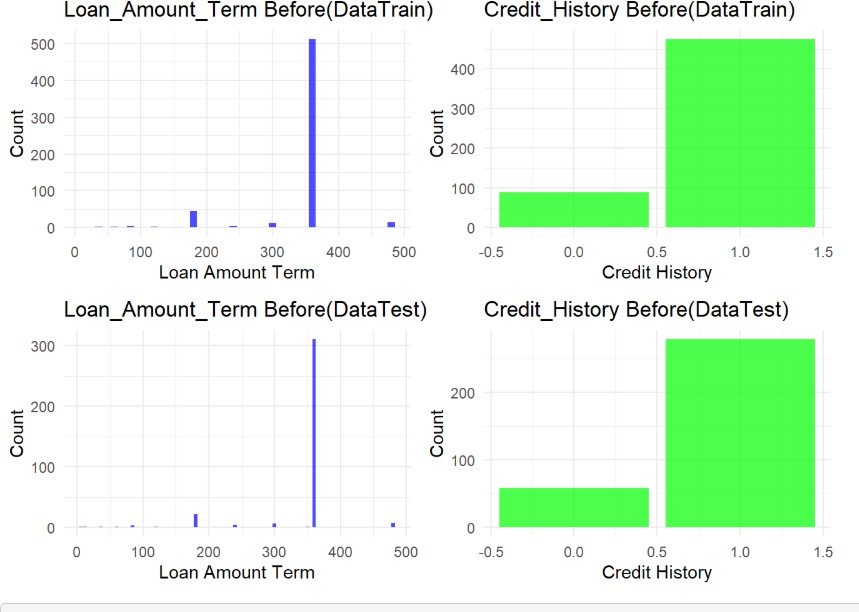
*Loan Report Figure 12*

Both histograms show a right-skewed distribution. This means that there are extreme values in leaning towards the lower end. When data is rightly skewed, it means that the mean is greater than the median because of outlier presence. Choosing mean

would therefore be ideal to maintain consistency within the distribution while also reflecting the average loan size. Choosing mean will also reflect the overall magnitude of Loan\_Amount.

The ‘handle\_missing\_values’ function is now created to handle the missing values in both of my datasets. It allowed both the use of mode and mean for the columns needing replacement of values. This approach will ensure integrity and ensure effective analysis despite having missing data.

For Credit\_History and Loan\_Amount\_Term, the need for visualization was important to understand the patterns and implications of the incomplete data. Visualization will quickly show the frequency distribution and identify the underrepresented values in the datasets. The comparisons are vital to predict the loan status effectively.



*Loan Report Figure 13*

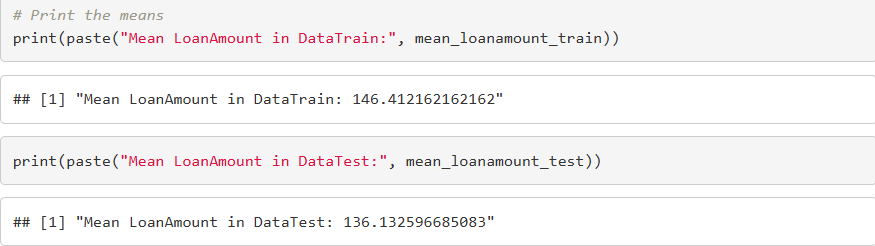
The graphs in figure 13 above indicate both the visuals of DataTrain and DataTest. To address the missing value, I will use the mode strategy. It is a logical choice because for discrete variables like Credit\_History and LoanAmount loan has one variable with the most count.

By using mode, I will preserve the integrity of my data which is the overall structure of the data. It is also necessary to use the same method for both the training and test data to prevent any discrepancies.

To calculate mode and median, I used codes to calculate the mode of a vector and compute the mean of the 2 datasets. This analysis is also important to identify the central tendency of the ‘LoanAmount’.

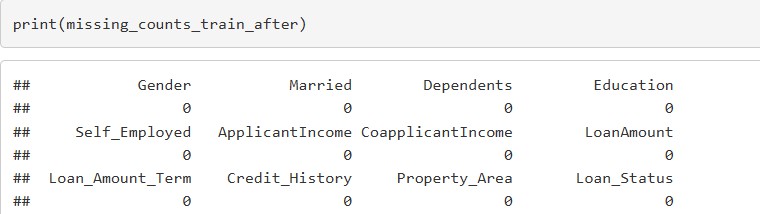
Figure 14 shows the mean of the LoanAmount in both DataTrain which is 146.4 and

136.13 for the DataTest.

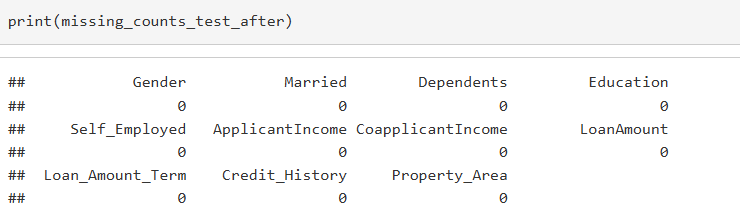


*Loan Report Figure 14*

After filling in the missing values, a validation is done across both datasets to ensure the data is now complete. It confirms that the data is ready for use and the data is now reliable for the next phases. Figure 15 and 16 confirms that no missing values are left.

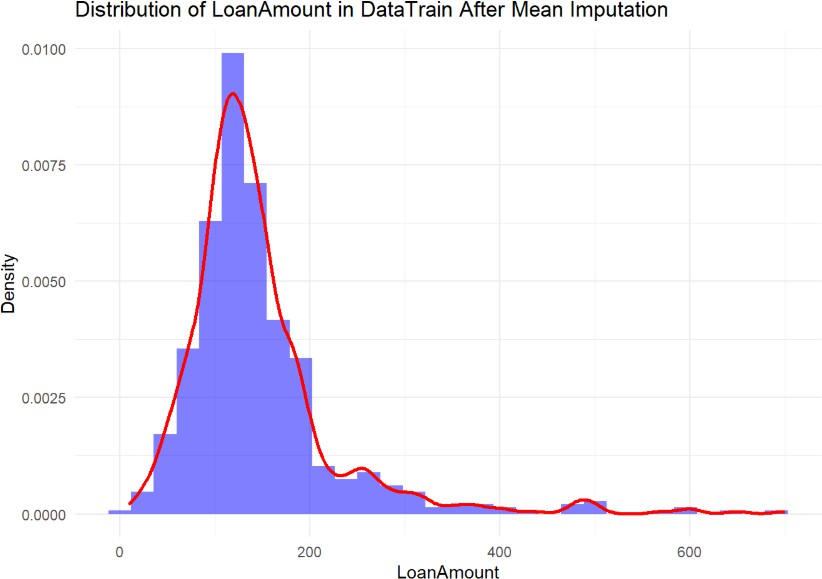


*Loan Report Figure 15*

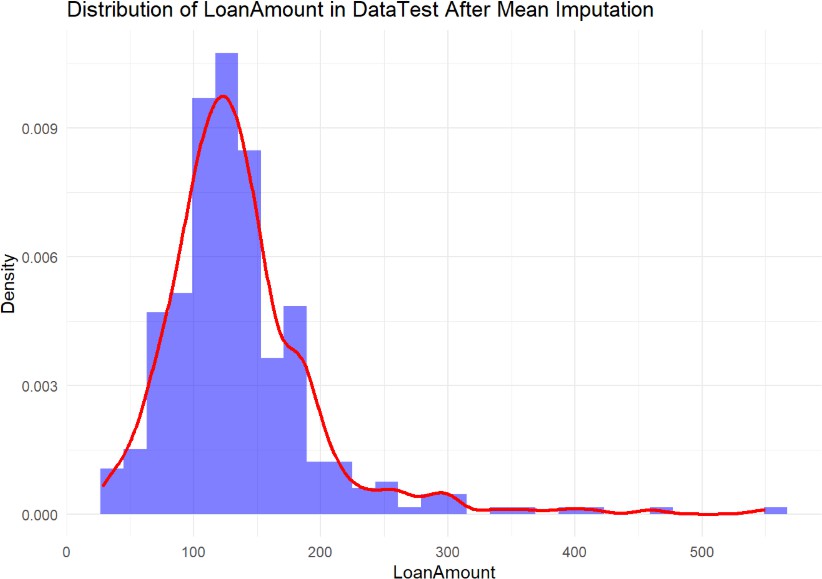
**

*Loan Report Figure 16*

After filling in missing values, it is important to create the visualizations again to see if the missing value changes the trend of the datasets. Introducing mean will cause some changes in the concentration of the values around the mean in the dataset and this will slightly change the distribution. If there are unusual spikes aft imputing the missing values, it means that bias has been introduced and this would affect the models.



*Loan Report Figure 17*

**

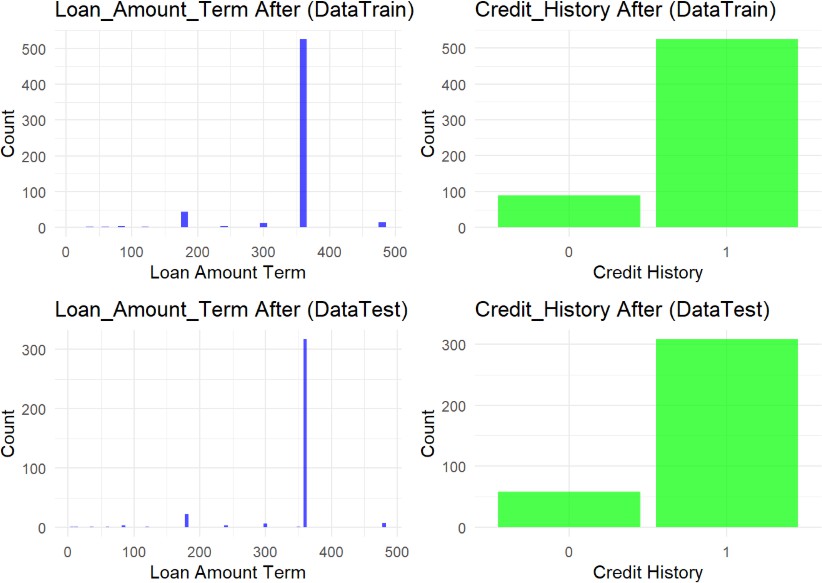
*Loan Report Figure 18*

For the LoanAmount in both datasets, the distribution of the LoanAmount demonstrates that the overall shape of the histograms remained consistent with the pre-imputation analysis. While the length of the bars in the concentrated areas changed slightly, this is an expected since the missing values that were filled with a value close to the central tendency.

Use of mean was therefore the most suitable as it allowed seamless handling of missing values and maintained the pattern of the datasets. This ensures its suitability for any statistical and modelling analysis.

Verification for the Loan\_Amount\_Term and Credit\_History is also vital. I used bar plots to assess the impact of the imputation process. Through verifying, there will be a clear intuitive representation of the data. Visualizing the columns also provides insight into whether the changes introduced in the columns caused any distortion or biasness in the data.

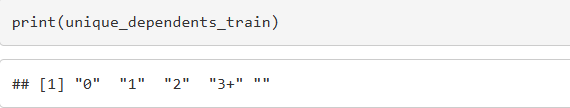
The visualizations in figures 17 and 18 reveal that the overall patterns remained constant in both datasets. For the Loan\_Amount\_Term, most values had a common term which was 360 months, and this is what we used to fill the missing values. Credit\_History retained its distribution with majority of entries being1, which means that most loan applicants have a credit history. This shows that the mode replacement was successful.



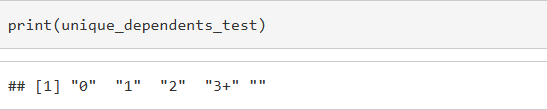
*Loan Report Figure 19*

#### Feature Engineering

‘Dependents’ column in both datasets has unique columns like 0, 1, 2 and 3+ as shown in figure 20 and 21. To do predictions easily, numerical values will work better hence the need to change 3+ to 3.

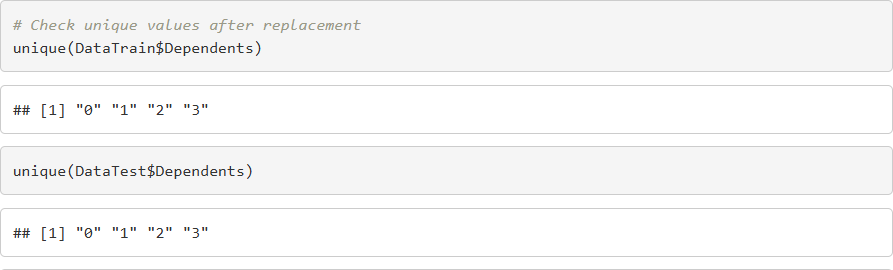


*Loan Report Figure 20*

**

*Loan Report Figure 21*

The change from 3+ to 3 maintains the nature of the value showing that higher value has greater significance. This step was important in preparing the data to ensure the insights derived are interpretable and accurate. After replacing the values in the ‘Dependents’ column, it is essential to check the unique values to ensure that the training and testing data have been modified correctly.



Loan Report Figure 22

All values have been correctly consolidated into category “3”. The data are now in a consistent which is essential for accurate analysis.

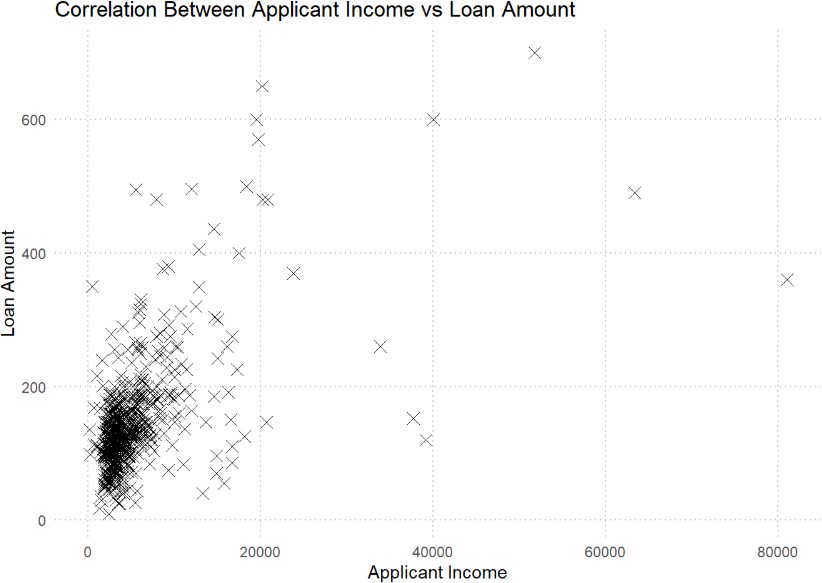
### Exploratory Data Analysis

This is a crucial step in any data analysis process. It involves visualization of key characteristics of the dataset to get a good understanding of the structure of the data. Through EDA, insights are derived that guide in selecting the appropriate modelling techniques.

The testing dataset, DataTest, even without performing its EDA, has undergone the same preprocessing steps as the training Dataset, DataTrain to ensure consistency.

**Applicant Income vs Loan Amount**

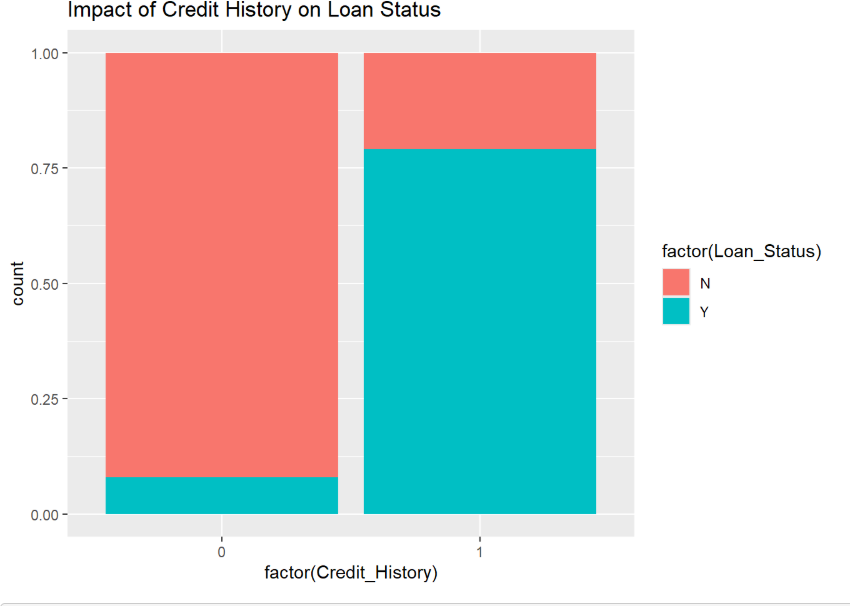
To analyse the relationship between Applicant Income and loan amount, a scatter plot was generated as shown in figure 23. It shows the relationship between applicant income and loan amount allowing an analysis of how these 2 variables correlate. A sizeable number of applicants have an income of between 0 to 20,000 and their loan application range between 0 to 400,000 which shows that many loan applicant fall within the lower income bracket. Very few individuals representing higher income earners have larger loan amounts which shows that there might be relationship between the amount of loan applied and the income of the applicant.



*Loan Report Figure 23*

**Credit history vs Loan Status**

Credit history is a record of your past debt. It shows how you have managed the repayments of your loans or even credit cards in the past. To understand how applicants with a credit history related with their loan status, a bar plot is created. The cast majority with a credit history received a loan as shown by the bar reaching almost 75% of the total applicants. About 10% of loan applicants got a loan approval without having a credit history. The graph indicates that having a credit history increases your chances of getting loan approval.



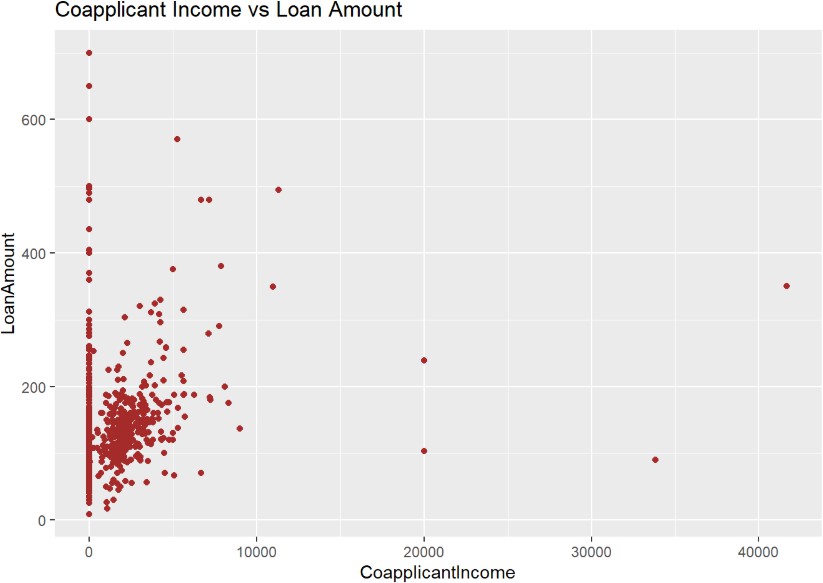
*Loan Report Figure 24*

**Coapplicant Income vs Loan Amount**

A co-applicant is somebody who participates in the loan underwriting and goes through the loan process with the main applicant. In home loan, most of the co-applicants are the partners(couple) or in some cases, a parent, daughter, or son. Understanding the relationship between a co-applicant and loan amount helps assess the risks associated with lending. Other than risks, it may show if loans are sought by households with dual or single people households.

The scatter plot in figure 25 visualizes the relationship between Co-applicant income and loan amount. A significant cluster of data point s are along the x-axis indicating several applicants applied for loans even without a co-applicant. They even applied for massive amounts of 400,000+ indicating that a co-applicant is not a mandatory

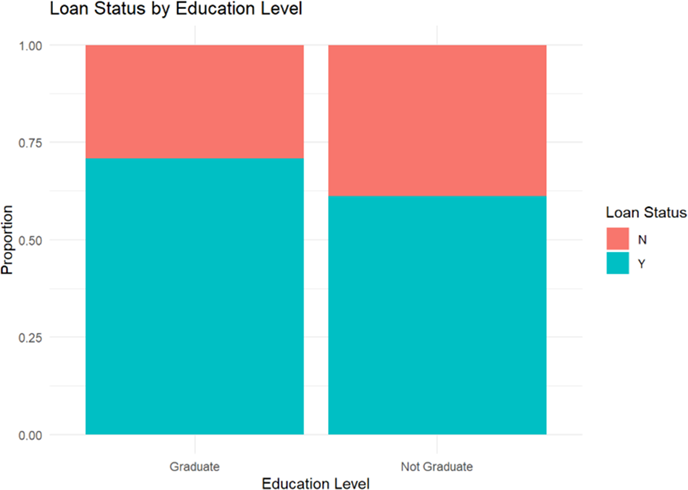
requirement to apply for a home loan. There is no strong correlation visible between a co-applicant and the amount of loan meaning that other variables might be more influential in a loan approval decision.



*Loan Report Figure 25*

**Education level vs loan status**

The bar plot in figure is done to visualize the relation between education and loan application is done. This helps understand the applicants’ profiles, understand if education influences an applicant’s financial behaviour and assess how education impacts access to financial services.



*Loan Report Figure 26*

**Property Area vs Loan Status**

In figure 27, it shows that the proportion of graduates who received a loan approval was about 70%. This could be an implication that an applicant at graduate level has higher chances of securing a loan since they may have a stable job and higher earning potential which makes them more attractive to lenders.

On the other side, 60% of non- graduates got loan approvals which means that non- graduates also have a reasonable chance of getting loan approvals. Further analysis will better understand the decision-making criteria.

In our data, we have a column showing property area. The entries are urban, semi- urban and rural areas. Understanding loan status on different geographical segments may indicate where applicants with financial stability are majorly located, financial literacy in different regions and where the finance housing loans are mostly considered.



*Loan Report Figure 27*

Key observation made in figure 27 show that approximately 75% of semiurban applicants got a loan approval indicating favourable conditions for loan approval. This may indicate stable incomes, moderate property values and reliable credit histories as well. Urban applicants rage between 63-67% which could also be attributed to financial stability and economic activities available in urban settings. Being lower than

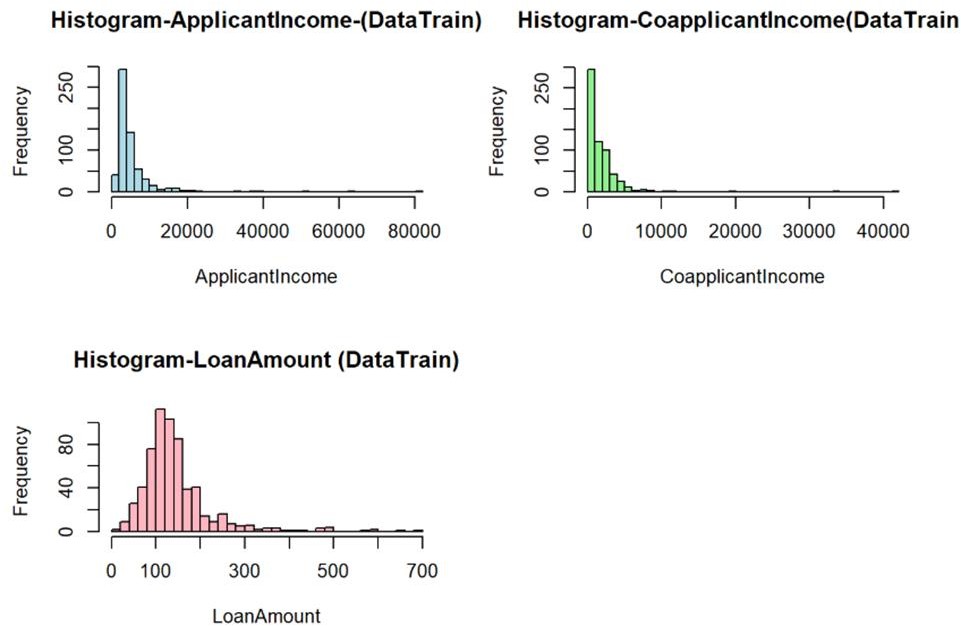
semiurban would be influenced by higher loan requests, higher cost of living and stricter credit checks. Lastly, the rural areas had the lowest loan approval of around 58%. Factors that contribute to this would be lack of financial literacy thus affecting their loan eligibility.

### Outliers

During the summary statistics and the EDA process, it was clear that there was skewed distribution of data. For example, our applicant income has a minimum value of 150 and maximum of 81,000 yet the median was 3,812 and the 75th percentile was 5,795. The minimum and maximum values were significantly higher than the interquartile range.

Similarly, the co-applicant income was ranging between 0 and 41,667 yet the median displayed as 1,188. This suggested data issues or zero meant that there were no co- applicants reported by the loan applicants.

Lastly, the loan amount also ranged between 9 and 700 yet the median was 128 and the 75th percentile was 168, this shows a big gap between 75th and maximum value indicating more outliers within the higher range.



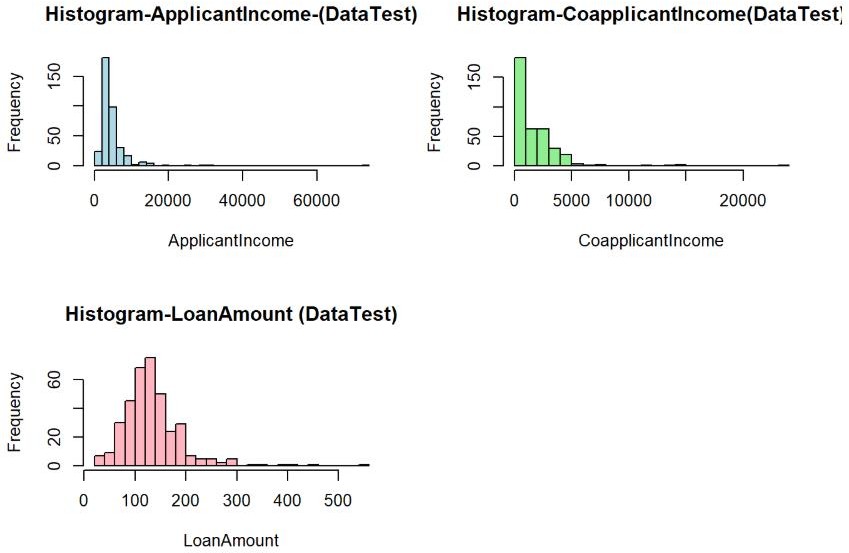
*Loan Report Figure 28*

I examined outliers using Histograms to visualize the outliers. Despite not having a clear guidance in how to deal with outliers properly as they are considered ‘problematic data, detecting them is important for better data understanding and identify if any extreme values show errors or anomalies. They show variability in data and sometimes do need further investigations.

In figure 28, the ApplicantIncome and Coapplicant show that most applicant incomes are at the lower levels. This means that low-income earners apply for housing finance more than the high-income earners. With the pressure of the rapid population worldwide that is estimated to be almost 9 billion by 2030, many urban poor may not afford housing without proper housing finance solutions, hence the need for affordable housing finance market (World Bank Group, 2023).

The outliers detect the high-paying jobs or entrepreneurs who also seek housing finance, and they are the minority. The LoanAmount Outliers are between 0 to 300,000. Considering the income of most applicants is low, it is logically sensible as to why the loan amount would be low as well.

Outliers in the DataTest are also checked using histograms in the same way and figure 29 shows that they have the same trend as the DataTrain.



*Loan Report Figure 29*

The decision to keep the outliers was influenced by the following reasons:

* In my Data, the outliers represent real, valid data. For example, in a case where a Loan Applicant has high income, it shows unique cases of high-income earners who still choose to use Housing Finance and high loan amounts reflect the needs of the loan applicants.
* Removing outliers might lead to the loss of valuable information of whether high loan amount earners get loan approvals or not. Does having a high loan increase or reduce your chances of getting a loan?
* Model robustness. Random Forest that is less sensitive to outliers because it treats all values that are on the same side equally. Ellis (2023). XGBoost that builds trees and corrects errors from them (Wang (2024). This minimizes the influence of the outliers.

# ENCODING

Encoding is a key step in data preprocessing especially when preparing categorical variables. Most machine learning algorithms cannot work directly on categorical data as they rely on mathematical operations that require numerical values. Some of the variables in my datasets such as Loan\_Status, Gender, Marital status (Married), employment status (Self\_employed) and education status(education) needed binary encoding(0,1). On the other hand, Property\_ Area was transformed through One-Hot encoding that allowed capturing distinct effects of each property without imposing an ordinal relationship and the original column was removed to avoid redundancy.

After encoding, it is essential to verify that the process was a success by inspecting the first few columns of both datasets. Figure 30 confirms that the encoding was successful for both DataTrain and DataTest.



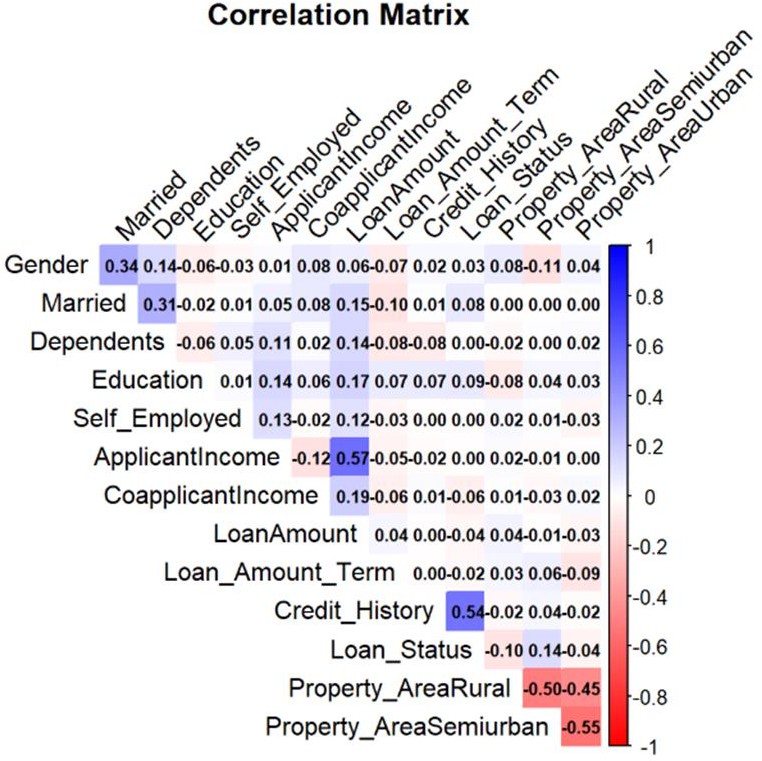
*Loan Report Figure 30*

**More EDA after encoding. Correlation heatmap.**

It provides a visual representation of the correlation coefficients between pairs of variables within a dataset. Relationships and dependencies can be easily examined which is crucial for data analysis and making predictions.

Some of the key observations include:

1. Positive correlations between
   * Credit history and loan status of 0.54 indicating that individuals with good credit history have high chances of getting loan approval.
   * Applicant Income and Loan Amount of 0.54 suggesting that applicants with higher income tend to request higher loan amounts due to their capacity to pay the loans.
2. Negative correlation among all property areas as they are all mutually exclusive.



*Loan Report Figure 31*

1. Weak Correlations between
   * Applicant income and co-applicant income of 0.19 suggesting that there is a relationship but not significant enough.
   * Loan\_Amount\_Term and Applicant Income of 0.01 which is completely negligible. This indicates that the loan duration is not at all influenced by the applicant’s income.

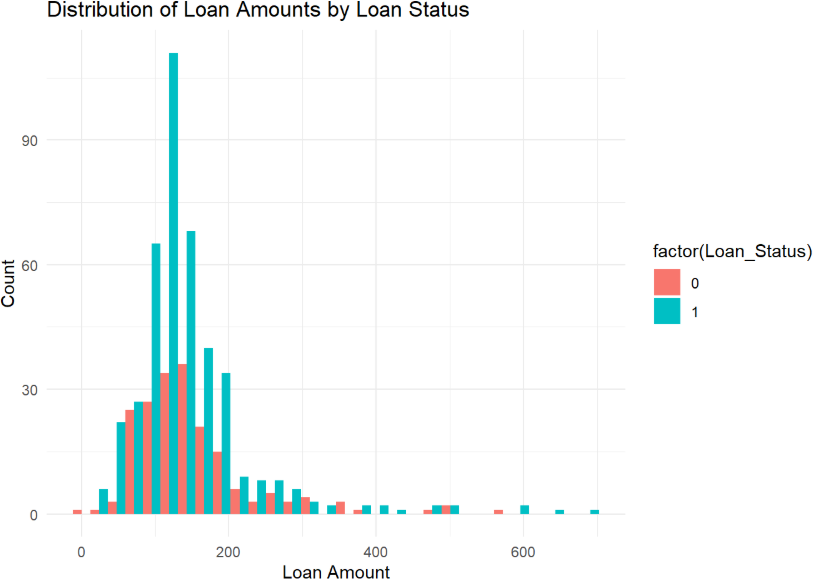
**Loan Amount vs Loan Status**

To understand how loan amounts and loan status relate, I made a histogram that will help understand how loan amounts vary between approved and rejected loans.

The histogram in figure 32 visualizes the loan status with approved vs rejected loans. The loans are dominantly between 100 and 250. There are barely any loans beyond

400. The teal bars show approved loans and are concentrated between 100 and 250.

The proportion of approved loans decreases with the increase in loan amounts, and this may be an indicator of a stricter approval criteria for higher loan amounts. The red



*Loan Report Figure 32*

bars represent unapproved loans, and they are present overall. The most frequent loans are between 150 and 200 and there are more approved than rejected.

**Loan Amount vs Applicant Income by loan status.**

In figure 33 shows a significant majority of data points at lower ranges of Applicant Income which is below 20,000 and loan amount of 250(000). Most approved loans are within the lower ranges whether approved or rejected. Rejected loans (red) are more dispersed and more common for higher loan amounts. Larger loans have higher rejection rates even for more moderate to high incomes.



*Loan Report Figure 33*

**EDA and Data Preprocessing Conclusion**

The EDA and preprocessing provided valuable insights and prepared the data for modelling. They have provided a clean and structured dataset laying a good foundation of doing the predictions using models. The next involves leveraging learning algorithms to predict loan status effectively based on the insights generated from this phase.

**MODELS TO DO PREDICTIONS**

In this study, machine learning models were implemented to predict the loan approval outcomes. The key factors influencing loan approval (Loan Status) are predicted as well. The models employed for this were Random Forest and XGBoost.

Random Forest builds multiple decision trees and aggregates their prediction to improve the accuracy of the outcomes. It is suitable for datasets with complex relationships, and it is also robust to outliers.

XGBoost is a high-performance gradient boosting algorithm that is known for its scalability and speed. It employs advanced regularization techniques to ensure optimum performance.

Both models handle non-linear relationships well and offer insights that aid in a comprehensive understanding of loan approval factors.

Before training the models, the data was pre-processed extensively to ensure its suitability for the analysis.

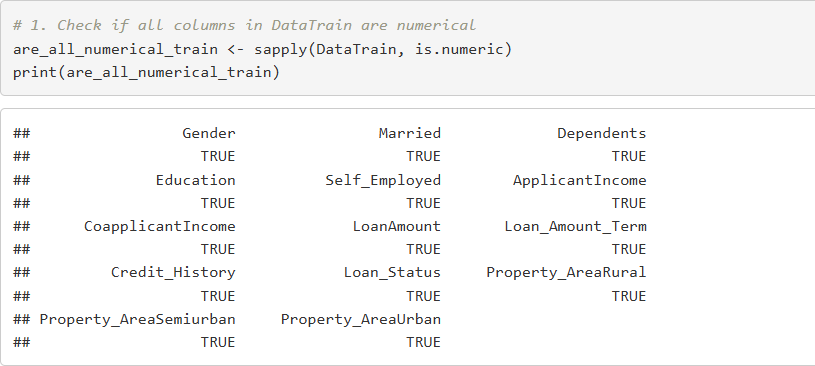
**Steps taken in using models for Prediction.**

1. **Loading the libraries**

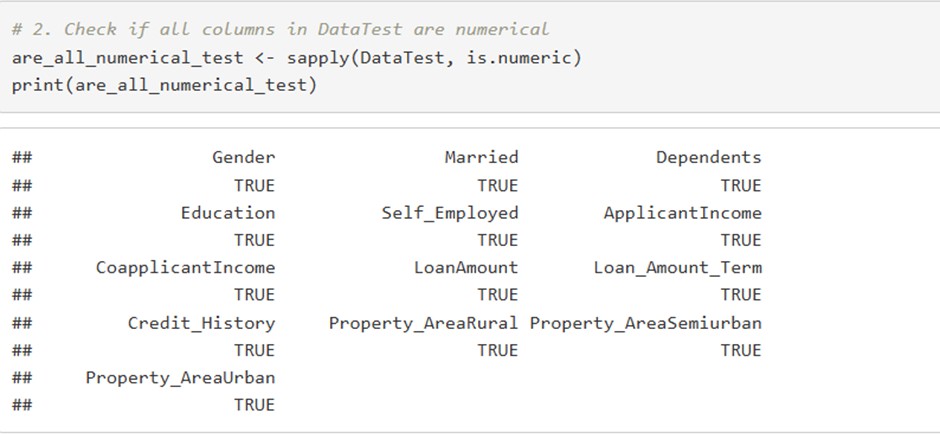
This loads the necessary libraries for the analysis. They are essential for training models, visualizing graphs, and streamlining the workflow.

1. **Confirm all columns are numeric.**

Since Random Forest and XGBoost use mathematical operations, all data must be in numeric for it to be interpretable**.**

****

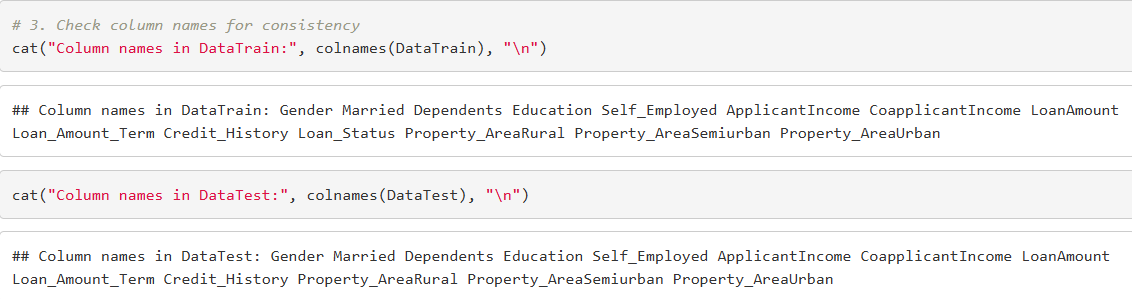
*Loan Report Figure 34*

**

*Loan Report Figure 35*

1. **Column name consistency**

The training and testing datasets need to have a consistency in their column names for the models to make predictions accurately.



*Loan Report Figure 36*

1. **Defining Features and Target.**

This section defines the target variable (Loan\_Status) and it selects the feature columns for training and testing. ‘Setdiff(names (DataTrain), exclude\_cols) identifies all the columns in DataTrain except for the target variable.

X\_train contains the feature columns while y-train contains the target variable that is converted to a factor because of classification.

X\_test is prepared the same way.

1. **Handling non-numeric columns.**

The purpose is to ensure all columns in the training dataset are numeric. If any non- numeric are found, one-hot coding is applied using ‘dummyVars’ converting any variable, incase there is any.



*Loan Report Figure 37*

### Random Forest

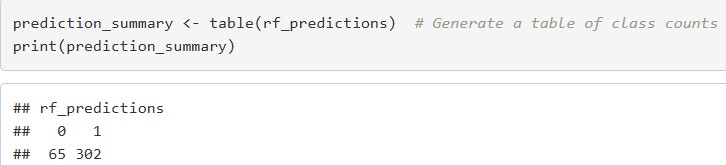
**A. Training Random Forest Model**

To predict loan approval (Loan\_Status), the Random Forest was trained using the pre- processed data. X\_train features were used, and y\_train was states as the target. The ntree = 100 means that 100 decision trees were built to form the forest to ensure robust predictions. Reproducibility of results was set using the random seed set. seed (42). The model is then used to make predictions on the test dataset (‘X\_test’). The code used establishes the groundwork for predicting loan status using Random Forest.

**b. Random Forest predictions**

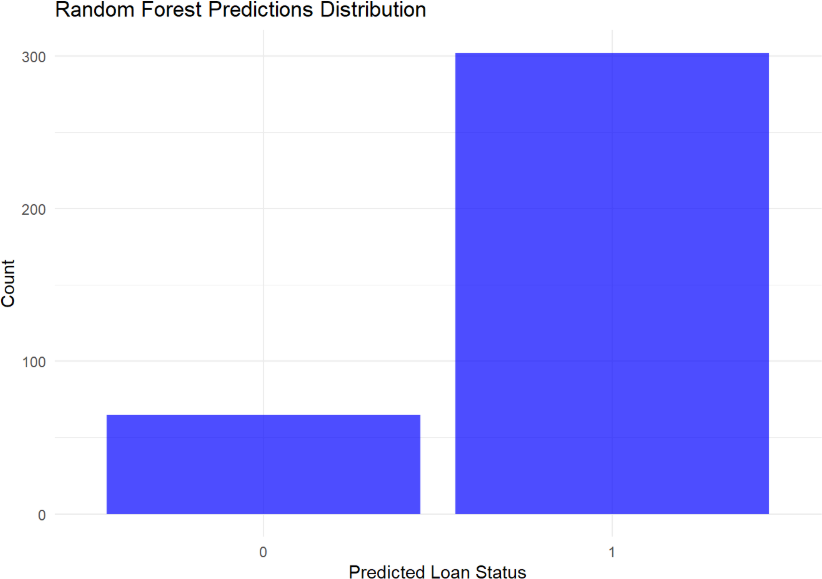
To understand the predictions made by Random Forest, the predictions are summarized in Figure 38. According to the table, 65 loans were predicted to NOT be approved, and 302 loans were approved.

This distribution provides insight into how the model is performing. There is a tendency to predict loan approvals than rejection as shown in figure 38.



*Loan Report Figure 38*

For a better understanding, figure 39 shows a visualization of the numbers. It helps identify the imbalances which is crucial for evaluating the model performance.

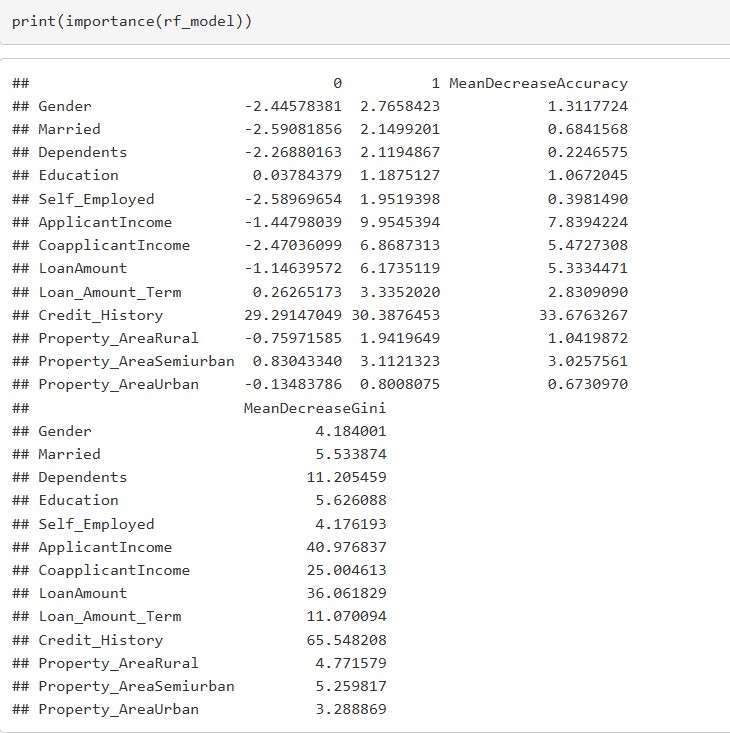


*Loan Report Figure 39*

**Random Forest Feature importance**

This section displays the importance of each feature used in the Random Forest model. Through this, variables with most influence on the models’ predictions are identified. Random Forest assesses the feature importance through 2 primary metrics.

* Mean Decrease Accuracy (MDA). It measures the decrease in model accuracy when a feature is randomly shuffled. A larger MDA means the feature plays a bigger role.
* Mean Decrease Gini (MDG). It involves calculating weighted impunity decrease for each feature across all Trees. Higher MDA suggests essential feature.



*Loan Report Figure 40*

According to figure 40, the top 3 influences are Credit\_History, ApplicantIncome and Loan\_Amount.

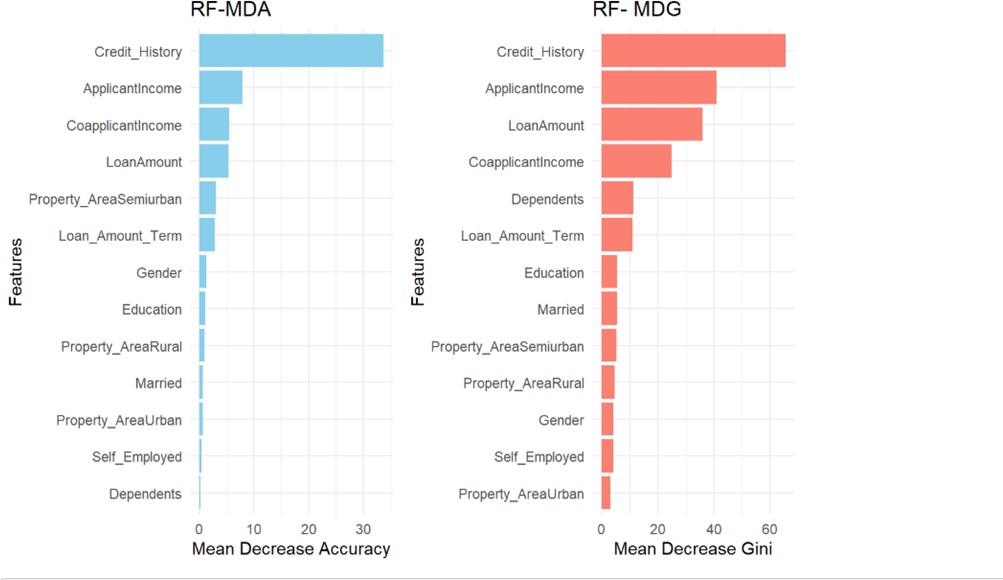
Credit History emerges as the MOST critical feature influencing loan approval as the MDA is 33.68 and the MDG is 65.55. This means that applicants with a credit history have a higher chance of receiving loan approvals.

ApplicantIncome is the second MOST important feature with an MDA of 7.84 and MDG of 40.98. The more income an applicant has, the higher the chances of securing an approval.

Loan Amount also impact the loan approval decisions with higher loans lowering the approval rate unless it is offset by other factors like a good credit history and a high income.

Gender, Property\_AreaRural, Property\_AreaUrban and Self\_Employed features are the least impactful features.

The Random Forest feature importance analysis provides invaluable insights into the key determinants of loan approval. When Dream house financing company focuses on Credit\_History, ApplicantIncome and LoanAmount, they can refine their lending strategies in ways that reduce credit risk.



*Loan Report Figure 41*

## XGBOOST

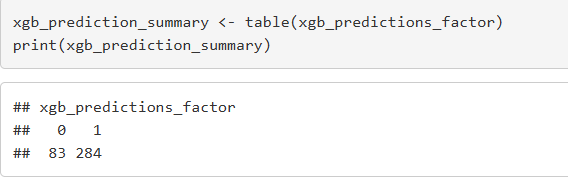
1. **XG Boost Training**

This part prepares the data for the XGBoost model by ensuring the target variable has been converted to numeric and adjusts it to be 0-based as needed by the model. I used the ‘`xgb.DMatrix(...)’ to create the Dmatrix objects for both training and testing datasets to improve performance and memory efficiency. I also defined the parameters of the model to show its multiclass and controlled the complexity of the model.

Preparing XGBoost involved focusing on preparing the data by converting y\_train to a numeric factor to ensure its zero-based. All feature matrices are also converted into a format optimized for XGBoost. The section also trains the XG Boost using specific parameters and training data.

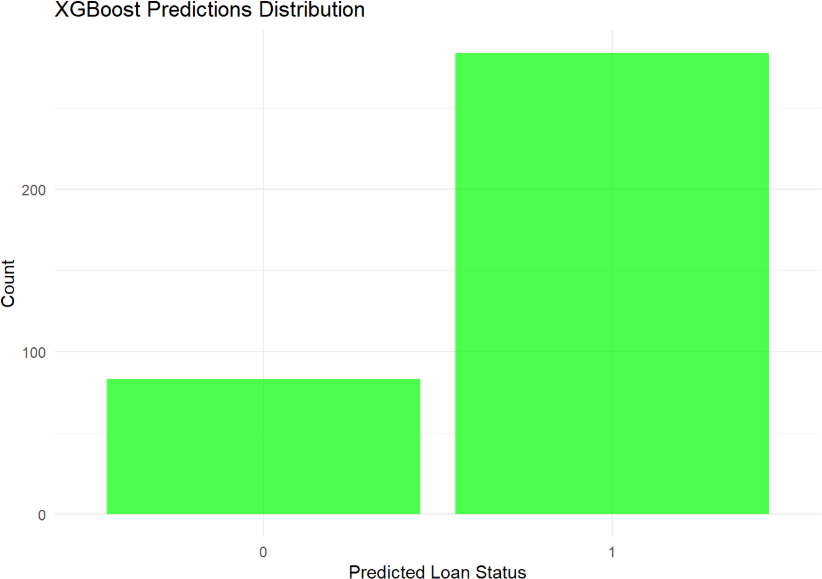
1. **XGBoost Predictions**

After training, the model’s capability was applied to test the dataset to forecast the loan approval outcomes. The model showed that 83 applicants were predicted as “0” (loan rejected) and 284 participants were predicted as “1” (loan approved).



*Loan Report Figure 42*

The bar graph in figure represents the XGBoost predictions and gives insight of the dataset’s trends and the model’s assessment of the applicants’ eligibility. It gives a quick visual snapshot making it easier for the stakeholders to understand the trends.



*Loan Report Figure 43*

**XGBoost Feature Importance**

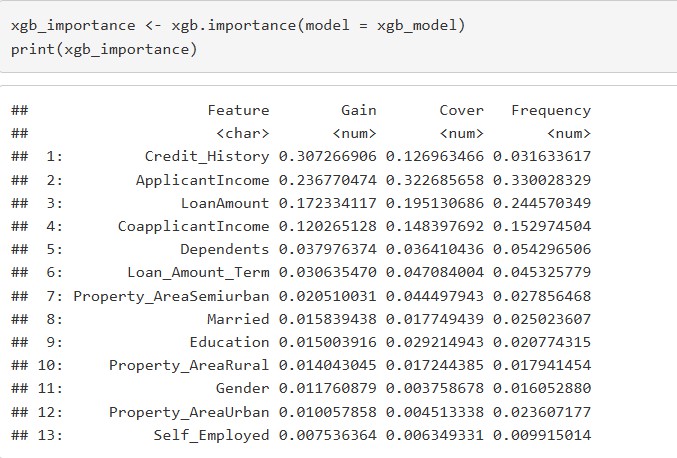
The feature importance in XGBoost is based on metrics such as Gain, Cover and frequency that provide insights into the contribution of each feature in model decision- making.

Credit\_History emerges as the most significant predictor of loan status with gain of 30.73%, Cover of 12.70% and frequency of 3.16%. gain shows that the model relies heavily on it for accuracy splits.

ApplicantIncome comes second with gain of 23.68%, cover of 32.27% and frequency of 33.00%. The cover percentage is evidence of how significant his portion is in the dataset.

The third top feature is LoanAmount with gain of 17.23%, cover of 19.51% and frequency of 24.46%. its combination of gain and frequency emphasizes its importance in model predictions.

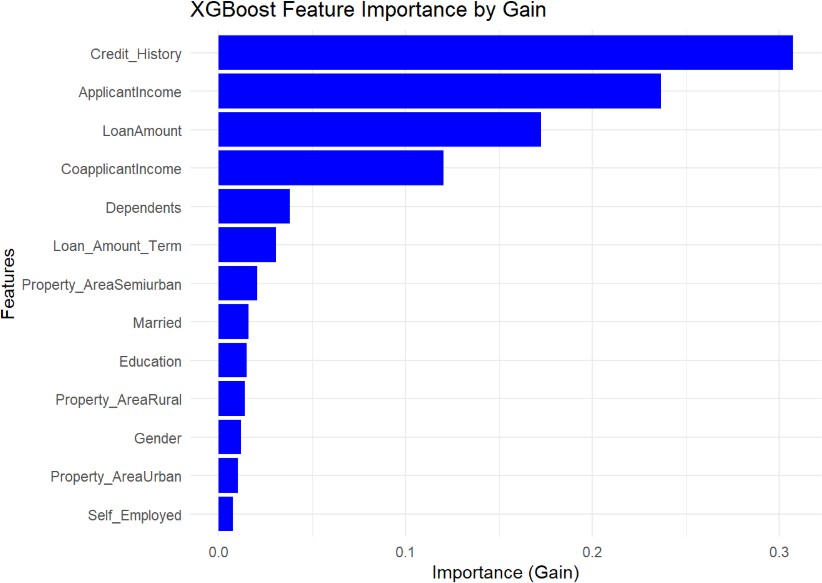
Features like Self-Employed, Property\_AreaUrban and Gender have the least impacts on loan status predictions showing that financial institutions prioritize financial and credit factors over employment status, property location and demographic characteristics.



*Loan Report Figure 44*

The figure 45 provides a quick overview of the most to the least influential factors in the loan approval process. By focusing on the high-gain features, lenders can easily

refine their evaluation criteria to ensure that their decisions align with the predicted insights.



## MODEL EVALUATION

*Loan Report Figure 45*

This is the process to assess the performance of both models. I used both cross validation and hyperparameter to evaluate the performance of Random Forest and Cross Validation.

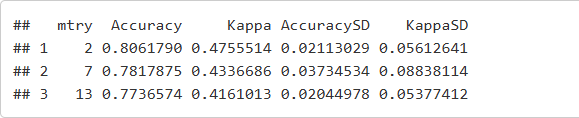
Cross validation is a robust statistical method used to evaluate the performance of machine learning models. It provides reliable performance evaluation; it mitigates the variance that may arise from single train-test and detects overfitting.

On the other hand, hyperparameter optimizes the model performance, avoids overfitting and underfitting and improves interpretability of the model.

In this report, we compare the performance of XGBoost and Random Forest using cross validation and hyperparameter tuning in predicting loan status. To achieve this, we need to assess the metrics of the models independently then compare their results.

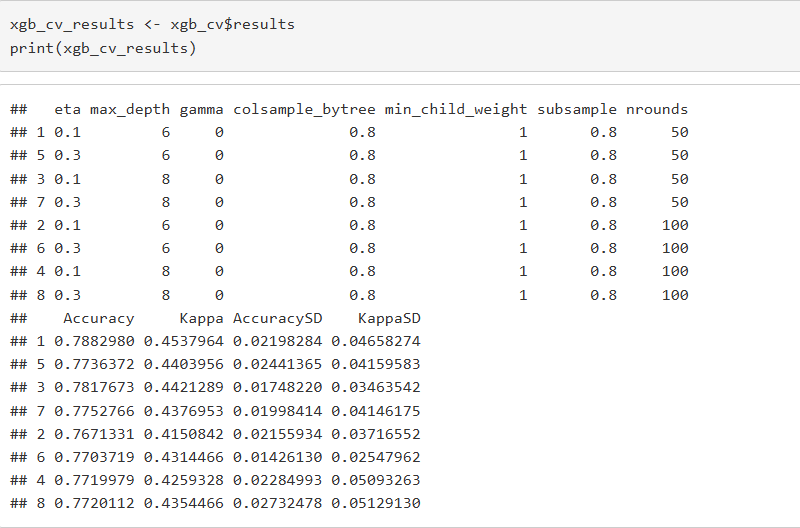
For Random Forest, under the cross validation, 5-fold cross-validation model is used to predict the loan status. The results from this will give insights into the model’s performance and will be used to compare the results with XGBoost.

The results in figure 46 shows Accuracy which is a proportion of correctly predicted instances and Kappa that provides insight beyond the mere accuracy.



*Loan Report Figure 46*

Cross-validation in conjunction with Hyperparameter tuning is also done in XG Boost to analyse its performance in predicting loan status. Hyperparameter tuning by boosting the number of rounds, the learning rate and maximizing the tree depth optimize the model’s performance. The results obtained as shown in figure 47 provides insights on how well the model performs in comparison to Random Forest.



*Loan Report Figure 47*

### Validation Results Explanation

The cross-validation results provide perfect profiling of Random Forest and XGBoost models based on their performance on different folds. Despite both models performing very well, it is evident that Random Forest performs better overall.

Based on the cross- validation results, the accuracy, kappa, Accuracy SD 6 and Kappa SD of both models are shown, and they help analyse which model is better for predicting loan status.

|  |  |  |
| --- | --- | --- |
|  | **Random Forest** | **XGBoost** |
| **Accuracy** | 80.62% | 78.83% |
| **Kappa** | 0.476 | 0.454 |
| **Accuracy SD** | ±2.11% | ±2.19% |
| **Kappa SD** | ±5.61% | ±4.66% |

The results indicate that Random Forest show a stable and high accuracy and shows a good agreement between the actual and predicted loan status measured by Kappa. The standard deviation across also supports the reliability of the model.

On the other hand, XGBoost also performed well despite having a lower accuracy than RF. It is also less consistent compared to RF making RF more dependable.

Random Forest outperforms XG Boost due to the following factors:

* Higher accuracy making it a reliable choice in predicting loan status.
* Better stability due to the lower standard deviation evidenced by Kappa.
* Ease of interpretation as its metrics (Mean Decrease Accuracy) and Mean Decrease Gini (MDG) both rank the most key features in loan status.

## BUSINESS INSIGHTS

The results prove that Random Forest is the preferred model for predicting loan status. Key takeaways from the analysis include:

**Credit History is the key determinant of loan approval**. Banks and lenders need to thoroughly check the credit history of the loan borrowers as it is an indicator their financial behaviour with loans. Lenders should also evaluate the credit scores of loan applicants during the loan approval process to assess how much risk is associated with every loan approval or each client.

**Applicant and CoApplicant Income significance.** The primary source of income of the loan applicant will directly impact their likelihood of getting a loan approval. It

assures that the lender that the applicant has a source of income and can meet their debt obligations. Coa-applicant Income enhances overall financial stability when the co-applicant (in most cases partner) has an income as well. It demonstrates a stronger financial base within the household.

**Evaluating Loan Amount** . This is the amount of money requested. This should align with the applicant’s financial capacity. Bigger loans mean bigger risks for the lenders hence higher scrutiny strategy to ensure they are lending to a dependable person. The lenders should also learn to educate their customers about borrowing within their capacity as it will reduce default rates in banks and the applicants can improve their finances to help them access bigger loans.

**Decision making based on the model results.** Random Forest ranked the key features. This should be used by lenders to influence their decision-making process. This will reduce potential bias that may be present when using traditional manual evaluations.

### Limitations

Retaining the outliers in our datasets as they represented high loan amounts and high incomes. Despite this decision being intentional, there were chances that it impacted the model stability despite using models that collaborate well with outliers.

The data was also limited to only one financial institution and this may not be a representation of the wider market. The dataset was small which means that it could lead to biased outcome.

Limitation of features in our datasets. Features such asset value of the applicant, employment details(part-time, full-time, or contractual), debt-to-income ratio which would provide their capacity to manage debt and collateral value. This would have provided a better detailed prediction.

### Future Work

Expansion of data to include a wider range of loan applicants. This would help reduce any bias in our work and create models that will better reflect real-world situations.

Feature engineering to improve on the model performance. For example, if there was available macroeconomic data and interest rates of the loans, they would improve the model’s ability to predict loan approvals.

Real-time prediction integration to enable faster decision-making and improving the costumer’s experience.

### Conclusion

Predictive models, Random Forest and XGBoost were developed to determine loan approval using the given test dataset.

Random Forest became the best- performing model revealing Credit History as the most significant feature in loan approval. These reveals the importance of good financial discipline and with features like loan amount and applicants’ income being significant, they also show the importance of financial stability.

These findings will provide a roadmap for lenders, especially Dream House Financing Company to streamline the loan approval processes. While the study has its limitations, it gives insight and motivation for future research to refine models that can approve loans in real time.

.

# REFERENCES

1. Kavaarpuo, G., Mintah, K. and Donkor-Hyiaman, K.A. (2023) 'Financing housing development in an underdeveloped financial market: Learning from developers’ financing adaptations,' *Housing Policy Debate*, pp. 1–28. Available at: <https://doi.org/10.1080/10511482.2023.2237004>.

1. Allied Market Research (2022) 'Financial services - analyzes credit and investment business functions,' *Allied Market Research*. Available at: [https://www.alliedmarketresearch.com/banking-financial-services-and-](https://www.alliedmarketresearch.com/banking-financial-services-and-insurance/financial-services-market-report) [insurance/financial-services-market-report](https://www.alliedmarketresearch.com/banking-financial-services-and-insurance/financial-services-market-report).
2. Doling, J. et al. (2013) 'Housing and housing finance: A review of the links to economic development and poverty reduction,' *ADB Economics Working Paper Series*. Asian Development Bank. Available at: <https://www.adb.org/sites/default/files/publication/30348/ewp-362.pdf>.
3. Yamparala, R. et al. (2022) 'Predicting loans using machine learning,' in *Advances in Intelligent Systems and Computing*, pp. 701–712. Available at: <https://doi.org/10.1007/978-981-19-3590-9_55>.
4. Fusefinance.com (2023) '13 benefits of loan automation for financial institutions,' *Fuse Finance*. Available at: <https://www.fusefinance.com/blog/benefits-of-loan-automation>.
5. Wingard, L. (2024) 'Top 10 banking industry challenges — and how you can overcome them,' *Hitachi Solutions*. Available at: [https://global.hitachi-](https://global.hitachi-solutions.com/blog/top-10-challenges-banking-financial-organizations-can-overcome/) [solutions.com/blog/top-10-challenges-banking-financial-organizations-can-](https://global.hitachi-solutions.com/blog/top-10-challenges-banking-financial-organizations-can-overcome/) [overcome/](https://global.hitachi-solutions.com/blog/top-10-challenges-banking-financial-organizations-can-overcome/).
6. Najafbagy, R. (2010) 'Banking automation in Iran: Its social and banking effects,' *Journal of Performance Management*.
7. MonJa (2023) '4 challenges of manual consumer lending and how automation solves them,' *MonJa*. Available at: [https://monjaco.com/blog/4-challenges-of-](https://monjaco.com/blog/4-challenges-of-manual-consumer-lending-and-how-automation-solves-them/#%3A~%3Atext%3D4%20Challenges%20of%20Manual%20Consumer%20Lending%201%201.%2C4.%20Wasted%20time%2C%20analyzing%20fairly%20simple%20cases%20) [manual-consumer-lending-and-how-automation-solves-](https://monjaco.com/blog/4-challenges-of-manual-consumer-lending-and-how-automation-solves-them/#%3A~%3Atext%3D4%20Challenges%20of%20Manual%20Consumer%20Lending%201%201.%2C4.%20Wasted%20time%2C%20analyzing%20fairly%20simple%20cases%20)

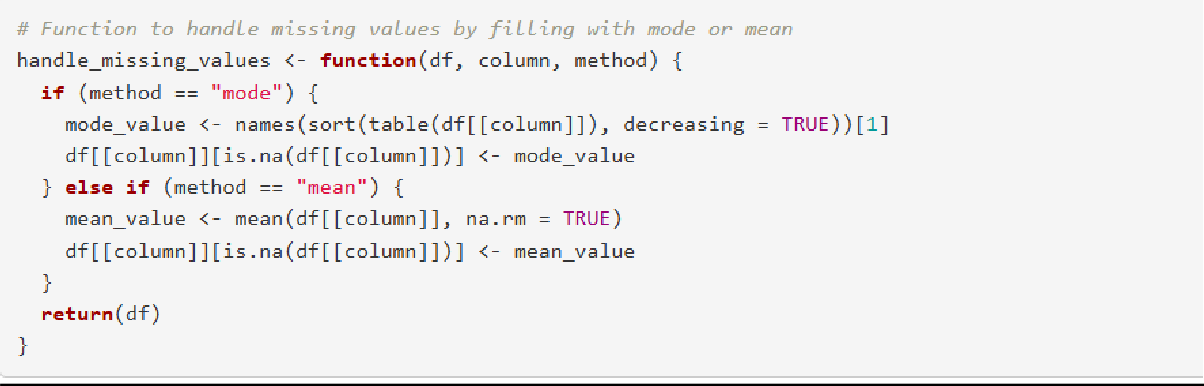
[them/#:~:text=4%20Challenges%20of%20Manual%20Consumer%20Lending](https://monjaco.com/blog/4-challenges-of-manual-consumer-lending-and-how-automation-solves-them/#%3A~%3Atext%3D4%20Challenges%20of%20Manual%20Consumer%20Lending%201%201.%2C4.%20Wasted%20time%2C%20analyzing%20fairly%20simple%20cases%20)

[%201%201.,4.%20Wasted%20time%2C%20analyzing%20fairly%20simple%](https://monjaco.com/blog/4-challenges-of-manual-consumer-lending-and-how-automation-solves-them/#%3A~%3Atext%3D4%20Challenges%20of%20Manual%20Consumer%20Lending%201%201.%2C4.%20Wasted%20time%2C%20analyzing%20fairly%20simple%20cases%20) [20cases%20](https://monjaco.com/blog/4-challenges-of-manual-consumer-lending-and-how-automation-solves-them/#%3A~%3Atext%3D4%20Challenges%20of%20Manual%20Consumer%20Lending%201%201.%2C4.%20Wasted%20time%2C%20analyzing%20fairly%20simple%20cases%20).

1. Emmanuel, T. et al. (2021) 'A survey on missing data in machine learning,' *Journal of Big Data*, 8(1). Available at: [https://doi.org/10.1186/s40537-021-](https://doi.org/10.1186/s40537-021-00516-9) [00516-9](https://doi.org/10.1186/s40537-021-00516-9).
2. Aguinis, H., Gottfredson, R.K. and Joo, H. (2013) 'Best-practice recommendations for defining, identifying, and handling outliers,' *Organizational Research Methods*, 16(2), pp. 270–301. Available at: <https://doi.org/10.1177/1094428112470848>.
3. World Bank Group (2023) 'Housing finance,' *World Bank*. Available at: <https://www.worldbank.org/en/topic/financialsector/brief/housing-finance>.
4. Wang, J. (2024) 'XGBoost: What, and how, and why?,' *Medium*, 16 November. Available at: [https://jimmy-wang-gen-ai.medium.com/xgboost-what-and-how-](https://jimmy-wang-gen-ai.medium.com/xgboost-what-and-how-and-why-e466b36e237c) [and-why-e466b36e237c](https://jimmy-wang-gen-ai.medium.com/xgboost-what-and-how-and-why-e466b36e237c).
5. MonJa (2023) '4 challenges of manual consumer lending and how automation solves them,' *MonJa*. Available at: [https://monjaco.com/blog/4-challenges-of-](https://monjaco.com/blog/4-challenges-of-manual-consumer-lending-and-how-automation-solves-them/) [manual-consumer-lending-and-how-automation-solves-them/](https://monjaco.com/blog/4-challenges-of-manual-consumer-lending-and-how-automation-solves-them/).
6. Baig, M., Sohoni, S. and Lhuer, X. (2024) 'Shifting to cloud-based platforms for banking efficiency,' *ResearchGate*. Available at: [https://www.researchgate.net/publication/362820053\_Shifting\_to\_Cloud-](https://www.researchgate.net/publication/362820053_Shifting_to_Cloud-Based_Platforms_for_Banking_Efficiency) [Based\_Platforms\_for\_Banking\_Efficiency](https://www.researchgate.net/publication/362820053_Shifting_to_Cloud-Based_Platforms_for_Banking_Efficiency).
7. Gogula, K. and Chattu, N. (2024) 'Loan eligibility prediction using machine learning,' *International Journal of Soft Computing and Engineering*, 14(4), pp. 12–15. Available at: <https://doi.org/10.35940/ijsce.c8144.14040924>.
8. Ellis, C. (2023) 'When to use random forests,' *Crunching the Data*. Available at: <https://crunchingthedata.com/when-to-use-random-forests/>.

#### 5.1.1 Appendix

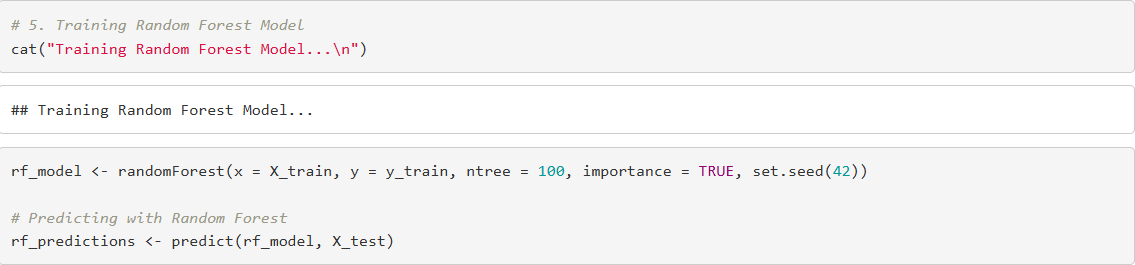
*Handling missing values for both Loan Amount Term and Credit history*

**

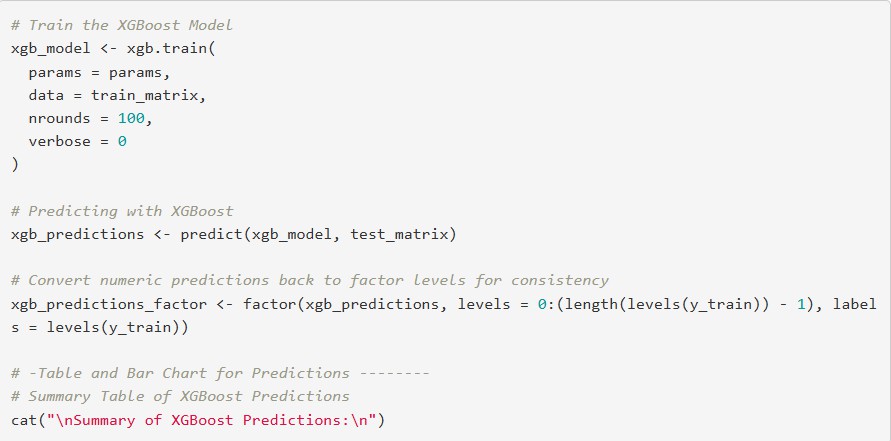
*Encoding codes*

**

*Training Random Forest*

**

*Training XGBoost*

**