***Mortgage Default Risk Analysis*:**

***Predictive Modeling for Residential Mortgage Default Risk Analysis***

Lavanya Govindarajan

Department of Computer and Information Sciences, Webster University

May 08, 2024

Table of Contents

[1. INTRODUCTION 3](#_Toc166067347)

[**1.1** **BUSINESS GOAL** 3](#_Toc166067348)

[**1.2** **ANALYTICS GOAL** 4](#_Toc166067349)

[**1.3** **ANALYTICS APPROACH** 4](#_Toc166067350)

[2. DATA PREPROCESSING 5](#_Toc166067351)

[**2.1** **TRANSFORMING DATASET** 10](#_Toc166067352)

[**2.2** **OUTLIER ANALYSIS** 11](#_Toc166067353)

[3. EXPLORATORY DATA ANALYSIS 12](#_Toc166067354)

[**3.1** **CATEGORICAL VARIABLE ANALYSIS** 12](#_Toc166067355)

[**3.2** **NUMERICAL VARIABLE ANALYSIS** 15](#_Toc166067356)

[4. PREDICTOR ANALYSIS AND RELEVANCY 20](#_Toc166067357)

[**4.1** **VARIABLE SELECTION** 23](#_Toc166067358)

[5. DIMENSION REDUCTION 25](#_Toc166067359)

[6. DATA PARTITIONING 25](#_Toc166067360)

[7. CLASSIFICATION MODEL SELECTION 25](#_Toc166067361)

[**7.1** **CLASSIFICATION TREE** 26](#_Toc166067362)

[**7.2** **LOGISTIC REGRESSION** 27](#_Toc166067363)

[**7.3** **NAÏVE BAYES** 28](#_Toc166067364)

[8. REGRESSION MODEL SELECTION 28](#_Toc166067365)

[**8.1** **MULTIPLE LINEAR REGRESSION:** 29](#_Toc166067366)

[**8.2** **REGRESSION TREES:** 30](#_Toc166067367)

[**8.3** **NEURAL NETWORK:** 31](#_Toc166067368)

[9. CLUSTER ANALYSIS 32](#_Toc166067369)

[10. MODEL EVALUATION 36](#_Toc166067370)

[11. PREDICTION 37](#_Toc166067371)

[12. CONCLUSION 39](#_Toc166067372)

[13. RECOMMENDATIONS 40](#_Toc166067373)

1. **INTRODUCTION**

The dream of owning a home is a common aspiration for many individuals, driving significant growth in the real estate sector. As a consequence, for the vast majority of individuals, purchasing a home often involves seeking financial assistance in the form of mortgage loans from banks and financial institutions. From the perspective of lenders, the decision to sanction a loan involves a thorough assessment of potential risks and rewards.

In particular, the possibility of a borrower defaulting poses a significant concern for banks, potentially resulting in substantial losses. Therefore, it is important for lending institutions to carefully evaluate the likelihood and potential impact of such risks, identifying borrowers who may present a higher risk of default. This allows for informed adjustments to pricing and loan approval criteria aimed at mitigating potential losses. Concurrently, while managing risks is critical, banks also seek to optimize profitability by attracting borrowers who demonstrate a high likelihood of timely loan repayment. Consequently, it becomes essential to maintain a portfolio that minimizes risk while maximizing profitability.

* 1. **BUSINESS GOAL**

For lenders, minimizing the risk of default is paramount as it directly impacts their financial stability. Defaulting borrowers pose a significant overhead to banks, leading to potential losses. Therefore, the business goal is to minimize default risks. This involves developing predictive models to identify borrowers at risk of mortgage payment defaults and how high of a risk. By identifying potential defaulters, lenders can proactively avoid extending loans to new high-risk individuals. As for the existing customers who have been identified as potential defaulters, the banks can implement proactive measures such as offering financial counseling, restructuring loan terms, or providing assistance programs to help them avoid default. This approach minimizes the likelihood of default, ensuring a more stable and secure mortgage portfolio for the bank in the long run.

* 1. **ANALYTICS GOAL**

***Identifying Defaulters:*** The first analytic goal is to accurately identify borrowers at risk of defaulting on their mortgage payments. It involves using a classification model to categorize the borrowers as either *default* or *non-default* based on their repayment behavior. This will enable lenders to proactively manage their loan portfolios, allocate resources efficiently, and tailor their strategies to mitigate default risks.

***Assessing Risk Levels (Predicting When Customers Might Default):*** To determine the likelihood of customers defaulting on loans, we use a regression model. This model helps us estimate the time period within which a default is likely to occur. We set a threshold at 25%, assuming that if customers default within the first quarter of the loan period, it's riskier for the bank as it might not even recover half of the interest. Therefore, customers falling below the 25% threshold are classified as high risk, while those above it are considered low risk.

***Understanding Default Customers’ Behavior:*** The goal is to use cluster analysis to group borrowers based on similarities in their attributes and behaviors to uncover distinct segments within the dataset. This approach helps identify common traits among borrowers, especially throwing light on the similar characteristics exhibited by the defaulters.

* 1. **ANALYTICS APPROACH**

The Analytics approach involves the following steps:-

1. ***Data Preprocessing***: Cleanse and preprocess the dataset by describing the various features, handling missing values, outliers, and encoding categorical variables as necessary.
2. ***Exploratory Data Analysis:*** Explore relationships, trends, patterns, and correlations among variables in the dataset.
3. ***Feature Selection:*** Identify the most relevant features that contribute significantly to the prediction task while reducing dimensionality and improving model performance. Techniques such as correlation analysis, feature selection techniques, and dimensionality reduction methods like Principal Component Analysis (PCA) may be employed.
4. ***Model Development:*** Develop classification models (e.g., logistic regression, k-NN, Naive Bayes, neural networks) to categorize the borrowers into default and non-default.
5. ***Model Evaluation:*** Assess model performance using appropriate metrics such as accuracy, sensitivity, specificity, and error rates to evaluate the effectiveness of the developed models.
6. ***Model Interpretation:*** This stage involves interpreting trained models to comprehend the influence of various features on loan borrowers, pinpointing the most influential factors that determine their classification into default and non-default categories.
7. ***Deployment and Recommendations:*** Deploy finalized classification models into operational systems or provide recommendations based on insights gained to assist banks and other financial institutions in making informed decisions regarding lending loans.
8. ***Conclusion and Further Work:*** Draw conclusions about findings and discuss further work to improve the model's robustness and accuracy, providing recommendations for future enhancements.
9. **DATA PREPROCESSING**

The mortgage dataset provides insights into the loan payment behaviors of residential property borrowers over a span of 60 periods. Comprising data from 50,000 borrowers, the dataset encompasses a total of 622,489 rows and 23 features. All features are numeric, with six fields indicating binary data, namely REtype\_CO\_orig\_time, REtype\_PU\_orig\_time, REtype\_SF\_orig\_time, investor\_orig\_time, default\_time, and payoff\_time. Additionally, the status\_time field categorizes borrowers into default, payoff, or non-default ongoing statuses. A detailed description of each field is provided in Figure 2.1.

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Value Representation** | **Description** |
| id | number | Borrower ID. Unique identifier for each borrower. |
| time | number | Time stamp of observation, indicating the moment when the observation was made. |
| orig\_time | number | Time stamp for origination, indicating when the loan was originated. |
| first\_time | number | Time stamp for first observation, indicating the first time the borrower's information was observed. |
| mat\_time | number | Time stamp for maturity, indicating when the loan matures or reaches the end of its term. |
| balance\_time | US Dollars($) | Outstanding balance at observation time. The amount of money still owed on the loan at the time of observation. |
| LTV\_time | Percentage(%) | Loan-to-value ratio at observation time, in %. This ratio indicates the percentage of the property's appraised value that is borrowed, while the remaining percentage is covered by the down payment or equity. A higher LTV ratio implies higher risk for the lender because the borrower has less equity in the property. |
| interest\_rate\_time | Percentage(%) | The interest rate on the loan at the time of observation, expressed as a percentage. |
| hpi\_time | number | House price index at observation time, given the index value of 100 for a base year. A higher HPI value than the base index value suggests that house prices have risen, while a lower value suggests a decline. |
| gdp\_time | Percentage(%) | Gross domestic product (GDP) growth at observation time, in %. A higher GDP growth rate suggests a growing economy, potentially indicating favorable conditions for borrowers in terms of employment opportunities and income growth. Conversely, a negative or low GDP growth rate might signal economic challenges, affecting borrowers' ability to repay loans due to factors like unemployment or reduced income. |
| uer\_time | Percentage(%) | Unemployment rate at observation time, in %. |
| REtype\_CO\_orig\_time | Binary(0,1) | Indicates whether the property type at origination time is a condominium (1) or not (0). |
| REtype\_PU\_orig\_time | Binary(0,1) | Indicates whether the property type at origination time is a planned urban development (1) or not (0). |
| REtype\_SF\_orig\_time | Binary(0,1) | Indicates whether the property type at origination time is a single-family home (1) or not (0). |
| investor\_orig\_time | Binary(0,1) | This variable indicates whether the borrower at the time of loan origination is an investor (1) or not (0). |
| balance\_orig\_time | US Dollars($) | This variable represents the amount of money owed on the loan at the time of origination. |
| FICO\_orig\_time | Percentage(%) | FICO score at origination time, in %  The overall FICO score range is between 300 and 850. In general, scores in the 670 to 739 range indicate “good” credit history, and most lenders will consider this score favorable. |
| LTV\_orig\_time | Percentage(%) | Loan-to-value ratio at origination time, in %. |
| Interest\_Rate\_orig\_time | Percentage(%) | Interest rate at origination time, in %. |
| hpi\_orig\_time | number | House price index at origination time, base year = 100. |
| default\_time | Binary(0,1) | Indicates whether the loan has defaulted (1) or not (0) at the observation time. |
| payoff\_time | Binary(0,1) | Indicates whether the loan has been paid off (1) or not (0) at the observation time. |
| status\_time | Ternary(0,1,2) | Represents the current status of the loan, with 1 indicating default, 2 indicating payoff, and 0 indicating non-default/non-payoff, at the observation time. |

**FIGURE 2.1** Attribute Description.

The data analysis was conducted using the R programming language, chosen for its widespread usage in analysis and its vast array of packages and applications. The skim function output, as shown in Figure 2.2, provides a comprehensive summary of the characteristics of the dataset. From the visualization of the missing values in Figure 2.3, only the column LTV\_time contains null values for 270 observations. The null values represent just 0.4% of the dataset, the decision was made to remove these missing values.

Among the 12 columns with zero values, 7 columns contain binary data, where zero values are acceptable. 323 observations have zero values for both *balance\_time* and *LTV\_time*, suggesting that loans associated with these records might have been fully paid off or transferred to another bank, resulting in zero balances. Similarly, approximately 89,805 observations have zero values for *Interest\_Rate\_orig\_time*, possibly indicating scenarios like promotional periods or subsidized loan programs.

A screenshot of a computer

Description automatically generated

**FIGURE 2.2** Skim function output.

A blue and red graph

Description automatically generated

**FIGURE 2.3** Missing values graphical representation.

* 1. **TRANSFORMING DATASET**

The dataset, initially comprising 622,489 rows, underwent a transformation to adopt a customer-centric approach by grouping entries based on unique customer IDs. This restructuring yielded a condensed dataset of 49,981 observations, where each row represents the latest available information for a particular customer. By focusing solely on the latest time period for each customer, historical loan installment data was omitted, streamlining the dataset for enhanced clarity and relevance. The reason behind this approach lies in the primary objective of identifying customers at risk of defaulting on their mortgage payments. Consequently, only customers categorized with a status of 1 or 2, indicating default or payoff, respectively, were segregated for further analysis. The remaining customers, representing current active borrowers, will constitute the holdout dataset on which the selected classification and regression models will be applied to proactively identify individuals susceptible to default and the level of risk associated.

Additionally, few calculated fields have been appended to the dataset, derived from existing columns. These fields are as follows:

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Value Representation** | **Description** |
| expected\_loan\_period | number | Represents the anticipated length of the loan period from origination to maturity(mat\_time –origin\_time). |
| observed\_loan\_period | number | Denotes the observed length of the loan period from origination to the current observation time (time – origin\_time). |
| observed\_loan\_period\_ptg | Percentage(%) | Represents the percentage of the loan period elapsed until the customer defaulted ([observed\_loan\_period / expected\_loan\_period] \* 100). |

* 1. **OUTLIER ANALYSIS**

Box plots were generated for the numerical attributes to detect outliers, as illustrated in Figure 2.2.1. Despite the presence of outliers across most attributes, it was decided to retain them due to their valuable insights into the distribution and behavior of the data. Rather than being anomalous data points, many of these outliers represent legitimate and meaningful observations that contribute to the richness and depth of the dataset. Notably, certain outliers, such as *balance\_orig\_time*, may be high for certain borrowers based on factors such as property value and financial capability or *mat\_time* could be high for borrowers who opt for longer loan closure periods.

These outliers offer valuable information about exceptional cases within the dataset, shedding light on unique features or characteristics of certain loan borrowers. By retaining these outliers, the comprehensive understanding of the dataset's diversity and variability is preserved, which is crucial for accurate analysis.

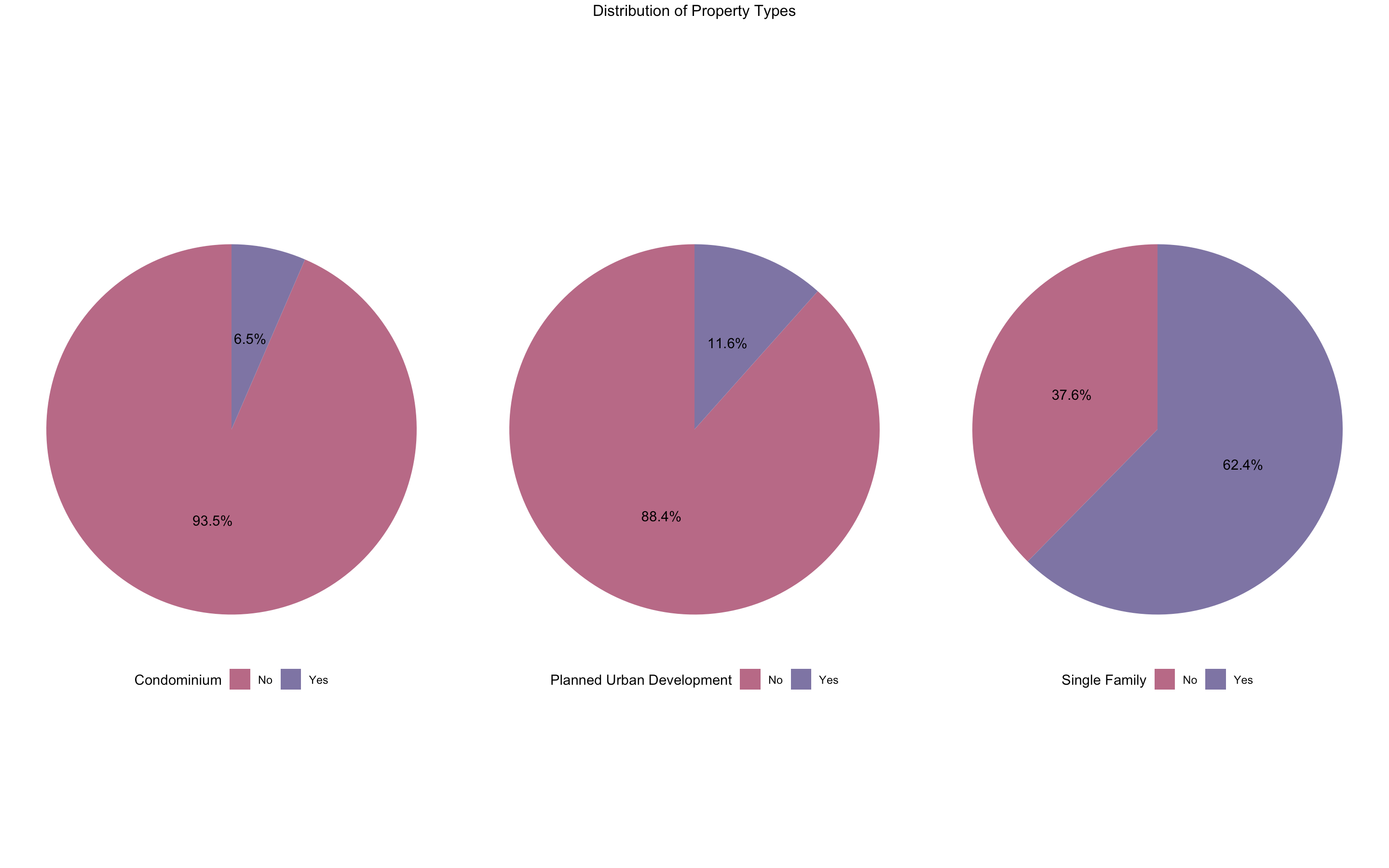
A group of graphs showing different sizes of data

Description automatically generated with medium confidence

**FIGURE 2.2.1** Box plot for numerical variables depicting outliers.

1. **EXPLORATORY DATA ANALYSIS**
   1. **CATEGORICAL VARIABLE ANALYSIS**

***Property Types :*** The chart depicted in Figure 3.1.1 illustrates the distribution of property types for mortgages. It is apparent from the data that the majority of borrowers, constituting approximately 62.4%, opt for Single Family properties. This could be because these homes typically offer more space, privacy, and a sense of ownership compared to other property types. Planned Urban Development emerges as the second most chosen property type, representing approximately 11.6% of borrowers, might be favored due to its community amenities and modern features, attracting those seeking a balance between urban living and suburban comforts. Conversely, Condominiums appear to be the least favored property type among mortgage borrowers, may be perceived as having limitations in space and freedom compared to standalone homes, making them less attractive to potential buyers.



**FIGURE 3.1.1** Distribution of Property types.

***Investor\_orig\_time :***  The bar chart in Figure 3.1.2 illustrates that the majority of borrowers are individual owners, with only a small proportion identified as investors. This indicates that Individuals prioritize owning a home over, indulging in real estate investment. Additionally, the capital requirements and financial responsibilities of real estate investment may act as barriers, dissuading many individuals from pursuing investment opportunities.

A screenshot of a graph

Description automatically generated

**FIGURE 3.1.2** Distribution of Investors.

***Status\_time :***  As depicted in Figure 3.1.3, a significant portion of borrowers successfully pay off their mortgages. However, it's notable that a substantial number of defaulters also exist and the rest are active ongoing customers.

A graph of a number of people

Description automatically generated with medium confidence

**FIGURE 3.1.3** Distribution of borrower status.

* 1. **NUMERICAL VARIABLE ANALYSIS**

***time, orig\_time, first\_time and mat time :*** From Figure 3.2.1, the distribution of the *time* variable shows that loans have a balanced range of durations, with a slight preference for longer repayment periods and very high no of customers with 60 payment periods. For *orig\_time*, most observations fall between 15 and 25 time periods, which could indicate loan transfers from other banks. The negative values may suggest loans originated before the observation period, possibly due to certain financial arrangements or refinancing situations. The distribution of *first\_time* suggests that most of the loans in the given dataset start at time periods between 15 - 30. The distribution of *mat\_time* indicates that most loans mature around the median time period of 142.A screenshot of a graph

Description automatically generated

**FIGURE 3.2.1** Histogram and summary statistics for time, orig\_time, first\_time and mat\_time.

***LTV, HPI, GDP and interest rates :*** Figure 3.2.2 provides insights into the distribution and characteristics of the variables related to loan-to-value ratio (LTV\_time), interest rate (interest\_rate\_time), house price index (hpi\_time), and gross domestic product growth (gdp\_time) at observation time. For LTV\_time on average, borrowers are financing around 79.31% of the property's appraised value through their mortgage loans indicating commonly made down payments are of approximately 15-30%, with the remainder financed.

A screenshot of a graph

Description automatically generated

**FIGURE 3.2.2** Histogram & summary statistics of LTV\_ratio, interest\_rate, hpi\_time, balance\_time.

Interest\_rate\_time exhibits a wide range of values, with a minimum of 0% and a maximum of 37.5%, indicating variability in the interest rates offered at observation time potentially influenced by factors such as creditworthiness, market conditions, and loan terms. The mean of 7.013% suggests that interest rates are typically moderate. The house price index (hpi\_time) demonstrates a range from 107.8 to 226.3. These values reflect changes in single-family house prices over time, with a positive skew towards higher values, indicating overall appreciation in house prices during the observation period. Balance\_time clearly indicates the presence of some outliers of very huge amounts( > 1 Million). However, most of the outstanding balances as of the present time are within $320000. The variable *balance\_time* evidently reveals the existence of significant outliers, with values exceeding $1 million. However, the majority of outstanding balances as of the present time are below $320,000.

A collage of red and pink bars

Description automatically generated

**FIGURE 3.2.3** Histogram & summary statistics of GDP, UER, FICO & LTV at origin time.

***GDP, UER, FICO, and LTV***: The summary statistics in Figure 3.2.3 provide insights into

various variables related to mortgage origination such as GDP, UER, FICO and LTV. GDP growth (gdp\_time) shows a range from -4.147% to 5.132%, with an average of 1.690%. These statistics indicate fluctuations in economic growth rates over time, with a slightly positively skewed distribution. The mean unemployment rate of 6.001 suggests that, on average, customers experienced moderate levels of unemployment during their loan repayment periods. However, the presence of outliers, such as the maximum rate of 10, indicates that some individuals encountered periods of significantly high unemployment. These spikes in unemployment could potentially have adverse effects on their ability to make timely loan payments, underscoring the importance of understanding and managing economic risks associated with mortgage lending.

For FICO\_orig\_time, the scores range from 400 to 840, with a mean of 661.3. This suggests a relatively good credit profile among borrowers, with most falling within the good credit range. The LTV with a mean of 79.71% indicates that, on average, borrowers obtained loans covering 79.71% of the property value at the time of origination.

A screenshot of a graph

Description automatically generated

**FIGURE 3.2.4** Histogram & summary statistics of interest rate, HPI, Balance at origin time and expected loan period.

***Interest rate, HPI, Balance at origin time and expected loan period:*** The summary statistics of interest\_rate indicate a wide range, with the majority of borrowers securing rates between 3.000% and 7.625%. The hpi\_orig\_time, representing the house price index at origination, ranges from 75.71 to 226.29. The median hpi\_orig\_time is 212.73, with a mean of 196.39, indicating a general upward trend in house prices over the base year. The balance at the origin time shows that the maximum amount the loan was obtained for is about $8 Million while the lowest being a few thousands indicating wide diversity of customers. Most of the loan period has a maturity of 120 periods which is about 10 years.

A graph of red and white bars

Description automatically generated with medium confidence

**FIGURE 3.2.5** Histogram & summary statistics of current observed loan periods.  
Based on Figure 3.2.5, the majority of current customers have progressed to only about 20% of their loan tenure, equivalent to roughly 24 months or 2 years out of the average 10-year loan duration. At this stage, some customers may be defaulting, while others are either making regular payments or paying off their loans entirely. Few customers are nearing loan maturity, indicating that most are still in the early stages of their repayment journey.

1. **PREDICTOR ANALYSIS AND RELEVANCY**

To comprehend the relationship between different variables, a correlation analysis was conducted. The correlation matrix illustrated both positive and negative correlations among the variables as shown in Figure 4.1. From the correlation matrix, several notable positive and negative correlations among the variables can be observed.

*Positive correlations:*

1. **orig\_time and hpi\_orig\_time:** The strong positive correlation (0.85) suggests that the original time of the loan is highly correlated with the Housing Price Index. It implies that loans originated at certain times tend to coincide with specific trends in the housing market.
2. **LTV\_time and hpi\_orig\_time:** The Loan-to-Value ratio is positively associated with changes in the Housing Price Index (0.64). It suggests that as housing prices increase, borrowers might be more likely to have higher LTV ratios.
3. **LTV\_time and mat\_time:** A positive correlation of 0.50 suggests that there is a moderate positive relationship between LTV ratio and the maturity time of the loan. As the maturity time of the loan increases, there tends to be a tendency for the LTV ratio to increase as well. Borrowers might be willing to accept higher LTV ratios for longer-term loans, as they may perceive less risk in repaying the loan over an extended period.

*Negative correlations:*

1. **hpi\_time** **and uer\_time:** The strong negative correlation (-0.78) suggests that the Housing Price Index is inversely related to the unemployment rate. It implies that during periods of economic downturns (high unemployment rates), housing prices tend to decline.
2. **FICO\_orig\_time and interest\_rate\_time:** The negative correlation (-0.43) suggests that as the FICO score at loan origination increases (indicating better creditworthiness), there is a slight tendency for the interest rate at a certain time to decrease, and vice versa. Lenders often use FICO scores as one of the factors to determine the interest rate for a loan. Higher credit scores may lead to lower perceived risk for the lender, resulting in a lower interest rate.
3. **LTV\_time and gdp\_time:** A negative correlation of -0.45 between LTV\_time and gdp\_time suggests a moderate inverse relationship between the Loan-to-Value ratio and Gross Domestic Product over time. During periods of economic growth (high GDP), lenders may be more willing to offer loans with lower LTV ratios, reflecting a healthier economic environment and increased borrower creditworthiness. Conversely, during economic downturns (low GDP), lenders may tighten lending standards and require higher LTV ratios to mitigate risk, leading to a higher LTV ratio.

A graph with numbers and text

Description automatically generated

**FIGURE 4.1** Simplified Correlation plot with numbers.

A screenshot of a computer

Description automatically generated

**FIGURE 4.2** Correlation plot for the output variable *default*\_time

The outcome variable of interest is *default\_time* which indicates if a customer has defaulted or not. From Figure 4.2, it can be observed that LTV ratio, hpi, gdp and unemployment rates play an important role in determining the default status of the loan, by exhibiting a high correlation with the *default\_time* variable.

* 1. **VARIABLE SELECTION**

Variable selection was performed using Boruta. According to Figure 4.11, Boruta identified 21 variables as significant. The box-and-whisker plots depict the distribution of importance scores assigned to each of the 24 features (excluding id) by Boruta. The length of the box represents the interquartile range (IQR), while the whiskers extend to the minimum and maximum values within 1.5 times the IQR from the lower and upper quartiles, respectively. Outliers beyond this range are indicated by individual points. The plots provide an overview of the variability and relative importance of each feature according to Boruta analysis, aiding in identifying potential significant predictors for further investigation. Since, the primary goal is to classify borrowers using classifiers, variable selection is not of high importance at this stage. The classifiers themselves will internally select the most relevant variables for the classification task, making explicit variable selection unnecessary.

A graph with colorful dots and text

Description automatically generated with medium confidence

**FIGURE 4.11** Boruta Box-and-whisker Plots across all 24 Features.

|  |  |
| --- | --- |
| ***Variable*** | ***Boruta*** |
| time | \* |
| orig\_time | \* |
| first\_time | \* |
| mat\_time | \* |
| balance\_time | \* |
| LTV\_time | \* |
| Interest\_rate\_time | \* |
| hpi\_time | \* |
| gdp\_time | \* |
| uer\_time | \* |
| Retype\_CO\_orig\_time | \* |
| Retype\_PU\_orig\_time |  |
| Retype\_SF\_orig\_time | \* |
| investor\_orig\_time |  |
| balance\_orig\_time | \* |
| FICO\_orig\_time | \* |
| LTV\_orig\_time | \* |
| Interest\_Rate\_orig\_time | \* |
| hpi\_orig\_time | \* |
| payoff\_time | \* |
| status\_time | \* |
| expected\_loan\_period | \* |
| observed\_loan\_period | \* |

**FIGURE 4.12** Boruta feature selection.

1. **DIMENSION REDUCTION**

Although Principal Component Analysis (PCA) is useful for dimension reduction, it may not be the most appropriate dimensionality reduction technique for the mortgage lending dataset. The dataset primarily consists of numeric features related to loan attributes, economic indicators, and borrower characteristics, which are already meaningful on their own and may not benefit significantly from the transformation offered by PCA. Moreover, PCA creates new variables that may lack interpretability in the context of mortgage lending, where understanding the factors influencing default and repayment behavior is crucial for informed decision-making. Instead, techniques like correlation analysis or classification trees might be more appropriate. Therefore, a deliberate decision has been made not to employ dimension reduction, instead variable selection techniques were employed.

1. **DATA PARTITIONING**

To facilitate model training and evaluation, the dataset was partitioned into two subsets: training and validation. This partitioning strategy, comprising 70% and 30% of the data, respectively, enables robust model development and assessment. The training set serves to train the models, the validation set aids in tuning model parameters, and the holdout set which in this case is the active customers dataset will be used for final model execution and prediction without introducing bias. For Regression, only the default customers were extracted from the dataset and partitioned in similar fashion.

1. **CLASSIFICATION MODEL SELECTION**

Classification models play a pivotal role in the mortgage lending business case by accurately categorizing borrowers as prospective defaulters or non-defaulters. By analyzing historical borrower data and repayment behaviors, these models provide insights into the likelihood of a borrower defaulting on their mortgage payments. This allows lenders to implement preemptive strategies to minimize default risks and optimize their loan portfolios. Additionally, classification models aid in assessing new loan applicants, helping lenders make informed decisions about whether to approve or deny loan requests based on the applicant's risk profile. Therefore, classification models serve as essential tools for lenders to maintain the stability and profitability of their mortgage portfolios while minimizing the impact of defaults.

The variable chosen as the target variable for all classifiers is *default\_time*. Additionally, *id*, which does not contribute meaningful information to the analysis, is excluded. Similarly, *payoff\_time* and *status\_time*, which are redundant for the classification task, are also removed from the dataset. Three classification models have been chosen: Classification Tree, Logistic Regression, and Naive Bayes.

* 1. **CLASSIFICATION TREE**

The decision to employ the Classification Tree model was driven by its simplicity and the interpretability of its results, which are presented as decision trees. These decision trees provide clear insights into the factors influencing the classification outcome, facilitating a deeper understanding of the classification process. As illustrated in Figure 7.11, the model identifies *LTV\_time*, *time, and FICO\_orig\_time* as the most significant predictors in distinguishing between defaulters and non-defaulters.

A screenshot of a computer

Description automatically generated

**FIGURE 7.11** Classification Tree

In terms of performance, the model achieved an accuracy of **0.7742**, indicating the proportion of correctly classified instances. Additionally, the sensitivity, which measures the model's ability to correctly identify defaulters, was found to be **0.6159**. On the other hand, the specificity, representing the model's capability to accurately classify non-defaulters, was notably higher at **0.8622**. This suggests that the model exhibits better performance in classifying non-defaulters compared to defaulters. Overall, the Classification Tree model offers a transparent and intuitive approach to borrower classification, shedding light on the key predictors driving the classification outcome. Despite its simplicity, the model demonstrates promising performance metric.

* 1. **LOGISTIC REGRESSION**

Logistic Regression is chosen because it is a powerful and widely used method for binary classification tasks. It provides interpretable coefficients that indicate the influence of each predictor variable on the outcome. Upon running the logistic regression model on the training data, it was observed that most predictors were statistically significant for classification. The logistic regression model demonstrated almost similar accuracy (**0.7747),** sensitivity(**0.6159**) and specificity (**0.8629**) compared to the Classification Tree model.

* 1. **NAÏVE BAYES**

Naïve Bayes was chosen as the third model to explore its potential for improving sensitivity in identifying defaulters. Despite its simplistic assumption of feature independence, Naïve Bayes has demonstrated effectiveness in various practical scenarios. Upon conducting classification, the model achieved an accuracy of **0.7586**, indicating its overall performance in correctly classifying instances. Notably, Naïve Bayes exhibited the highest sensitivity of **0.7089** in correctly identifying defaulters, showcasing its strength in this aspect. Additionally, the model maintained a specificity of **0.7863**, ensuring a balanced trade-off between sensitivity and specificity. This performance suggests that Naïve Bayes may offer advantages in accurately detecting defaulters while maintaining a reasonable level of specificity compared to the other models.

1. **REGRESSION MODEL SELECTION**

In the context of assessing risk levels, regression models prove invaluable in predicting when customers might default on loans. This dataset aims to achieve this goal by estimating the time period within which a default is likely to occur for each customer. The target variable chosen is *observed\_loan\_period\_ptg* to understand the percentage period of the loan duration at which customers tend to default . Thus, *observed\_loan\_period\_ptg* serves as a crucial metric in evaluating risk, offering insights into how far into the loan term defaults typically occur, aiding in risk classification and mitigation strategies. Three regression models have been chosen: Multiple Linear Regression, Regression Tree, and Neural Networks.

* 1. **MULTIPLE LINEAR REGRESSION:**

Multiple Linear Regression is the first regression model utilized in this analysis to predict default probabilities for loans. This model aims to capture the relationship between multiple predictor variables and the target variable, *observed\_loan\_period\_ptg*, which represents the percentage of time taken for a customer to default on their loan. The model's performance is evaluated based on several metrics. Notably, the adjusted R-squared value of approximately 0.9508 indicates that around 95.08% of the variability in the target variable is explained by the predictor variables included in the model. Additionally, the F-statistic is substantial (1.078e+04), with a very low p-value (< 2.2e-16), indicating that the overall model is statistically significant.

Interpreting the histogram of residuals from Figure 8.11, we expect a central peak around zero, reflecting the symmetric distribution of prediction errors. However, the presence of outliers, particularly the maximum value of 54.86349, may result in a long right tail in the histogram, indicating instances where the model substantially overestimated default probabilities. Conversely, bins with high frequencies are likely centered around zero, representing residuals closer to the mean prediction.

A graph of a bar graph

Description automatically generated

**FIGURE 8.11** Residual Distribution

Furthermore, assessing predictive accuracy, the correlation coefficient between the predicted and actual default payoff percentages is approximately 0.976, indicating a strong positive linear relationship. The mean absolute error (MAE) is approximately **0.5688**, suggesting that, on average, the model's predictions deviate from the actual values by about 0.5688 percentage points. Additionally, the root mean squared error (RMSE) is approximately **1.7292**, which provides another measure of the model's prediction accuracy. Overall, these metrics indicate that the Multiple Linear Regression model provides reasonably accurate predictions of default probabilities, offering valuable insights for risk assessment in loan management.

* 1. **REGRESSION TREES:**

Regression Tree is the next regression model utilized in this analysis for predicting default probabilities in loan management. The regression tree also shows the significant predictors at the top of the tree, which are *time, mat\_time, hpi\_orig\_time, uer\_time,* and *orig\_time* as seen in Figure 8.21. Additionally, this model predicted with an MAE of **2.2422** suggesting that, on average, the model's predictions deviate from observed spending by approximately 2.24 units from the normalized used price and an RMSE of **3.3602**.

A diagram of a company structure

Description automatically generated**FIGURE 8.21** Regression Tree

* 1. **NEURAL NETWORK:**

The next regression model employed in this analysis is a neural network. Neural networks are powerful machine learning algorithms capable of capturing complex relationships between input variables and output targets. In this particular implementation, a neural network model with four hidden layers as illustrated in Figure 8.31 is trained using the training data, aiming to predict default payoff percentages based on various predictor variables.

A diagram of a network

Description automatically generated**FIGURE 8.31** Neural Network Model Topology.

After training the model, its predictive accuracy was evaluated using MAE, resulting in a value of **5.7057**. This indicates the average absolute difference between the predicted and actual values of the target variable.

1. **CLUSTER ANALYSIS**

Cluster analysis is employed to gain insights into the behavior of borrowers and uncover any significant characteristics within the dataset. By grouping borrowers based on similarities in their attributes and behaviors, cluster analysis facilitates the identification of distinct borrower segments or clusters. This approach helps to unveil patterns and trends that may not be apparent through traditional analysis methods. For instance, clusters may reveal common traits among borrowers who default on their mortgage payments, such as high loan-to-value ratios or specific property types. Additionally, cluster analysis enables the identification of borrower segments with different risk profiles, allowing lenders to tailor their strategies and interventions accordingly. Overall, cluster analysis offers a holistic perspective on customer behavior, providing valuable insights for risk assessment and decision-making in mortgage lending.

***k-means:***

Utilizing the k-means algorithm, the dataset was segmented into three clusters, as depicted in Figure 8.1.

A graph with different colored lines

Description automatically generated with medium confidence

**FIGURE 8.1** Bar plot of k-means cluster analysis for 3 clusters.

***Cluster 1: Active Customers***

Cluster 1 encompasses active customers who are in the process of repaying their loans. These individuals exhibit a commendable track record of loan management, evident from their low current balances and conservative LTV ratios. They maintain stability in their repayment behavior, benefiting from a favorable GDP environment and low unemployment rates. Their preference for properties with lower LTV ratios suggests a cautious approach to property acquisition. Overall, this cluster represents a group of financially prudent individuals who prioritize loan repayment.

***Cluster 2: Pay-off Customers***

Cluster 2 represents a segment characterized by a customers known for loan repayment. These customers exhibit a pattern of high payoff rates, which is reflective of their responsible borrowing behavior. Despite moderately high loan amounts, their current balances remain comparatively low, suggesting efficient management of their financial obligations. Notably, they maintain lower LTV ratios and are subject to higher interest rates, indicating a deliberate effort to minimize risk exposure. Furthermore, this cluster demonstrates a preference for single-family properties, a choice often associated with stability and long-term residency. Absence of investor involvement and relatively shorter loan periods further emphasize their commitment to timely repayment.

***Cluster 3: Default Customers***

Cluster 3 comprises customers who exhibit characteristics indicative of high-risk borrowers. These individuals have a propensity for defaulting, often associated with high loan amounts and pending balances. Their loan-to-value (LTV) ratios are notably elevated, suggesting substantial financial exposure. Concurrently, they are influenced by economic factors such as housing price index (HPI) and gross domestic product (GDP), albeit in a manner that does not deter their high default tendencies. Moreover, the cluster is marked by a prevalence of properties classified as condominiums (CO) and planned urban developments (PU), underscoring potential investment or speculative motivations.

Further, the cluster analysis highlights a clear relationship between property type, economic conditions, and borrower behavior. Those opting for condominiums, often single individuals seeking risk, demonstrate a higher propensity for default compared to those preferring single-family homes, typically families seeking stability. Moreover, economic conditions exert a significant influence on repayment ability. Cluster 3, reflecting harsh economic environments, exhibits increased default rates, while Cluster 1 and 2, indicative of flourishing economies, showcases more stable borrowers, with early loan payoffs or consistent installment payments.

***Default Customer Behavior*  
Large Loan Amounts**: Defaulters often borrow significant sums, suggesting they might be overextending themselves financially or seeking high-value properties.

**High Loan-to-Value (LTV) Ratio**: The high LTV ratio implies that defaulters are investing in properties with a substantial portion financed by debt, amplifying the risk in case of property devaluation or market downturns.

**Sensitivity to Economic Indicators**: Economic downturns marked by poor GDP performance, high unemployment rates, and low housing price indices (HPI) serve as warning signs for default risk. This indicates that defaulters may struggle to meet mortgage obligations during economic hardships.

**Preference for Condominiums or Planned Urban Developments**: Defaulters often gravitate towards these property types, possibly due to their attractiveness for investment or lifestyle preferences. The communal nature of these properties may also pose unique financial challenges in times of economic stress.

**Investment Orientation**: Some defaulters may view property ownership as an investment opportunity rather than solely for residential purposes. This investment mindset may lead them to take on more risk or leverage than typical homeowners.

**Relatively Good FICO Scores**: Despite their default status, defaulters often possess decent FICO scores, indicating that traditional credit scoring metrics alone may not accurately predict mortgage repayment capability. Other factors such as income stability and debt-to-income ratios might contribute to their creditworthiness assessment.

1. **MODEL EVALUATION**

From Figure 10.1, it is clear that among the three classification models evaluated, Naive Bayes is selected as the preferred model due to its relatively good sensitivity and balanced trade-off between sensitivity and specificity. While all models demonstrate similar overall accuracy levels, Naive Bayes exhibits the highest sensitivity, indicating its capability to correctly identify positive instances (in this context, likely loan defaults). Additionally, its specificity, although slightly lower than that of the other models, remains reasonably high. This suggests that Naive Bayes strikes a favorable balance between correctly identifying both positive and negative instances, making it a suitable choice for the classification task at hand.

A table with numbers and text

Description automatically generated**FIGURE 10.1** Classification Model Evaluation Metrics.

Figure 10.2 illustrates a comparison of three regression models, revealing that Multiple Linear Regression exhibits the lowest Mean Absolute Error (MAE) and RMSE values. Following closely is Regression Trees. However, the Neural Network is eliminated as it shows relatively higher MAE and RMSE. Consequently, ***Multiple Linear Regression*** is deemed the optimal regression model, as it not only yields a favorable MAE but also explains a linear relationship between predictors and the target variable. A table with numbers and text

Description automatically generated**FIGURE 10.2** Regression Model Evaluation Metrics.

1. **PREDICTION**

The naive Bayes classifier is applied to the active customer dataset, treated as the holdout set, to classify customers as defaulters and non-defaulters. The model identifies approximately **39.12%** of customers as defaulters as shown in Figure 11.1, indicating the proportion of customers who pose a risk of defaulting.

A graph of a graph

Description automatically generated with medium confidence

Subsequently, the Multiple Linear Regression model is applied specifically to these identified default customers to estimate the default time period. Figure 11.2 illustrates the distribution of default time periods for these customers. Notably, with a threshold set at 25%, no customers are identified as posing a severe default risk. However, it's worth noting that a significant portion (93.43%) of defaulters tend to default within the 25% to 35% range of the loan period.

A graph of a number of people

Description automatically generated with medium confidence**FIGURE 11.2** Distribution of Default time periods.

1. **CONCLUSION**

This study aimed to elevate mortgage lending practices by employing advanced analytics to predict and manage default risks effectively. Through rigorous model evaluations and analysis techniques, the objectives were to gain deep insights into borrower behavior, improve risk assessment accuracy, and develop proactive strategies for mitigating default risks.

Beginning with classification model evaluations, Naive Bayes emerged as the preferred choice due to its superior sensitivity and balanced trade-off between sensitivity and specificity. This model facilitated accurate identification of potential loan defaulters, enabling proactive risk mitigation strategies. In parallel, regression model evaluations highlighted Multiple Linear Regression as the optimal choice for predicting the loan periods for default, demonstrating the lowest Mean Absolute Error and Root Mean Squared Error values.

Application of the Naive Bayes classifier to the active customer dataset allowed for the classification of customers as defaulters and non-defaulters, with approximately 39.12% identified as defaulters. Subsequent use of Multiple Linear Regression specifically on these identified default customers facilitated the estimation of default time periods. Notably, the analysis revealed that a significant portion of defaulters tend to default within specific time ranges of 25 – 35% of the total loan duration, offering actionable insights for risk management and intervention planning.

Moreover, cluster analysis uncovered meaningful patterns and behaviors among borrowers, facilitating the identification of distinct borrower segments with varying risk profiles. By grouping borrowers based on similarities in attributes and behaviors, cluster analysis provided holistic insights into customer behavior, enabling tailored risk assessment and decision-making strategies.

In summary, the integration of advanced analytics techniques has translated the analytical goals into actionable insights for enhancing mortgage lending practices. By leveraging models and analysis methods, a deeper understanding of borrower behavior, improved risk assessment accuracy was gained. These findings serve as invaluable tools for lenders in optimizing lending strategies, minimizing default risks, and fostering a more resilient and sustainable mortgage lending ecosystem.

1. **RECOMMENDATIONS**
2. ***Avoidance:*** Avoid high-risk new customers by implementing stringent screening measures. During economic downturns, refrain from onboarding even low-risk customers to mitigate potential defaults.
3. ***Reduction:*** Implement early warning systems to identify customers nearing the critical default risk period (25-35% of loan tenure). Intervene proactively through policy revisions, insurance guidance, or interest rate adjustments to reduce default risks for ongoing customers.
4. ***Transference:*** Transfer high-risk mortgages to other financial institutions to liquidate, reallocating funds to lend to potential pay-off customers.
5. ***Use factors other than FICO:*** to assess the capability of loan repayments, by assessing their income stability, debt-to-income ratios, by requiring collaterals or using insurance to secure the loan.