Project-3 Mortgage Payback Analytics

Table of Contents

1 Introduction

In the world of mortgages, where people borrow money to buy homes, understanding how borrowers behave and predicting what will happen with their loans is crucial for banks and lenders. This project is all about digging into a big pile of data about mortgages in the United States to learn as much as we can. Using Machine learning algorithms to find out things like who's borrowing, when they're borrowing, and what happens to their loans over time.

By looking closely at all this data, we hope to discover some useful information that can help banks make better decisions about lending money for mortgages. Our goal is to uncover secrets hidden in the data that can help lenders avoid risks and make smarter choices, ultimately making the mortgage process better for everyone involved.

2 *Business Goal*

The primary business goal of this project is to enhance the decision-making process for banks and lending institutions in managing mortgage portfolios. By leveraging data analytics, the aim is to reduce risks associated with mortgage lending, improve loan performance, and optimize profitability. Ultimately, the goal is to facilitate informed decision-making that leads to more efficient and effective mortgage lending practices, ensuring financial stability and customer satisfaction.

3 Analytical Goal

3.1 **Goal 1 Classification of Active Borrowers**

Objective

To predict whether active borrowers will either pay off their mortgage or default.

Approach

Utilize historical mortgage data of active borrowers, including borrower and loan characteristics, to develop machine learning models capable of classifying borrowers into two categories paid off or default.

Importance

This goal helps in assessing the current risk level of existing loans in the portfolio and aids in managing delinquencies effectively.

3.2 Goal 2 Classification of New Customers

Objective

To forecast whether new customers applying for a mortgage will pay off their loans or default.

Approach

Utilize applicant data, including borrower characteristics and loan details, to develop predictive models that classify new customers into potential outcomes loan paid or default.

Importance

This goal assists lenders in evaluating credit risk associated with new borrowers and making informed decisions during the loan approval process, thereby optimizing loan portfolio performance, and minimizing potential losses.

4 *Data Preprocessing*

4.1 Dataset Description

The dataset "Mortgage" comprises 622,489 observations across 23 variables, representing origination and performance data for residential U.S. mortgage borrowers. Each observation includes unique identifiers such as borrower ID ("id") and timestamps for different stages of the loan lifecycle, including origination, maturity, and observation periods ("time," "orig_time," "mat_time"). Key features include loan balance, loan-to-value ratio ("LTV"), interest rate, house price index ("hpi"), gross domestic product ("gdp"), and unemployment rate ("uer").

4.2 Attributes Definition

1. id Borrower ID

The "id" variable in our dataset serves as the Borrower ID, uniquely identifying each borrower included in the dataset. This identifier enables us to track the performance and characteristics of individual borrowers in relation to their mortgages. Each entry in the dataset corresponds to a borrower who has already obtained a loan, and their associated data includes information such as loan amount, interest rate, and other relevant attributes pertaining to their mortgage. By analyzing the data using this identifier, we can gain insights into the behavior, performance, and specific attributes of each borrower within the context of their mortgage agreements.

2. time Time stamp of observation

The "time" variable in our dataset serves as a timestamp for each observation, indicating the specific point in time when the data was recorded. With this information, we can track changes in borrower characteristics and loan performance over time, allowing us to analyze trends and patterns that may emerge across different periods. In the context of our dataset, where observations span across 60 periods, the "time" variable enables us to analyze the dynamics of borrower behavior and mortgage performance over the entire duration of the dataset.

3. orig_time Time stamp for origination

The "orig_time" variable denotes the timestamp when the mortgage loan originated or initiated. It serves as a critical starting point for assessing both loan performance and borrower behavior throughout the loan lifecycle. Notably, observations span from -40 to 60, indicating that loans were initiated at different time points. The presence of negative values suggests that some borrowers obtained their loans forty periods before the observed timeframe.

4. first_time Time stamp for first observation

The "first_time" variable represents the timestamp of the borrower's first observation in the dataset. It signifies the initial recording of the borrower's information, aiding in understanding the time elapsed between loan origination and the first recorded observation.

5. mat_time Time stamp for maturity

The "mat_time" variable denotes the timestamp representing the maturity date of the mortgage loan. This signifies the end of the loan term or the anticipated date by which the borrower is expected to fully repay the loan amount.

6. balance_time Outstanding balance at observation time

The "balance_time" variable signifies the remaining principal balance owed by the borrower at the time of observation. It offers crucial insights into the borrower's repayment behavior and current loan status.

7. LTV_time Loan-to-value ratio at observation time, in %

The "LTV_time" variable represents the loan-to-value (LTV) ratio at the observation time, expressed as a percentage. This ratio compares the loan amount to the appraised value of the property, serving as an indicator of risk associated with the loan and reflecting the borrower's equity in the property.

8. interest_rate_time Interest rate at observation time, in %

The "interest_rate_time" variable signifies the annual interest rate charged on the mortgage loan at the time of observation, expressed as a percentage. This rate directly impacts the borrower's monthly mortgage payments and overall affordability.

9. hpi_time House price index at observation time

Measures the change in residential house prices relative to a base year. It provides insights into the housing market trends and influences the property's value and collateral.

10. gdp_time Gross domestic product (GDP) growth at observation time, in %

Indicates the rate of economic growth as measured by the change in GDP over time. It reflects the overall economic conditions and impacts borrowers' income and employment stability.

11. uer_time Unemployment rate at observation time, in %

Represents the percentage of unemployed individuals in the labor force at the time of observation. It reflects the health of the labor market and affects borrowers' ability to repay their mortgages.

12. REtype_CO_orig_time Real estate type condominium at origination time (Binary Variable)

Indicates whether the property type is a condominium at the time of loan origination. It categorizes the type of residential real estate securing the mortgage loan.

13. REtype_PU_orig_time Real estate type planned urban development at origination time (Binary Variable)

Indicates whether the property type is a planned urban development at the time of loan origination. It categorizes the type of residential real estate securing the mortgage loan.

14. REtype_SF_orig_time Single-family home at origination time

Indicates whether the property type is a single-family home at the time of loan origination. It categorizes the type of residential real estate securing the mortgage loan.

15. investor_orig_time Investor borrower at origination time

Indicates whether the borrower is an investor (as opposed to an owner-occupant) at the time of loan origination. It distinguishes between borrowers who purchase properties for investment purposes.

16. balance_orig_time Outstanding balance at origination time

Represents the initial principal balance of the mortgage loan at the time of origination. It serves as the starting point for tracking changes in the loan balance over time.

17. FICO_orig_time FICO score at origination time, in %

Represents the borrower's creditworthiness based on their FICO credit score at the time of loan origination. It assesses the borrower's likelihood of default and influences loan approval and interest rates.

18. LTV_orig_time Loan-to-value ratio at origination time, in %

Calculates the ratio of the loan amount to the appraised value of the property at the time of origination. It assesses the initial risk associated with the loan and reflects the borrower's equity in the property at the start of the loan term.

19. Interest_Rate_orig_time Interest rate at origination time, in %

Represents the annual interest rate charged on the mortgage loan at the time of origination. It influences the borrower's initial mortgage payments and affordability.

20. hpi_orig_time House price index at origination time

Measures the residential house prices relative to a base year at the time of loan origination. It reflects the property's initial value and market conditions at the start of the loan term.

21. default_time Default observation at observation time

Indicates whether the mortgage loan defaulted (1) or remained current (0) at the time of observation. It serves as the target variable for default prediction models and indicates the borrower's payment delinquency status.

22. payoff_time Payoff observation at observation time

Indicates whether the mortgage loan was paid off (1) or remained outstanding (0) at the time of observation. It serves as the target variable for payoff prediction models and reflects the borrower's repayment behavior.

23. status_time Default, payoff, and nondefault/nonpayoff observation at observation time (Categorical variables represented as integer)

The target variable in this dataset is the status_time variable. It represents the current status of the mortgage loan at the time of observation and categorizes observations into three distinct categories

1. Default (1) Indicates that the mortgage loan has defaulted at the time of observation, meaning the borrower has failed to meet their repayment obligations, resulting in delinquency or foreclosure.

2. Payoff (2) Indicates that the mortgage loan has been paid off in full at the time of observation, meaning the borrower has successfully repaid the entire outstanding balance of the loan.

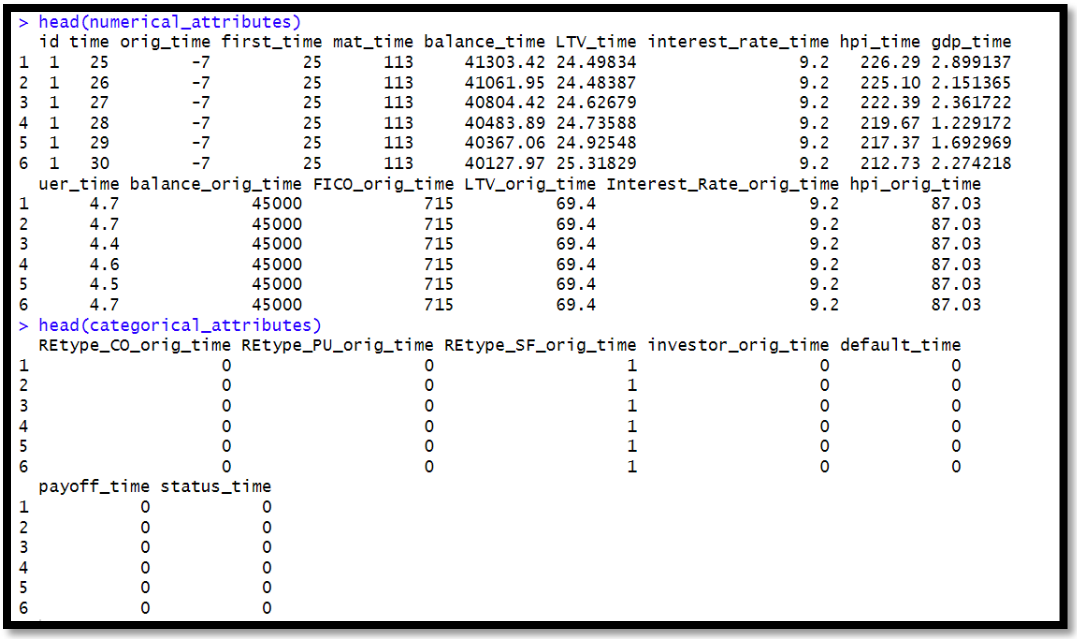
3. Nondefault/Nonpayoff (0) Indicates that the mortgage loan is neither in default nor fully paid off at the time of observation. This category includes observations where the loan is current, partially paid, or exhibits other non-default/non-payoff statuses.

The status_time variable serves as the target for predictive modeling tasks, such as default prediction or payoff prediction. The goal is to develop models that can accurately predict whether a mortgage loan will default, pay off, or remain in a non-default/non-payoff status based on various borrower and loan characteristics observed over time.

Sample Data Consists of Numerical and Categorical data.

Figure 1 illustrates a dataset containing a combination of numerical and categorical attributes, offering a comprehensive insight into mortgage transactions. Numerical attributes such as \`time\`, \`orig_time\`, \`first_time\`, \`mat_time\`, \`balance_time\`, \`LTV_time\`, \`interest_rate_time\`, \`hpi_time\`, \`gdp_time\`, \`uer_time\`, \`balance_orig_time\`, \`FICO_orig_time\`, \`LTV_orig_time\`, \`Interest_Rate_orig_time\`, and \`hpi_orig_time\` likely capture various quantitative aspects of the transactions, including timing, loan amounts, loan-to-value ratios, interest rates, and economic indicators.

On the other hand, categorical attributes like \`REtype_CO_orig_time\`, \`REtype_PU_orig_time\`, \`REtype_SF_orig_time\`, \`investor_orig_time\`, \`default_time\`, \`payoff_time\`, and \`status_time\` provide categorical information such as property type, investor type, default status, and transaction status. Together, this dataset offers a rich source of information for analyzing and modeling mortgage-related phenomena.

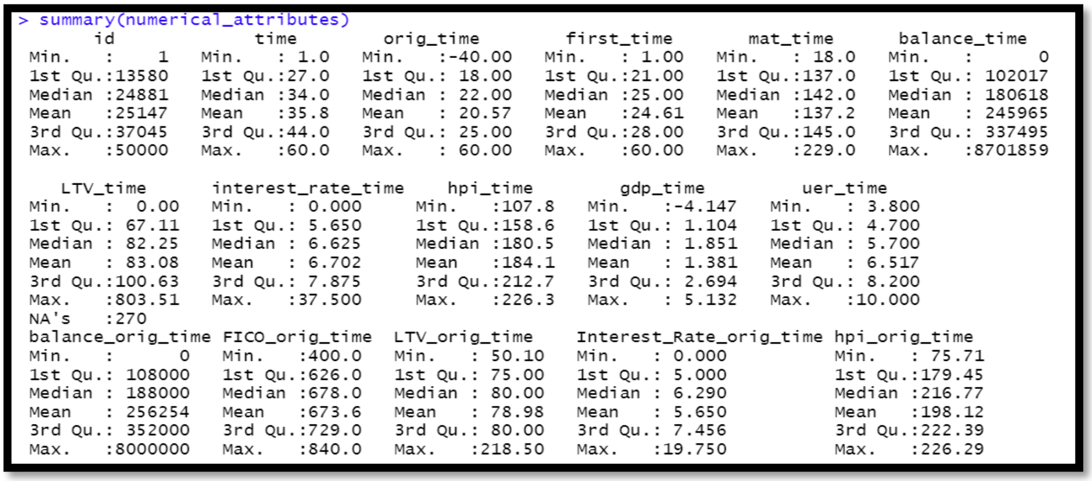


```
> head(numerical_attributes)
  id time orig_time first_time mat_time balance_time LTV_time interest_rate_time hpi_time gdp_time
1  1 25          -7         25      113     41303.42 24.49834                9.2   226.29 2.899137
2  1 26          -7         25      113     41061.95 24.48387                9.2   225.10 2.151365
3  1 27          -7         25      113     40804.42 24.62679                9.2   222.39 2.361722
4  1 28          -7         25      113     40483.89 24.73588                9.2   219.67 1.229172
5  1 29          -7         25      113     40367.06 24.92548                9.2   217.37 1.692969
6  1 30          -7         25      113     40127.97 25.31829                9.2   212.73 2.274218
  uer_time balance_orig_time FICO_orig_time LTV_orig_time Interest_Rate_orig_time hpi_orig_time
1      4.7             45000            715          69.4                     9.2         87.03
2      4.7             45000            715          69.4                     9.2         87.03
3      4.4             45000            715          69.4                     9.2         87.03
4      4.6             45000            715          69.4                     9.2         87.03
5      4.5             45000            715          69.4                     9.2         87.03
6      4.7             45000            715          69.4                     9.2         87.03
> head(categorical_attributes)
  REtype_CO_orig_time REtype_PU_orig_time REtype_SF_orig_time investor_orig_time default_time
1                   0                   0                   1                  0            0
2                   0                   0                   1                  0            0
3                   0                   0                   1                  0            0
4                   0                   0                   1                  0            0
5                   0                   0                   1                  0            0
6                   0                   0                   1                  0            0
  payoff_time status_time
1           0           0
2           0           0
3           0           0
4           0           0
5           0           0
6           0           0
```

Figure 1 Sample Data

4.3 Summary Statistics

The summary statistics reveal a wide range of values across the dataset's numerical attributes. Observation periods range from 1 to 60, with median origination and first observation times around 20 and 24, respectively. Balance amounts at observation time vary significantly, with a median of 180,618. Loan-to-value ratios cluster around 82.24%, while interest rates display a median of 6.625%. Economic indicators such as HPI and GDP show typical fluctuations, and unemployment rates range from 3.8% to 10%, with a median of 5.7%.



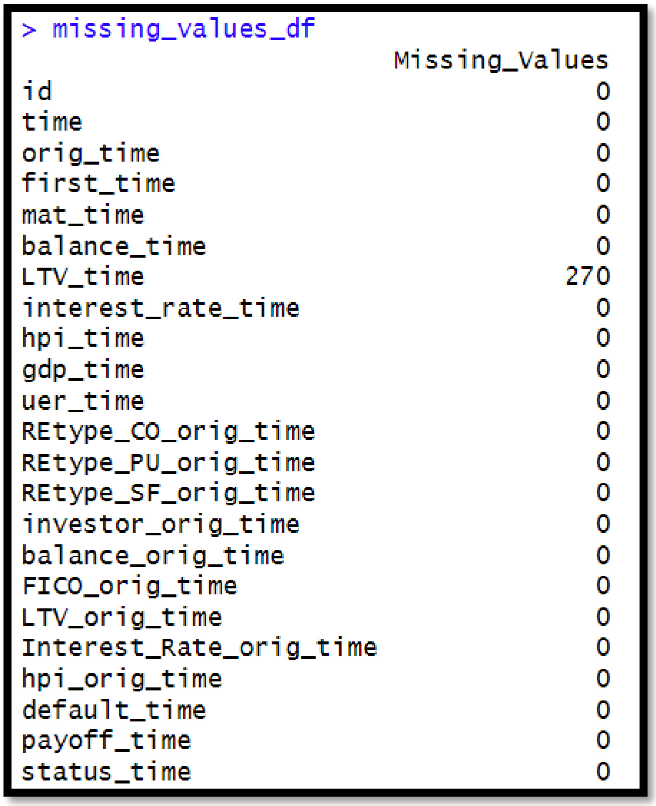
```
> summary(numerical_attributes)
       id             time         orig_time        first_time       mat_time      balance_time
 Min.   :    1   Min.   : 1.0    Min.   :-40.00    Min.   : 1.00    Min.   : 18.0    Min.   :      0
 1st Qu.:13580   1st Qu.:27.0    1st Qu.: 18.00    1st Qu.:21.00    1st Qu.:137.0    1st Qu.: 102017
 Median :24881   Median :34.0    Median : 22.00    Median :25.00    Median :142.0    Median : 180618
 Mean   :25147   Mean   :35.8    Mean   : 20.57    Mean   :24.61    Mean   :137.2    Mean   : 245965
 3rd Qu.:37045   3rd Qu.:44.0    3rd Qu.: 25.00    3rd Qu.:28.00    3rd Qu.:145.0    3rd Qu.: 337495
 Max.   :50000   Max.   :60.0    Max.   : 60.00    Max.   :60.00    Max.   :229.0    Max.   :8701859

    LTV_time      interest_rate_time    hpi_time        gdp_time         uer_time
 Min.   :  0.00   Min.   : 0.000     Min.   :107.8   Min.   :-4.147   Min.   : 3.800
 1st Qu.: 67.11   1st Qu.: 5.650     1st Qu.:158.6   1st Qu.: 1.104   1st Qu.: 4.700
 Median : 82.25   Median : 6.625     Median :180.5   Median : 1.851   Median : 5.700
 Mean   : 83.08   Mean   : 6.702     Mean   :184.1   Mean   : 1.381   Mean   : 6.517
 3rd Qu.:100.63   3rd Qu.: 7.875     3rd Qu.:212.7   3rd Qu.: 2.694   3rd Qu.: 8.200
 Max.   :803.51   Max.   :37.500     Max.   :226.3   Max.   : 5.132   Max.   :10.000
 NA's   :270
 balance_orig_time FICO_orig_time  LTV_orig_time    Interest_Rate_orig_time hpi_orig_time
 Min.   :      0   Min.   :400.0   Min.   : 50.10   Min.   : 0.000          Min.   : 75.71
 1st Qu.: 108000   1st Qu.:626.0   1st Qu.: 75.00   1st Qu.: 5.000          1st Qu.:179.45
 Median : 188000   Median :678.0   Median : 80.00   Median : 6.290          Median :216.77
 Mean   : 256254   Mean   :673.6   Mean   : 78.98   Mean   : 5.650          Mean   :198.12
 3rd Qu.: 352000   3rd Qu.:729.0   3rd Qu.: 80.00   3rd Qu.: 7.456          3rd Qu.:222.39
 Max.   :8000000   Max.   :840.0   Max.   :218.50   Max.   :19.750          Max.   :226.29
```

Figure 2 Summary Statistics

5 *Data exploration*

5.1 Check for missing values

After analyzing, it was observed that 270 entries have missing values in the `LTV_time` variable.

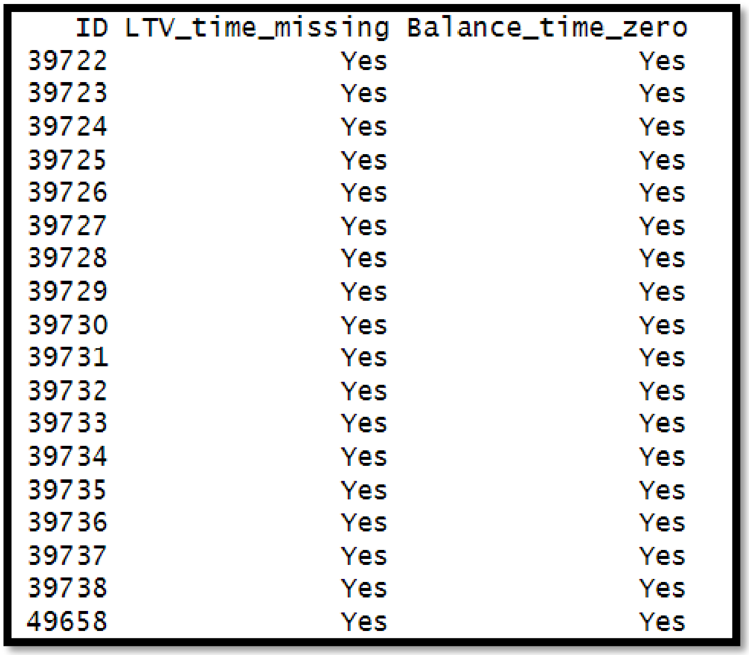


```
> missing_values_df
                         Missing_Values
id                                    0
time                                  0
orig_time                             0
first_time                            0
mat_time                              0
balance_time                          0
LTV_time                            270
interest_rate_time                    0
hpi_time                              0
gdp_time                              0
uer_time                              0
REtype_CO_orig_time                   0
REtype_PU_orig_time                   0
REtype_SF_orig_time                   0
investor_orig_time                    0
balance_orig_time                     0
FICO_orig_time                        0
LTV_orig_time                         0
Interest_Rate_orig_time               0
hpi_orig_time                         0
default_time                          0
payoff_time                           0
status_time                           0
```

Table 1 Missing Values

5.2 Approach for Handling Missing Values

Table 2 reveals that when the "balance_orig_time" variable contains zero values for certain IDs, the corresponding entries in the "LTV_time" column are missing. Considering that the LTV ratio is typically calculated based on the loan balance and property value, and if the balance is reported as zero, it implies there's no outstanding loan amount, resulting in an LTV ratio of zero. Therefore, when encountering missing values in LTV_time corresponding to zero balances, replacing these instances with zeros seems to be a logical approach.

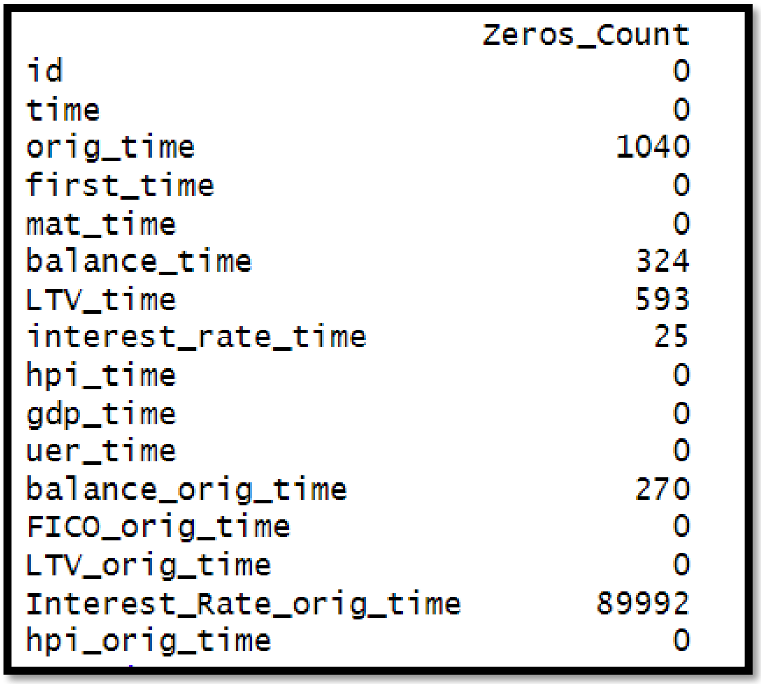


```
   ID LTV_time_missing Balance_time_zero
39722              Yes               Yes
39723              Yes               Yes
39724              Yes               Yes
39725              Yes               Yes
39726              Yes               Yes
39727              Yes               Yes
39728              Yes               Yes
39729              Yes               Yes
39730              Yes               Yes
39731              Yes               Yes
39732              Yes               Yes
39733              Yes               Yes
39734              Yes               Yes
39735              Yes               Yes
39736              Yes               Yes
39737              Yes               Yes
39738              Yes               Yes
49658              Yes               Yes
```

Table 2 IDs Having Zero Balance

5.3 Check for Zeros

Check missing values only for numerical attributes and identify zeros across various variables in the dataset.



```
                         Zeros_Count
id                                 0
time                               0
orig_time                       1040
first_time                         0
mat_time                           0
balance_time                     324
LTV_time                         593
interest_rate_time                25
hpi_time                           0
gdp_time                           0
uer_time                           0
balance_orig_time                270
FICO_orig_time                     0
LTV_orig_time                      0
Interest_Rate_orig_time        89992
hpi_orig_time                      0
```

Table 3 Zero Values

5.4 Variables having Zeros

Orig_time

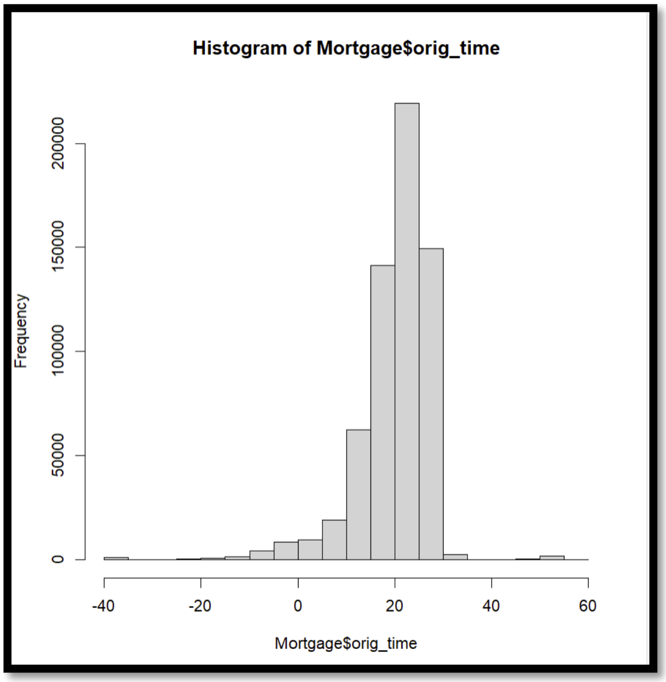


Figure 3 Distribution of Origination time.

The variable "orig_time" represents the time stamp indicating when each mortgage loan was initiated or originated. This timestamp marks the point at which the borrower closed on the property and signed the mortgage deed, officially beginning the loan agreement.

The "orig_time" variable in this dataset ranges from -40 to 60, covering both negative values and zeros. The negative values indicates that the observation period for the mortgage loans begins before the actual origination date of the loans., suggest that the dataset includes information not only about loans originated during the observation period but also about loans that were originated before the observation period began. While zeros could represent loans that originated at the beginning of the observation period or loans with missing origination timestamps.

Balance_time

Zeros in the "balance_time" variable indicate that the borrower has paid off their loan. A zero balance suggests that the outstanding balance on the mortgage loan has been fully repaid, resulting in no remaining debt owed by the borrower.

LTV_time

The presence of zeros in the "LTV_time" variable suggests that borrowers associated with those IDs have paid off their loans entirely, resulting in a zero outstanding balance. Since the Loan-to-Value (LTV) ratio is calculated by dividing the outstanding loan amount by the appraised value of the property, a zero balance implies that there is no loan amount outstanding, resulting in an LTV ratio of zero.

Interest_rate_time

The presence of zeros in the "interest_rate_time" variable may indicate various scenarios within the dataset. These could include instances where loans are associated with promotional or introductory periods featuring zero or extremely low-interest rates, commonly found in certain loan programs.

Balance_orig_time

The zeros in the "balance_orig_time" variable suggests that borrowers may have fully paid off their loans, resulting in a zero outstanding balance. This observation indicates that these borrowers have successfully completed their loan payments and no longer owe any remaining principal amount.

Interest_Rate_orig_time"

The zeros in the "Interest_Rate_orig_time" variable suggests that interest rates were not recorded for certain mortgage loans at the time of origination. This could indicate either missing data or cases where interest rate information was not available or documented during the loan origination process.

6 *Data Preparation and Transformation*

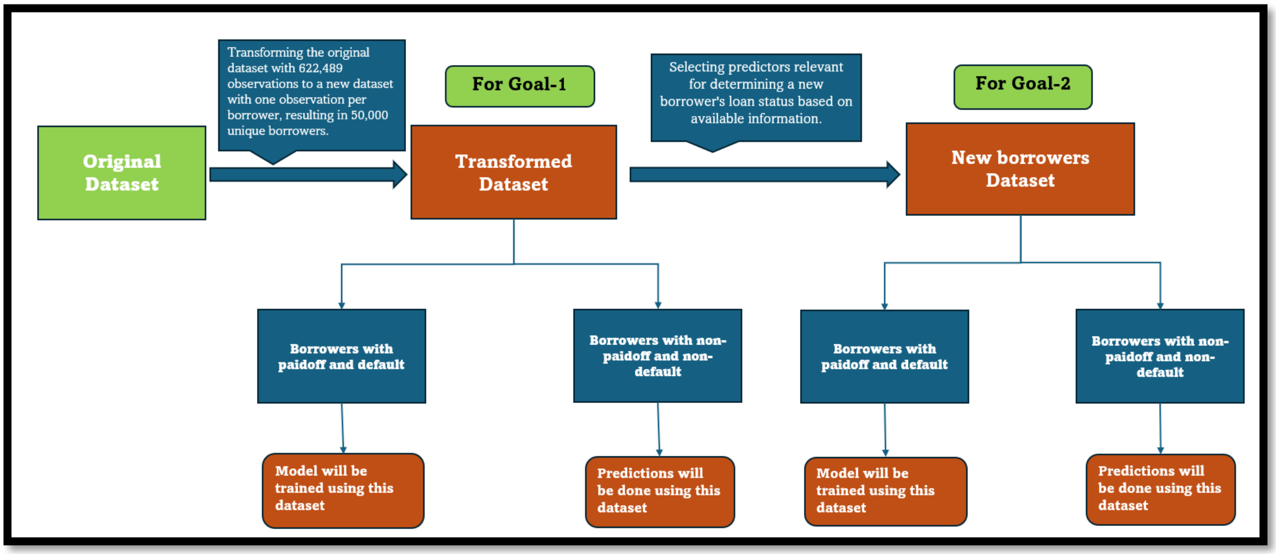

Transforming the original dataset with 622,489 observations to a new dataset with one observation per borrower, resulting in 50,000 unique borrowers.
For Goal-1
Selecting predictors relevant for determining a new borrower's loan status based on available information.
For Goal-2
Original Dataset
Transformed Dataset
New borrowers Dataset
Borrowers with paidoff and default
Borrowers with non-paidoff and non-default
Borrowers with paidoff and default
Borrowers with non-paidoff and non-default
Model will be trained using this dataset
Predictions will be done using this dataset
Model will be trained using this dataset
Predictions will be done using this dataset

Figure 4 Data Transformation

Time

Time represents the timestamp of each recorded observation. The figure illustrates varying transaction counts per ID, indicating distinct borrowing behaviors. To facilitate mortgage analysis, I propose adding a new column named 'Number of transactions,' containing the count of transactions for each ID. IDs with only one transaction may not contribute meaningfully to mortgage analysis, especially if the status is 'paidoff' or 'default.' Such instances could represent non-mortgage loans or incidental transactions like home remodeling. Hence, I intend to exclude IDs with a single transaction if their status is 'paidoff' or 'default.' However, transactions with non-'paidoff' or non-'default' statuses will be retained for validation purposes.

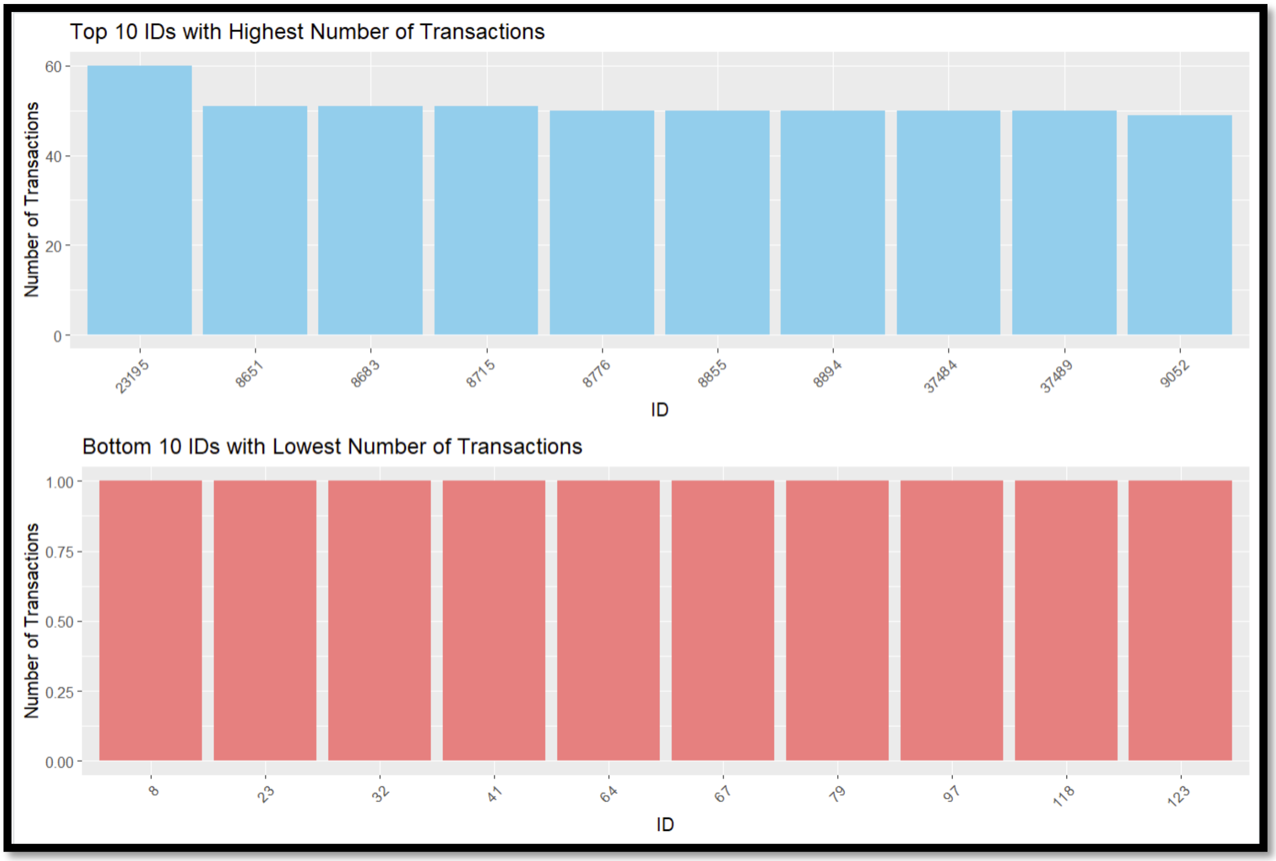
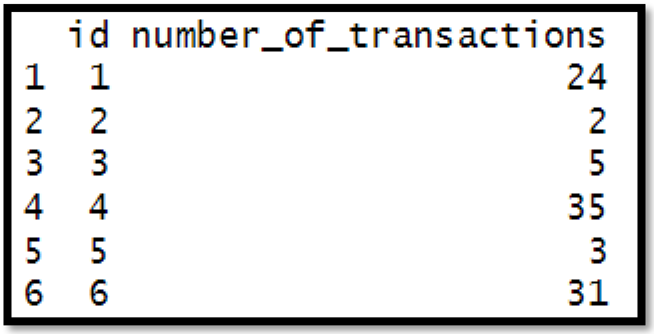


Figure 5 Number of transactions for each id

Number of transactions

Calculated the number of transactions for each unique ID and used to create a new column in the dataset, storing the number of transactions for each ID. Additionally, I filtered out IDs with only one transaction and where the `status_time` variable was either 1 or 2. Finally, I removed the time column from the dataset. These steps were aimed at enhancing the dataset's structure for further analysis, ensuring that it captures relevant information while eliminating unnecessary temporal details.



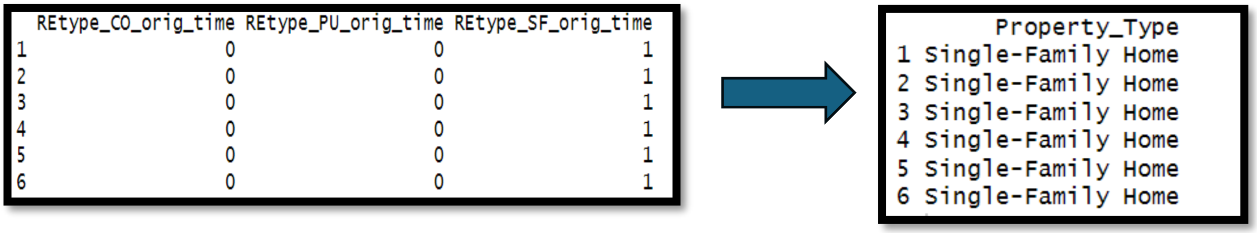
```
  id number_of_transactions
1  1                     24
2  2                      2
3  3                      5
4  4                     35
5  5                      3
6  6                     31
```

Figure 6 Number of transactions.

Combining Property Types

Combined the three columns `REtype_CO_orig_time`, `REtype_PU_orig_time`, and `REtype_SF_orig_time` into a single column `Property_Type` to simplify the representation of the property type for each borrower. These three columns originally represented different property types Condominium, Planned Urban Development (PU), and Single-Family Home (SF). By combining them into one column, we streamline the dataset and make it easier to analyze and interpret the property type information. This transformation helps in creating clearer visualizations and conducting analyses where property type is a factor without the need to consider multiple columns.

Original Columns ***Transformed Column***

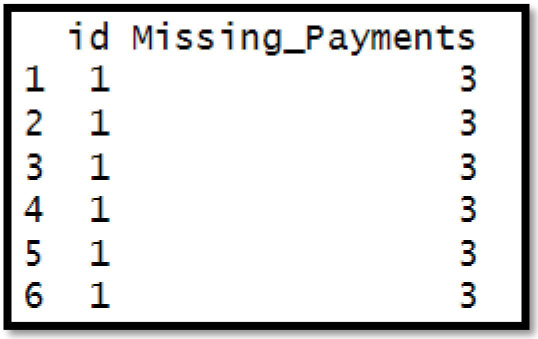


```
  REtype_CO_orig_time REtype_PU_orig_time REtype_SF_orig_time
1                   0                   0                   1
2                   0                   0                   1
3                   0                   0                   1
4                   0                   0                   1
5                   0                   0                   1
6                   0                   0                   1
```

```
       Property_Type
1 Single-Family Home
2 Single-Family Home
3 Single-Family Home
4 Single-Family Home
5 Single-Family Home
6 Single-Family Home
```

Balance_time

In the process of analyzing the 'balance_time' variable within the Mortgage dataset, I calculated the difference between consecutive rows for each unique ID, sorted by 'id' and 'time'. This was achieved using the `mutate` function in combination with `lag` to calculate the difference. Subsequently, I computed the sum of instances where the difference equaled zero, indicative of missing payments, across all rows within each ID group. This sum was then stored in a newly created column named 'Missing_Payments'. The aim was to quantify and record the count of missing payments for each ID, facilitating further analysis of payment patterns and delinquency metrics.

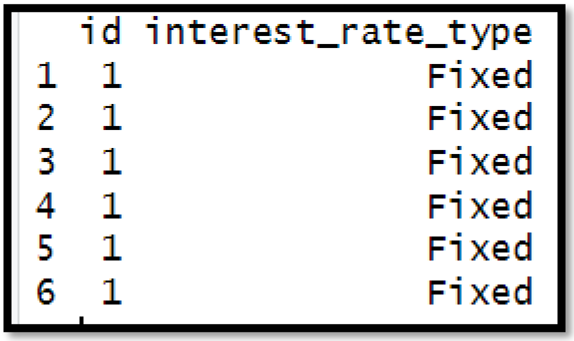


```
  id Missing_Payments
1  1                3
2  1                3
3  1                3
4  1                3
5  1                3
6  1                3
```

Figure 7 Missing Payments

Categorizing Interest Rate Types

To categorize the type of interest rate for each borrower in the Mortgage dataset, I first calculated the standard deviation of the interest rate for each borrower. Subsequently, I established a threshold value to differentiate between fixed and variable interest rates. Based on this threshold, if the standard deviation of the interest rate for a borrower was below the threshold, I classified it as "Fixed"; otherwise, it was categorized as "Variable". This classification was applied to create a new categorical column named 'interest_rate_type' for each borrower ID. This column now denotes whether the interest rate for a borrower is fixed or variable, providing valuable insight into the variability of interest rates across the dataset.



```
  id interest_rate_type
1  1              Fixed
2  1              Fixed
3  1              Fixed
4  1              Fixed
5  1              Fixed
6  1              Fixed
```

Figure 8 Interest Rate Types

7 *Data Analysis*

7.1 For Numerical Attributes

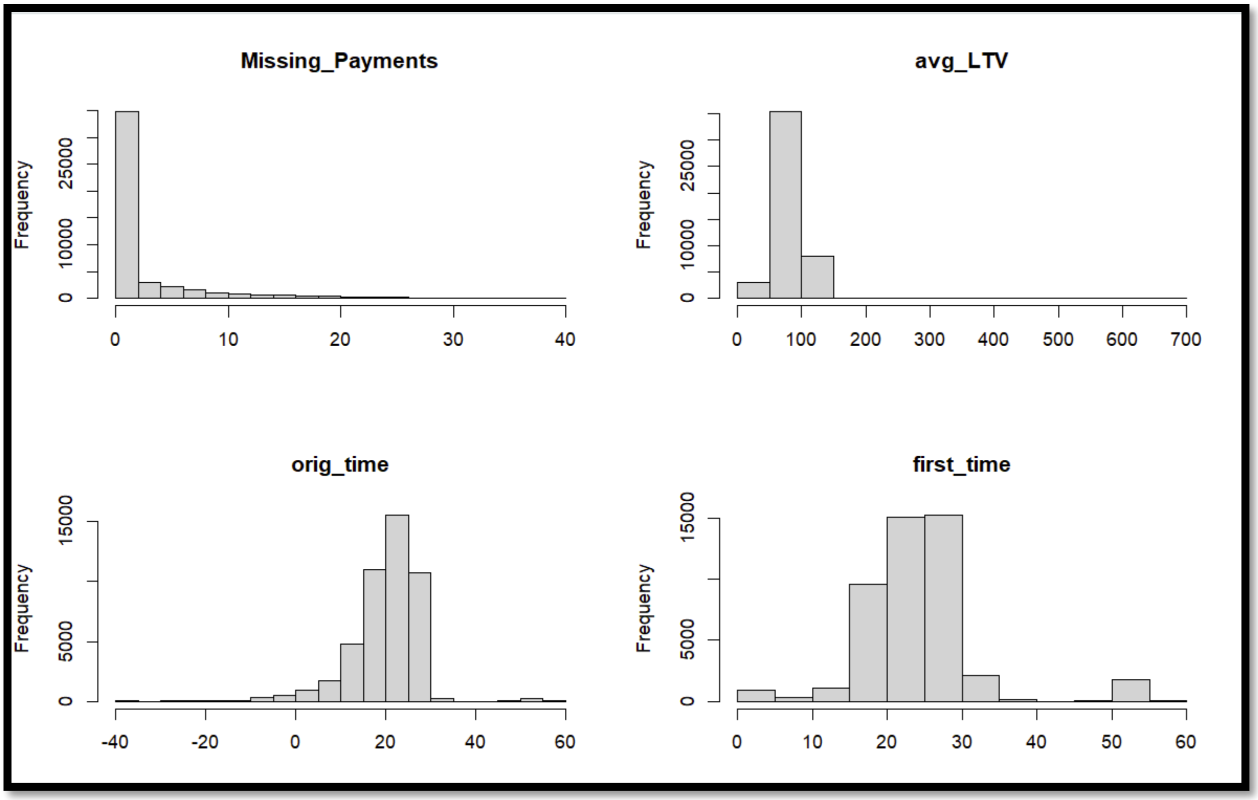


Figure 9 Distribution of First four Numerical Attributes

Number of Missing payments

The plot indicates a higher frequency of 1 to 3 missing payments compared to the range of 5 to 10 missing payments. This observation suggests that a significant proportion of borrowers experienced relatively fewer instances of missed payments, while occurrences of 5 to 10 missed payments were less prevalent within the dataset.

Orig_time The histogram plot reveals that a significant portion of loans originated during periods 15 to 30. This suggests that most borrowers obtained their loans within this timeframe.

First_time During the time period from 15 to 30, the majority of observations mark the first instance of borrower information being recorded in the dataset. This indicates that a significant number of borrowers had their details documented for the first time within this timeframe.

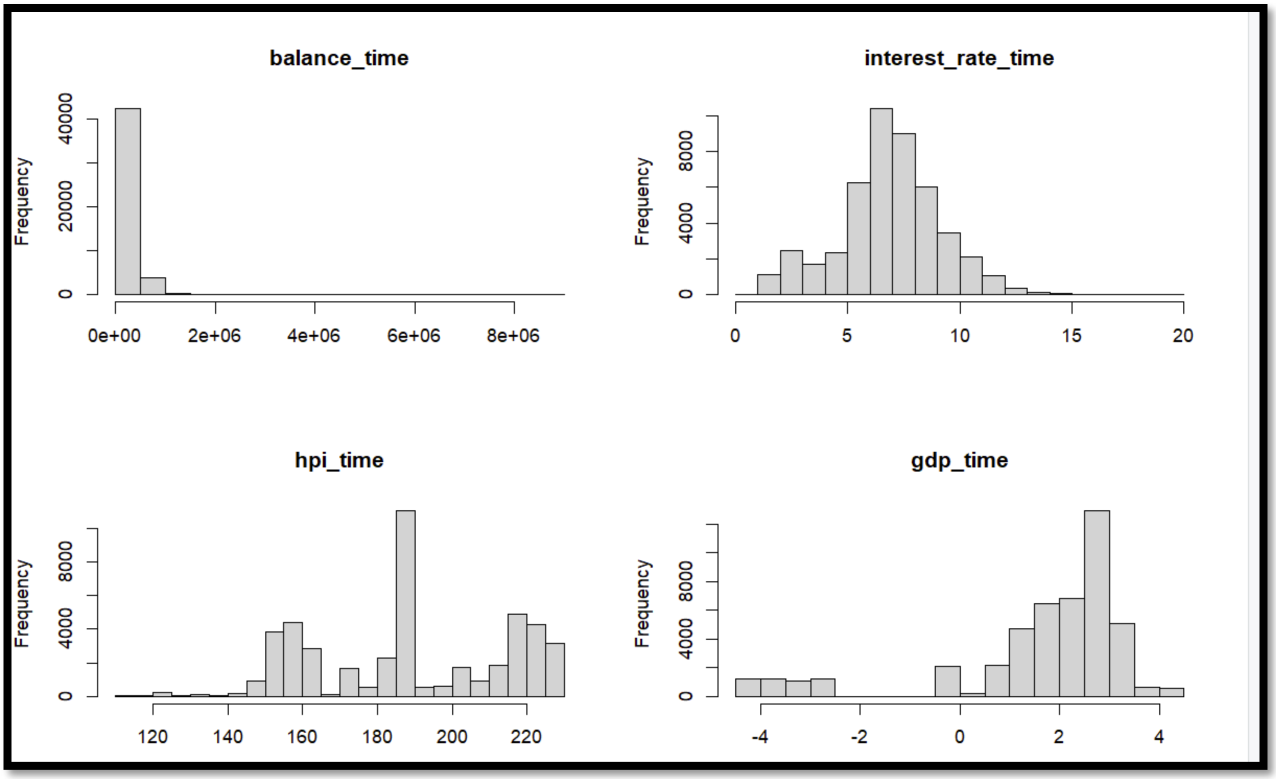


Figure 10 Distribution of second four Numerical attributes.

Balance_time Observing the data distribution, it's notable that the outstanding balances for the majority of borrowers typically fall within the range of one million units. This suggests that most borrowers have outstanding balances around this amount at the time of observation.

Interest_rate_time Observing the data distribution, it is notable that the interest rates for the majority of borrowers typically range from 6% to 8%. This indicates that most borrowers have mortgage loans with interest rates falling within this range at the time of observation.

Gdp_time

It appears that the GDP growth at observation time is expressed as a percentage and ranges from -4% to +4%. It's noteworthy that most of the GDP values fall between 2% to 3.5%, indicating moderate to high economic growth during those periods. This information is valuable as it gives insight into the prevailing economic conditions, which can influence borrowers' income levels, employment stability, and overall financial health.

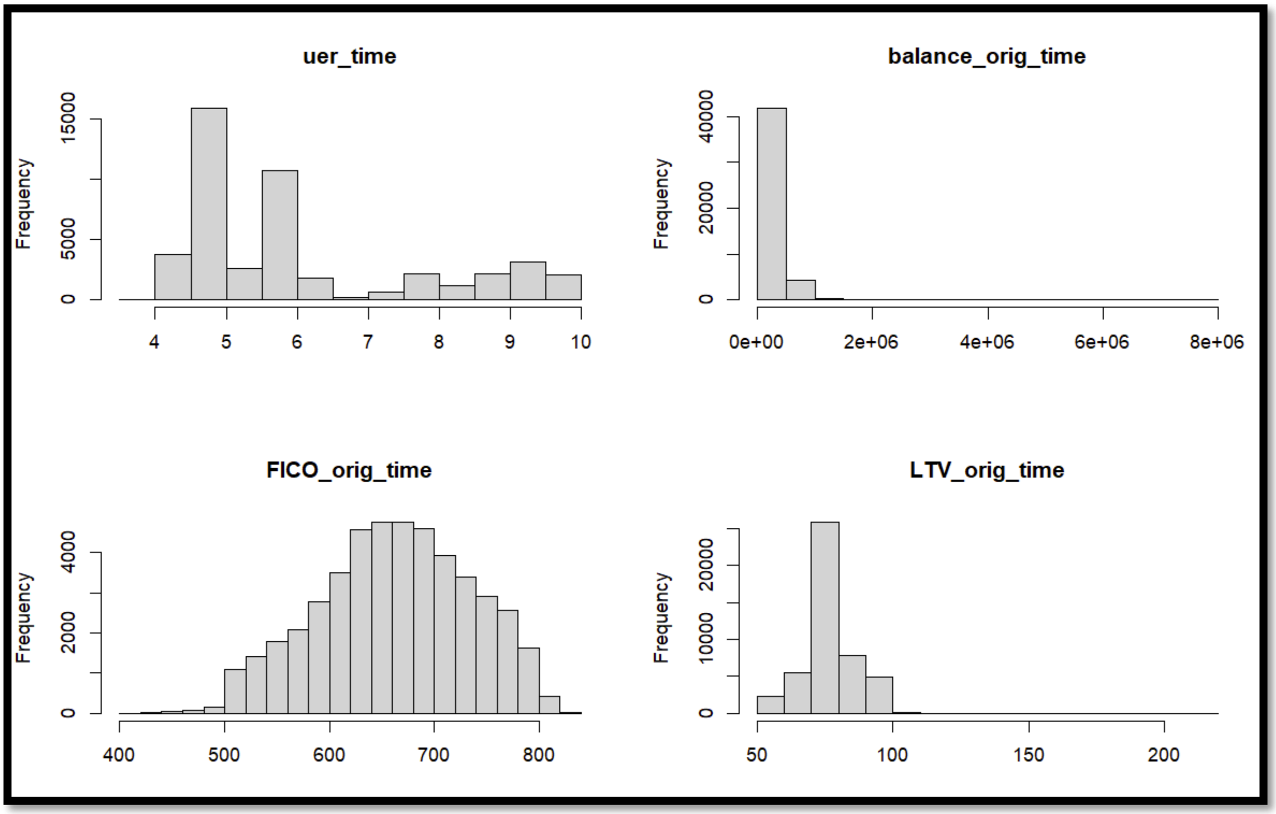


Figure 11 Distribution of Numerical Attributes

Uer-time

Based on the histogram plot, it is observed that the most frequent unemployment rate falls within the range of 4.5% to 5%. This range indicates a notable proportion of unemployed individuals relative to the labor force during those periods.

Fico_orig_time

Fico_orig_time observations predominantly cluster within the range of 600 to 800 based on the histogram plot analysis. This suggests that borrowers' original FICO scores tend to concentrate within this interval.

LTV_Orig_time

From the below histogram, it is noted that most of the LTV values are around 75%. This suggests that, on average, borrowers obtained loans that represented 75% of the appraised value of their properties at the time of origination. Understanding the distribution of LTV ratios is important as it provides insights into the level of leverage assumed by borrowers and the potential risk exposure of lenders.

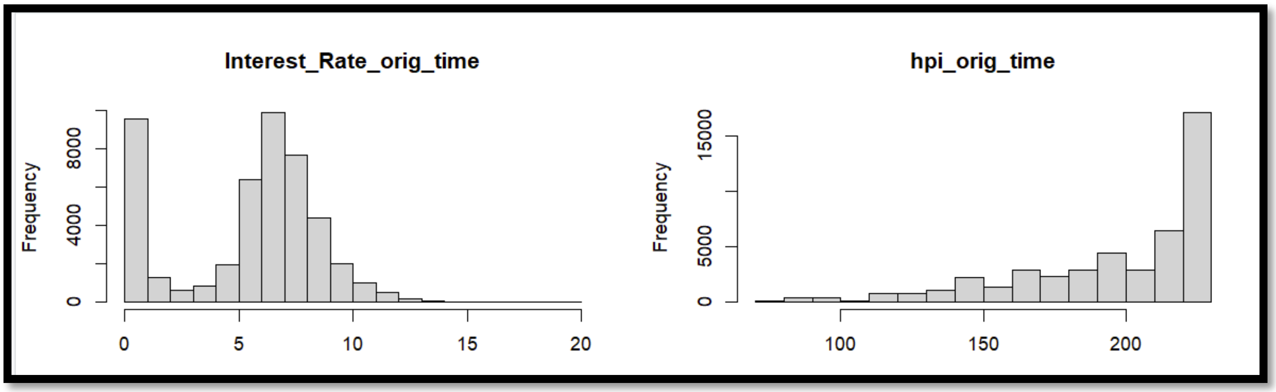


Figure 12 Distributions of Numerical Attribute

Interest rate Classification

From figure 7, bar plot illustrating the distribution of interest rate types, we observe that 30.66% of borrowers have a variable interest rate, while 69.34% have a fixed interest rate. This classification was derived by calculating the standard deviation of interest rates for each borrower, where a threshold of 0.01 was used to classify whether the interest rate is fixed or variable. Subsequently, the frequencies of fixed and variable interest rates were counted, and the percentages were calculated based on the total number of observations. The bar plot, supplemented with percentages above each bar, provides a clear visualization of the interest rate classification among borrowers.

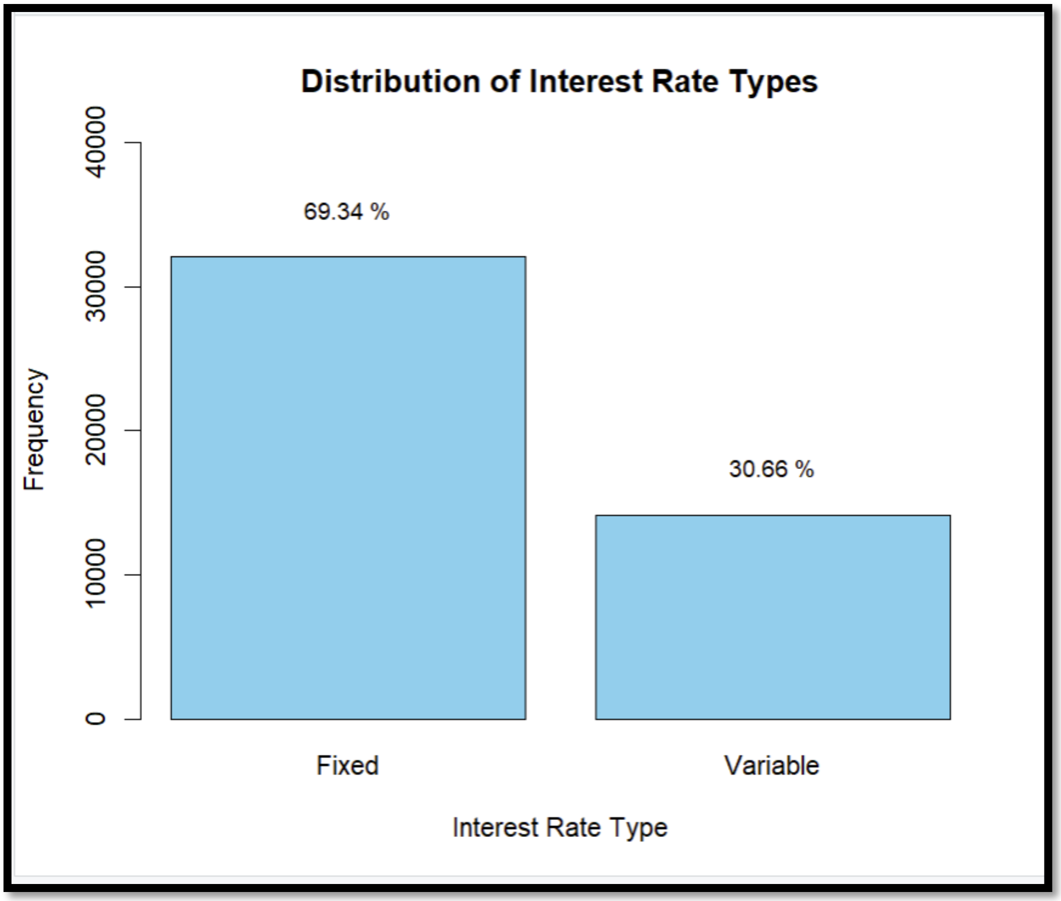


Figure 13 Distribution of Interest Rate Type

7.2 For Categorical Attributes

Figure 8, illustrating the property types of borrowers, it's clear that a substantial number of borrowers possess single-family homes. Conversely, the significant presence of the "Other" category suggests instances where borrowers may not have specified a particular property type or may not own any property at all. This observation underscores the prevalence of single-family home ownership among borrowers while highlighting the need for further examination of the "Other" category to understand whether it signifies a lack of property ownership or other nuanced factors influencing property classification.

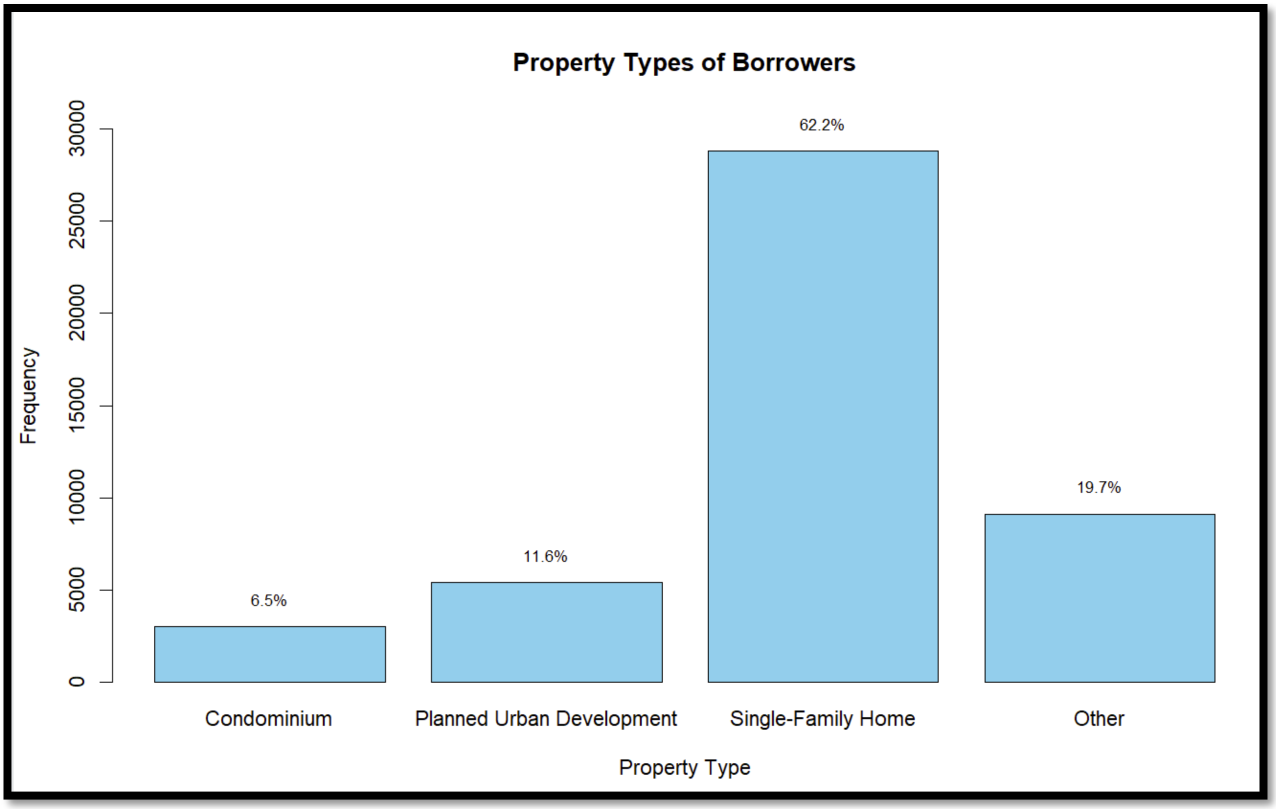


Figure 14 Bar plot for Property Type of Borrowers

Investor_orig_time Investor at origination time

The bar plot representing investor types at origination time illustrates that the majority of borrowers are not classified as investors at the time of loan origination. This observation suggests that a significant portion of borrowers are seeking mortgages for primary residence purposes rather than investment ventures.

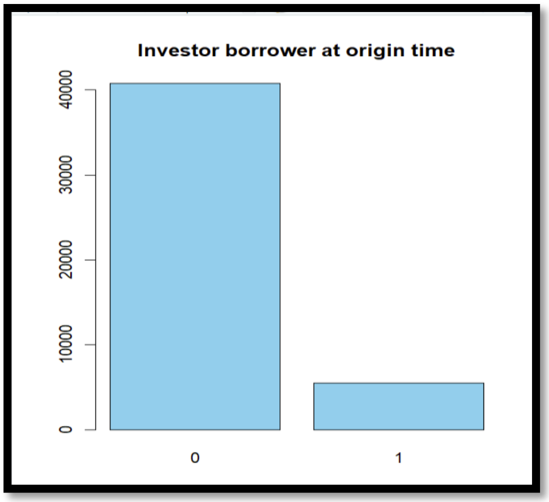


Figure 15 Bar plot for the investor at origination time.

7.3 Target Variable Status_time

The analysis reveals a majority of borrowers in the "paidoff" (2) category, indicating significant mortgage repayments. This underscores ongoing mortgage obligations and highlights the need for predictive models to understand and forecast borrower repayment behavior, crucial for risk assessment in lending institutions.

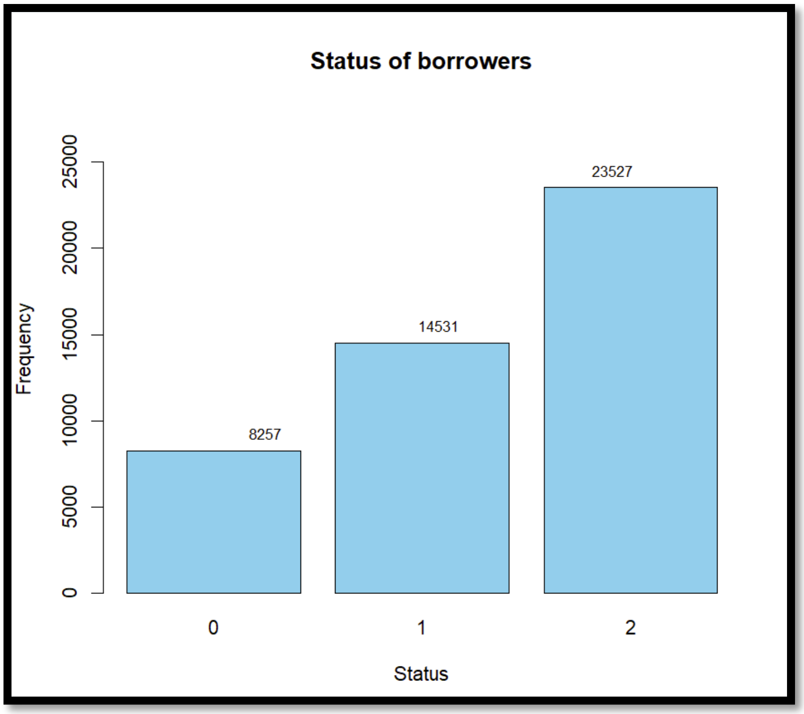


Figure 16 Bar plot of Borrower Status

8 *Feature Selection Methods*

8.1 Correlation Analysis

From the correlation plot, it's evident that there is a strong negative correlation between the unemployment rate (`uer_time`) and the housing price index (`hpi_time`), with a correlation coefficient of -0.84. This suggests that as the unemployment rate increases, the housing price index tends to decrease, and vice versa.

Additionally, we observe a strong positive correlation between several variables, such as `orig_time` and `first_time`, with a correlation coefficient of 0.84. This indicates a high degree of association between the origination time of the mortgage and the first observation time.

Furthermore, `LTV_time` and `hpi_time` exhibit a moderately negative correlation of -0.49. This suggests that as the loan-to-value ratio (`LTV_time`) increases, the housing price index tends to decrease, indicating a potential relationship between mortgage loan risks and fluctuations in housing prices.

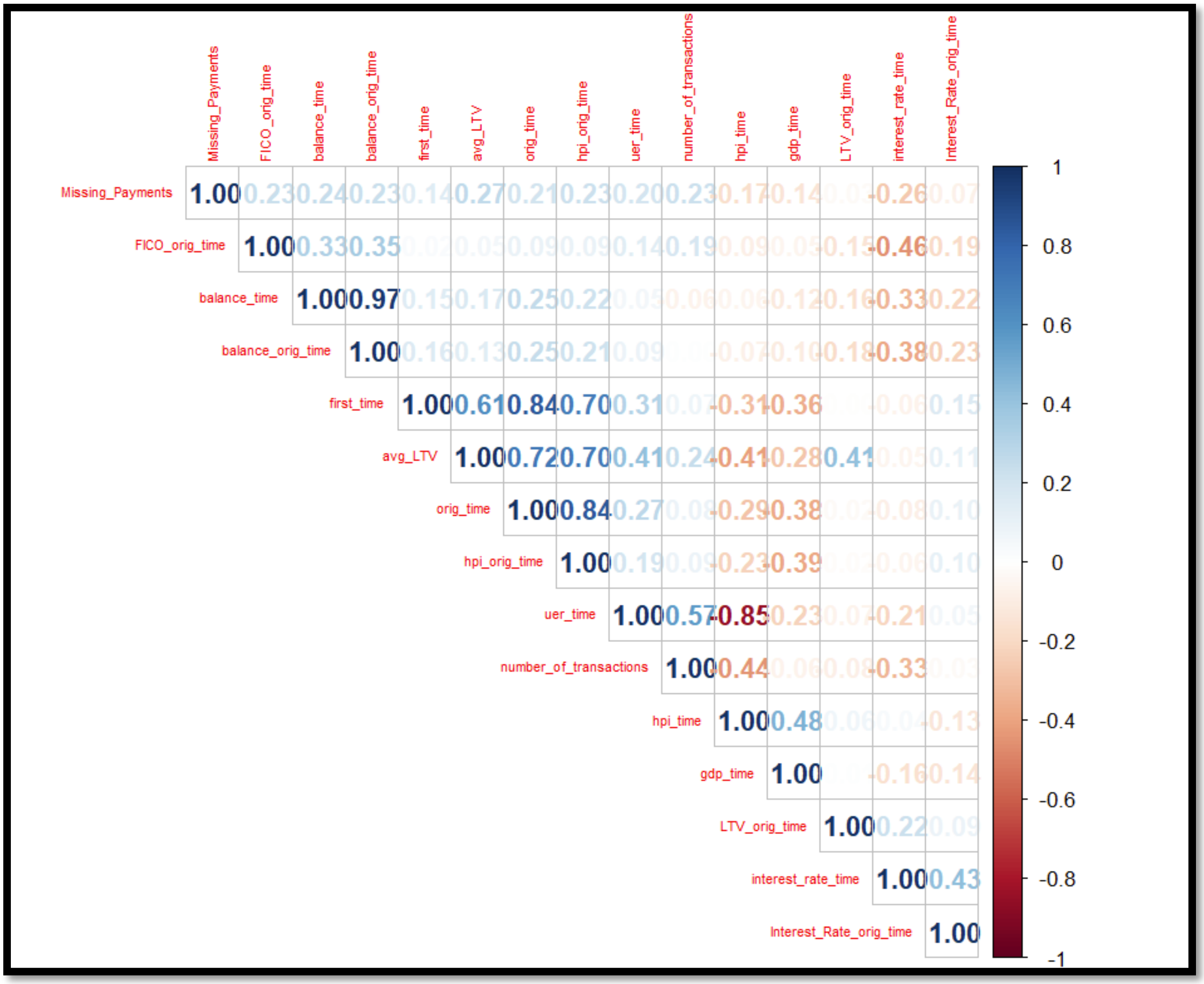
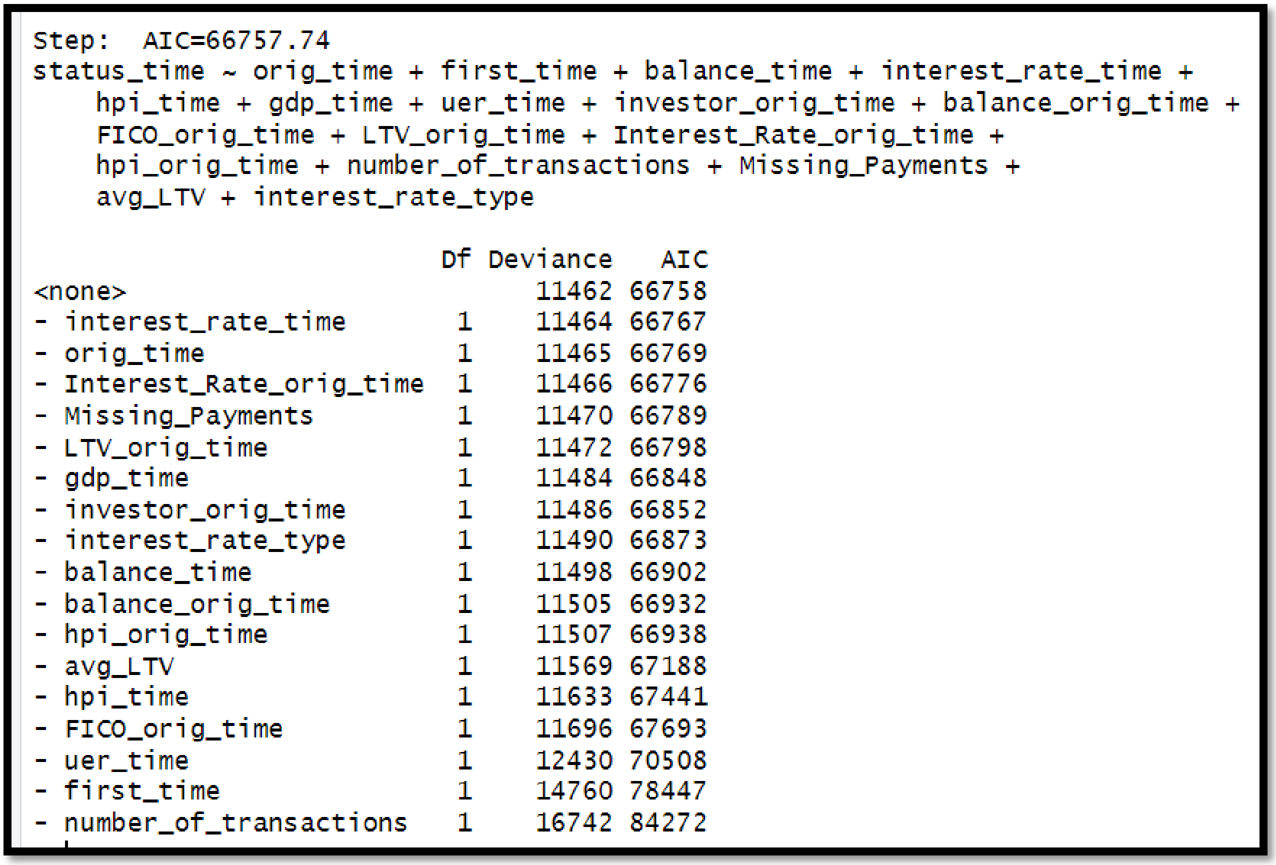


Figure 17 Correlation Analysis for Numerical Attributes

8.2 Logistic Regression with stepAIC

The stepwise backward selection method was applied to a generalized logistic model (GLM) with the aim of identifying the most relevant predictors for predicting the `status_time` variable. This iterative process sequentially removes predictors that contribute the least to the model's performance, as assessed by the Akaike Information Criterion (AIC). Upon completion of the stepwise backward selection, the final model retained all predictors, indicating that no further improvement in AIC was achieved by removing additional predictors. The AIC of the final model was 66,758, suggesting it strikes a reasonable balance between model complexity and goodness of fit. Notably, variables such as `hpi_orig_time`, `Property_Type`, `gdp_time`, and `investor_orig_time` exhibited relatively higher AIC values when omitted from the model, indicating their significance in predicting the `status_time` variable.



```
Step:  AIC=66757.74
status_time ~ orig_time + first_time + balance_time + interest_rate_time +
    hpi_time + gdp_time + uer_time + investor_orig_time + balance_orig_time +
    FICO_orig_time + LTV_orig_time + Interest_Rate_orig_time +
    hpi_orig_time + number_of_transactions + Missing_Payments +
    avg_LTV + interest_rate_type

                          Df Deviance   AIC
<none>                          11462 66758
- interest_rate_time       1    11464 66767
- orig_time                1    11465 66769
- Interest_Rate_orig_time  1    11466 66776
- Missing_Payments         1    11470 66789
- LTV_orig_time            1    11472 66798
- gdp_time                 1    11484 66848
- investor_orig_time       1    11486 66852
- interest_rate_type       1    11490 66873
- balance_time             1    11498 66902
- balance_orig_time        1    11505 66932
- hpi_orig_time            1    11507 66938
- avg_LTV                  1    11569 67188
- hpi_time                 1    11633 67441
- FICO_orig_time           1    11696 67693
- uer_time                 1    12430 70508
- first_time               1    14760 78447
- number_of_transactions   1    16742 84272
```

Figure 18 Logistic Model Using StepAIC

Exploring model performance with both important predictors identified through feature selection methods and with all predictors will provide valuable insights into the impact of variable selection on model accuracy and interpretability. This comparative analysis will inform optimal feature selection strategies for enhancing predictive modeling outcomes in mortgage loan classification tasks.

8.3 Dimension reduction

Reducing dimensions by removing the default_time and payoff_time columns simplifies the dataset and eliminates redundant information. Since the status_time column already encapsulates the default and payoff statuses, keeping these individual columns would not provide additional predictive power and could potentially introduce multicollinearity issues in predictive modeling.

Reducing dimensions makes the dataset simpler, speeds up calculations, and makes it easier to understand how the remaining variables relate to the main outcome. This helps create better models for tasks like predicting defaults or estimating when loans will be paid off.

9 *Data partitioning*

The total data set under consideration contains 50,000 rows. To focus on the borrowers with definitive outcomes, we have filtered out the borrowers who have a status of "paidoff" and "default", while excluding those who are still active or have ongoing loans. As a result, the non-payoff or non-default borrowers have been placed in the holdout set, which is used for model evaluation.

For model building and validation, I have partitioned the dataset into a training set and a validation set. The training set consists of 22,834 rows, while the validation set contains 15,223 rows. This partitioning of the data allows us to build a model using the training data and evaluate its performance using the validation data. It is important to note that the holdout set, which contains the active borrowers i.e., 8257 is separate from both the training and validation sets and is used exclusively for model evaluation.

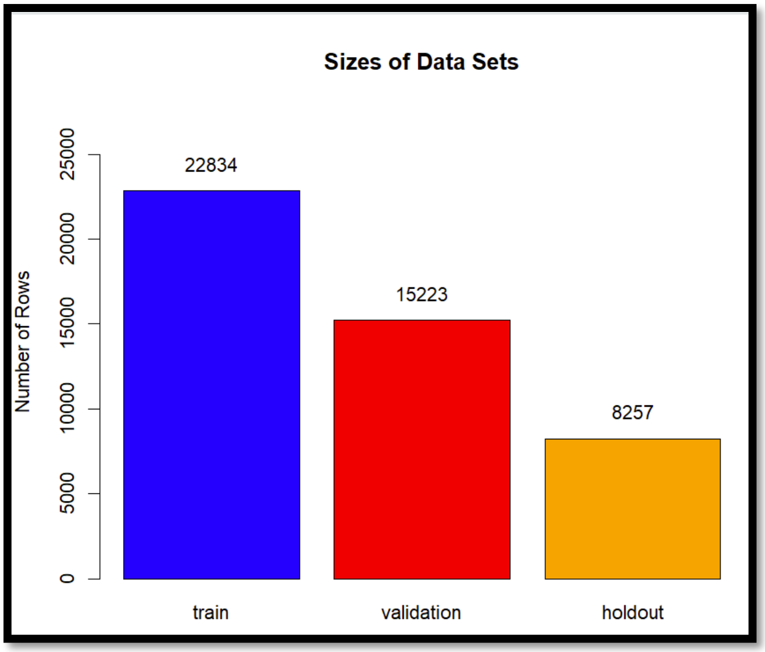


Figure 19 Data Partitioning

10 *Model Selection*

10.1 Analytical Goal 1 Classification of Active Borrowers

To predict whether active borrowers or ongoing borrowers will either pay off their mortgage or default.

To achieve this goal, I would like to build two machine learning models.

10.1.1 Model-1 Logistic Regression

Logistic regression is a popular machine learning algorithm for classification problems, and it is well-suited for this task because it can handle both numerical and categorical data. The model will be trained on the training data, and its performance will be evaluated on the validation data. We will use various metrics, such as accuracy, precision, recall, and F1 score, to assess the model's performance.

10.1.2 Model-1 Performance

The logistic regression model was trained on the training data and evaluated on the validation data. The model's coefficients suggest that several variables, such as time, original term, first payment date, maturity term, balance, loan-to-value ratio, interest rate, HPI, GDP, unemployment rate, and original FICO score, are significant predictors of whether a borrower will pay off their mortgage or default. The model achieved an accuracy of 78.23% on the validation data, with a sensitivity of 72.63% and a specificity of 81.27%. These results suggest that the logistic regression model is a reasonable fit for the data and can be used to predict the outcome for active borrowers. However, there is still room for improvement, and we will compare the performance of this model to a decision tree model to determine which model performs better.

Accuracy	**78.23%**
Sensitivity	**72.63%**
Specificity	**81.27%**

10.1.3 Model -2 Decision Tree

To build an alternative model for classifying active borrowers and predicting whether they will pay off their mortgage or default, we will use a decision tree algorithm. Decision trees are a type of machine learning algorithm that can be used for both regression and classification tasks. They are easy to interpret and visualize, making them a good choice for this task. The decision tree model will be trained on the training data and evaluated on the validation data using the same metrics as the logistic regression model.

10.1.4 Model-2 Performance

The model had a complexity parameter (CP) of 0.34592502 at the root node, and the variable importance analysis showed that LTV_time, time, hpi_time, mat_time, and orig_time were the top five important variables. The model achieved an accuracy of 77.37% (95% CI 0.7662, 0.781) on the validation data, with a sensitivity of 70.18% and a specificity of 81.11%.

The positive predictive value was 83.93%, and the negative predictive value was 65.94%. The balanced accuracy was 75.65%. The Kappa statistic was 0.5051, indicating moderate agreement between the predicted and actual class labels.

Accuracy	**76.25%**
Sensitivity	**68.82%**

Specificity	**80.63%**

10.1.5 Model Comparison

Model	***Accuracy***	***Sensitivity***	***Specificity***
Logistic Regression	***78.23%***	***72.63%***	***81.27%***
Classification Tree	***76.25%***	***68.82%***	***80.63%***

Based on the comparison of the two models, logistic regression exhibits a slightly higher accuracy of 78.23% compared to the classification tree model, which achieved an accuracy of 76.25%. Furthermore, logistic regression demonstrates a superior sensitivity of 72.63% compared to the classification tree model's sensitivity of 68.82%, indicating its ability to better identify the positive class. However, logistic regression displays a lower specificity of 81.27% compared to the classification tree model's specificity of 80.63%, suggesting that the classification tree model excels at discerning the negative class. In summary, logistic regression effectively classifies active borrowers into default or paid-off categories, showcasing good performance overall.

In the context of predicting borrower loan repayment, the cost of misclassifying a borrower who will default as paid off is high, as it can result in significant financial losses for the lender.

10.2 Analytical Goal 2 Classification of New Customers

To forecast whether new customers applying for a mortgage will pay off their loans or default.

Using the above two models to achieve Goal-2.

10.2.1 **Data preprocessing for this Goal-2**

To prepare the data for the second analytical goal, which is to classify new customers as to whether they will pay off their loans or default, we will remove several columns from both the trained data and validation data. Specifically, we will remove the time, orig_time, first_time, mat_time, balance_time, Ltv_time, Interest_rate_time, and hpi_time columns. After removing these columns,number of transactions, avg_ltv, missing payments, we will build a new model and evaluate its performance using the validation data.

10.2.2 **Model-1 Logistic Regression**

The logistic regression model for goal-2 achieved an accuracy of 74.17% in predicting whether new customers will pay off their loans or default. The sensitivity of the model was 75.93%, indicating its effectiveness in identifying borrowers who will pay off their loans compared to those who will default. However, the specificity of the model was 70.30%, suggesting that there is still room for improvement in correctly identifying borrowers who will default. Overall, the model performs reasonably well. Since the goal is to focus on borrowers who will pay off their loans, the model's sensitivity remains a more crucial metric than its specificity.

Accuracy	**74.17%**
Sensitivity	**75.93%**
Specificity	**70.30%**

10.2.3 **Model -2 Classification Tree**

The classification tree model achieved an accuracy of 74.33% in predicting loan repayment for new customers. The sensitivity of the model was 76.03%, indicating its effectiveness in identifying borrowers who will pay off their loans compared to those who will default. However, the specificity of the model was 70.56%, suggesting that there is still room for improvement in correctly identifying borrowers who will default. Overall, the model performs reasonably well, particularly in identifying borrowers who will pay off their loans. Since the goal is to focus on borrowers who will pay off their loans, the model's sensitivity remains a more crucial metric than its specificity.

Accuracy	**74.33%**
Sensitivity	**76.03%**
Specificity	**70.56%**

10.2.4 Model comparison

Model	*Accuracy*	*Sensitivity*	*Specificity*
Logistic Regression	***74.17%***	***75.93%***	***70.30%***
Classification Tree	***74.33%***	***76.03%***	***70.56%***

Both the logistic regression and classification tree models demonstrate reasonable performance in predicting whether new customers will pay off their loans or default. The logistic regression model achieved an accuracy of 74.17% and a sensitivity of 75.93%, slightly higher than the classification tree model's accuracy of 74.33% and sensitivity of 76.03%. This suggests that the logistic regression model is marginally better at identifying borrowers who will pay off their loans. However, both models exhibit similar specificities, with the logistic regression model at 70.30% and the classification tree model at 70.56%. Therefore, if the primary objective is to focus on borrowers who will pay off their loans, the logistic regression model may offer a slight advantage.

11 *Prediction on holdout set of selected models for Goal-1 and Goal-2*

11.1 For Goal-1

Based on the logistic regression model's predictions on the holdout set, which comprised 8170 active borrowers, the model forecasted that 5916 borrowers would repay their loans, while 2253 borrowers were anticipated to default. This suggests that the model has identified a notable proportion of active borrowers who are likely to

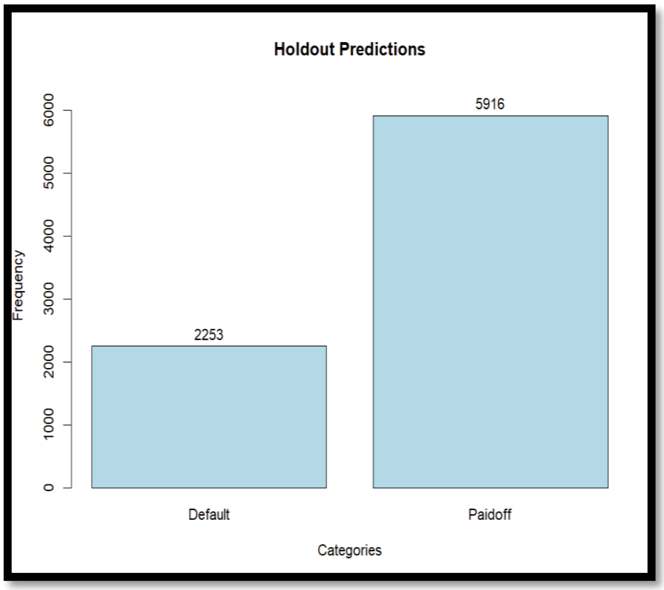


Figure 20 Classified as default & Paidoff

default, enabling the lending institution to implement proactive measures to address the risk of default. Overall, the logistic regression model has demonstrated encouraging outcomes in predicting the loan repayment behavior of active borrowers, offering valuable insights for lending institutions to effectively manage their credit risk.

Holdout prediction for active borrowers	
Default	Paid off
2253	5916

11.2 **For Goal-2**

The logistic regression model predicted that out of the 8257 new customers, 7238 are likely to repay their loans, while 1019 are not. This prediction can help the lending institution to make informed decisions about approving new mortgages and manage their credit risk.

Prediction of new customers	
Default	Paid off
1019	7238

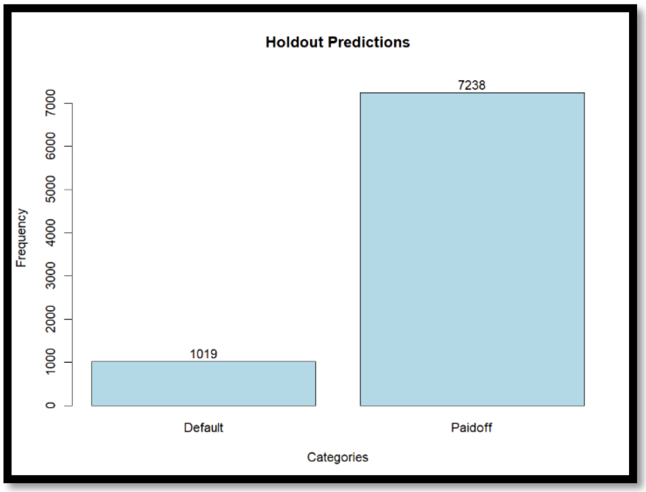


Figure 21 Classified as default & Paidoff

In Analytical Goal 1, it's crucial to identify those who might not pay back their loans among existing borrowers. This helps the bank take steps to prevent losses, like offering advice or changing repayment plans.

In Analytical Goal 2, the main aim is to figure out who will pay back their loans among new customers. This allows the bank to approve loans confidently, grow its customer base, and avoid risky loans that might not be paid back.

12 *Conclusion*

In conclusion, the logistic regression model has demonstrated promising predictive capabilities for both Analytical Goal 1 and Goal 2, providing valuable insights for lending institutions in managing credit risk. For active borrowers, the model accurately identified a significant portion likely to default, enabling proactive risk mitigation measures. Similarly, for new customers, the model's predictions offer crucial guidance for loan approval decisions, aiding in effective credit risk management. While acknowledging the model's imperfections and the need for ongoing refinement, its performance underscores its utility as a valuable tool in the lending industry's arsenal for assessing and managing loan repayment behavior.

13 *Business recommendations*

1. Help those who might struggle By leveraging predictions from Goal 1's model, we can identify borrowers who may face challenges in repaying their loans. Armed with this insight, we can proactively reach out to these borrowers, offering them tailored advice or adjusting their repayment plans to prevent them from falling behind on payments. This proactive approach not only helps borrowers stay on track with their loans but also minimizes the risk of defaults, ultimately fostering stronger borrower-lender relationships.

2. Make smarter loan decisions Utilizing the predictive power of Goal 2's model enables us to make more informed decisions when approving new loans. By confidently approving loans for individuals who are likely to repay them, we mitigate the risk of defaults and ensure a healthier loan portfolio. Additionally, by identifying applicants who may struggle to repay their loans, we can avoid extending loans to them, thereby safeguarding the financial interests of both the lending institution and the borrowers. This approach promotes responsible lending practices and contributes to overall financial stability.

Future Work

Future work will encompass refining the existing models through continued optimization techniques to enhance their predictive accuracy. Additionally, integrating cost matrices will be crucial in quantifying the financial implications of misclassifications, thereby guiding decision-making processes more effectively. Furthermore, recommendations based on cost matrix analysis, such as adjusting model thresholds or prioritizing certain classification outcomes, will be incorporated to better align model predictions with the financial objectives of lending institutions.

14 *Executive Summary*

Sri Vamshi Polela

April 24, 2024

Business Opportunities

The project is designed to be a valuable asset for banks and lending institutions seeking to improve their mortgage portfolio management decisions. By harnessing the power of data analytics, the primary aim is to minimize risks inherent in mortgage lending, optimize loan performance, and ultimately bolster profitability. At its core, the project strives to ensure that mortgage lending decisions are not only well-informed but also efficient, thereby fostering financial stability and customer contentment.

Solution

Our project offers a strategic solution tailored to assist banks and lending institutions in refining their mortgage portfolio management practices through the application of data analytics. Through meticulous data preprocessing, exploration, and transformation, we've unearthed crucial insights into borrower profiles and loan dynamics.

Leveraging predictor analysis and dimension reduction techniques, we've streamlined the dataset, enabling the selection of logistic regression and decision trees as optimal classification models. By segmenting the data and rigorously assessing model performance, our project provides stakeholders with actionable insights to mitigate risks, enhance loan performance, and drive profitability. In essence, our solution serves as a catalyst for fostering financial stability and elevating customer satisfaction within the mortgage lending industry.