Hema Lakshmi Prasanna Sanaka

Webster University

CSDA 6010 DATA ANALYTICS PRACTICUM

SPRING 2024

FINAL PROJECT (Case-3)

JP WANG

MORTGAGE PAYBACK ANALYTICS

**TABLE OF CONTENTS**

1. INTRODUCTION……………………………………………………………………………………………...6
2. BUSINESS GOAL……………………………………………………………………………………………….6

3. ANALYTICAL APPROACH…………………………………………………………………………..………6

1. DATA PREPROCESSING…………………………………………………………………………………….7
   1. ATTRIBUTE DEFINITIONS……………………………………………………..……………… 7
   2. DATA EXPLORATION…………………………………………………………….……………….8
   3. HANDLING CATEGORICAL VARIABLES………………………………………………….12
   4. CHECKING FOR MISSING VALUES………………………………….…………………….12
   5. HANDLING MISSING VALUES……………………………………..……………………….13
   6. CHECKING FOR ZERO’S…………………………………………….………………………….14
   7. ADDING NEW COLUMN WHICH IS LOAN\_PERFORMANCE…….……………….15
   8. ADDING NEW COLUMN LENGTH\_LOAN…………………………………………………15
   9. ADDING NEW COLUMN REPAYMENT………………………………………………………15
2. PREDICTOR ANALYSIS…………………………………………………………………………………………16

5.1 CORRELATION FOR ALL NUMERICAL COLUMNS…………………………………………….30

6. DIMENSIONAL REDUCTION……………………………………………………………………………………31

7. DATA TRANSFORMATION……………………………………………………………………………………….32

8. DATA PARTITIONING METHOD……………………………………………………………………………….32

9. CLASSIFIER MODEL SELECTION……………………………………………………………………………….33

9.1.1 LOGISTIC REGRESSION………………………………………………………………………...34

9 .1.2 FORWARD STEP WISE LOGISTIC REGRESSION………………………………….…36

9.2 CLASSIFICATION TREE USING RPART () ……………………………………………………....37

9.2.2 CLASSIFICATION TREE USING C5.0 MODEL…………………………………….…41

10. BEST CLASSIFER MODEL SELECTION…………………………………………………………….45

11. REGRESSION MODEL SELECTION………………………………………………………..……….45

11.1 MULTIPLE LINEAR REGRESSION……………………………………………….………….46

11.2 FORWARD STEPWISE LINEAR REGRESSION………………………………………….47

11.3 BACKWARD STEPWISE LINEAR REGRESSION……………………………………….49

11.4 REGRESSION TREE……………………………………………………………………………….50

11.5 BEST REGRESSION MODEL SELECTION………………………………………………….54

12. HOLDOUT MODEL PERFORMANCE………………………………………………………………….54

13. CONCLUSION AND FUTURE SCOPE……………………………………………………………………56

**EXECUTIVE SUMMARY**

* SBS Bank, initially focused on gold and student loans, is now offering mortgage loans to its customers. To ensure responsible lending, the bank is implementing two predictive models to assess customer risk and set appropriate interest rates.
* The first model is a forward stepwise logistic regression, designed to classify customers' repayment performance. This model allows the bank to quickly identify high-risk borrowers and take action to reduce loan default risk.
* The second model is a regression tree used to predict the interest rate for individual customers. This approach accounts for credit history and other factors, enabling the bank to offer interest rates that reflect customer risk profiles.
* By using these models, SBS Bank aims to make informed lending decisions, mitigating risk, and enhancing profitability. This strategy is crucial as the bank expands its services to mortgage lending. Accurate risk assessment and tailored interest rates will help SBS Bank maintain financial stability and better serve its customers.

**1.INTRODUCTION:**

SBS Bank, which initially dealing with gold and student loans, is now planning to offer mortgage loans to its customers. Instead of offering mortgage loans to all applicants, the bank has decided to lend to those who are likely to repay. Additionally, they aim to set interest rates that adequately compensate for the increased risk of default. As the bank has no historical data for the mortgage loan the bank has decided to use the data gathered from various sources. This information will assist SBS Bank in offering mortgage loans to specific customers and set interest rates the compensate well.

**2.BUSINESS GOAL:**

SBS Bank aims to optimize decision-making processes and enhance profitability in mortgage loans. By leveraging predictive analytics, the goal is to accurately predict the interest rates that can be offered to customers based on key factors such as credit score, House Price Index (HPI), and Loan-to-Value (LTV) rate. It is crucial to determine the appropriate interest rates for different customers based on their background. For instance, customers with good credit scores typically receive lower interest rates due to their perceived low risk, whereas those with lower credit scores may receive higher interest rates as they are considered higher risk borrowers. Predicting interest rates allows banks to make informed decisions while adjusting margins for different customers, thereby mitigating risks.

Classifying loans based on how well customers repay them is crucial. It allows banks to quickly decide whether to give a loan by examining past loan repayment behaviour. Analysing loan performance helps banks spot ways to improve, reduce risks, and create better strategies to boost profits. The goal is to offer attractive mortgage options to customers while keeping risks low and returns high for the bank.

**3.ANALYTICS APPROACH:**

In this analysis, exploratory data analysis will be conducted to understand loan performance trends and the relationship between features to enhance predictive modelling. Predictive models will be developed to accurately estimate interest rates based on borrower background and market factors such as FICO score, LTV ratio, HPI index, GDP, property value, and outstanding balance amount. Additionally, classification models will be developed to classify loan performance as good or bad using predefined metrics. If the loan is defaulted the customer's performance is considered bad; whereas if the loan is not defaulted, paid off and then the customer's performance is considered good and other factors. Based on this classification, the bank can get to know about the high-risk customers, and they can take the certain actions according to the agreement of the loans. For bank it’s important to know about the default customers so, that they can avoid the risk.

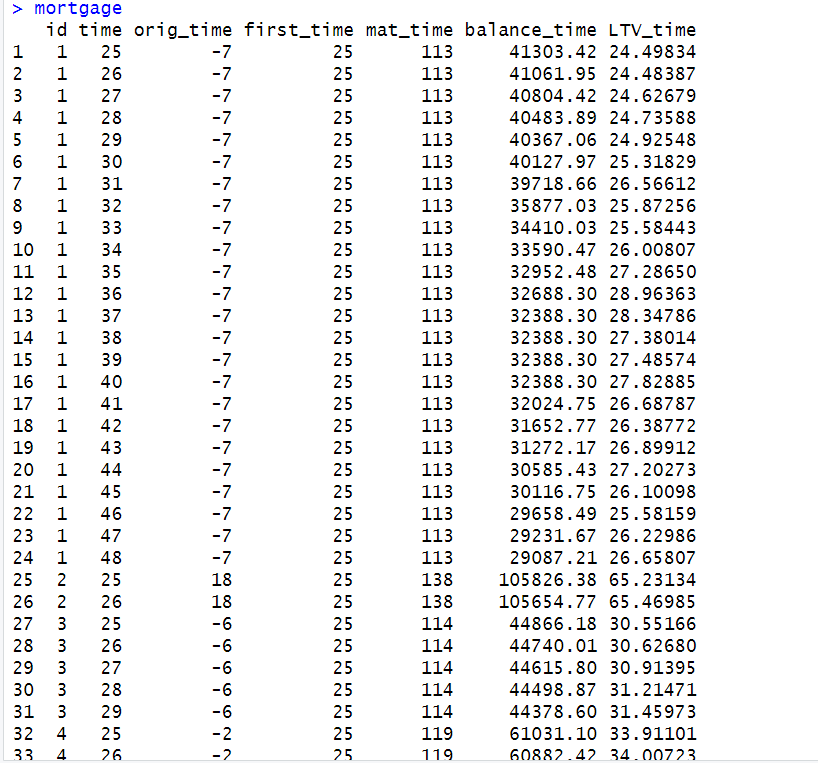
.**4.DATA PREPROCESSING:**

**4.1. ATTRIBUTE DEFINITIONS:**

1. Id: Id which indicates the barrower id.
2. Time: Which indicates the time stamp of the observation.
3. Orig\_time: It represent the time when a loan or mortgage was originated or initiated.
4. First\_time: It represent the initial point in time when information about a specific borrower or loan was recorded.
5. mat\_time: It indicates time stamp for maturity which means that when the loan is scheduled fully paid off.
6. Balance\_time: It indicates that how much amount remains to be repaid at specific point of time.
7. LTV Ratio: Loan to value ratio which represents the ration of the loan amount to the appraised value which is in terms of percentage, and it helps to assess the risk associated with the loan to value of the collateral.
8. interest\_rate\_time: It indicates that Interest rate associated with the loan which is in terms of percentages.
9. hpi\_time: It represents measures the movement of single-family house prices relative to a base year. It provides insights into trends in property values.
10. gdp\_time: It represents growth at observation time represents the rate of growth of the economy, which can influence various factors such as employment and consumer confidence.
11. uer\_time: It represents the unemployment rate at observation time, which in terms of percentages.
12. REtype\_CO\_orig\_time: It represents the real estate type condominium or not.
13. REtype\_PU\_orig\_time: It represents the real estate type planned urban development or not.
14. REtype\_SF\_orig\_time: It represents the property type single-family home or not.
15. investor\_orig\_time: Indicates whether the borrower is an investor (1) or not (0) at origination time.
16. balance\_orig\_time: Represents the outstanding balance of the loan at origination time.
17. FICO\_orig\_time: The FICO score at origination time reflects the borrower's creditworthiness, with higher scores indicating better credit history.
18. LTV\_orig\_time: The Loan-to-Value (LTV) ratio at origination time represents the ratio of the loan amount to the appraised value of the property at the time of loan origination.
19. Interest\_Rate\_orig\_time: The interest rate at origination time represents the initial interest rate associated with the loan.
20. hpi\_orig\_time: The House Price Index (HPI) at origination time measures property value relative to a base year at the time of loan origination.
21. default\_time: Indicates whether the loan has defaulted (1) or not (0) at the observation time.
22. payoff\_time: Indicates whether the loan has been paid off (1) or not (0) at the observation time.
23. Status\_time: Default (1), payoff (2), and nondefault/nonpayoff (0) observation at observation time.

**4.2 DATA EXPLORATION:**

In this dataset have the historical information of the barrower’s along with other useful information in structured data. Initially we have total 622489 rows and 23 columns and can observe that have multiple records on the same ID.Let’s have a look on the data



**Figure 4.2.1 Mortgage Data Before Aggregation**

By examining the figure 4.2.1, it is evident that multiple records exist for each ID. To enhance analysis, it's essential to aggregate the IDs and replace the other variables with appropriate values such as max, min, sum, or average. Each variable will be replaced with the value that best aligns with its characteristics. Specifically, variables like 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, default\_time, payoff\_time, and status\_time will be replaced accordingly. Based on the values and according to business goal aggregate the data based on id with the most recent values which will be more useful in the further analysis. Now, let’s have a look on the aggregated data column wise

ID: In this data have multiple observation on same ID’s so, Aggregate the data by using groupby ID to minimize the data.

Time: For further analysis would like to keep the latest time for each ID while aggregating each ID.

Orig\_time: For further analysis would like to keep the latest orig\_time for each ID while aggregating each ID which represents the loan was initiated.

First\_time: For further analysis would like to keep the latest first time for each ID while aggregating each ID which represents when the repayment was paid first time.

Mat\_time: For further analysis would like to keep the latest maturity time for each ID while aggregating each ID which represents the duration of the loan.

Balance\_time: For further analysis would like to keep the maximum amount for each ID while aggregating each ID which represents most recent value of the total amount of the loan approved from the bank.

LTV\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents most recent value of the ratio of the loan amount.

Interest\_rate\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents most recent value of the interest rate time.

Hpi\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represent the most recent house price index value.

GDP\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents the most recent value of the GDP.

UER\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents the most recent unemployment rate.

REtype\_CO\_orig\_time, REtype\_PU\_orig\_time, REtype\_SF\_orig\_time: For further analysis would like to keep the maximum occurred for each ID while aggregating each ID which represents the maximum occurred house type.

Investor\_orig\_time: For further analysis would like to keep the maximum occurred for each ID while aggregating each ID which represents weather the investor or not.

balance\_orig\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents the remaining amount of the loan.

FICO\_orig\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents the credit score of the customer.

LTV\_orig\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents the ratio of the LTV when the loan was initiated.

Interest\_Rate\_orig\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents interest rate when the loan was initiated.

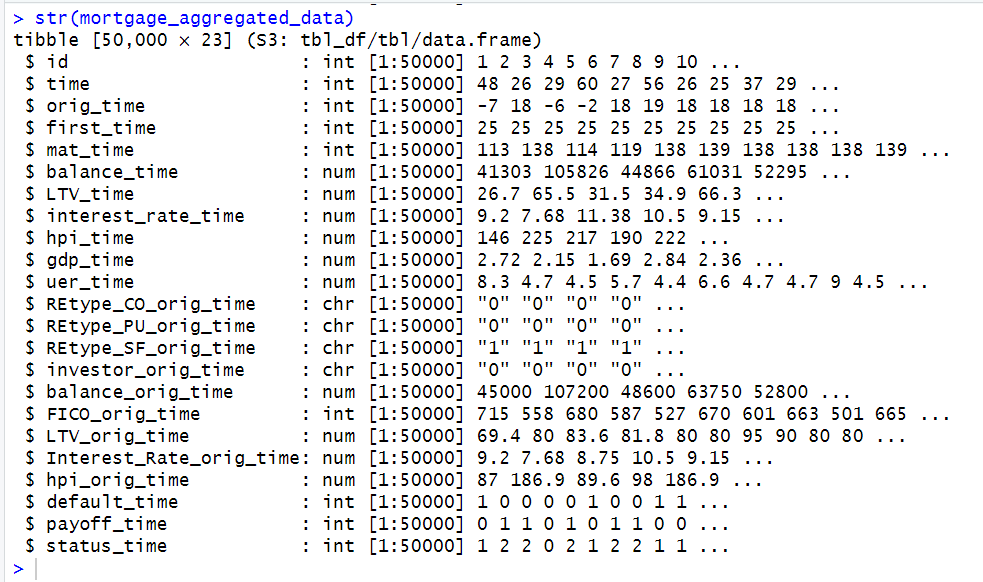
hpi\_orig\_time: For further analysis would like to keep the latest value for each ID while aggregating each ID which represents house price index value when the loan was initiated.

default\_time: For further analysis would like to keep the maximum occurred value for each ID while aggregating each ID.

payoff\_time: For further analysis would like to keep the maximum occurred value for each ID while aggregating each ID.

Status\_time: For further analysis would like to keep the maximum occurred value for each ID while aggregating each ID.

Let’s have a look on the structure of the aggregated data



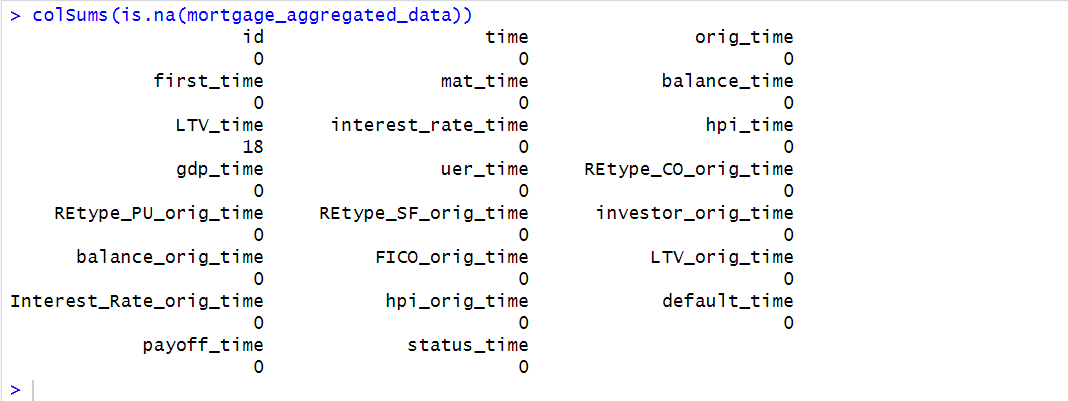
**Figure 4.2.2 Structure of Mortgage Aggregated Data**

**4.3 HANDLING CATEGORICAL VARIABLES:**

By examining the above figure 4.2.2, it is evident that the "char" datatype is present. If necessary, during modelling, it needs to be converted into factors or dummy variables based on the specific requirements of the models being used. Linear and logistic regression models automatically handle categorical columns during the modelling process. However, when employing K-nearest neighbours (KNN), it is necessary to convert categorical columns into dummy variables. On the other hand, decision trees handle categorical columns automatically without the need for additional conversions.

**4.4 CHECKING FOR MISSING VALUES:**

Let’s check for the missing values in the data to get to know how many missing values in the entire dataset. Let’s have a look on the missing values in the data

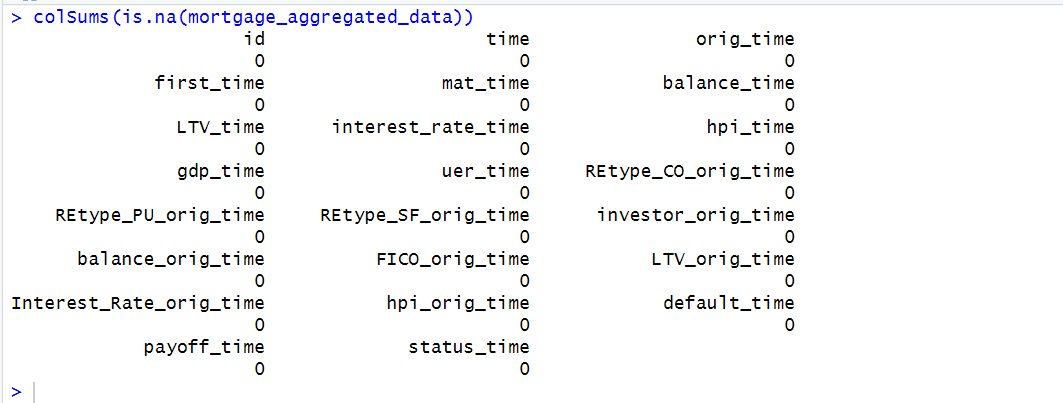


**Figure 4.4.1 Missing values**

By looking at above figure 4.4.1 can be seen that have 18 missing values in the LTV\_time.

**4.5 HANDLING MISSING VALUES:**

To address the missing values in the data used knn imputation method. For this installed VIM package. In this method will replace the missing values based upon the distance between the points.

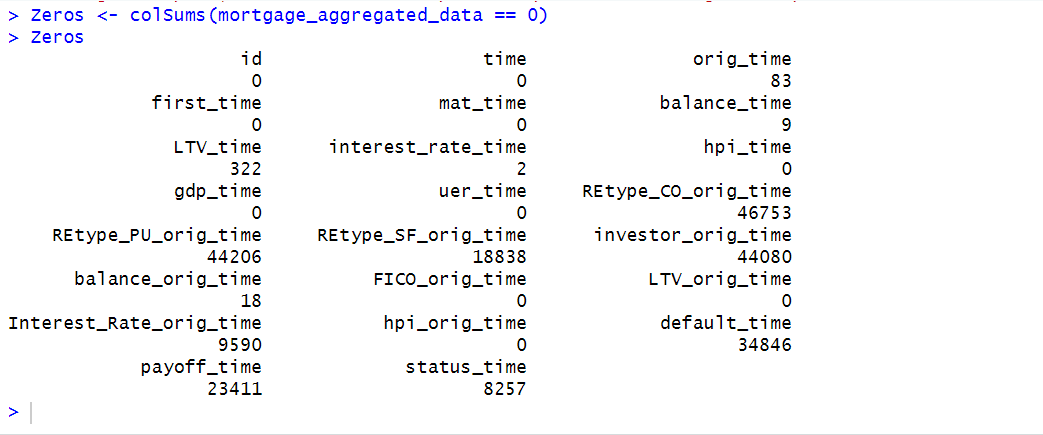


**Figure 4.5.1 After Replacing the Missing values**

In the aggregated mortgage data, missing values have been filled in using the kNN method. This method identifies the k nearest neighbour’s observations with similar characteristics based on other variables in the data, such as the LTV (loan-to-value) ratio. The algorithm then calculates an average or weighted average of these neighbours to replace the missing values, effectively using the information from similar data points to fill in the gaps.

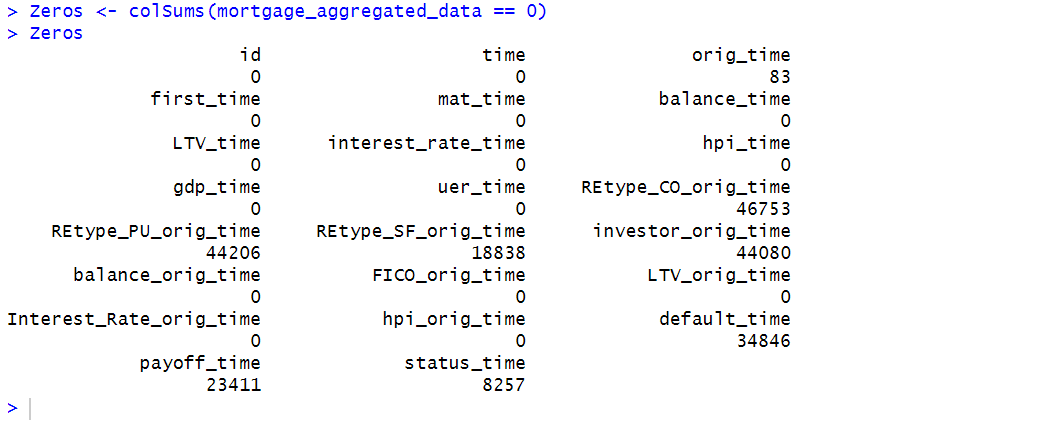
**4.6 CHECKING FOR ZERO’S:**

Now, checking for zeros in the data to get to know how many zero values in the entire data in column wise let’s have a look on the zero’s

****

**Figure 4.6.1 Zero’s in Mortgage Aggregated Data**

By looking at the above figure 4.6.1 can be seen that have zeros in different columns those are 83 zeros’ in orig\_time, 322 zeros in LTV\_time, 2 zeros in interest\_rate\_time, 18 zeros in balance\_orig\_time, 9590 zeros in interest\_rate\_orig\_time. Need to replace these zero values with the appropriate values.



**Figure 4.6.2 After Replacing Zero’s in mortgage Aggregated Data**

Replaced zero values in the LTV\_time, if zero values present it replaces it with the corresponding values from the "LTV\_orig\_time" variable which is initial variable. Likewise replaced values in the balance\_time, if zero values present it replaces it with the corresponding values from the balance\_orig\_time variable which is the initial value. Replaced zeros in interest\_rate\_time with the mean of intereset-rate\_time based on the previous data. Replaced zeros in interest\_rate\_orig\_time with mean of interest\_rate\_orig\_time based on the previous data.

**4.7 ADDING NEW COLUMN IS LENGTH OF LOAN:**

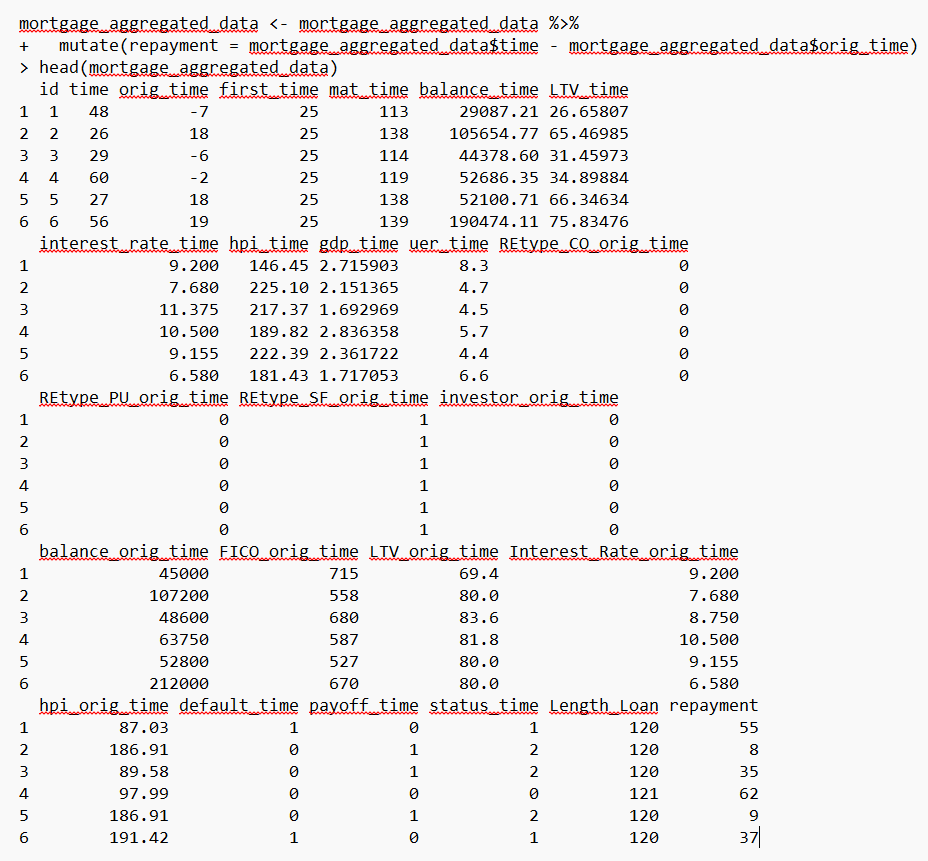
Creating a new column indicating the length of the loan based on the difference between maturity\_time and orig\_time. If maturity\_time - orig\_time - orig\_time equals to the length of the loan, it signifies the total duration of the loan.

**4.8 ADDING NEW LOAN PERFORMANCE COLUMN:**

Creating a new column which tells us performance of the loan either good or bad based on the status time. If the customer is defaulted, then can be able to say that performance of the customer loan is bad performance because they are not repaying loan amount on time. If the customer is not defaulted, then can be able to say that performance of the customer loan is good performance because they are repaying the loan amount on time.

**4.9 ADDING NEW COLUMN IS REPAYMENT:**

Creating a new column repayment based on the difference between orig\_time and time which tells us about the initial repayment of the customer.

****

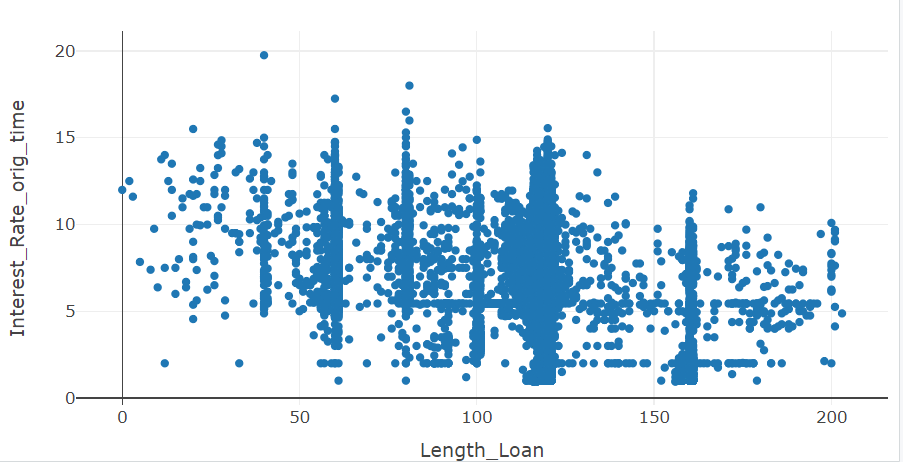
**Figure4.7.1 AFTER ADDING ALL COLUMNS TO MORTGAGE AGGREGATED DATA**

In further analysis going to use this as a target column for the classification model. Model will classify weather the performance of loan is either good or bad.

**5. PREDICTOR ANALYSIS:**

By doing the predictor analysis can get more detailed information about the variables and relationship between the variables. Can observe the which variables more impact on the target column. Let’s have a look on the different plots for more insights from the data.

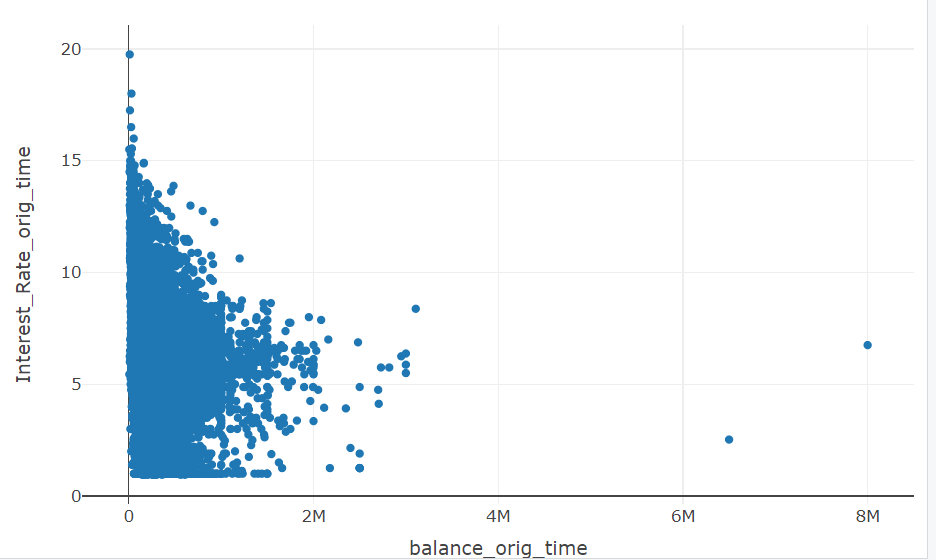
**Scatter plot for Length\_loan and interest\_orig\_time:**

****

**Figure 5.1 Scatter plot for mat\_time and interest\_orig\_time**

By looking at the above figure 5.1 can be able to say that less length of loan indicates that higher interest\_rate\_orig\_time and maximum length of loan indicates that lower interest\_rate\_orig\_time.

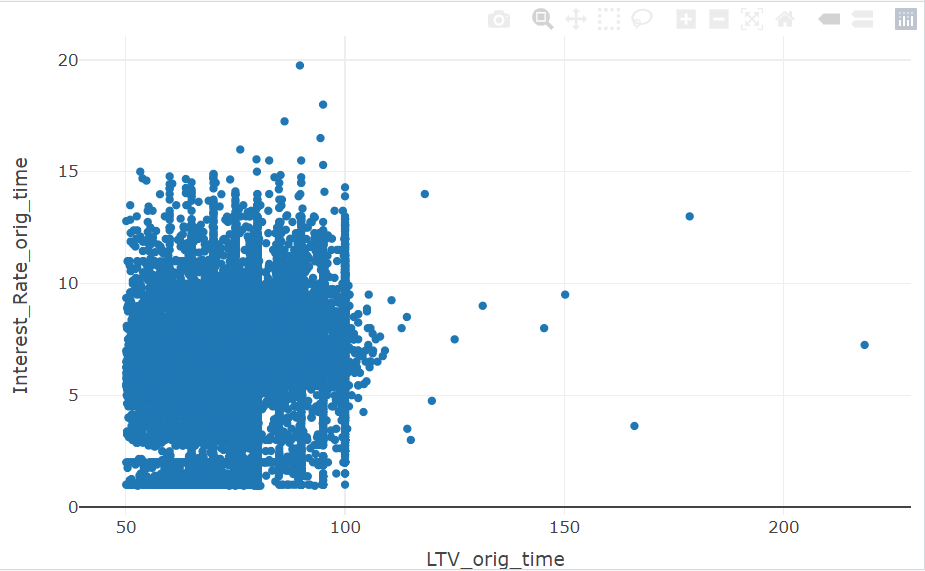
**Scatter plot for balance\_orig\_time and interest\_orig\_time:**

****

**Figure 5.2 Scatter plot for** **balance\_orig\_time and interest\_orig\_time**

By looking at the above figure 5.2 can be able to say that low balance\_orig\_time customers having the high interest rate for their loans and on the other hand high balance\_orig\_time customers having the low interest rate for their loan so, to get more profits bank need to focus on customers looking for less balance amount for loan.

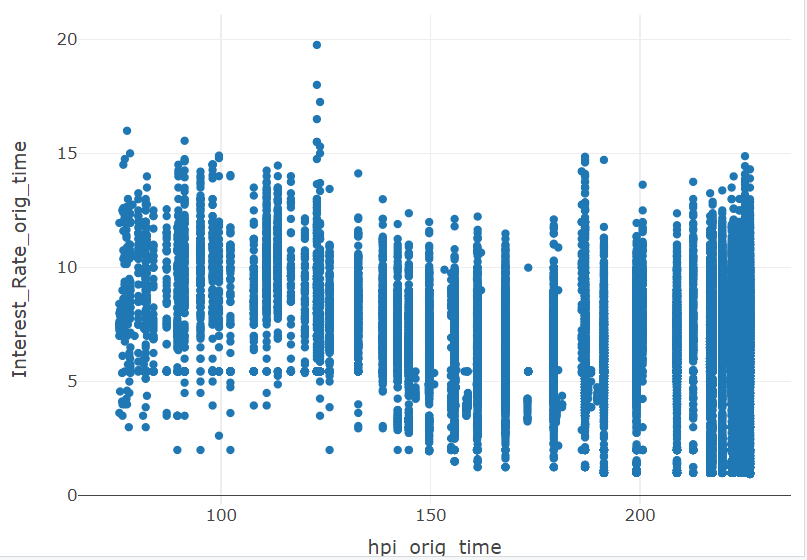
**Scatter plot for LTV\_orig\_time and interest\_rate\_orig\_time:**

****

**Figure 5.3 Scatter plot for LTV\_orig\_time and interest\_rate\_orig\_time**

By observing the above Figure 5.3, it appears that both variables exhibit a low correlation. As the LTV\_orig\_time increases, then the interest\_rate\_orig\_time also appears to decrease. To mitigate risk, banks may opt to offer lower interest rates to customers with higher LTV ratios.

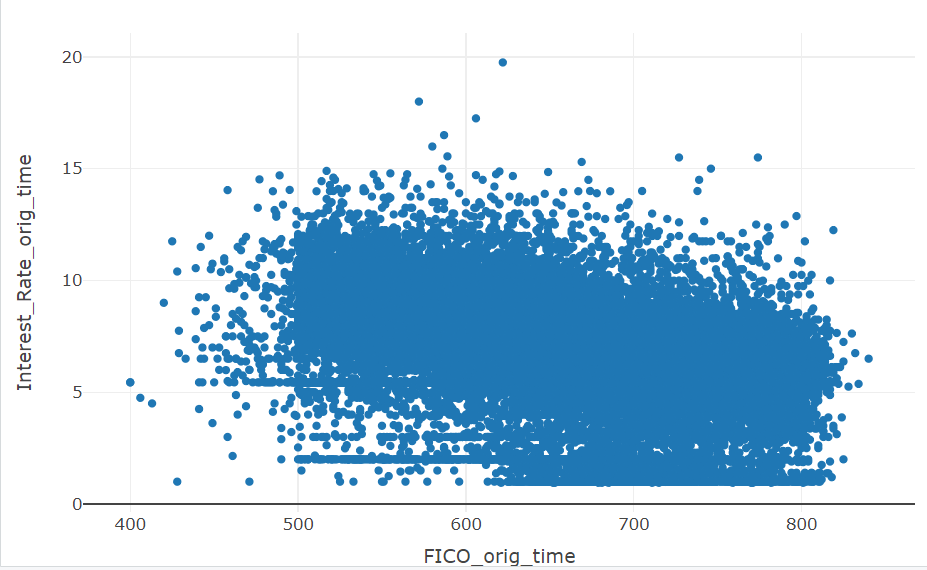
**Scatter plot for hpi\_orig\_time and interest\_rate\_orig\_time:**

****

**Figure 5.4 hpi\_orig\_time and interest\_rate\_orig\_time**

By observing Figure 5.4 above, it appears that there is not a significant relationship between hpi\_orig\_time and interest\_rate\_orig\_time. It may be concluded that interest rates can vary based on hpi\_orig\_time over time. So, it might not be a useful predictor for predicting the interest\_rate\_orig\_time.

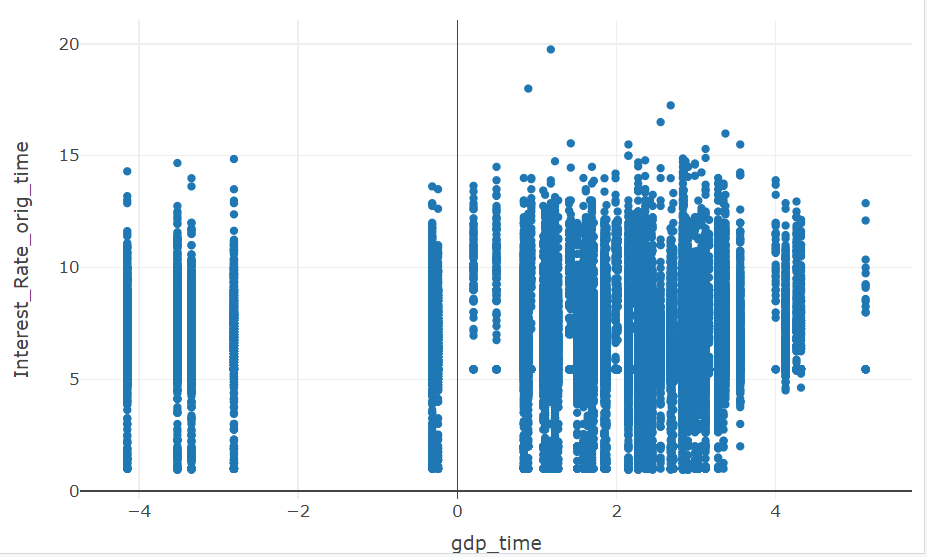
**Scatter plot for FICO\_orig\_rate and interest\_rate\_orig\_time:**

****

**Figure 5.5 Scatter plot for FICO\_orig\_rate and interest\_rate\_orig\_time**

Observing Figure 5.5 above, it is evident that customers with low FICO scores tend to have higher interest rates, while customers with high FICO scores tend to have lower interest rates. Specifically, customers with FICO scores less than 500 have higher interest rates, whereas those with FICO scores greater than 600 have lower interest rates.

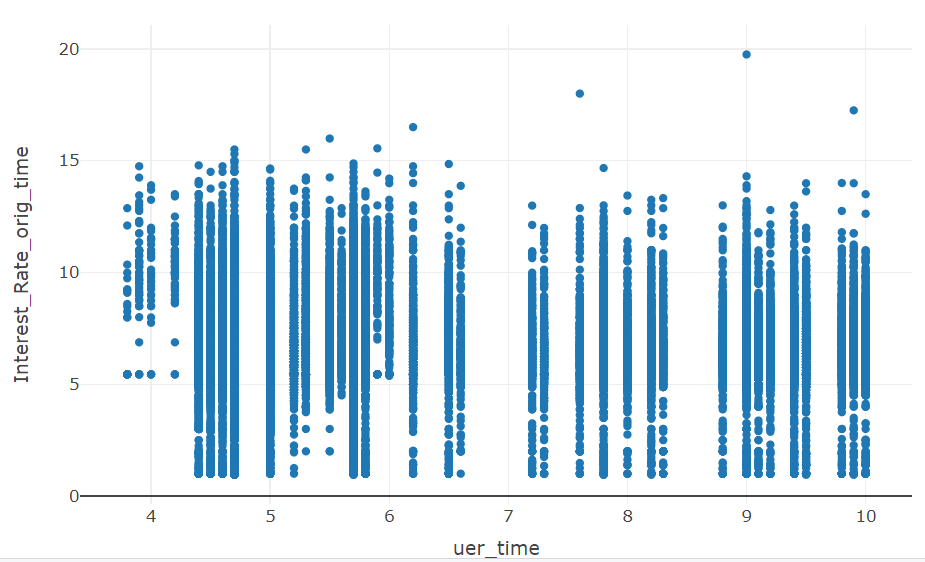
**Scatter plot for gdp\_time and interest\_rate\_orig\_time:**

****

**Figure 5.6 Scatter plot for gdp\_time and interest\_rate\_orig\_time**

By looking at the Figure 5.6 above, it appears that increasing gdp\_time may lead to decrease in interest\_rate\_orig\_time. The relationship between gdp\_time and interest\_rate\_orig\_time is influenced by financial market conditions; if gdp\_time increases, then interest rates may decrease. Specifically, based on financial market conditions, gdp\_time and interest\_rate\_orig\_time may either increase or decrease. Bank may offer less interest rate while gdp\_time increasing.

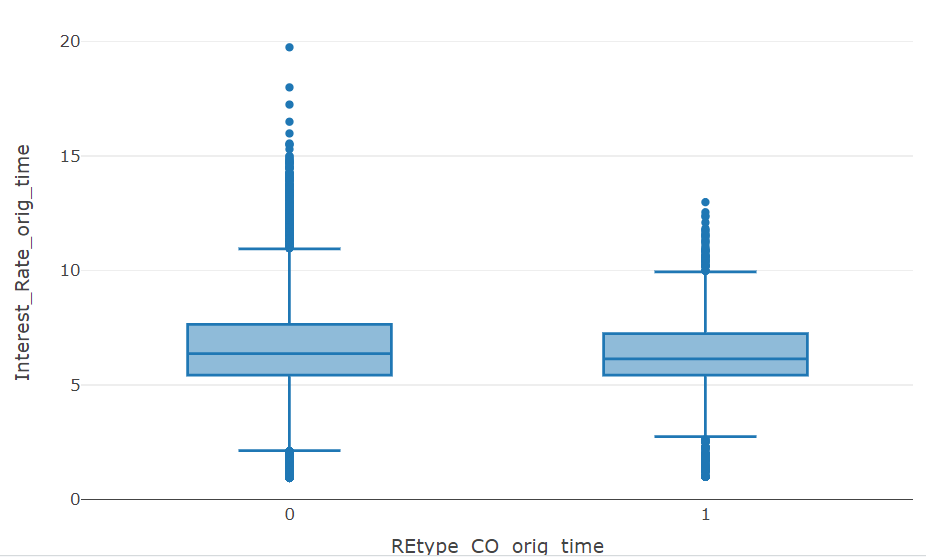
**Scatter plot for uer\_time and interest\_rate\_orig\_time:**

****

**Figure 5.7 Scatter plot for uer\_time and interest\_rate\_orig\_time**

Observing Figure 5.7 above, it appears that when the unemployment rate is less than 4%, the interest\_rate\_orig\_time is above 5%. However, it is noted that when the unemployment rate is more than 5%, the interest rate is not stable. Therefore, it can be concluded that the unemployment rate does not have a significant impact on the interest rate.

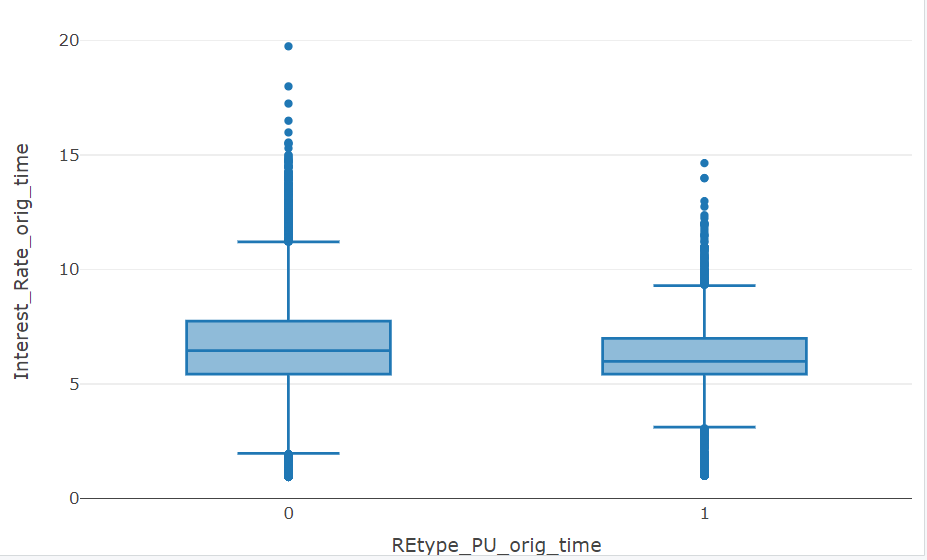
**Box plot for Retype\_co\_orig\_time and interest\_rate\_orig\_time:**



**Figure 5.8 Box plot for Retype\_co\_orig\_time and Interest\_rate\_orig\_time**

Observing Figure 5.8 above, it seems that condominium type houses may have lower interest rates compared to other types of houses.

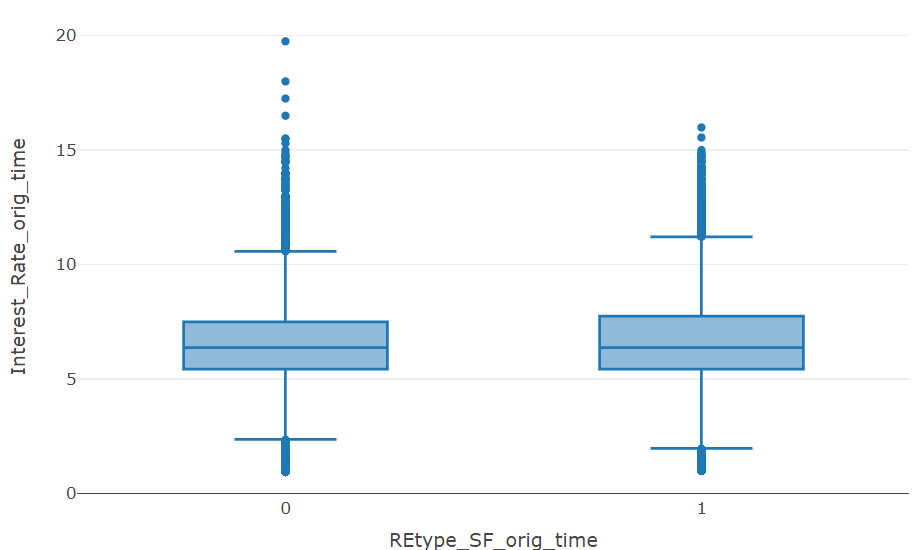
**Box plot for Retype\_pu\_orig\_time and interest\_rate\_orig\_time:**

****

**Figure 5.9 Box plot for Retype\_pu\_orig\_time and interest\_rate\_orig\_time**

By examining Figure 5.9 above, it appears that banks offer low interest rates to planned urban development houses, while other types of houses may have higher interest rates.

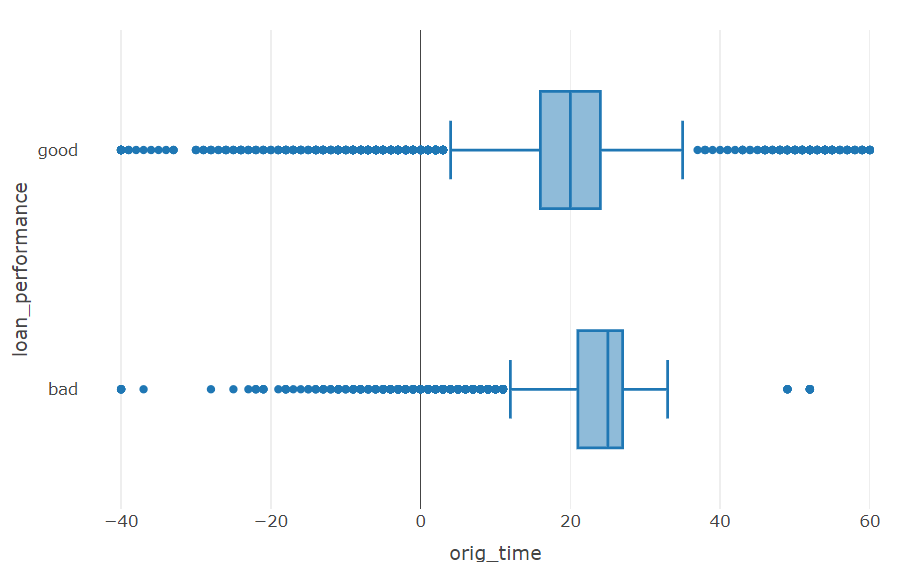
**Box plot for retype\_sf\_orig\_time and interest\_rate\_orig\_time:**

****

**Figure 5.10 Box plot for Retype\_SF\_orig\_time and interest\_rate\_orig\_time**

By examining Figure 5.10 above, it seems that single-family type houses have the highest interest rate compared to other types of houses. Therefore, to maximize profits, banks will need to focus on single-family type houses while providing the loan to the customers.

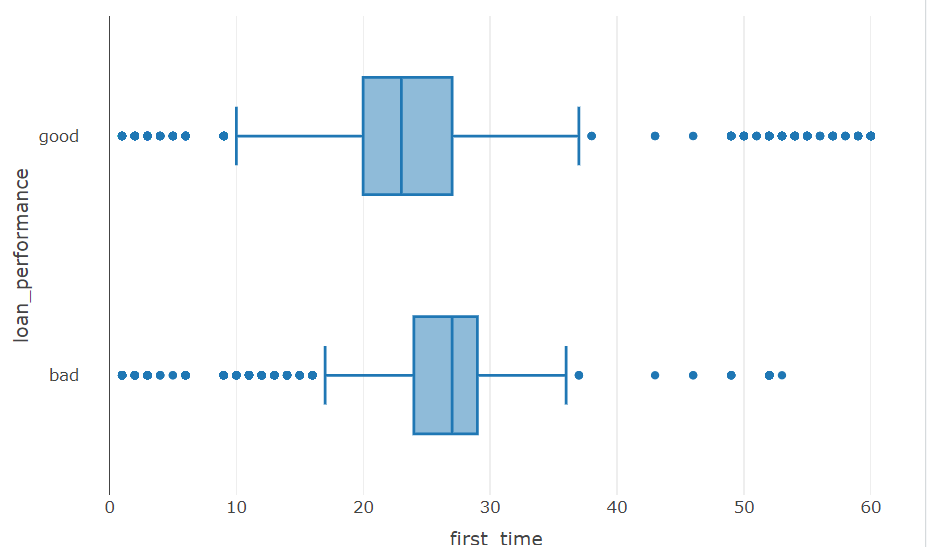
**Box plot for orig\_time and loan\_performance:**

**X54**

**Figure 5.11 Box plot for orig\_time and loan\_performance**

By examining Figure 5.11 above, it appears that when the loan was initiated, most of the customers' performance is good, indicating that most customers have not defaulted and are paying on time according to their loan agreement.

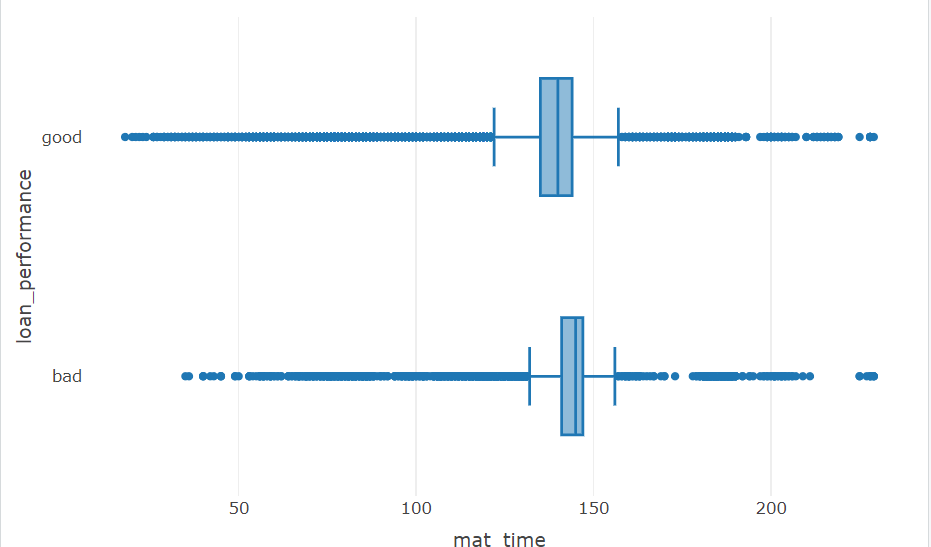
**Box plot for first\_time and loan\_performance:**

****

**Figure 5.12 Box plot for first\_time and loan\_performance**

Observing Figure 5.12 above, it can be said that most customers do not default when repaying their loans according to the loan agreement. Thus, the performance of most customers is good during their initial repayment.

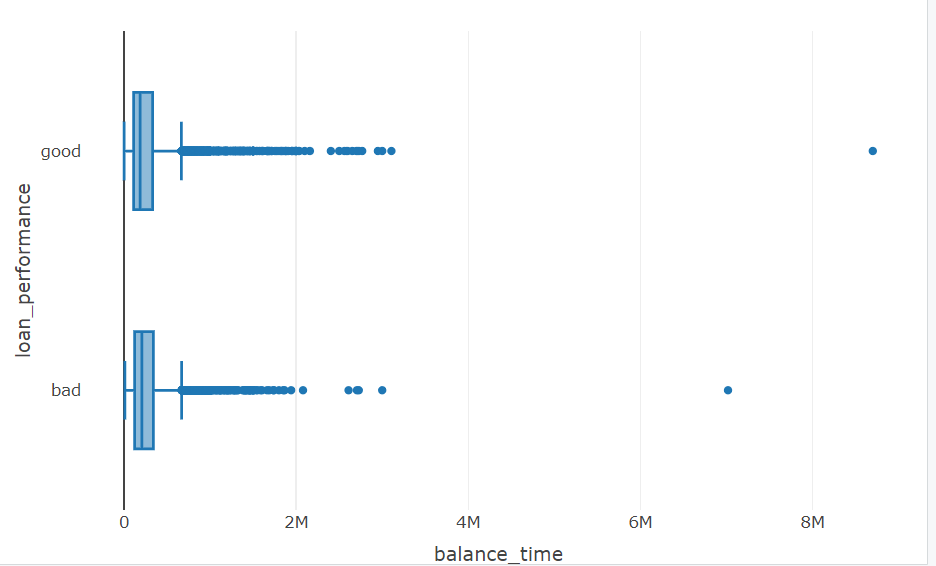
**Box Plot for mat\_time and loan\_performance:**

****

**Figure 5.13 Box Plot for mat\_time and loan\_performance**

Observing Figure 5.13 above, it appears that the performance of most customers is good between mat\_time 100 and 150.

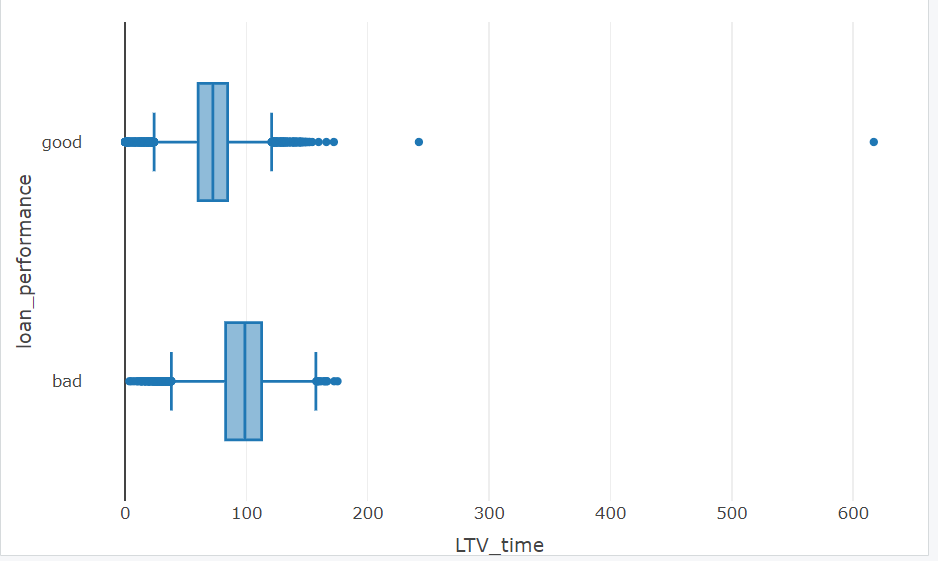
**Box plot for balance\_time and loan\_performance:**

****

**Figure 5.14 Box plot for balance\_time and loan\_performance**

Observing Figure 5.14 above, it can be seen that customers who took out a balance amount below 2M some customers with good performance and some customers with bad performance.

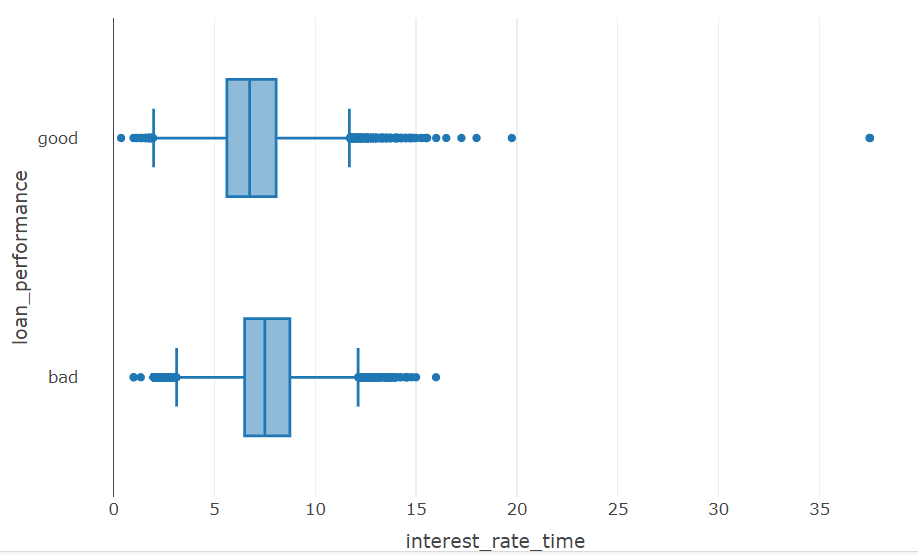
**Box Plot for LTV\_time and loan\_performance:**

****

**Figure 5.15 Box plot for LTV\_time and loan\_performance**

Observing Figure 5.15 above, it can be inferred that when LTV\_time is below 100, the performance of most customers is good, whereas when LTV\_time is above 100, the performance of customers is bad.

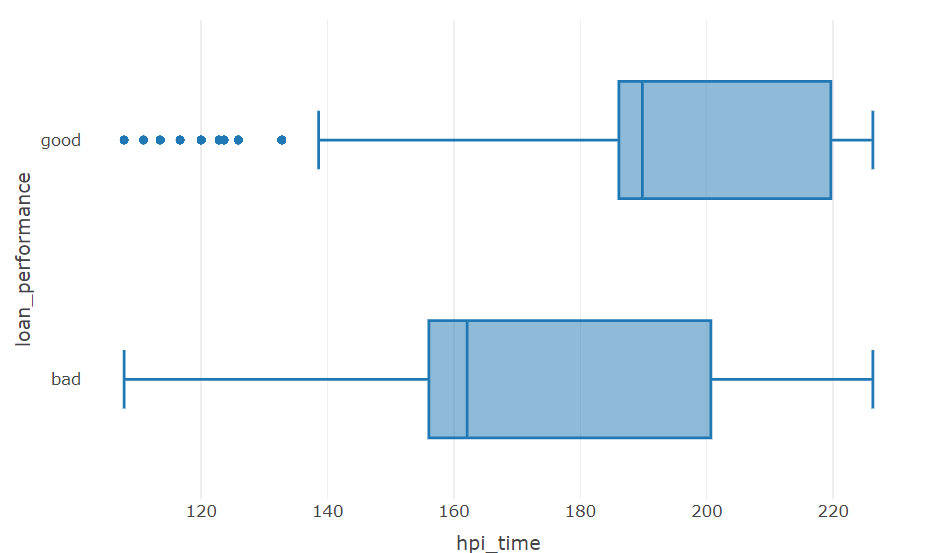
**Box plot for interest\_rate\_time and loan\_performance:**

****

**Figure 5.16 Box plot for interest\_rate\_time and loan\_performance**

Observing Figure 5.16 above, can be seen that customers with low interest rates have good performance and less risk to the bank, whereas customers with high interest rates have poor performance and more risk to the bank.

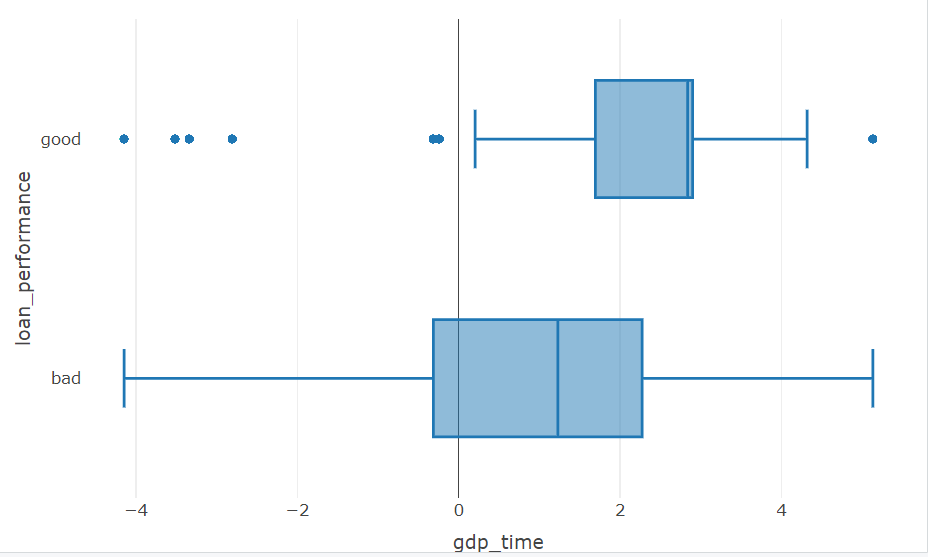
**Box plot for hpi\_time and loan\_performance:**



**Figure 5.17 Box plot for hpi\_time and loan\_performance**

Observing Figure 5.17 above, it can be seen that when the hpi\_time value is greater than 200, the performance of the customers is good, whereas when the hpi\_time value is less than 200, the performance of the customers is bad.

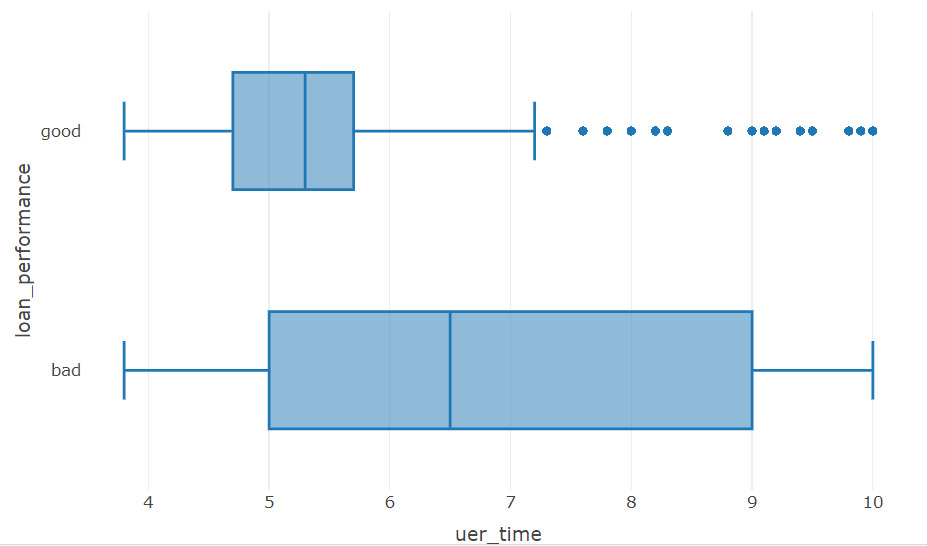
**Box plot for gdp\_time and loan-performance:**

****

**Figure 5.18 Box plot for gdp\_time and loan\_performance**

Observing Figure 5.18 above, when the gdp\_time is greater than 2%, the performance of most customers is good, whereas when the gdp\_time is between 0% and 2%, the customers performance is bad. Thus, the gdp\_time is impacting the performance of the customer loan.

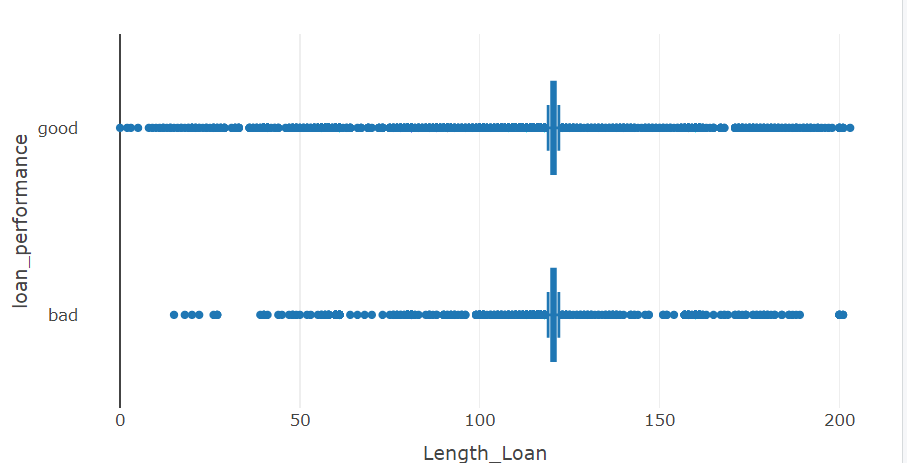
**Box plot for uer\_time and loan\_performance:**

****

**Figure 5.19 Box plot for uer\_time and loan\_performance**

By looking at the above figure 5.19 can be seen that when the unemployment rate is below 6% then the performance of the customer loan is good and when the unemployment rate is more than 6% then most of the customers loan performance is bad. Can clearly say that unemployment rate is impacting the loan performance.

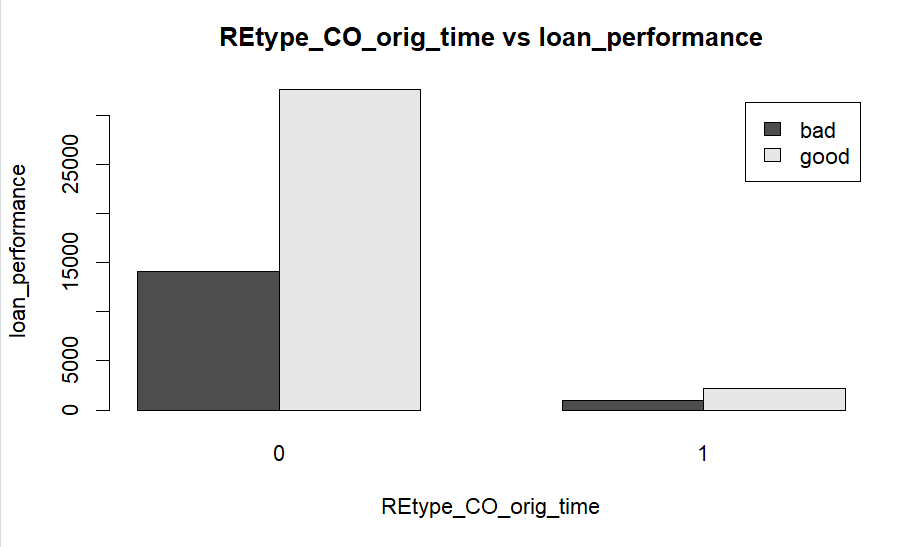
**Box Plot for length\_Loan and loan\_performance:**

****

**Figure 5.20 Box Plot for Length\_Loan and loan\_performance**

By looking at the figure 5.20 can be able say Loans with a length of over 100 typically exhibit both good and bad performance customer loans.

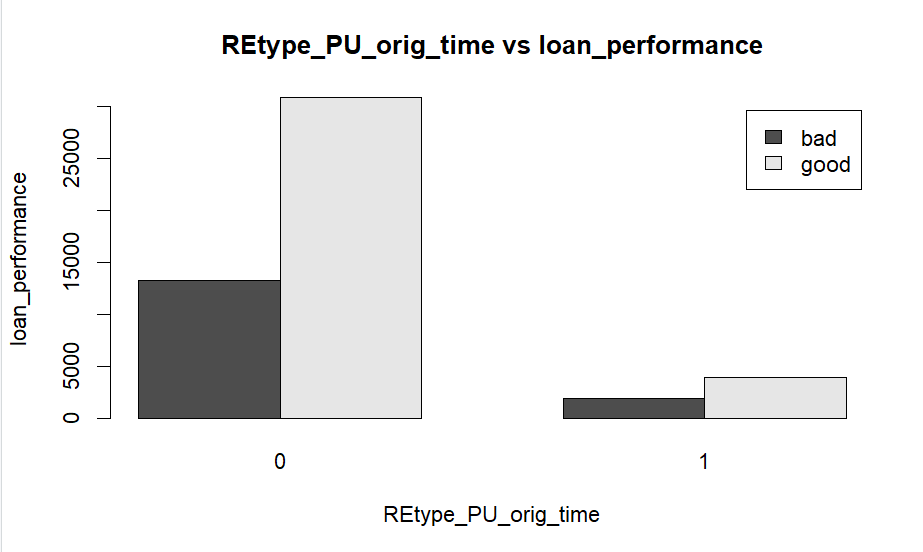
**Bar plot for Retype\_CO\_orig\_time and loan\_performance:**



**Figure 5.21 Bar plot for Retype\_CO\_orig\_time and loan\_performance**

By looking at the above figure 5.21 can be seen that customers who having the condominium type property their loan performance is good when compared with other type of the property.

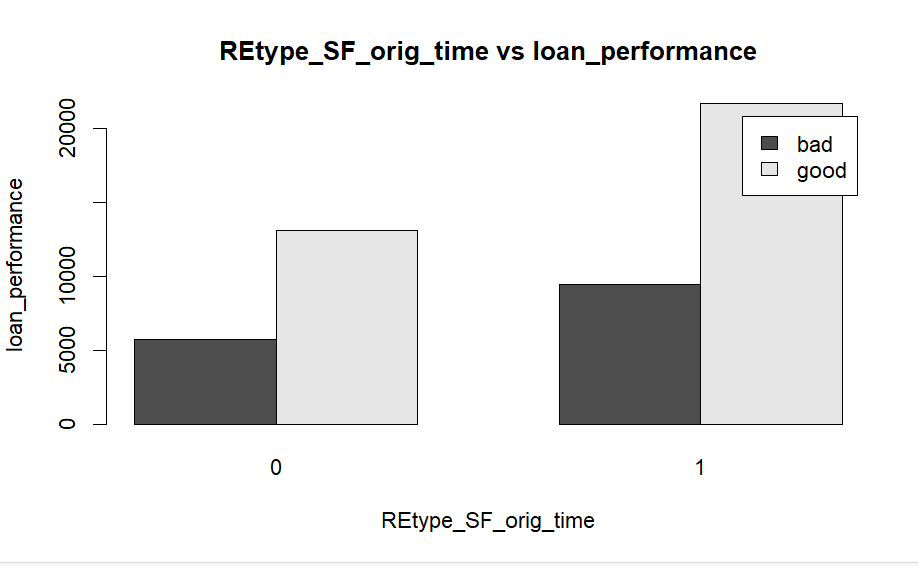
**Bar plot for Retype\_CO\_orig\_time and loan\_performance:**

****

**Figure 5.22 Bar plot for Retype\_PU\_orig\_time and loan\_performance**

By looking at the above figure 5.22 can be seen that customers who having the planned urban development type property their loan performance is good when compared with other type of the property.

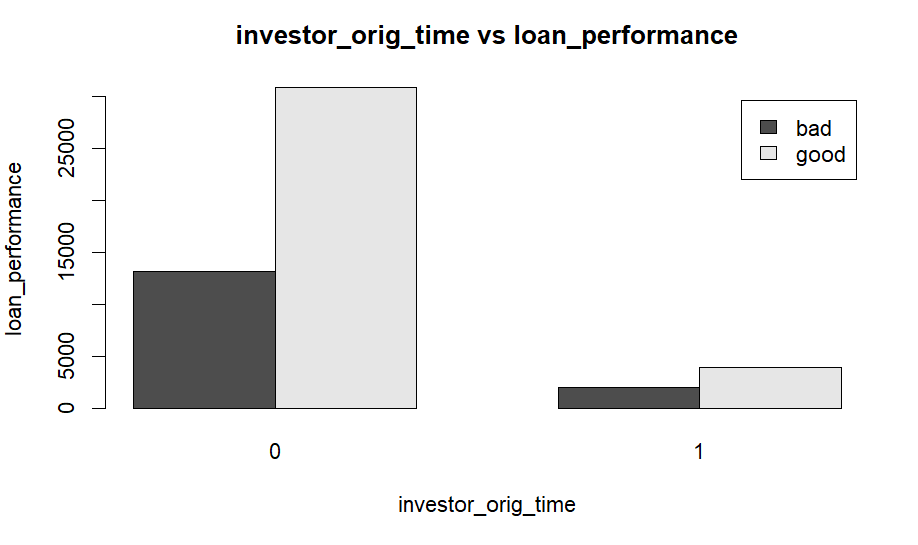
**Bar plot for Retype\_PU\_orig\_time and loan\_performance:**



**Figure 5.23 Bar plot for Retype\_PU\_orig\_time and loan\_performance**

By looking at the above figure 5.23 can be seen that customers who having the single-family type of property their loan performance is good when compared with other type of the property.

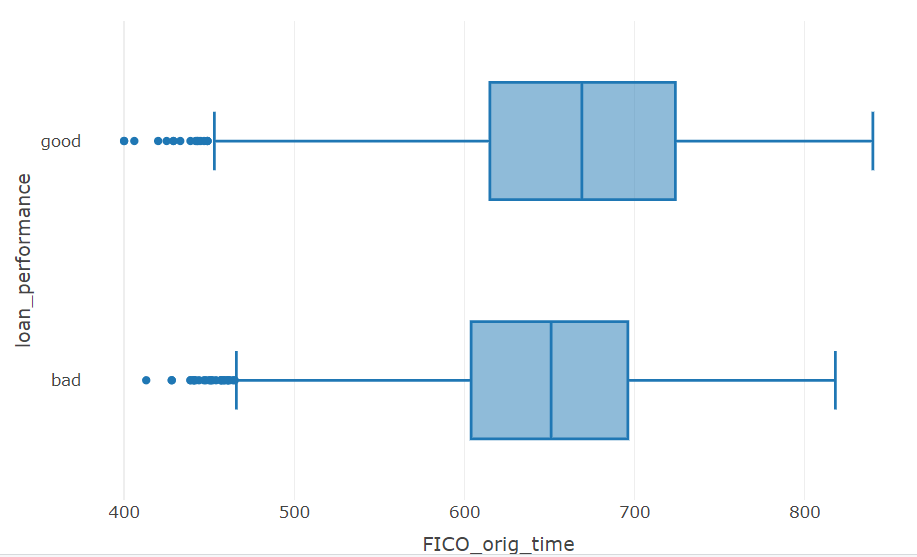
**Bar plot for investor\_orig\_time and loan\_performance:**



**Figure 5.24 Bar plot for investor\_orig\_time and loan\_performance**

Observing Figure 5.24 above, it is evident that investors have good loan performance compared to non-investors. Conversely, non-investors exhibit poor loan performance.

**Box plot for FICO\_orig\_time and Loan\_performance:**



**Figure 5.25 Box plot for FICO\_orig\_time and loan\_performance**

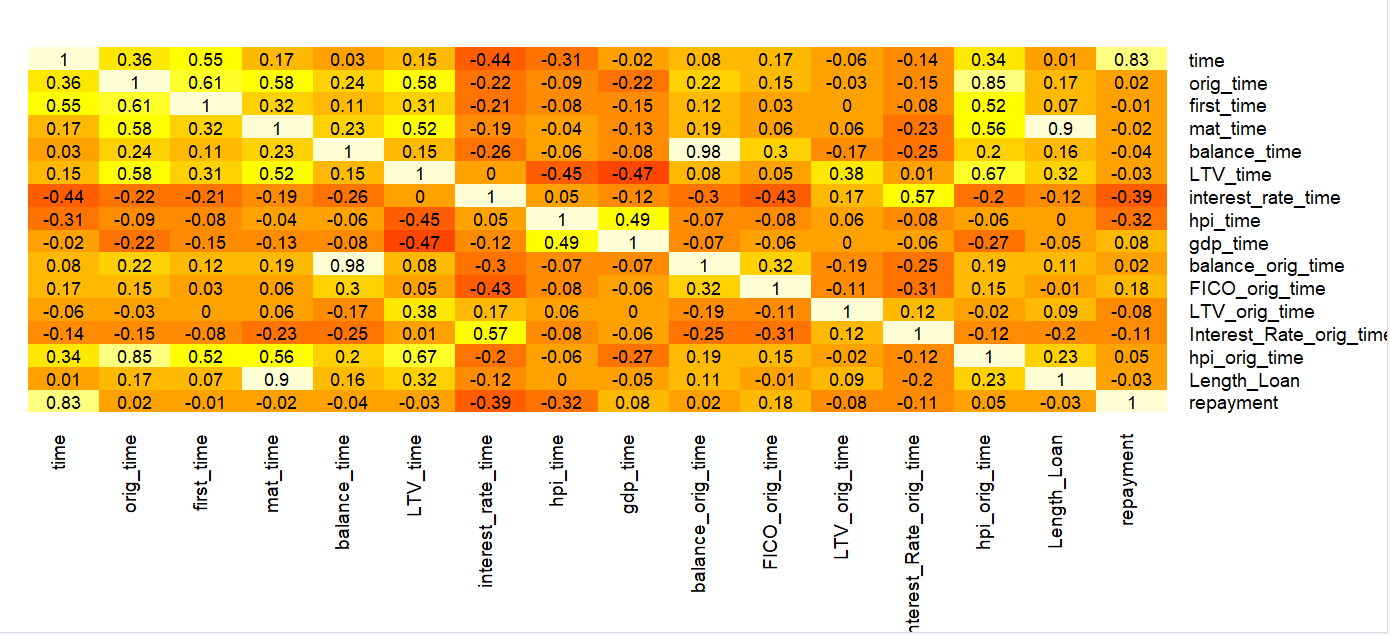
By looking at the above figure 5.25 can be seen that customers who having the high FICO score their loan performance is good and customers who having the low FICO score their loan performance is bad.

After conducting the predictor analysis, it was observed how each variable impacts the outcome variable, and scatter plots provided useful information about the variables and their correlation with the target columns. Able to identify how each variable varies over time. These visualizations collectively contribute to a comprehensive understanding of the data and its key patterns. By looking at the bar plots, the performance of the customers at each variable can be seen.

After conducting exploratory data analysis, outliers were identified in the dataset. However, it was decided not to remove these outliers because each outlier represents the unique characteristics or behaviour of individual customers. As such, removing outliers could potentially distort the true nature of the data and affect the accuracy of the analysis. Therefore, the decision was made to retain the outliers in the dataset.

**5.1 CORRELATION BETWEEN THE NUMERICAL COLUMNS:**

Let’s have a look on the correlation between the numerical columns



**Figure 5.1.1 Correlation between the numerical columns**

By looking at the above figure 5.1.1 can be seen that correlation between the variables. Orig\_time and hpi\_orig\_time is showing the high positive correlation with each other, LTV\_time, first\_time and mat\_time with hpi\_orig\_time is showing the positive correlation with each other. Observing that there is no correlation between the interest\_rate\_time and LTV\_time. From the correlation plot can say that there is a weak correlation with all attributes in the data.

**6.DIMENSIONALITY REDUCTION:**

The dataset encompasses all relevant variables related to the target column, and each column holds its significance. Given the low-dimensional nature of the data, there is no need to employ dimensionality reduction techniques.

**For regression task:**

It is sensible to select only a subset of variables that are expected to have a significant impact on predicting the interest rate.

The chosen variables, such as balance\_orig\_time, LTV\_orig\_time, hpi\_orig\_time, FICO\_orig\_time, gdp\_time, uer\_time, retype\_co\_orig\_time, retype\_pu\_orig\_time, and retype\_sf\_orig\_time, length\_loan is likely to provide valuable insights into the relationship with the target variable. Based on these factors will predict the interest rate while approving the loan. Remaining columns not required to predict the interest because that information related while repaying the loan. For predicting interest rate to new customer this information is enough.

**For classification task:**

For classification task the chosen variables such as balance\_orig\_time, LTV\_orig\_time, hpi\_orig\_time, FICO\_orig\_time, gdp\_time, uer\_time, retype\_co\_orig\_time, retype\_pu\_orig\_time, and retype\_sf\_orig\_time, length\_loan, Repayment, loan\_performance, balance\_time, LTV\_time, hpi\_time, interest\_rate\_time, and interest\_rate\_orig\_time will be the target column. In a classification task, will exclude ongoing customers. Without proper information on their status, can't classify whether their performance is good or bad. They're still making repayments, but don't know if they'll default or pay off their debt at some point. In the future, if those ongoing customers pay off their loan on time, they will be considered as good performance customers; otherwise, they will be considered as bad performance customers.

Excluding the remaining variables is logical as they do not directly contribute to determining loan performance and may not provide relevant information for classification. To classify the performance loan these variables information is enough so choose only these variables for classification.

**7. DATA TRANSFORMATION:**

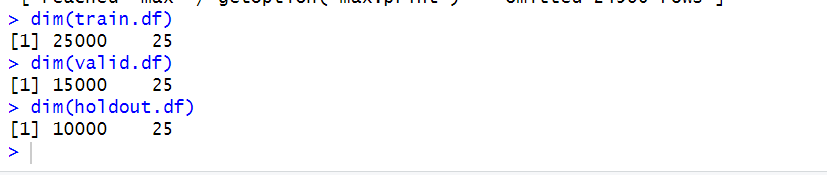
Data transformation is a crucial step in preparing data for machine learning models, and the nature of these transformations can vary based on the requirements of the selected model. For example, when applying a k-nearest neighbours (KNN) model, it may be necessary to scale features and create dummy variables. Conversely, when using logistic regression, converting categorical columns into factors might be essential before constructing the model. Therefore, the data preparation process is customized to meet the specific needs and assumptions of the chosen model. Further analysis has created two new columns one is length of loan derived from the maturity time and orig\_time. Another column is loan\_performance based on the default, paid off and not paid off. If the customer is default then consider as performance of the customer loan is bad and if the customer is paid off, not defaulted then consider as performance of the customer loan is good. Another column created based on the difference between orig\_time and time which tells us initial loan repayment of the customers.

**8. DATA PARTITIONING METHODS:**

To choose the best model that performs optimally in classifying the outcome variable of interest with the available data, it is crucial to avoid introducing optimism bias. This bias arises when the model is developed and assessed using the same dataset, potentially leading to issues in real-time scenarios. To mitigate overfitting problems, the data is divided into three partitions: train, validation, and holdout.

The training partition, being the largest, contains data used to build models for examination. This same training data is utilized to develop multiple models. The validation partition assesses the predictive performance of each model, allowing for model comparison and selection of the best one. In some algorithms, the validation partition may also be used to tune and improve the model. The holdout partition is then employed to assess the performance of the chosen model with new data.

The data is split randomly into three parts: train, validation, and holdout partitions, with the training partition comprising 50%, the validation partition 30%, and the holdout partition 20%. After this partitioning, the records in each set should be distinct. The training data consists of 25,000 observations and 26 columns, the validation data contains 15,000 observations and 26 columns, and the holdout data comprises 10,000 observations and 26 columns.



**8.1 Data Partitioning**

**9. CLASSIFIER MODEL SELECTION:**

Will going to construct a classification predictive model to assess whether the performance of a customer's loan is classified as either good or bad. The model's output will facilitate informed decisions regarding loan approvals for existing customers seeking re-loans. Additionally, notifications will be sent to customers demonstrating poor loan performance, prompting adherence to the loan agreement terms.

The target variable, loan performance, indicates whether a customer's loan performance is deemed good or bad based on their repayment behaviour. balance\_orig\_time, LTV\_orig\_time, hpi\_orig\_time, FICO\_orig\_time, gdp\_time, uer\_time, retype\_co\_orig\_time, retype\_pu\_orig\_time, retype\_sf\_orig\_time, length\_loan, repayment, loan\_performance, balance\_time, LTV\_time, hpi\_time, interest\_rate\_time and interest\_rate\_orig\_time is likely to provide valuable insights into the relationship with the target variable. These variables play a major role to classify the performance of the loan performance.

The model's performance will be evaluated using the confusion matrix, with a specific focus on accurately identifying the class associated with high resale value in the market. To achieve this classification objective, logistic regression, classification tree, KNN, and neural networks will be explored.

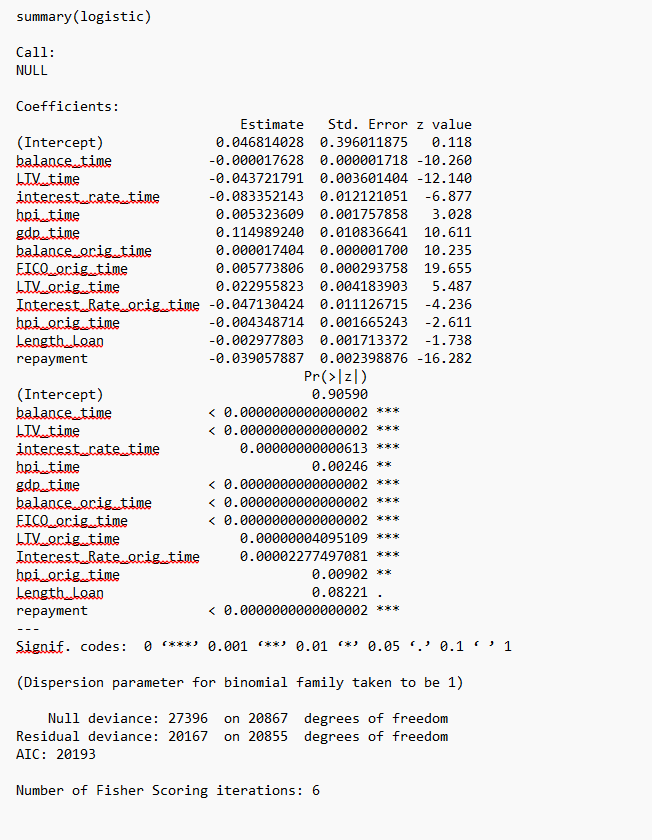
For classification models not going to use entire data and excluding the ongoing customers because don’t know their status either default and paid off and how it will be in future. So, it is difficult to classify those customers.

**9.1 LOGISTIC REGRESSION:**

Logistic regression is the most popular classification model. In this model, use the glm() function. To access this function, need to install and import the “Car” and “Caret” packages. Logistic regression extends the idea of linear regression to situations where the outcome variable is categorical. Logistic regression estimates probabilities, indicating the likelihood of each class. It uses a threshold value to classify the record into each class.

**9.1.1 PREPARING THE DATA FOR LOGISTIC REGRESSION**

Before applying logistic regression, will encode the factors for categorical columns and then proceed to fit the model. First, will fit the model with the training data and subsequently evaluate its performance with the validation data. Let's examine the summary of the logistic regression in Figure 9.1.1, which is presented in the form of linear output, to be later converted into probability. Based on the threshold value, it will assign the class.

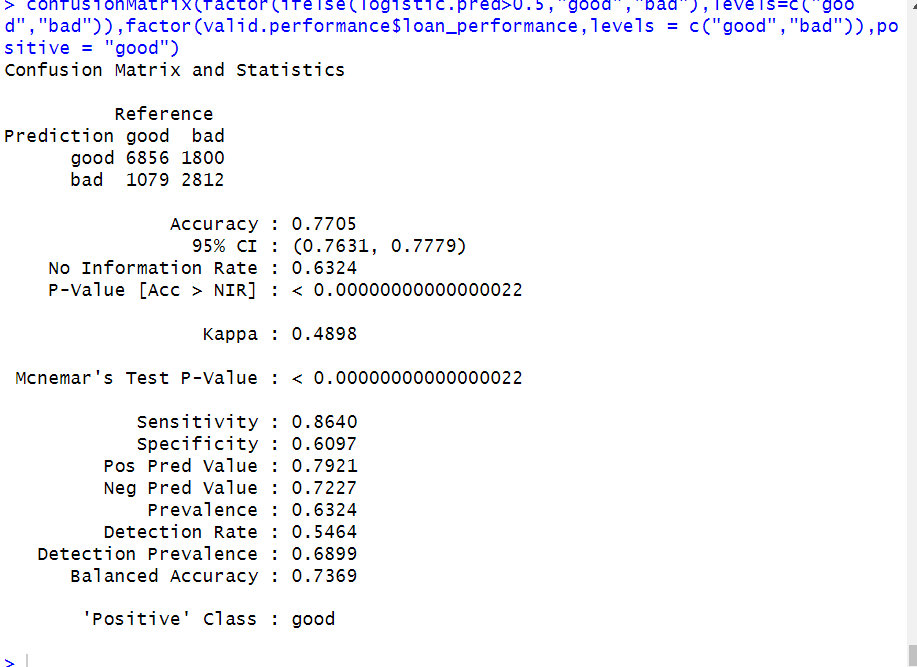
****

**9.1.1 Summary of the logistic regression**

From the summary of the logistic regression will get the estimated equation and can be able to see the significance of the variables and which variables are most important to predict the performance of the loan. The greater number of this “\*” symbol indicates the how much important those variables to classify the loan performance. The lower number AIC value indicates the good fitting models. The number of fishers scoring iterations is 6. By using that estimated equation can be able to classify the new customer performance.

**9.1.2 EVALUATING THE PERFORMANCE:**

After fitting the model, we will evaluate its performance using a confusion matrix, which shows how accurately the model predicts classifications based on a given threshold. The confusion matrix helps identify the number of correct and incorrect predictions made by the classifier. In this context, a threshold of 0.5 is used, meaning that predictions with probabilities above 0.5 are classified as 'good' customer loan performance, while those with probabilities below 0.5 are classified as 'bad' performance. For SBS Bank, the key focus is on identifying 'good' performance customers, as this helps mitigate risk and avoid high-risk customers when deciding on offering loans. By using this classification, the bank can better determine which types of customers are more likely to repay loans, allowing them to make informed lending decisions in the future.

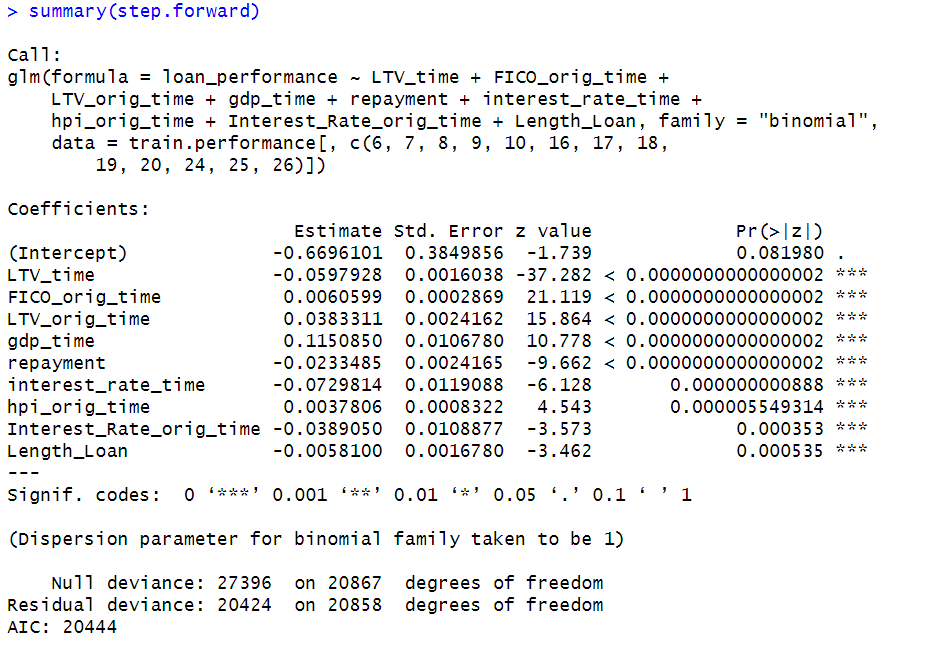


**Figure 9.1.2 Confusion matrix of the Logistic Regression**

By examining the confusion matrix in Figure 9.1.2 for logistic regression, it reveals an accuracy of 77.05%, sensitivity of 86.40%, and specificity of 60.97%. Evaluating the model's performance based on sensitivity, can conclude that there is an 86.40% chance of correctly classifying the positive class. So, this model suggests that there is high possibility of predicting the good performance of the customers to make more profits and mitigate the risk.

**Improving model performance by using forward step wise logistic regression:**

In forward step wise logistic regression, it starts with no predictors and add predictors one by one. Let’s have a look on the summary of the forward stepwise logistic regression

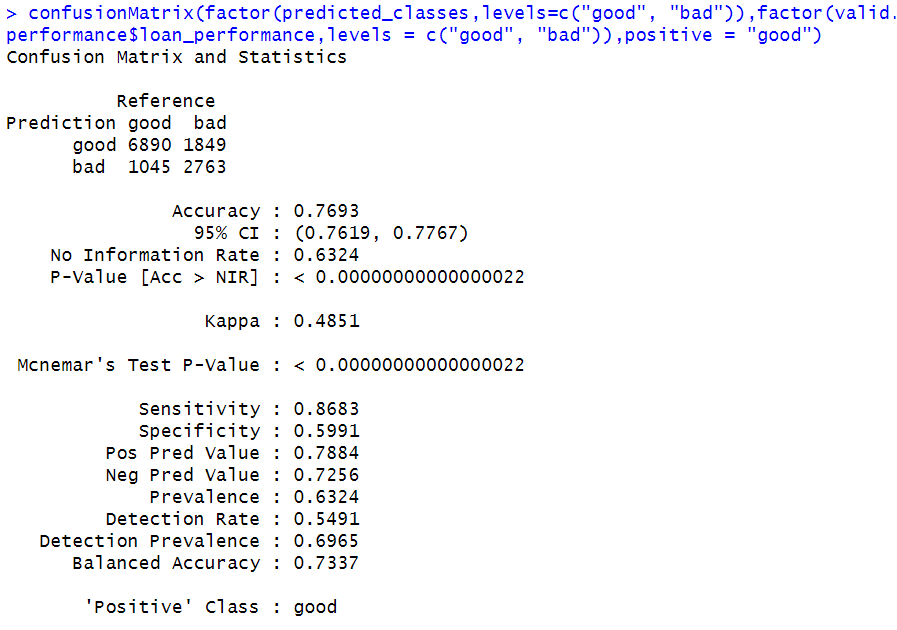


**Figure 9.1.3 Estimated equation of forward stepwise logistic regression**

From this summary of forward stepwise logistic regression can be able to see the most important predictors to classify the performance of the loan and shows the significance of the variables. The number of this “\*” indicates how much important those variables to classify the performance. The lower AIC indicates the good fitting of the model. Forward step wise logistic regression will generate estimated equation by using that can be able to classify the new customer.

**Evaluating the model performance:**

Let’s have a look on the confusion matrix



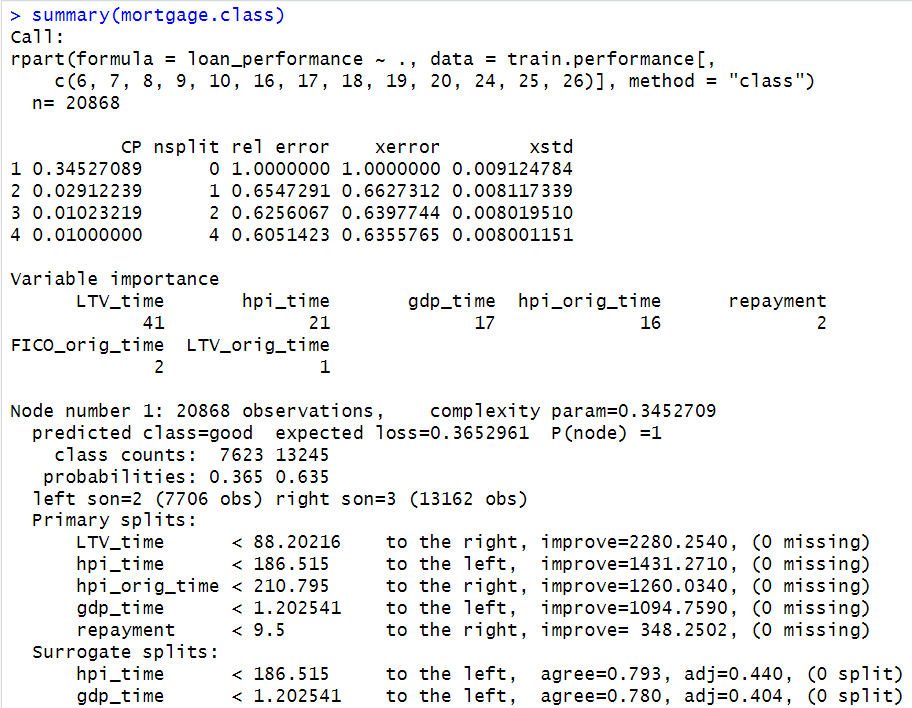
**Figure 9.1.4 Confusion matrix of forward step wise logistic regression**

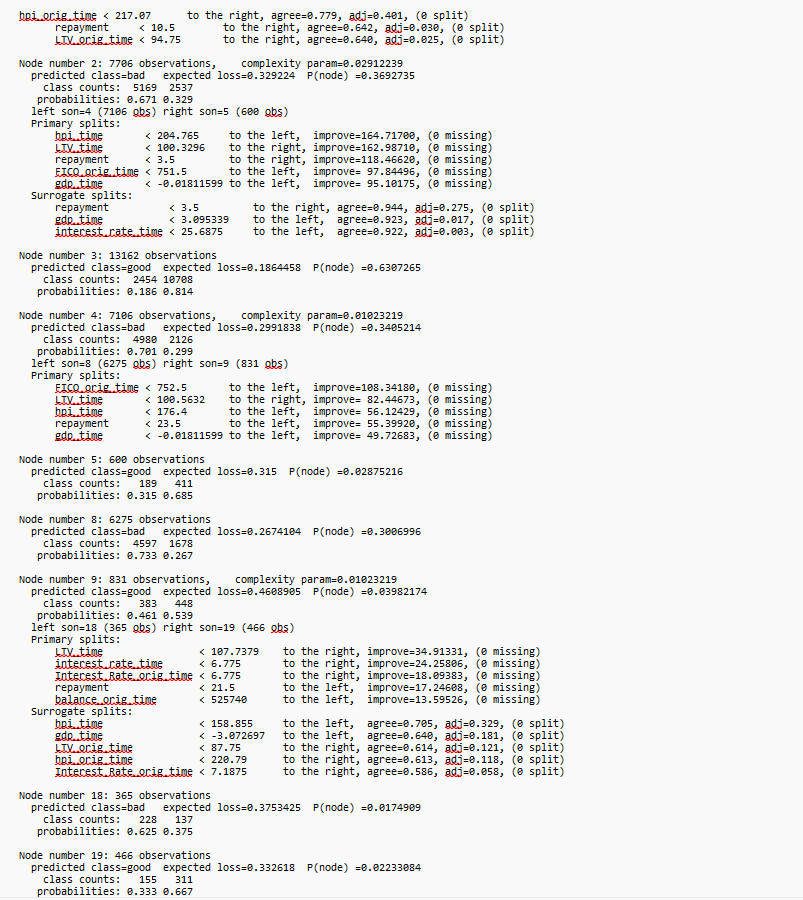
By examining the confusion matrix in Figure 9.1.4 for logistic regression, it reveals an accuracy of 76.93%, sensitivity of 86.83%, and specificity of 59.91%. Evaluating the model's performance based on sensitivity, can conclude that there is an 86.83% chance of correctly classifying the positive class. So, this model suggests that there is high possibility of predicting the good performance of the customers to make more profits and mitigate the risk.

**9.2 CLASSIFICATION TREE:**

The classification tree is a widely used method for categorizing record outputs based on the rules represented on the tree. It considers most votes when classifying a new record. The process involves recursively splitting the dataset into rectangles, with each split determined by the most important variable at that point, aiming to make each rectangle as homogeneous or pure as possible.

To implement this, utilize the 'rpart' and 'rpart.plot' packages in R to construct and visualize the classification tree. Initially, the model is built using the training data, and subsequently, the interpretation of the model output is examined.

****

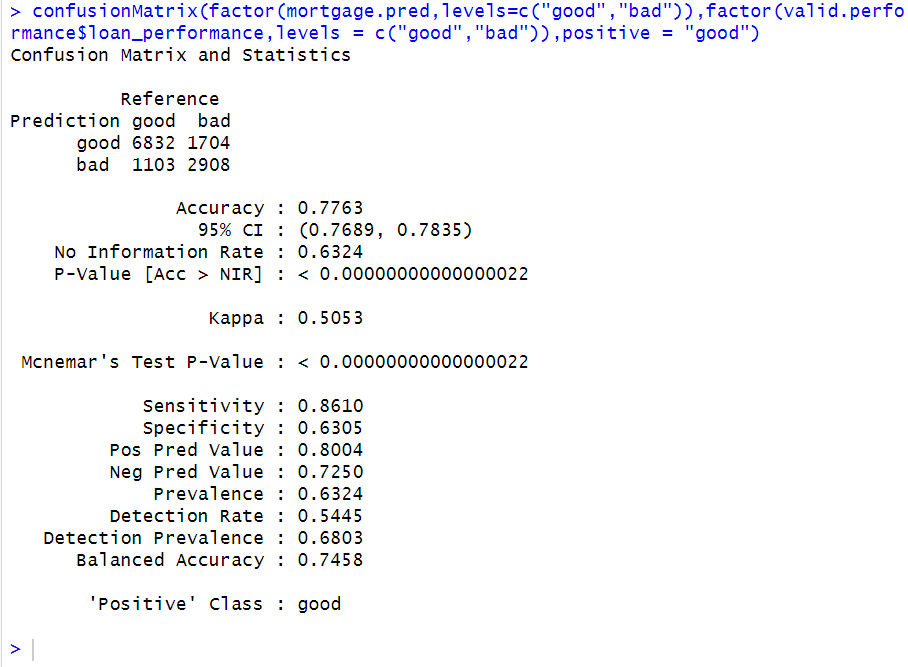
****

**Figure 9.2 Summary of the classification tree**

By looking at the above figure 9.2 summary of the classification tree can understand each leaf node represent the classification rule. Based on this will get to know how it partitions the data based on predictor variables and makes predictions for each partition.

**9.2.1 EVALUATING THE PERFORMANCE:**

After constructing the model using the training data, the performance will be evaluated with the validation data using the confusion matrix. Assessing the model's performance is based on metrics from the confusion matrix, particularly focusing on sensitivity. Sensitivity measures the model's capability to correctly classify the positive class, signifying a higher ability to predict good performance customer loans. Here the positive class is good, it is very important to bank to classify the good performance customers and they can avoid the risky customers. By relying on this, SBS bank can make decisions regarding which types of customers they can provide the reloan when they are in need to avoid the risk factor, allowing them to concentrate more on those and maximize profits. Now, let's examine the confusion matrix.

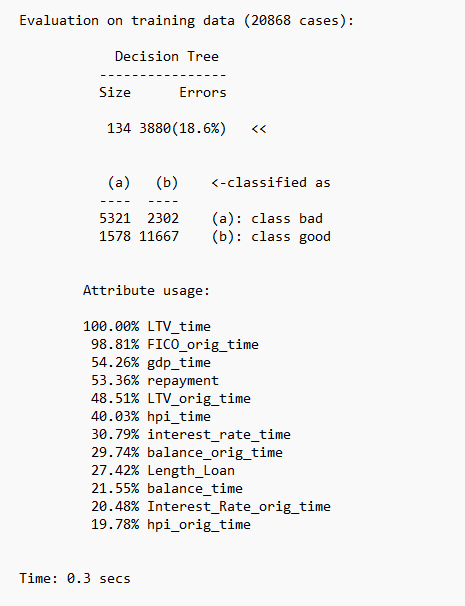


**Figure 9.2.1 Confusion matrix for Classification Tree**

By examining the confusion matrix in Figure 9.2.1 for the classification tree, it can be observed that the accuracy is 77.63%, sensitivity is 86.10%, and specificity is 63.05%. By considering the sensitivity, it can be stated that there is an 86.10% chance of correctly classifying the positive class, indicating the likelihood of correctly classifying the good performance of the customer loans. Consequently, based on this result, the SBS can make informed decisions regarding customer loans based on their performance while repaying the loan amount.

**9.2.2Classification Tree by using C5.0:**

This is another approach to in classification tree by using the C5.0. First will build the model with the training data then will interpret the results of the model

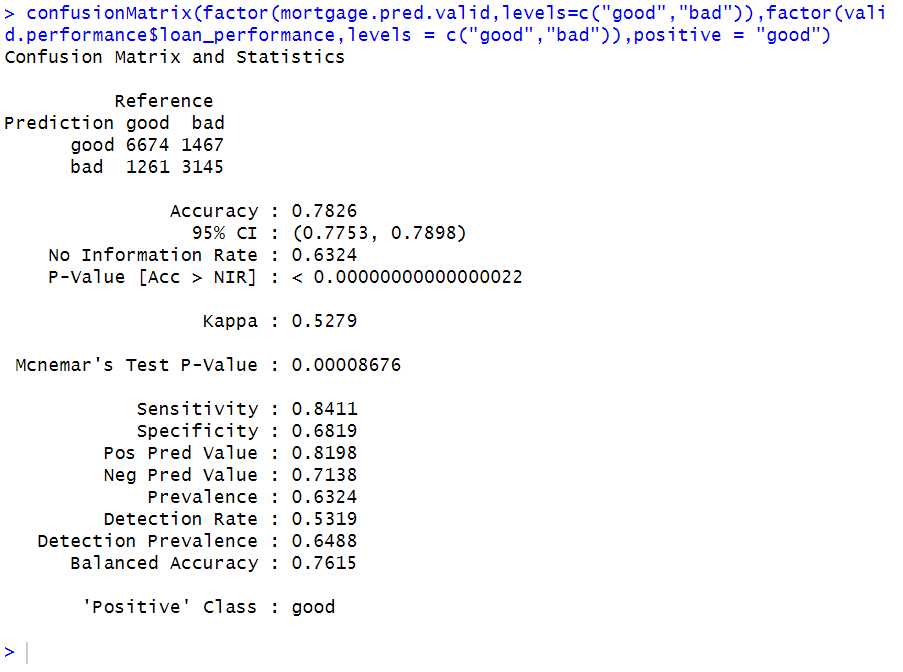


**Figure 9.2.2 evaluation on training data by using C5.0 model**

By looking at the above figure 9.2.2 evaluation of the classification tree using c5.0 can understand which attributes used for training. Based on this will get to know which attributes more important for prediction and makes predictions for each partition. In future it will be helping to bank people instead of collecting all the data they concentrate on important attributes.

**9.2.3 EVALUATE THE PERFORMANCE USING C5.0:**

Now evaluate the performance with the validation data through the confusion matrix. Based on the confusion matrix metrics will evaluate the performance of the model. Let’s have a look on the confusion matrix.

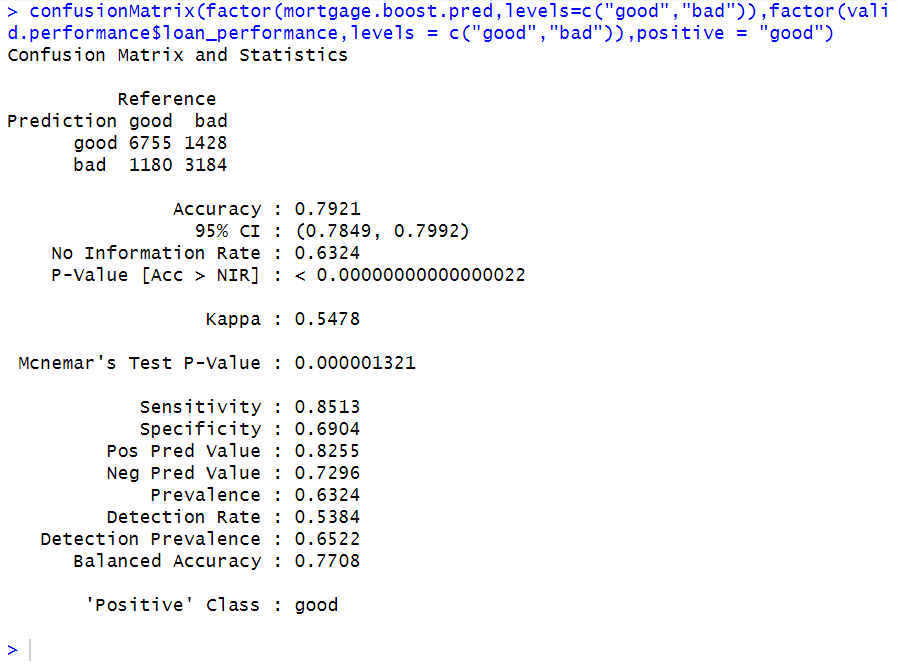


**Figure 9.2.3 Confusion Matrix for classification tree using C5.0**

By examining the confusion matrix in Figure 9.2.3 for the C5.0 model, it can be observed that the accuracy is 78.26%, sensitivity is 84.11%, and specificity is 68.19%. Based on the sensitivity results, it can be stated that there is an 84.11% chance of correctly classifying the good performance of the customer loans. So, the SBS can make informed decisions based on the performance of the customers and take the action according to the terms and conditions of the agreement. Identifying customers with good performance is crucial for SBS, as it helps them manage risk. The bank can focus on these customers while also taking specific measures to address those with bad performance, ultimately aiming to mitigate risk.

**Improving model performance by using trial option:**

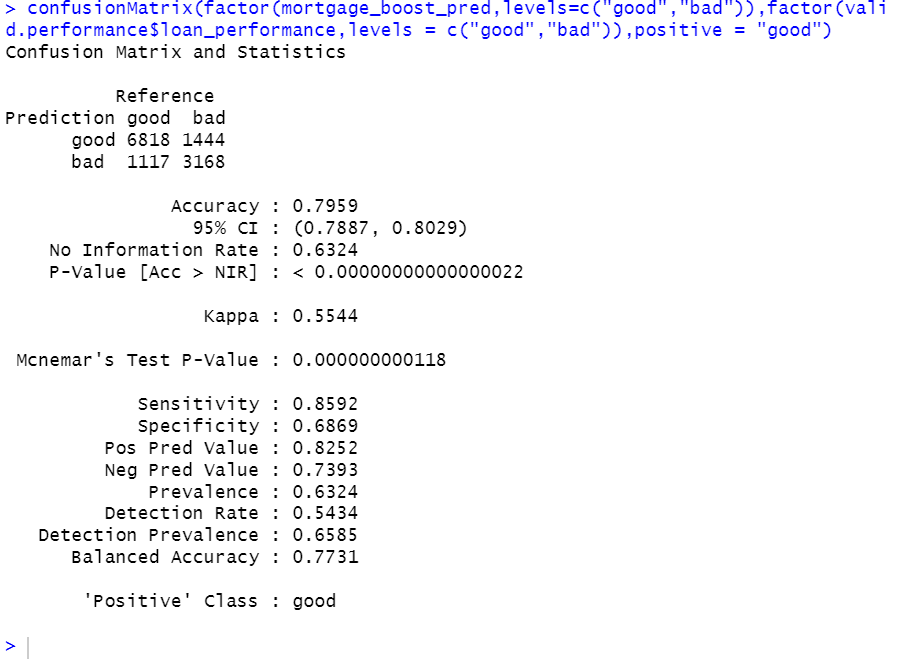
By using this trial options can be able to increase the model performance. Let’s have a look on the confusion matrix of the C5.0 model with trial 6



**Figure 9.2.4 confusion matrix of the C5.0 model with trial 6**

By examining the confusion matrix in Figure 9.2.4 for the C5.0 model, the accuracy is 79.29%, sensitivity is 85.13%, and specificity is 69.04%. The sensitivity results indicate that there is an 85.13% chance of correctly classifying customer loans with good performance. This allows SBS to make informed decisions about customer performance and take appropriate action based on the bank's terms and conditions. Identifying customers with good performance is crucial for SBS, as it helps them manage risk. The bank can focus on these customers while also taking specific measures to address those with bad performance, ultimately aiming to mitigate risk.

Let’s have a look on the confusion matrix of the C5.0 model with trial 15



**Figure 9.2.5 Confusion matrix of the C5.0 model with trial 15**

By examining the confusion matrix in Figure 9.2.5 for the C5.0 model, we can see that the accuracy is 79.59%, sensitivity is 85.92%, and specificity is 68.69%. Given these sensitivity results, there is an 85.92% chance of correctly classifying customer loans with good performance. In this context, good performance refers to non-defaulting customers who have paid off their loans, while bad performance refers to customers who have defaulted. This information allows SBS to make informed decisions based on customer performance, taking appropriate actions according to the bank's terms and conditions. Identifying customers with good performance is crucial for SBS, as it helps manage risk. The bank can focus on these customers while also taking specific measures to address those with bad performance, ultimately aiming to mitigate risk.

So, by observing all model’s performance using C5.0 can say that trial 15 model has the better performance among other models using C5.0 model.

**10. BEST CLASSIFIER MODEL SELECTION:**

|  |  |  |
| --- | --- | --- |
| **Model Selection** | **Accuracy** | **Sensitivity** |
| Logistic regression | 77.05% | 86.40% |
| Forward Step wise logistic regression | 76.93% | 86.83% |
| Classification tree using rpart() | 77.63% | 86.10% |
| Classification tree using C5.0 | 78.26% | 84.11% |
| Classification tree using C5.0 with trial 15 | 79.59% | 85.92% |

**Table 10 Best Classifier model**

In selecting the best model, emphasis was placed on achieving high sensitivity. Upon reviewing the comparison table 10, it is evident that Forward logistic regression is the best classifier model with highest sensitivity rate with 86.83% because based upon our business requirement forward logistic regression correctly predict the probability of belonging to a particular class which will give us an information about the higher probability indicates that high chance of becoming a good performance of customer loan so, SBS bank will more focus on the good performance customers and maximize their profits and mitigate the risk.

**11. REGRESSION MODEL SELECTION:**

Now, regression models will be built to predict the estimated interest\_rate for the different customers. Based on different factors, regression models will be constructed to predict the Interest\_rate\_orig\_time. By building predictive regression models, the best performing model can be selected based on metrics.

**PREPARING THE DATA:**

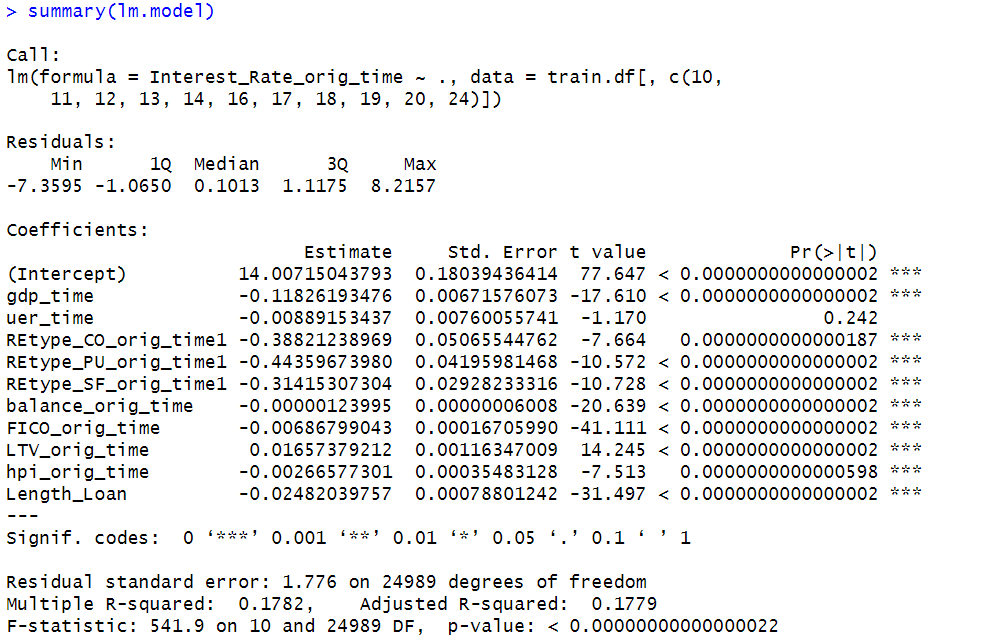
To build the predictive models, the data has already been divided into three partitions. Using these partitions, the model will be built. First, the data will be trained using the training data, and the model's performance will be tuned and evaluated using validation data to select the best regression models. In the final step, the model's performance will be evaluated with holdout data, which is new or unseen to the model.

**11.1 MULTIPLE LINEAR REGRESSION:**

Multiple linear regression model represents the relationship between the target and predictor variables, and it will assume the linear relationship between the target and predictor variables. By considering some features as a predictors and interest\_rate\_orig\_time is a target variable. In linear regression no need to create the dummy variables in pre-processing while building the model it will automatically create. The estimated equation for the multiple linear regression model is

Interest\_rate\_orig\_time = b0+b1.x1+b2.x2+b3.x3+…………………bn.xn

Will use the lm() to build the linear model. First will train the data with the training model then will evaluate the model with the validation model. Let’s have a look on the lm model equation



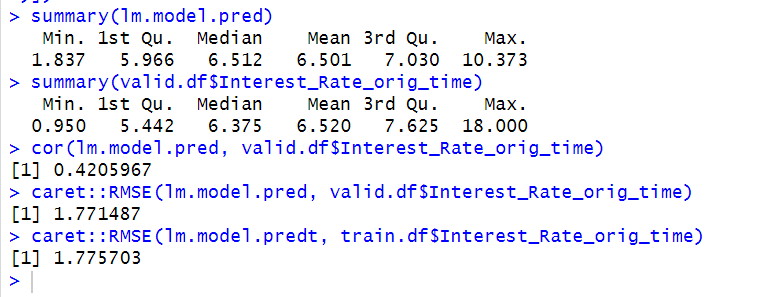
**Figure 11.1 Summary of the linear model**

The estimated equation for the linear model is

14.00715043793 + gdp\_time \*-0.11826193476 + uer\_time \* -0.00889153437

+ REtype\_CO\_orig\_time1\* -0.38821238969 + REtype\_PU\_orig\_time1\* -0.44359673980 + REtype\_SF\_orig\_time1 \* -0.31415307304 + balance\_orig\_time \* -0.00000123995 + FICO\_orig\_time \* -0.00686799043 + LTV\_orig\_time \* 0.01657379212 + hpi\_orig\_time\* -0.00269368063 + Length\_Loan \* -0.02482039757

By utilizing this equation can be able to predict the interest\_rate\_orig\_time for each individual customer. So, SBS bank will get an idea about the interest\_rate\_orig\_time based on the features to offer the individual customers.

****

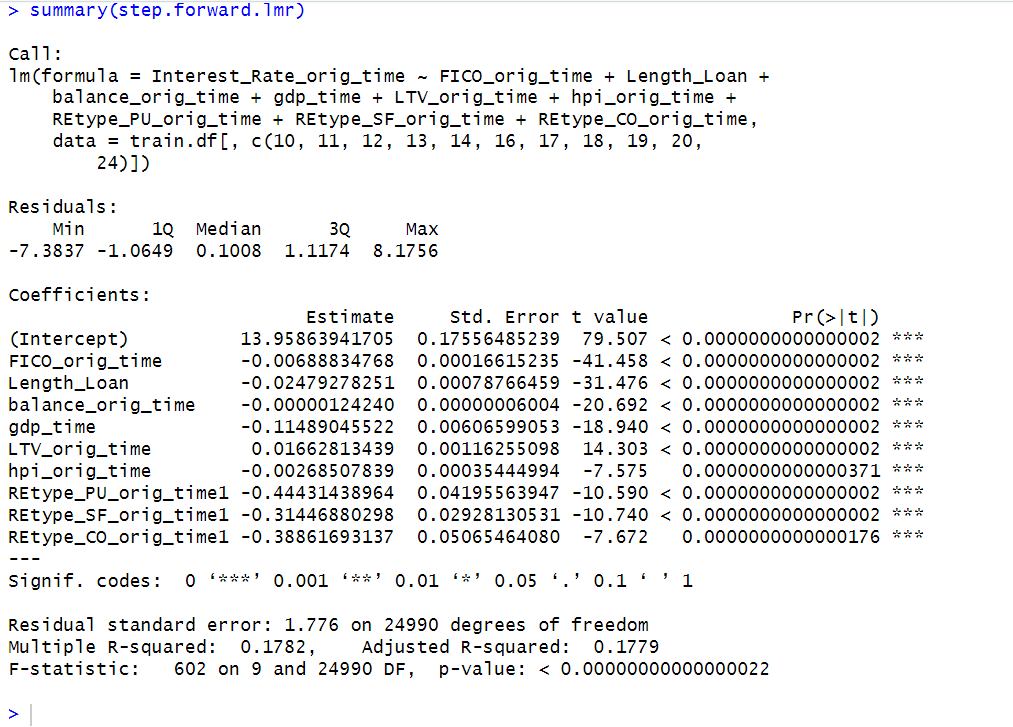
**Figure 11.2 Summary of Actual and Predicted Values**

By looking at the above figure 11.2 can be seen those summaries of actual and predicted values of the Interest\_rate\_orig\_rate do not have much difference between them so, the predictive power is relatively more, and data is distributed uniformly. The correlation between the actual and predicted values is 0.4 and the RMSE value is 1.7714.

**11.2 FORWARD STEP-WISE LINEAR REGRESSION:**

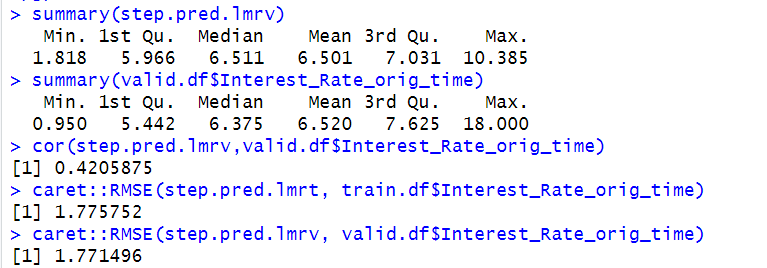
To improve the performance of the linear regression going to perform the forward stepwise regression in this it will starts with 0 predictors, and it will add the one-by-one predictors to get the accurate results.

Let’s have a look on the summary of the forward stepwise linear regression model



**Figure 11.3 Summary of the forward stepwise linear model**

By looking at the above figure 11.3 can be seen that estimated equation for the forward step wise linear regression and most significant predictors to predict the interest\_rate\_orig\_time. The multiple R-squared is 0.1782 and Adjusted R-squared is 0.1779. By using this estimated equation can predict the interest rate

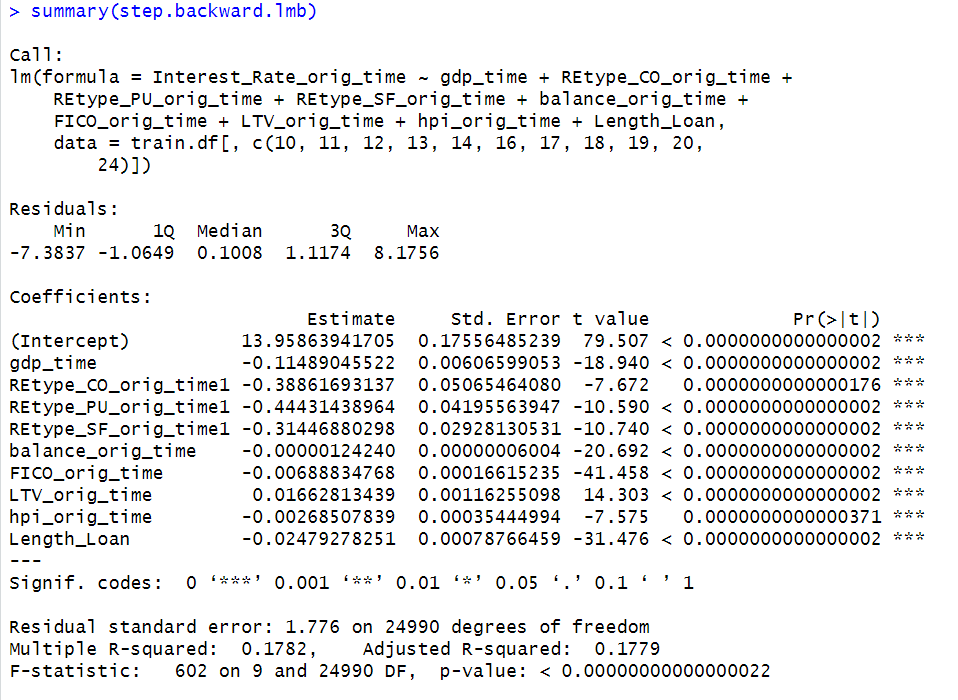


**Figure 11.4 Summary of the Actual and Predicted Values**

By looking at the above figure 11.4 can be seen those summaries of actual and predicted values of the Interest\_rate\_orig\_rate do not have much difference between them so, the predictive power is relatively more, and data is distributed uniformly. The correlation between the actual and predicted values is 0.4 and the RMSE value is 1.7714.

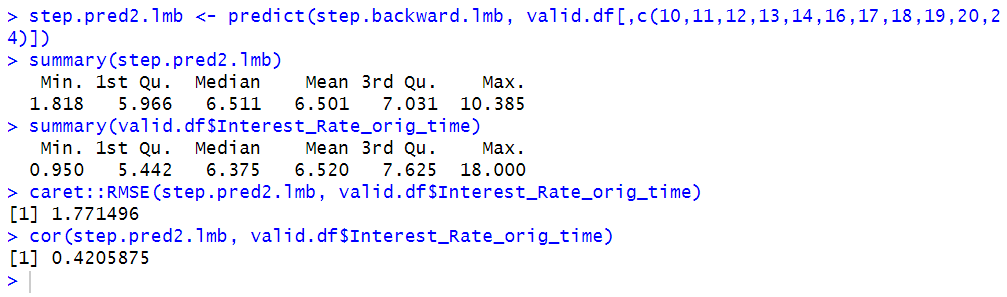
**11.3 BACKWARD STEP-WISE LINEAR REGRESSION:**

To improve the performance of the linear regression going to perform the backward stepwise regression in this it will starts with all predictors, and it will reduce the one-by-one predictors to get the accurate results. Let’s have a look on the summary of the backward stepwise linear regression model



**Figure 11.5 Summary of the Backward Stepwise Linear Regression**

By looking at above figure 11.5 can be seen that estimated equation for the backward stepwise linear regression and most significant predictors.



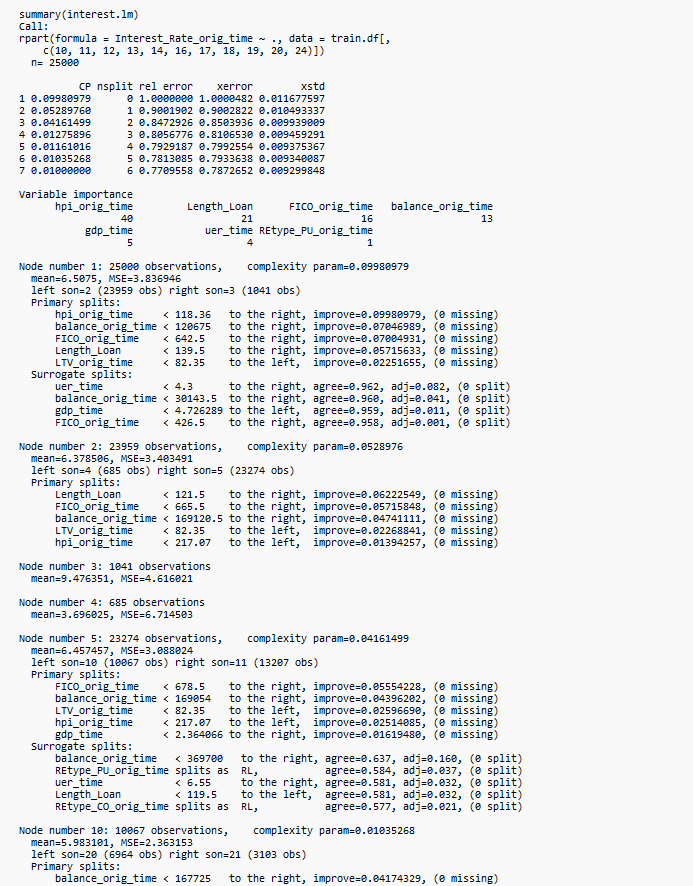
**Figure 11.6 Summary of Actual and Predicted Values of Backward**

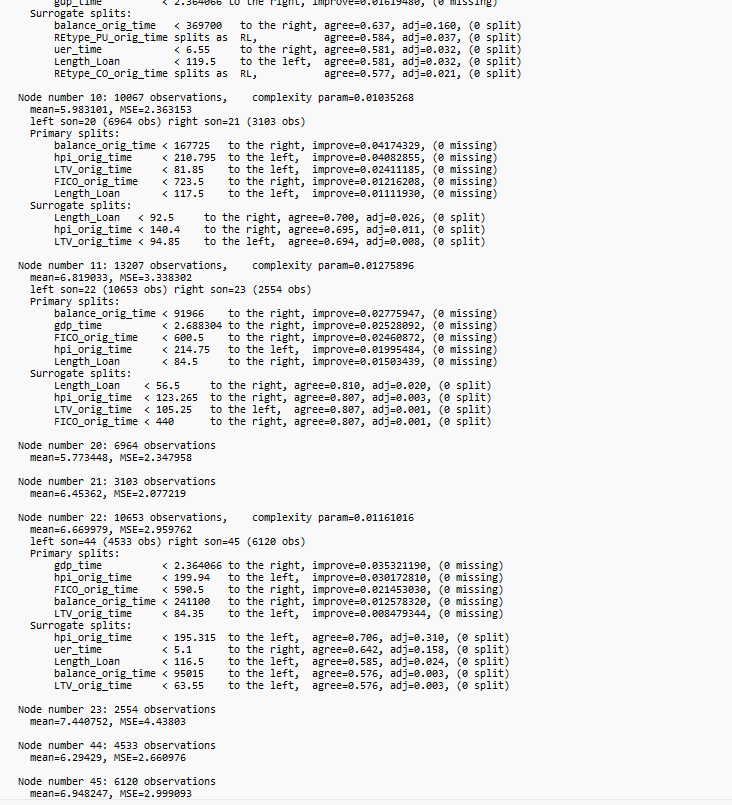
By looking at the above figure 11.6 can be seen those summaries of actual and predicted values of the Interest\_rate\_orig\_rate do not have much difference between them so, the predictive power is relatively more, and data is distributed uniformly. The correlation between the actual and predicted values is 0.4 and the RMSE value is 1.7714.

**11.4 Regression Tree**

In regression tree prediction for new data are based on the rules represented on the tree in regression trees predictions are obtained by averaging the outcome values in the nodes. To access this regression tree use r.part() function. It is based on the rules of the tree. In regression-tree prediction, node homogeneity is measured by various statistics such as variance, standard deviation, or absolute deviation from the mean.

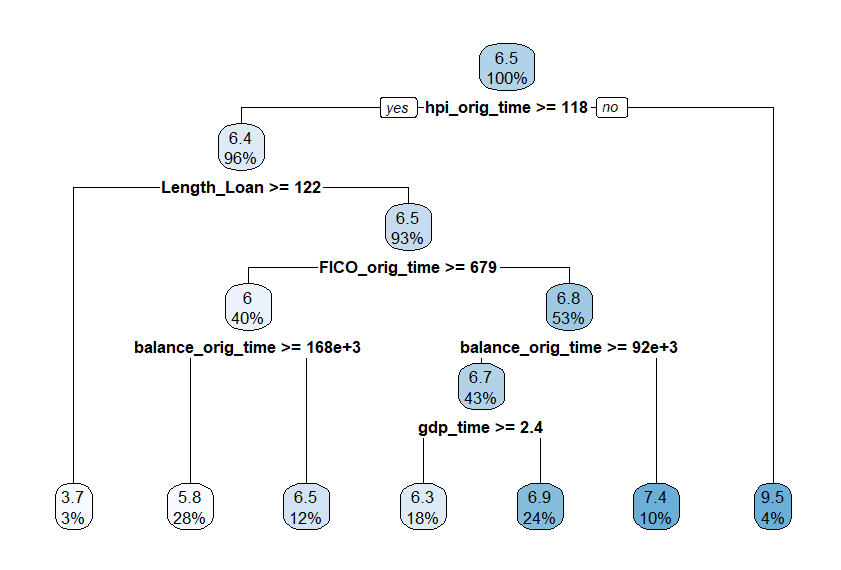
Let’s have a look on the summary of the model

****

****

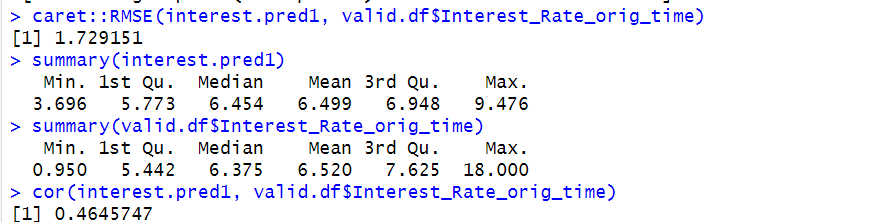
**Figure 11.7 Summary of the regression tree model**

By looking at the above figure 11.7 can be seen that variable importance and the number of splits. Let’s have a look on the plotted tree



**Figure 10.8 Plotted Tree**

By looking at the above figure 10.8 can be seen that splits are based on the rules. Let’s have a look on the summaries of the actual and predicted values



**Figure 11.9 Summary of Actual and Predicted Values**

By looking at the above figure 11.9 can be seen those summaries of actual and predicted values of the Interest\_rate\_orig\_rate do not have much difference between them so, the predictive power is relatively more, and data is distributed uniformly. The correlation between the actual and predicted values is 0.4 and the RMSE value is 1.7291.

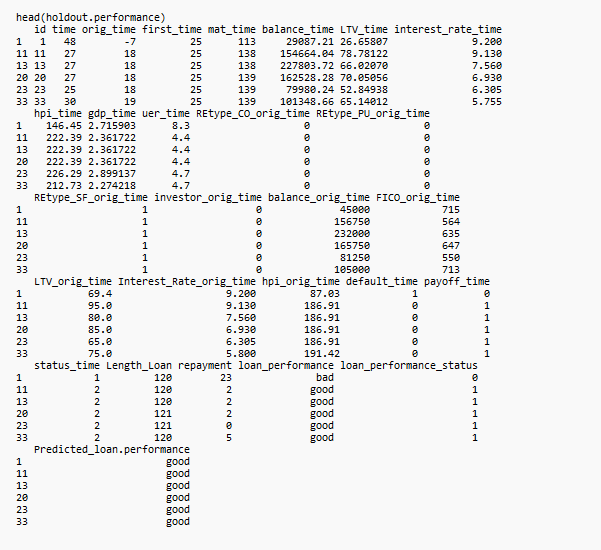
**11.5 COMPARISION OF REGRESSION MODELS:**

|  |  |
| --- | --- |
| REGRESSION MODELS | RMSE |
| Multiple Linear Regression | 1.7714 |
| Forward Stepwise Linear Regression | 1.7714 |
| Backward Stepwise Linear Regression | 1.7714 |
| Regression Tree | 1.7291 |

Based on the lowest RMSE value, regression tree model emerges as the best regression model among all others. This model selectively chooses important variables to achieve accuracy and sensitivity while reducing unimportant ones. In the future, this approach will aid business owners in data gathering. Rather than collecting all variables, they can focus on the important ones, saving time and obtaining accurate results. Regression tree will provide more accurate results to predict the interest\_rate\_orig\_time this will help to bank how can offer interest rate to different customers based on the different factors. For SBS, bank it is very important to predict the interest for individual customers based on their background. So, they can provide the high interest rate for high risk customers and they can mitigate the risk.

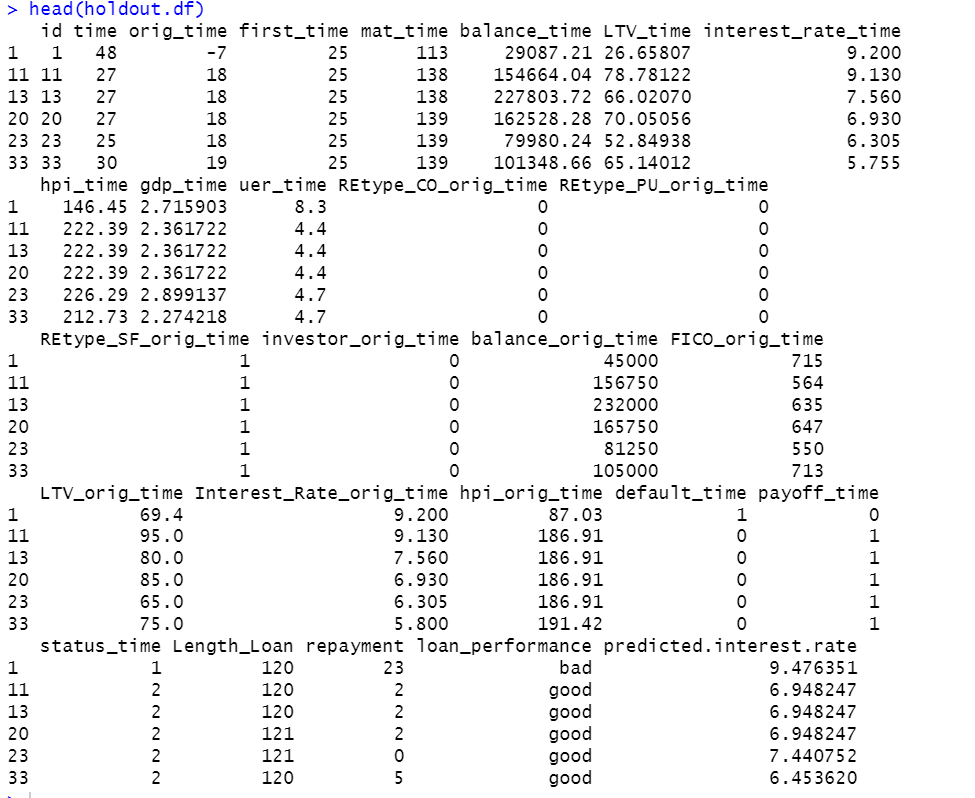
**12. HOLDOUT MODEL PERFORMANCE:**

In this step, the model performance will be evaluated with the holdout data. This analysis will assist the SBS bank how to offer the interest rate based on the history of the individual customer. By providing inputs to this model, it will generate predicted interest rate, predict loan\_performance based on their repayment performance. This will give them an idea of which type of customers improve their business and gain an overview of the different customers behaviour. To assess the performance of the models, the holdout data will be applied by adding the predicted columns. Now, will add the predicted interest\_rate.



**Figure 12.1 Added Predicted loan performance**

By looking at the above figure 12.1 can observe that new predicted loan performance by evaluating with the holdout data which give provide the better understanding about the prediction of loan performance of the customers.



**Figure 12.2 Added Predicted interest rate**

By looking at the above figure 12.2 can be seen that predicted interest rate by using the holdout data which will give better understanding about the predictions on interest rate. This will help to SBS bank how to offer interest rate for different customers based on their credit history and background and they can mitigate the risk.

**13. CONCLUSION AND FUTURESCOPE:**

* In conclusion, the analysis on the data has been conducted, exploring the relationships between variables. Data partitions were established to build predictive models, focusing on classifying customer loans with good performance. Among the classification models explored, forward step wise logistic regression is the best model, demonstrating high sensitivity and the greatest likelihood of correctly identifying positive class instances. Regression models were utilized to predict interest rates tailored to individual customer backgrounds. This information aids the bank in determining appropriate interest rates for different customers, thereby mitigating risk. Among the regression models examined, regression tree emerged as the most effective, promising accurate results in interest rate prediction. By using these models, SBS Bank aims to make informed decisions, mitigating risk, and enhancing profitability. This strategy is crucial as the bank expands its services to mortgage lending. Accurate risk assessment and interest rates will help SBS Bank maintain financial stability and better serve its customers. As it is starting newly offering mortgage loans by using this analysis they can mitigate the risk and they can approve the loans to customers wisely.