**Addressing Bank's Financial Instability**

Rajesh Dasari Timmisetty

**Executive Summary**

Bank is currently facing low financial stability due to two primary reasons: first, a significant number of loans are defaulting because the bank has been approving loans for customers who are unable to repay them. The approved loan amounts are not sustainable for these customers. Second, the bank is unable to estimate which actively paying customers will default in the future. To overcome these challenges and improve financial stability, Bank has decided to utilize predictive analytics on historical data containing the loan progression data of 50,000 customers. This dataset has been preprocessed by aggregating data, removing null and zero values, and eliminating features based on domain knowledge modeling.

Classification models have been built to determine whether new customers will pay off their loans or default. Among all the models considered, the backward selection logistic regression has been chosen as the best model for this purpose. Additionally, a regression model has been developed to estimate the appropriate loan amount to offer new customers who are classified as likely to pay off their loans. The backward linear regression model has been identified as the best model for this task. Finally, another classification model has been built to predict whether ongoing customers will pay off their loans or default. The backward selection model has been deemed the best model for this analysis.

With these models, Bank can offer appropriate loan amounts to new customers who are likely to pay off their loans, thereby reducing the risk of defaults. Moreover, the bank can identify which ongoing customers are at risk of defaulting, allowing for timely intervention to enhance financial stability.

**Table of Contents**

1. Introduction …………………………………………………………………………….5
   1. Business Background …………………………………………………………………..5
   2. Business Goal…………………………………………………………………………...5
   3. Analytical Approach ……………………………………………………………………5
   4. About Dataset…………………………………………………………………………...6
2. Exploring and preprocessing the data…………………………………………………..8
3. Dimension Reduction…………………………………………………………………...40
4. Partition the Data………………………………………………………………………..42
5. Modelling For Classification Task……………………………………………………...43
   1. Decision Tree……………………………………………………………………………43
      1. Measuring Performance…………………………………………………………….44
      2. Improving Model Performance……………………………………………………..45
   2. Logistic Regression……………………………………………………………………..46
      1. Measuring Performance…………………………………………………………….48
      2. Improving model performance……………………………………………………...48
6. Classifier Model Selection……………………………………………………………...50
7. Modelling For Regression Task…………………………………………………………51
   1. Linear Regression Model……………………………………………………………….51
      1. Measuring Performance…………………………………………………………….52
      2. Improving Model Performance……………………………………………………..52
   2. Decision Tree for Regression(regression tree)………………………………………….54
      1. Measuring Performance…………………………………………………………….55
      2. Improving Model Performance……………………………………………………..55
8. Regression Model Selection…………………………………………………………….56
9. Classification Model for classifying on-going customers...…………………………….56
   1. Decision Tree……………………………………………………………………………57
      1. Measuring Performance………………………………………………………….....58
      2. Improving Model Performance……………………………………………………..58
   2. Backward Selection Logistic Regression……………………………………………….60
      1. Measuring Model Performance……………………………………………………..61
10. Ongoing Customer classification Model Selection……………………………………..62
11. Predicting New Data………………………………………………………………........62
12. Conclusion………………………………………………………………………………63
13. **Introduction**
    1. **Business Background**

Bank, which offers mortgage loans and operates numerous branches across the United States, is currently experiencing low financial stability. This is mainly due to two reasons: first, a significant number of loans are defaulting because the bank has been approving loans for customers who are unable to repay them. The approved loan amounts are not sustainable for these customers. Second, the bank is unable to estimate which actively paying customers are going to default.

* 1. **Business Goal**

The primary goal is to increase the financial stability of the bank by employing predictive analysis on historical data. This analysis aims to accurately estimate appropriate loan amounts for new customers and determine whether a new customer is likely to repay the loan or default. By accurately assessing the risk of default and determining sustainable loan amounts, Bank can offer loans more selectively, thereby reducing the incidence of loan become defaults. Additionally, predictive analysis can help identify which customers currently making payments are at risk of defaulting. This early identification enables the bank to mitigate risks, possibly through the acquisition of collateral, thereby enhancing its financial stability.

* 1. **Analytics Approach**

The main objective is to utilize predictive analysis on historical data to accurately estimate appropriate loan amounts for new customers and to determine whether a new customer is likely to repay the loan or default. Additionally, the aim is to identify which customers, currently making payments, are at risk of defaulting. To achieve these goals, a classification model is first build to classify whether a new customer is likely to repay the loan or default. This enables Bank to approve loans for selected customers instead of approving loans for all applicants, thereby reducing the risk of loans becoming defaults. Furthermore, a regression model is employed to estimate the appropriate loan amount that the bank can offer to new customers who are likely to pay off the loan. This approach aids in providing sustainable loans to customers and reduces the likelihood of defaults. Another classification model is build to classify which actively paying customers will become default or pay off their loans. Utilizing this model allows the bank to identify risks early. If an actively paying customer is classified as likely to default, the bank can take the required action on the properties of those customers to improve financial stability.

* 1. **About Dataset**

The dataset is historical data containing mortgage loan information for 50,000 different customers. It encompasses the overall loan progression from start to end for each loan at a particular time. The dataset comprises a total of 622490 rows and 23 columns. Now, let's explore the definition of each column.

* id: Borrower ID
* time: Time stamp of observation
* orig\_time: Time stamp for origination which indicates the timestamp that loan actually initiated.
* first\_time: Time stamp for first observation
* mat\_time: A timestamp for maturity, indicating the month when the final payment is due.
* balance\_time: Remaining loan amount that borrower needs to pay at observation time
* LTV\_time: Loan-to-value ratio at observation time, in %

LTV ratio is calculated by dividing the amount borrowed by the appraised value of the property, expressed as a percentage. For example, if you buy a home appraised at $100,000 for its appraised value, and make a $10,000 down

* interest\_rate\_time: Interest rate at observation time, in %
* hpi\_time: House price index at observation time, base year = 10

The FHFA HPI is a broad measure of the movement of single-family house prices. The FHFA HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties.

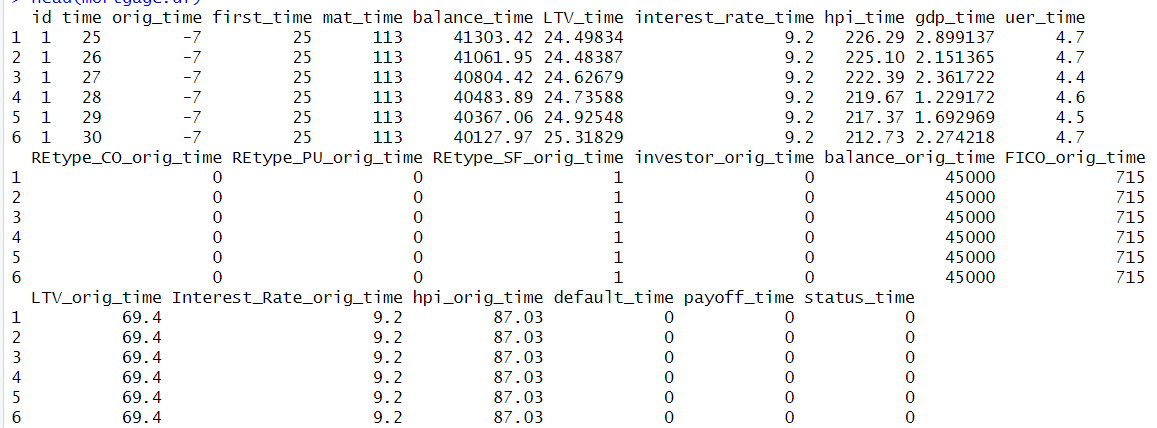
* gdp\_time: Gross domestic product (GDP) growth at observation time, in %
* uer\_time: Unemployment rate at observation time, in %
* REtype\_CO\_orig\_time: Real estate type which is a binary column condominium = 1, otherwise = 0
* REtype\_PU\_orig\_time: Real estate type planned urban development = 1, otherwise = 0
* REtype\_SF\_orig\_time: Single-family home = 1, otherwise = 0
* investor\_orig\_time: which tells if customer is investor or not at the time of borrowing loan. Investor borrower = 1, otherwise = 0
* balance\_orig\_time: Outstanding balance at origination time which indicates the loan amount that bank offered to the customer. Which is considered as the target column for regression model.
* FICO\_orig\_time: FICO score at origination time, in %

The overall FICO score range is between 300 and 850. In general, scores in the 670 to 739 range indicate “good” credit history, and most lenders will consider this score favorable

* LTV\_orig\_time: Loan-to-value ratio at origination time, in %
* Interest\_Rate\_orig\_time: Interest rate at origination time, in %
* hpi\_orig\_time: House price index at origination time, base year = 100
* default\_time: Default observation at observation time.Will be 1 if loan is defaulted, otherwise Zero.
* payoff\_time: Payoff observation at observation time Will be 1 if loan is paid off, otherwise Zero.
* status\_time: Default (1), payoff (2), and nondefault/nonpayoff (0) at observation time which is considered as target for classification tasks at hand.

1. **Exploring and preprocessing the data**

In this section the data is explored and preprocessed to make the data ready for building the classification and regression models. First will look the first 6 rows of the data.



**Figure 2.1. first 6 rows of the data**

From Figure 2.1, it is evident that there are multiple rows for each ID. Therefore, the data needs to be aggregated for classification and regression models.

Time: While aggregating this column, the last value is considered for each id because it is important to keep the latest timestamp for each customer ID to represent the most recent data.

Mat\_time: For this column, the last value is considered for each id to represent the latest loan term for each individual customer id.

Balance\_time: For this column, the last value is considered for each id to represent the remaining balance that each customer needs to pay to clear the loan.

Ltv\_time: For this column, the last value is considered for each id to represent latest LTV ratio value.

Interest\_rate\_time: For this column, the last value is considered for each id to represent the latest interest rate that the customer is paying for the loan.

hpi\_time : For this column, the last value is considered for each id to represents the latest hpi\_time.

gdp\_time : For this column, the last value is considered for each id to represents the latest gdp

uer\_time: For this column, the last value is considered for each id to represents the latest unemployment rate

default\_time : For this column, the last value is considered for each ID to represent if the customer is defaulted or not.

payoff\_time : For this column, the last value is considered for each id to represents if the customer is paid the loan or not.

status\_time : For this column, the last value is considered for each ID to represents the latest status of the customer.

orig\_time : As per the column definition there should be only one orig\_time value for each id but there are some customers who has different origin\_time values so instead of choosing first value maximum occurred value for each id is considered to represents the time when the loan was initiated.

first\_time : For this column, the first value is considered for each ID to represents the Time stamp for first observation

retype\_co\_orig\_time : As per the column definition there should be only one retype\_co\_ orig\_ time value for each id but there are some customers who has different retype\_co\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered.

retype\_pu\_orig\_time: As per the column definition there should be only one retype\_pu\_ orig\_ time value for each id but there are some customers who has different retype\_pu\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered.

retype\_sf\_orig\_time : As per the column definition there should be only one retype\_sf\_ orig\_ time value for each id but there are some customers who has different retype\_sf\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered. investor\_orig\_time : As per the column definition there should be only one investor\_orig\_time value for each id but there are some customers who has different investor\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered to represents if customer is investor or not.

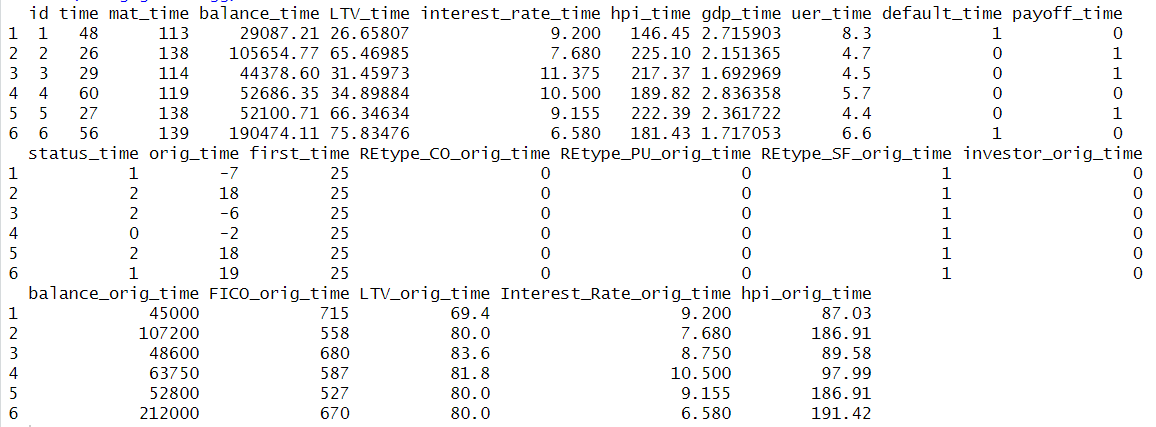
balance\_orig\_time: As per the column definition there should be only one balance\_orig\_time value for each id but there are some customers who has different balance\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered to represents the loan amount approved by bank for customer.

fico\_orig\_time : As per the column definition there should be only one fico\_orig\_time value for each id but there are some customers who has different fico\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered to represents the fico score for customer.

ltv\_orig\_time: As per the column definition there should be only one ltv\_orig\_time value for each id but there are some customers who has different ltv\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered to represents the ltv ratio value for the customer property at the origination of the loan.

interest\_orig\_time : As per the column definition there should be only one interest\_orig\_time value for each id but there are some customers who has different interest\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered to represents the interest rate for the customer at the origination of the loan.

hpi\_orig\_time : As per the column definition there should be only one hpi\_orig\_time value for each id but there are some customers who has different hpi\_orig\_time values so instead of choosing first value maximum occurred value for each id is considered to represents the hip value of the customer property at origination of the loan.



**Figure 2.2 first 6 rows after aggregating the data**

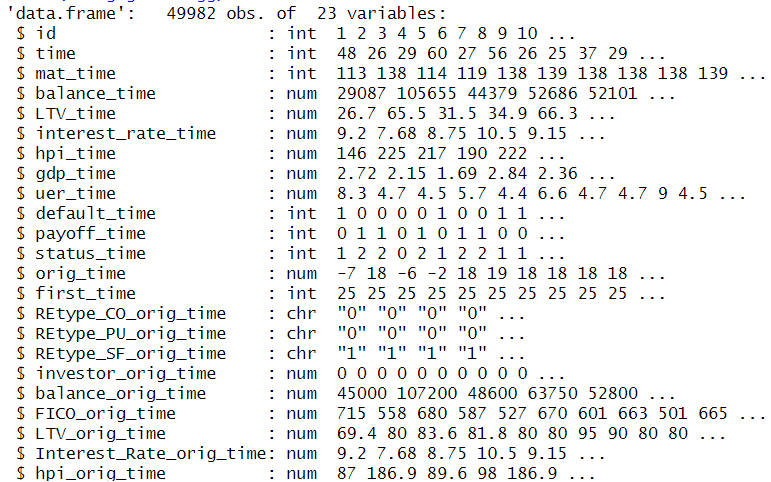
**Missing Values**

Using colsums(is.na()) in R helps identify null values in each column. There are 18 null values in lvt\_time column There are various methods for replacing null values, the most common ways to replace them with the mean, median, or to simply delete them. However, these traditional approaches are generalized, and complex patterns in the data are not considered while replacing the values, which will potentially impact model performance. To address this issue, the k-nearest neighbor imputation method is used by assuming similar properties will have same lvt\_time and k-nearest neighbor imputation method tries to replace with the values that are closest to the null values.

**Zero Values**

Number of zero values in each column can be identified by using colSums(data == 0, na.rm = TRUE) in R. Other than categorical and dummy columns, there are 323 zero values in the balance\_time column, 322 in LTV\_time, 2 in interest\_rate\_time, 83 in orig\_time, 10,134 in Interest\_Rate\_orig\_time, and 18 in balance\_orig\_time. Based on the column definitions, zero values in orig\_time can be used as they are because there may be situations where a loan can be initiated on the same day as the application. The zero values in interest\_rate\_time and Interest\_Rate\_orig\_time will be replaced with the mean. Zero values in balance\_time can be used as it is because of the assumptions that the customer has cleared their loan. There are zero values in balance\_orig\_time, indicating that the customer has not borrowed any loan, so customers with balance\_orig\_time zero can be removed. The zero values in LTV\_time can be replaced with LTV\_origin\_time because by assuming if the LTV\_time is zero then the property LTV is not changing from the LTV\_origin\_time.

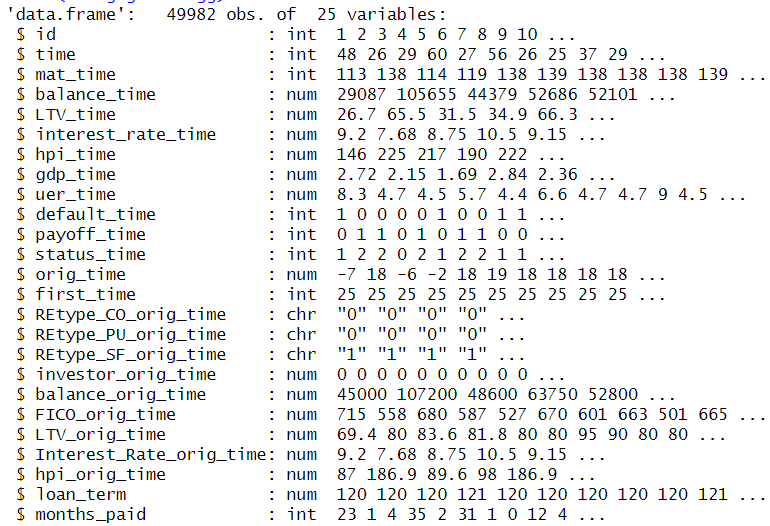
Now let’s see the structure of the data



**Figure 2.3 structure of aggregated data**

**Creating new columns**

Based on the structural analysis of the data shown in Figure 2.3, it is evident that no columns indicate the actual loan term or the number of months paid by a customer. Given domain knowledge, both the loan term and the number of months paid will be important predictors in determining whether an ongoing customer will become default or payoff. Therefore, a new column named loan\_term is created using the formula: Loan Term = Mat\_time - Orig\_time. Additionally, a months\_paid column is created using the formula: Months Paid = Time - First\_time.



**Figure 2.4 Structure of the data after adding new column**

**Removing Unwanted data**

In the dataset, some customers are identified with a status\_time as neither default nor payoff, indicating that these are ongoing customers. These customers cannot be utilized for modeling purposes because they have not been categorized as either default or payoff. Consequently, such customers are considered either new data or unwanted data for the purposes of modeling and are not included in further analysis.

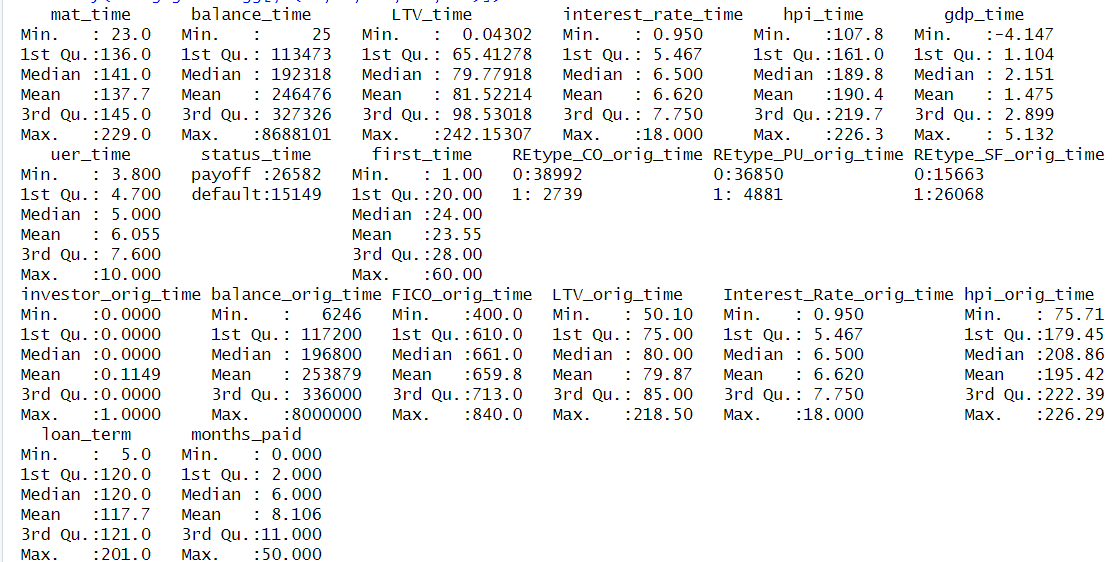
**Handling Categorical Variables**

Based on the structure analysis of the data shown in figure 2.4. Converting all the categorical columns like REtype\_CO\_orig\_time, REtype\_PU\_orig\_time, REtype\_SF\_ orig\_time, status\_time into factors is appropriate to handle them correctly by the models.

**Feature elimination based on domain knowledge**

Several columns are unnecessary for the task at hand and should be removed to prevent unnecessary complexity in the analysis. The columns id, default time, time, orig\_time and pay\_off\_time are candidates for removal. The id column serves solely as a unique identifier, while the time column represents timestamps in the dataset and orig\_time represents timestamp when loan was initiated which is not much useful for goals at hand. Additionally since the target column status\_time is derived from pay\_off\_time, and default\_time, these columns can also be excluded from further analysis.

Now let’s analyze the data of each feature and how it is impacting the target columns status\_time and balance\_orig\_time.

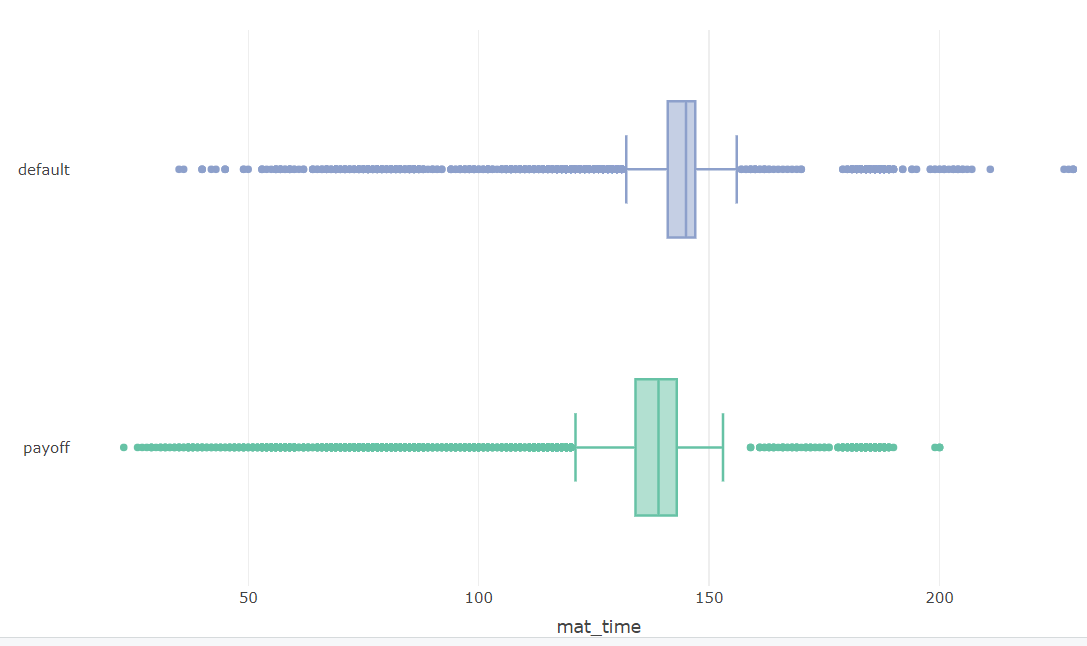
****

**Figure 2.5 summary of the data**

From figure 2.5 it is clear that the different variables have very different range of values. All the null values are replaced and there are no extreme values in the data.

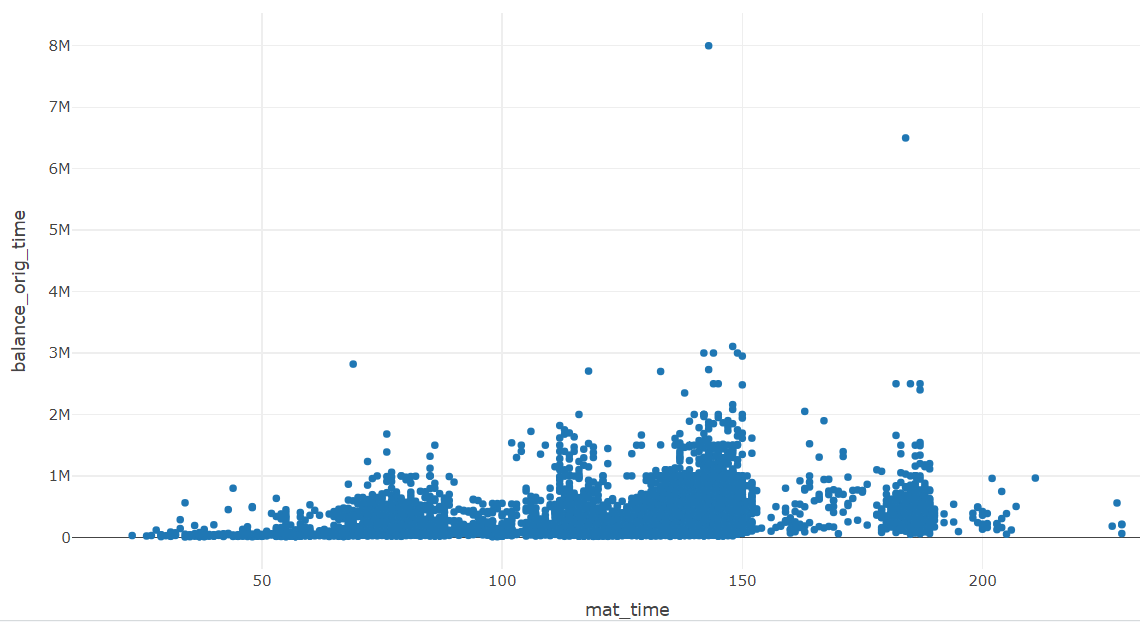
Now let’s analyze each column how it is impacting the target columns.

**Mat\_time:** Mat\_time is a timestamp column in months, indicating the number of the months the final payment due. Let’s try to understand how the column is impacting the status\_time column and the balance\_orig\_time column

****

**Figure 2.6 mat\_time vs status\_time column**

Figure 2.6 shows that most of the customers who have the maturity term between 143 to 156 months will be classified as default and most of the customer who have the maturity term between 121 to 143 months will be considered as payoff.

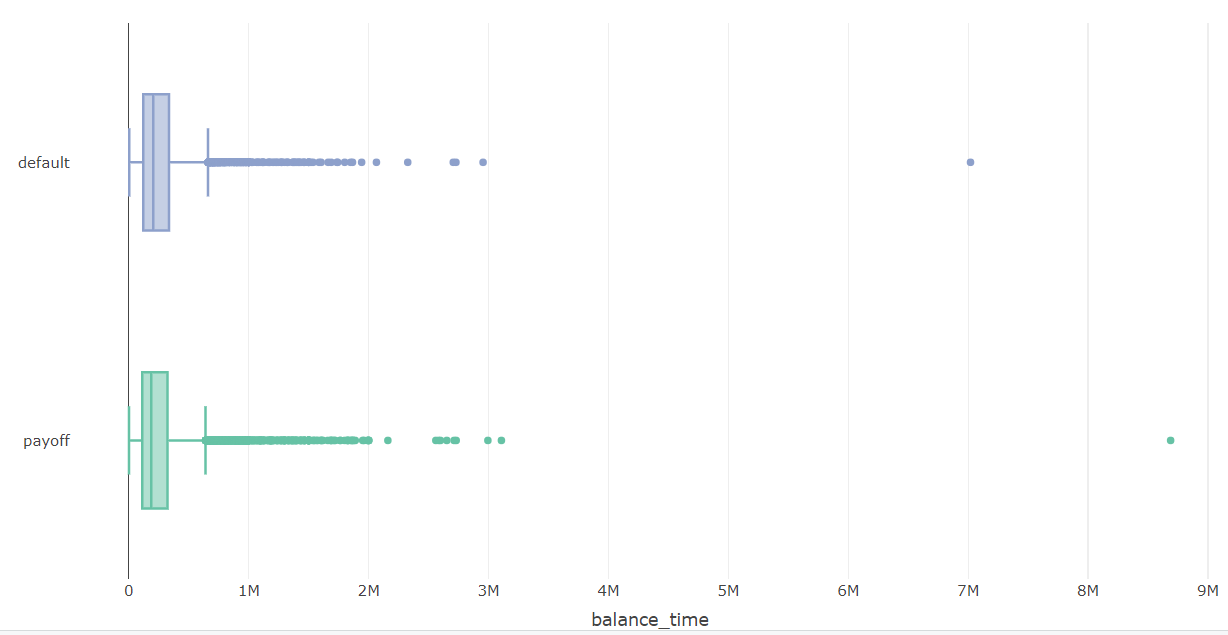
****

**Figure 2.7 mat\_time vs balance\_orig\_time**

From Figure 2.7, it is clear that loan terms between 100 to 200 months have higher loan amounts. Therefore, customers who accepted loan terms within this range can be offered higher loan amounts.

**Balance\_time**

Balance\_time which represents the remaining amount to pay to clear the loan. Now lets try to see how the Balance\_time impacting the target columns.

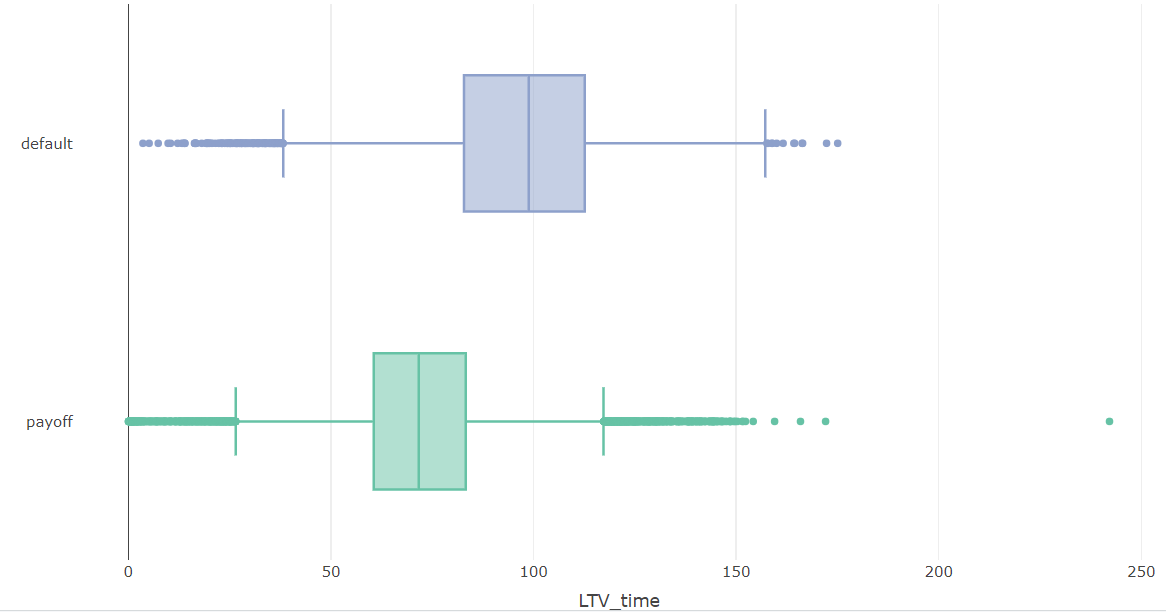
****

**Figure 2.8 balance\_time vs status\_time column**

From the figure 2.8 the box plot distribution looks similar for both payoff and default customers so balance\_time might not be important to classify if the customer is going to be payoff or default.

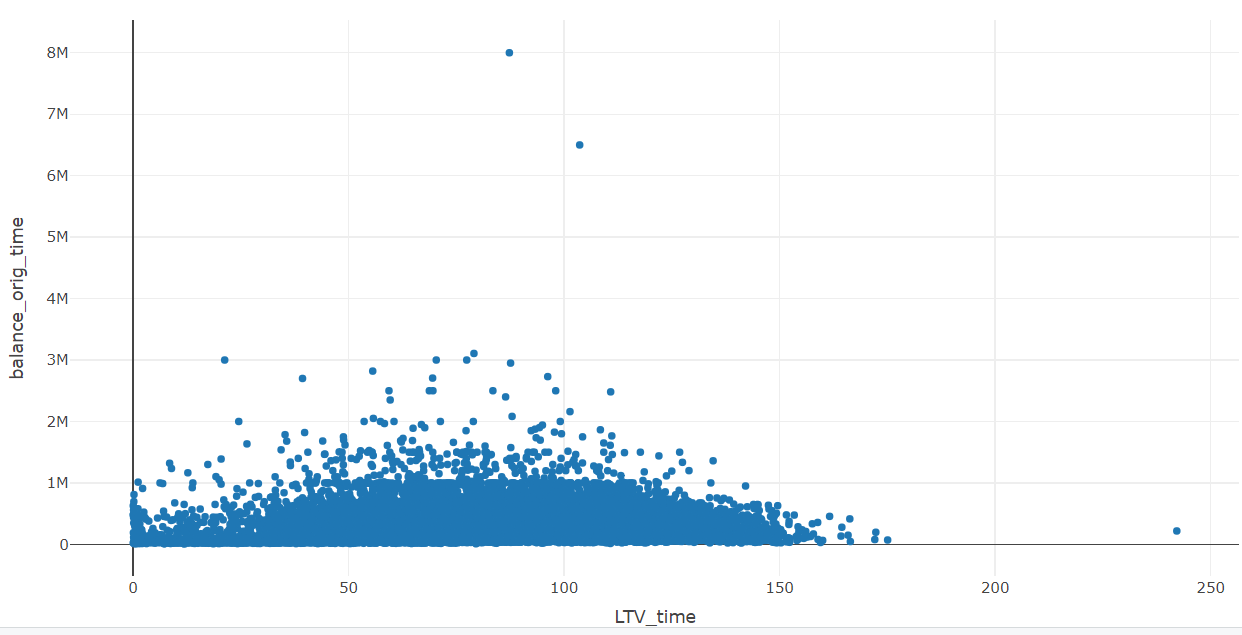
**LTV\_time**

LTV\_time which represents the loan to value ratio of the property. Let’s try to understand how the column is impacting target columns.

****

**Figure 2.9 ltv\_time vs status\_time**

Based on Figure 2.9, it is evident that most customers with an LTV ratio of less than 80% are classified as payoff customer.

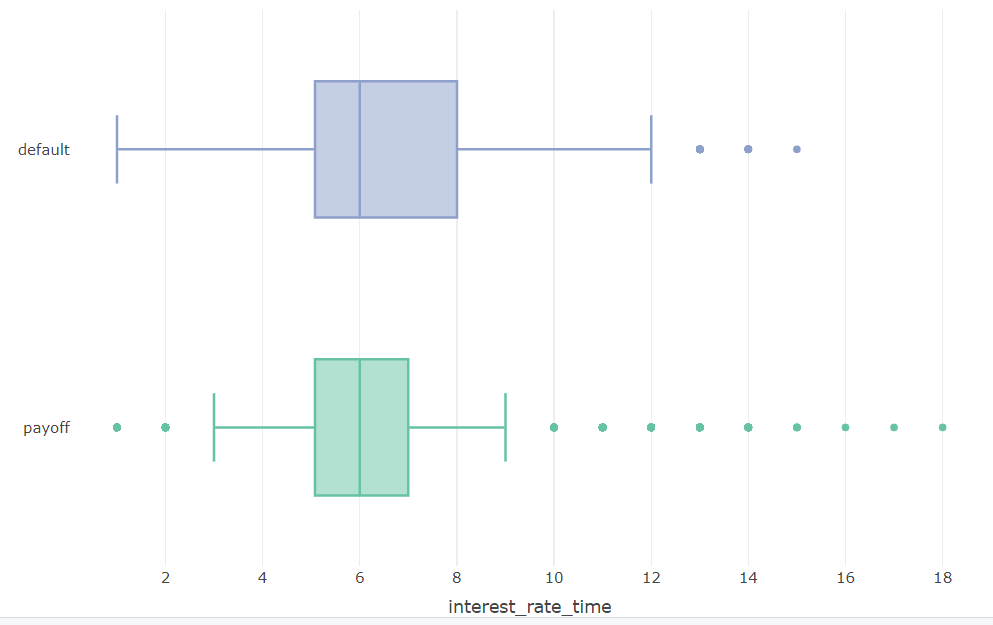
****

**Figure 2.10 LTV\_time vs balance\_orig\_time**

According to Figure 2.10, customers with properties having an LTV ratio of less than 100% are approved for higher loan amounts. Therefore, the bank can offer higher loan amounts to customers whose property's LTV ratio is less than 100%.

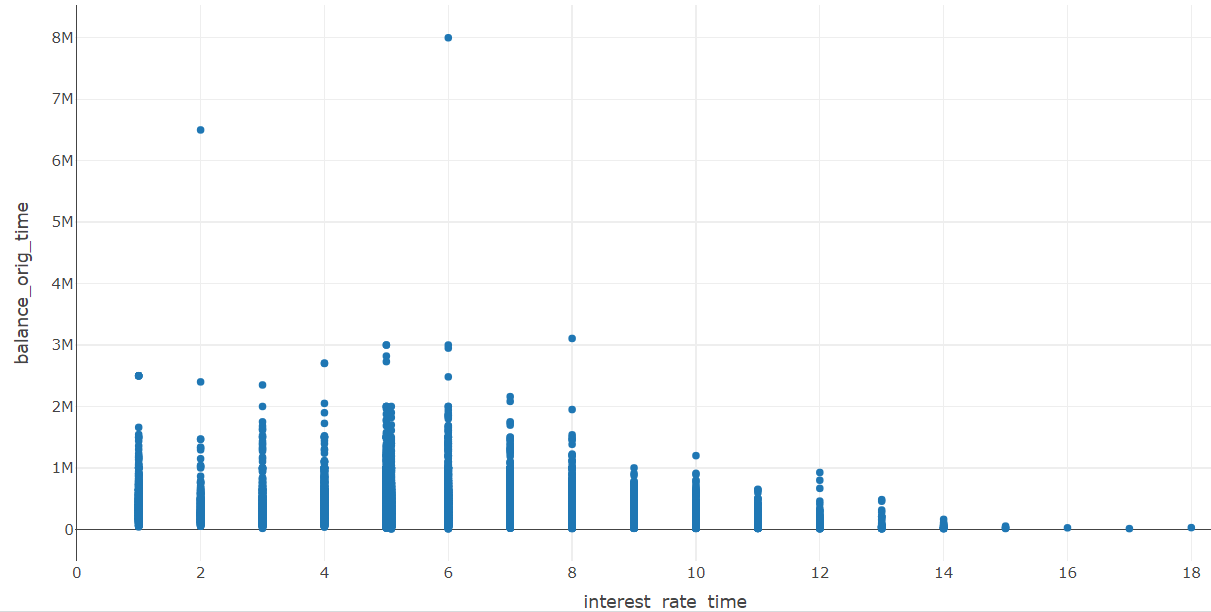
**Interest\_rate\_time**

Interest\_rate\_time which represents the interest rate the customer is paying. Let’s try to understand how the column is impacting target columns

****

**Figure 2.11 interest\_rate\_time vs status\_time**

Based on Figure 2.11, it is evident that most customers paying an interest rate greater than 7% are classified as default.

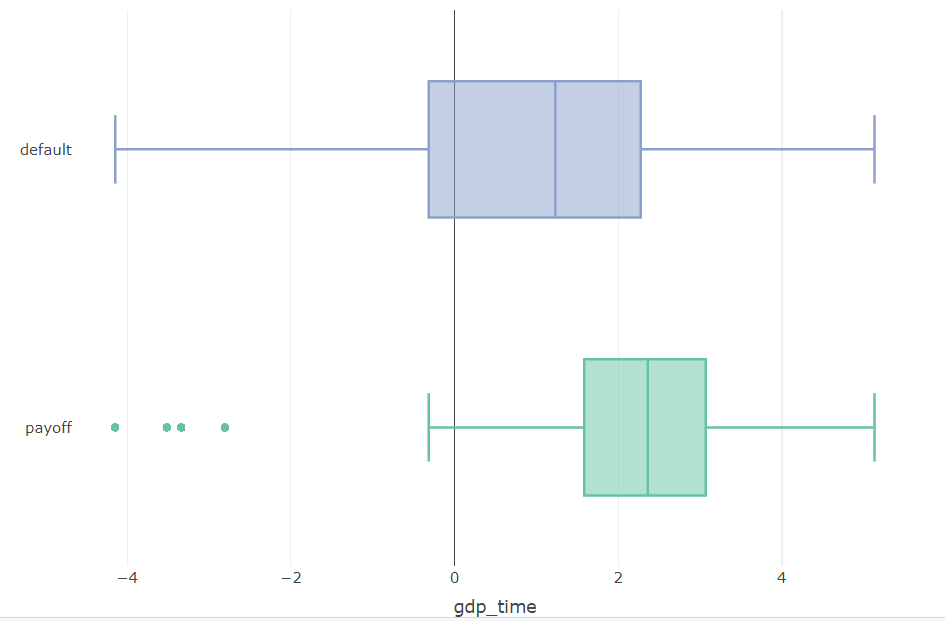
****

**Figure 2.12 interest\_rate\_time vs balance\_orig\_time**

From Figure 2.12, it is evident that most customers are paying an interest rate less than 8%. Customers who are paying an interest rate greater than 8% has less loan amount. Therefore, the bank can offer a higher interest rate for customers who borrow less amount.

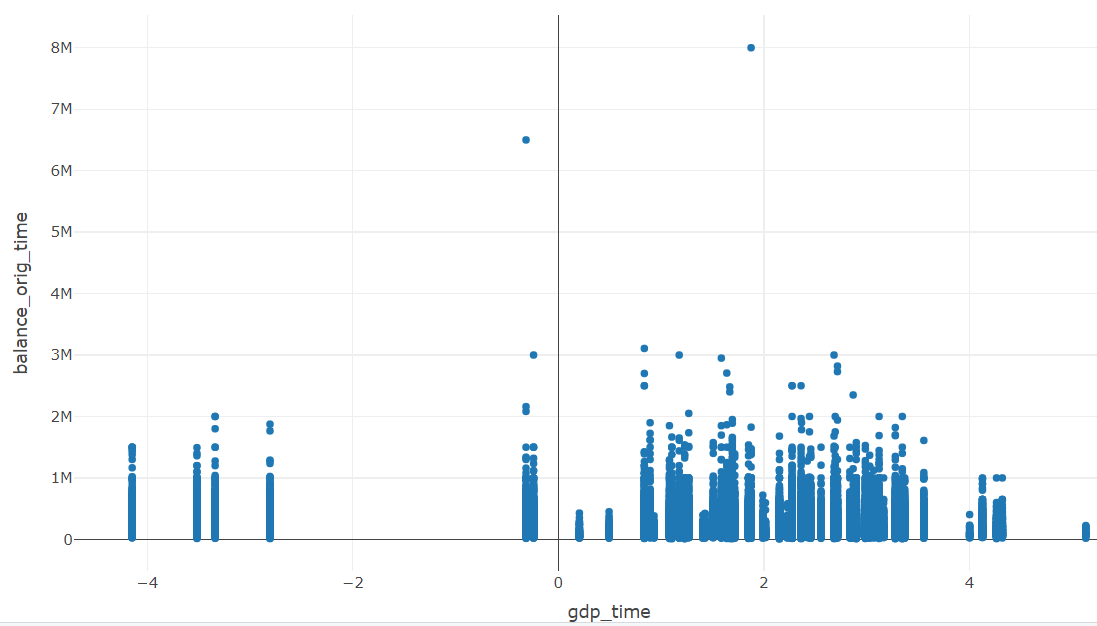
**Gdp\_time**

This column represents the GDP rate. Let’s try to understand how this column is impacting the target columns.



**Figure 2.13 Gdp\_time vs status\_time**

Based on Figure 2.13, it is evident that almost all customers who took out loans when the GDP was negative are classified as default. Additionally, most of the customers who took out loans when the GDP rate was greater than 1.9% are classified as payoff.

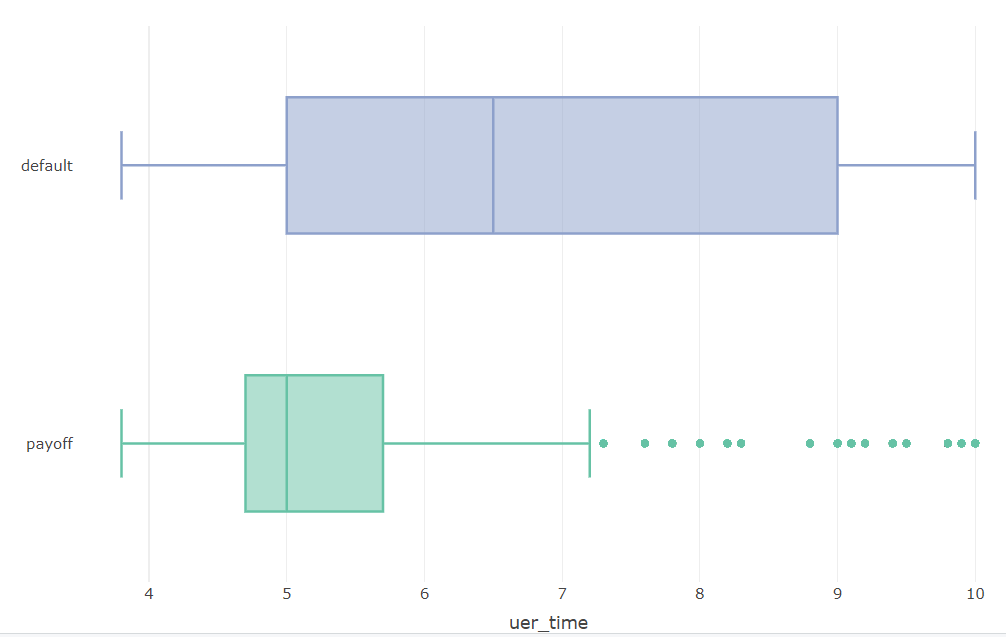
****

**Figure 2.14 gdp\_time vs balance\_orig\_time**

Based on figure 2.14, it is evident that customers who borrowed loans when the GDP rate was between 0.5% to 3.7% had higher loan amounts. Therefore, the bank can consider offering higher loan amounts to customers who applied for loans when the GDP rate was in this range.

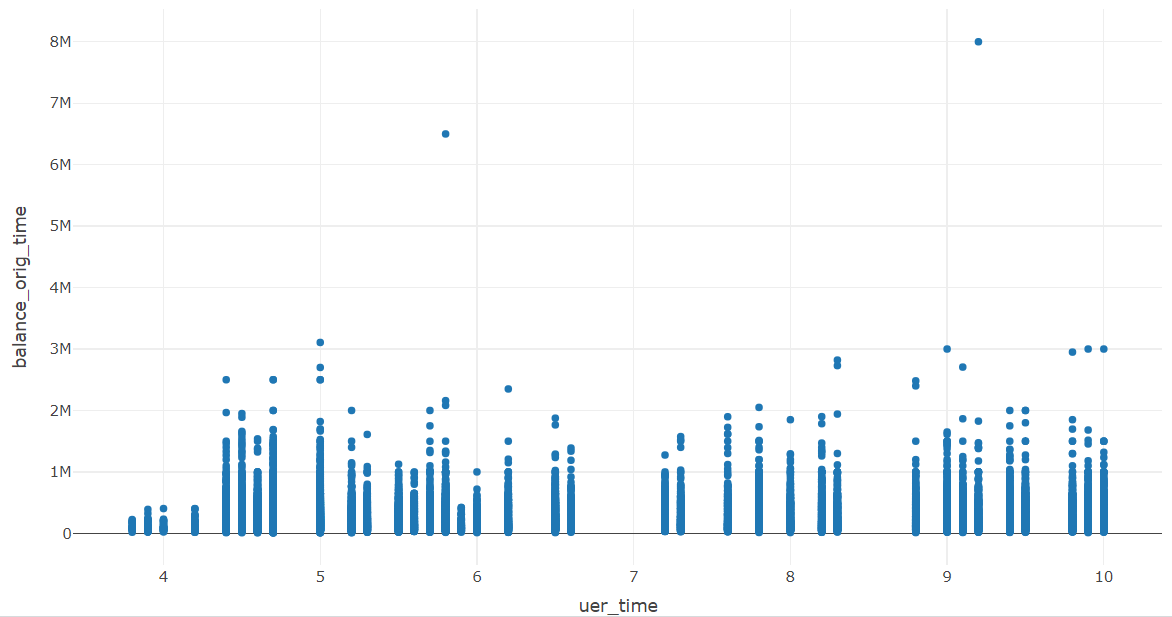
**uer\_time**

Which represents unemployment rate. Let’s try to understand how the column is impacting target columns.



**Figure 2.15 uer\_time vs status\_time**

Based on Figure 2.15, it is evident that most of the customers who took out loans when the unemployment rate was less than 6% are classified as payoff.

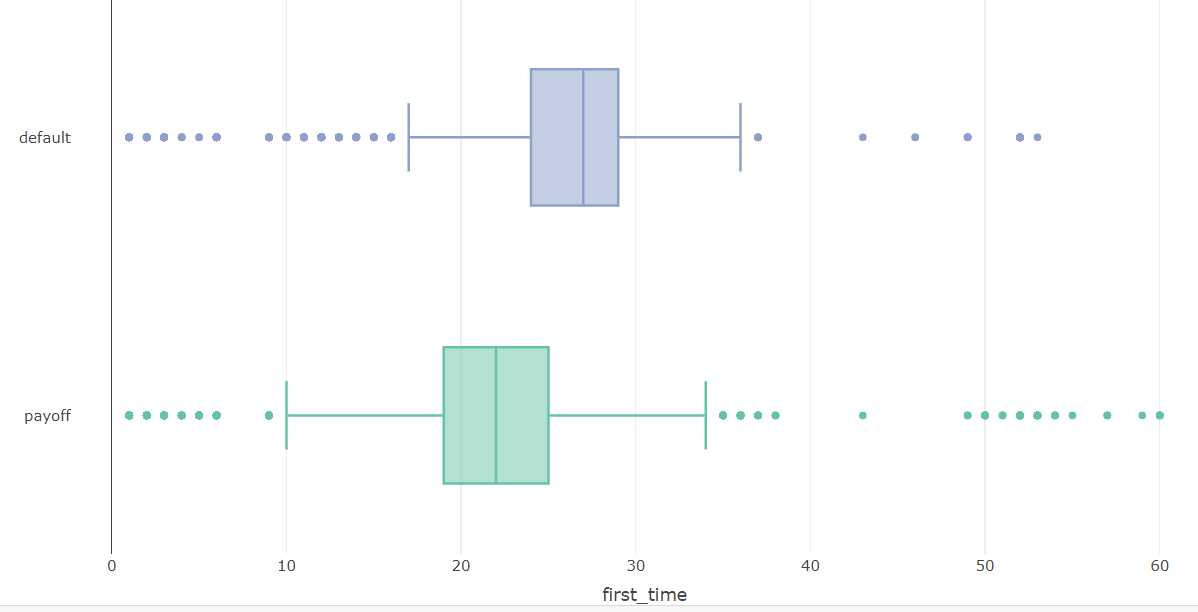
****

**Figure 2.16 uer\_time vs balance\_orig\_time**

Based on Figure 2.16, it is evident that customers who took out loans when the unemployment rate was less than 4% have less loan amount. Therefore, Bank can give lesser loan amount for customers who are applying for a loan when the unemployment rate is less than 4% .

**first\_time**

This column represents the time in months when the customer made his first payment. Let’s try to understand how this column is impacting the target columns.

****

**Figure 2.17 first\_time vs status\_time**

Based on Figure 2.17, it is evident that most of the customers who started their loan payment between 10 to 12 months are classified as payoff.

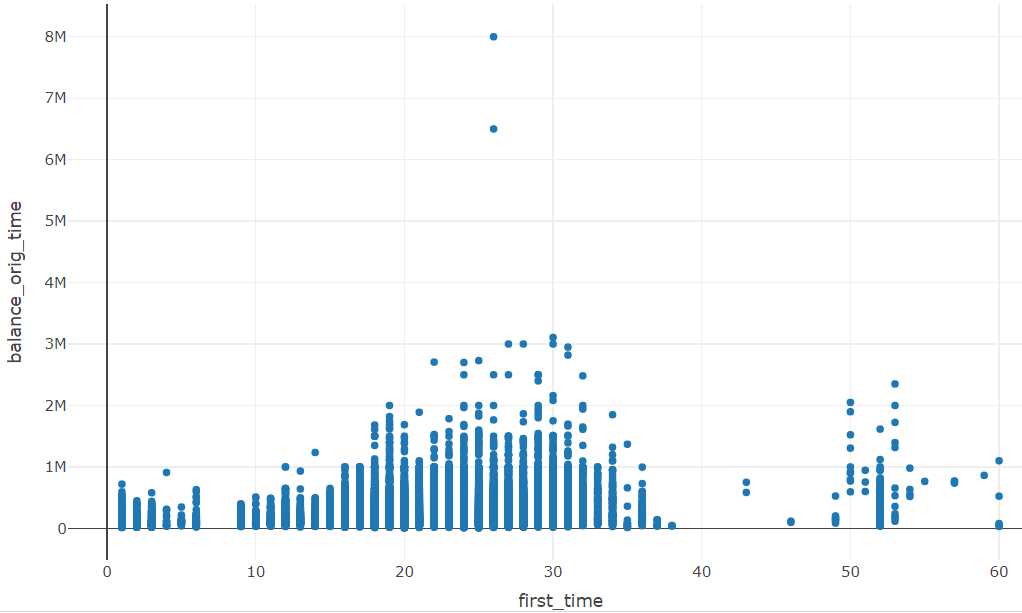
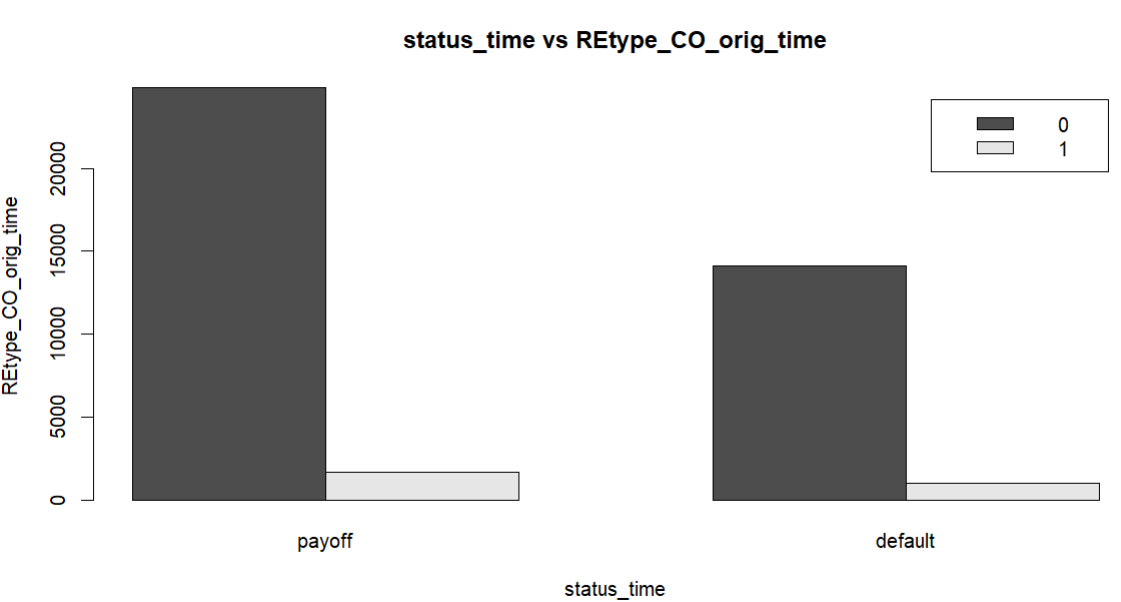


Figure 2.18 **first\_time vs balance\_orig\_time**

Based on Figure 2.18, it is evident that customers who borrowed high loan amounts started paying between 20 and 35 months.

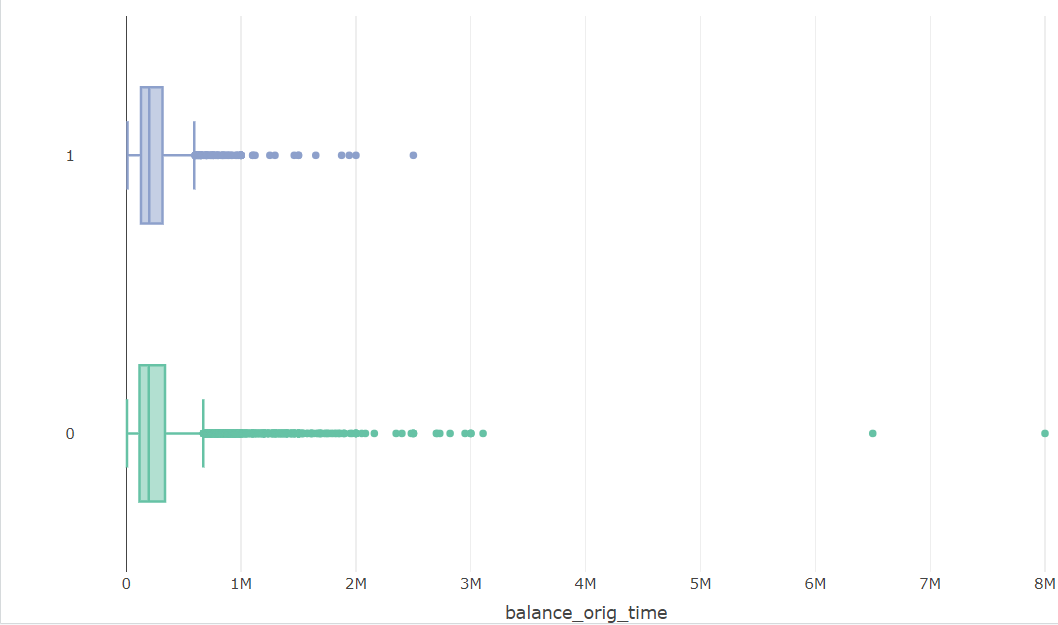
**REtype\_CO\_orig\_time:**

This column represents whether the real estate type is a condominium or not. Let’s try to understand how this column is impacting the offer\_top\_up column and the balance\_orig\_time.

****

**Figure 2.19 REtype\_CO\_orig\_time vs status\_time**

Based on Figure 2.19, it is evident that most of the customers whose property is not a condominium are classified as payoff.

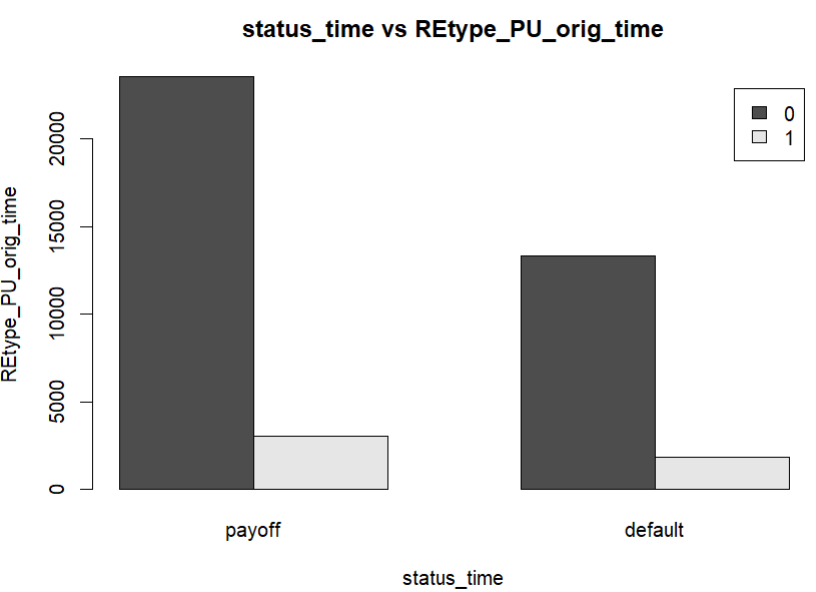
****

**Figure 2.20 REtype\_CO\_orig\_time vs balance\_orig\_time**

Based on Figure 2.20, it is evident that customers whose estate is not a condominium are approved for more loan amount compared to customers whose estate is a condominium. However, the bank may consider offering higher loan amounts to customers whose property is categorized as condominium.

**REtype\_PU\_orig\_time**

This column represents if the customer estate is planned as urban development. Let’s try to understand how the column is impacting offer\_top\_up column and balance\_orig\_time.

****

**Figure 2.21 REtype\_PU\_orig\_time vs status\_time**

Based on Figure 2.21, it is evident that most of the customers whose estate is not planned as urban development are classified as payoff.

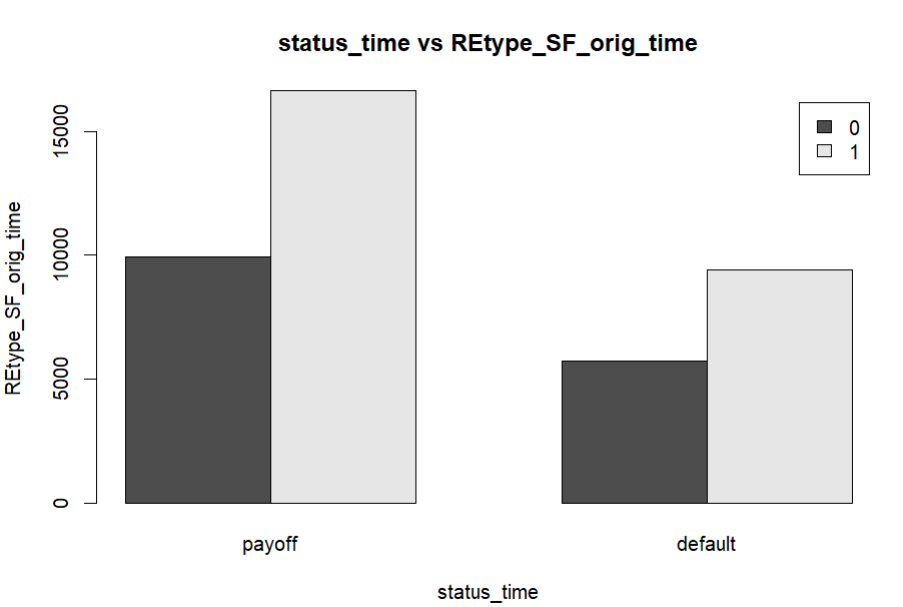
****

**Figure 2.22 REtype\_PU\_orig\_time vs balance\_orig\_time**

Based on Figure 2.22, it is evident that customers whose real estate is planned as urban development have higher loan amounts. Therefore, the bank may consider providing higher loan amounts to customers whose property is planned as urban development.

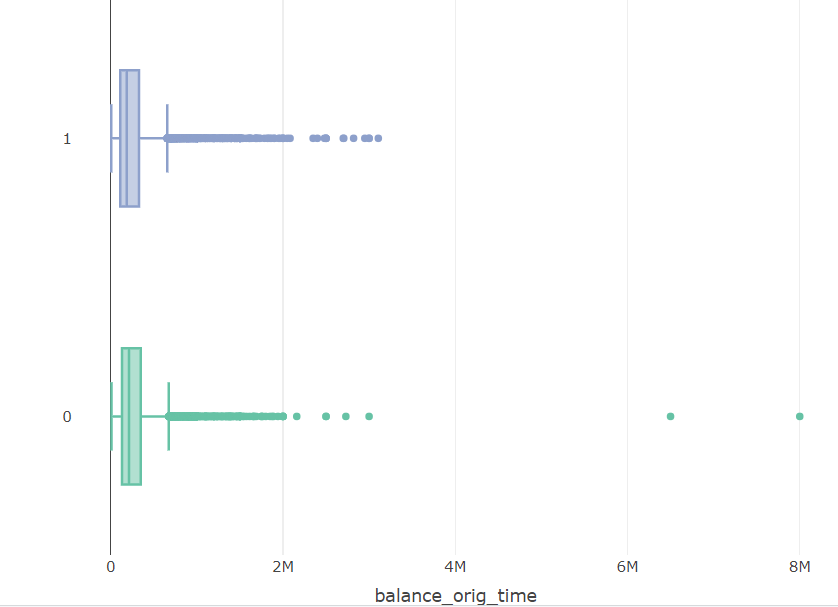
**REtype\_SF\_orig\_time**

Which represents if the customer estate is single family home. Let’s try to understand how the column is impacting offer\_top\_up column and balance\_orig\_time.

****

**Figure 2.23 REtype\_SF\_orig\_time vs status\_time**

Based on Figure 2.23, it is evident that a higher number of customers with non-single-family homes are classified as payoff.

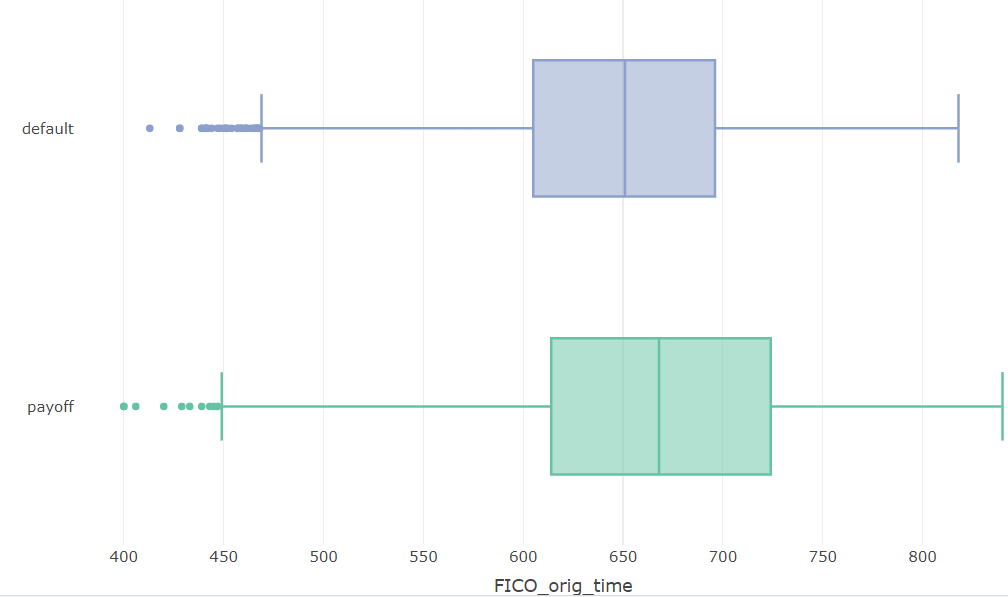
****

**Figure 2.23 REtype\_SF\_orig\_time vs balance\_orig\_time**

Based on Figure 2.23, it is evident that the box plot distribution is similar for customers with single-family homes and those without single-family homes. Therefore, it may not be useful for estimating the balance\_orig\_time.

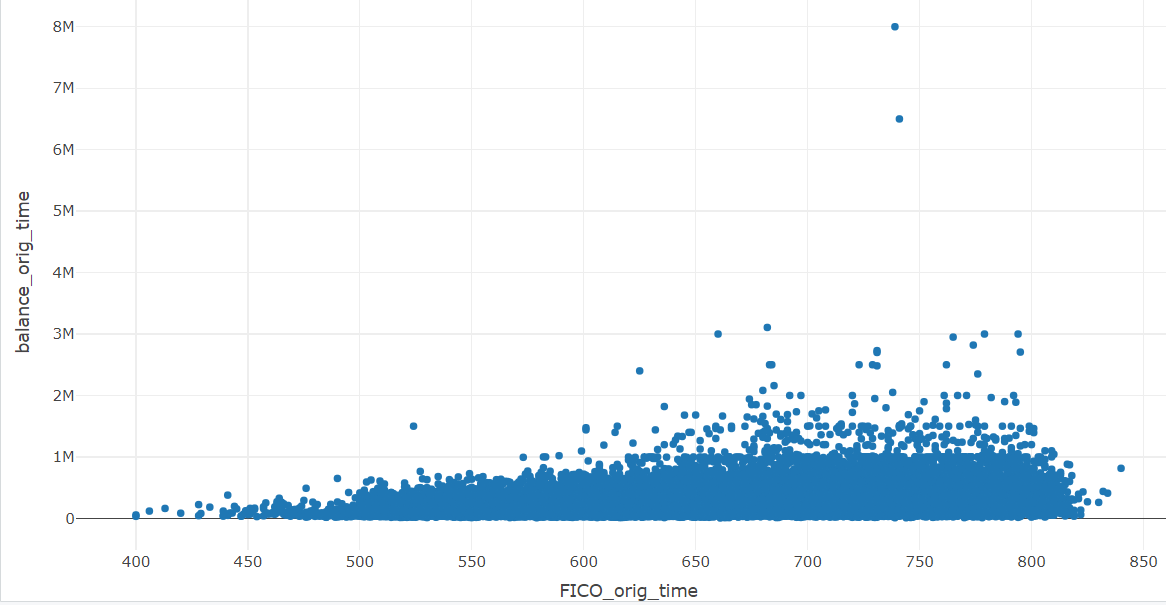
**FICO\_orig\_time**

This columns represents the credit score of the customer. Let’s try to understand how the column is impacting offer\_top\_up column and balance\_orig\_time

****

**Figure 2.24 FICO\_orig\_time vs status\_time**

Based on Figure 2.24, it is evident that most of the customers with a FICO score of more than 700 are classified as payoff.

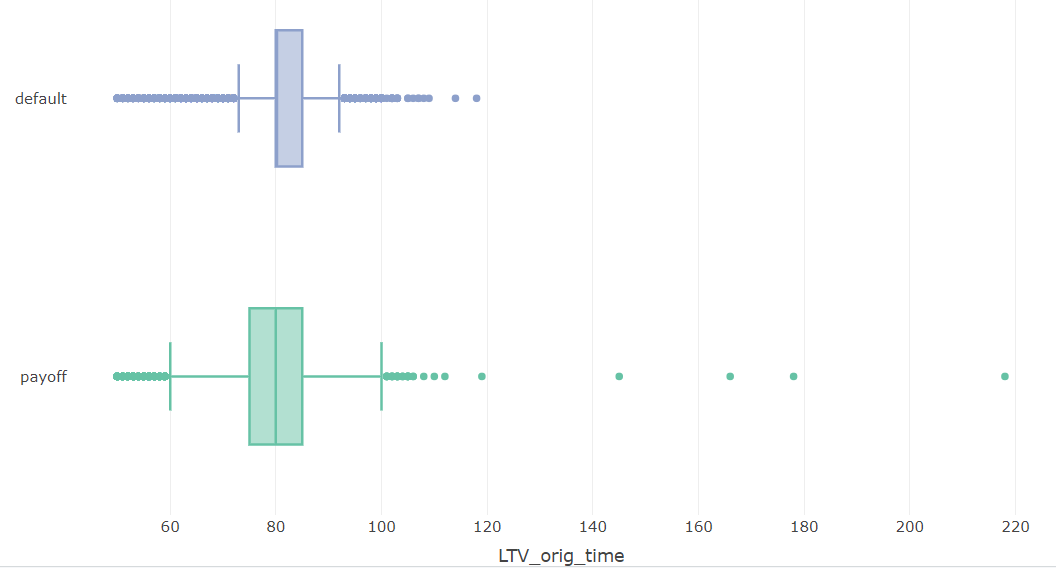
****

**Figure 2.25 FICO\_orig\_time vs balance\_orig\_time**

Based on Figure 2.25, it is evident that customers with a FICO score greater than 600 have higher loan amounts. Therefore, the bank may consider offering higher loan amounts to customers with FICO scores exceeding 600.

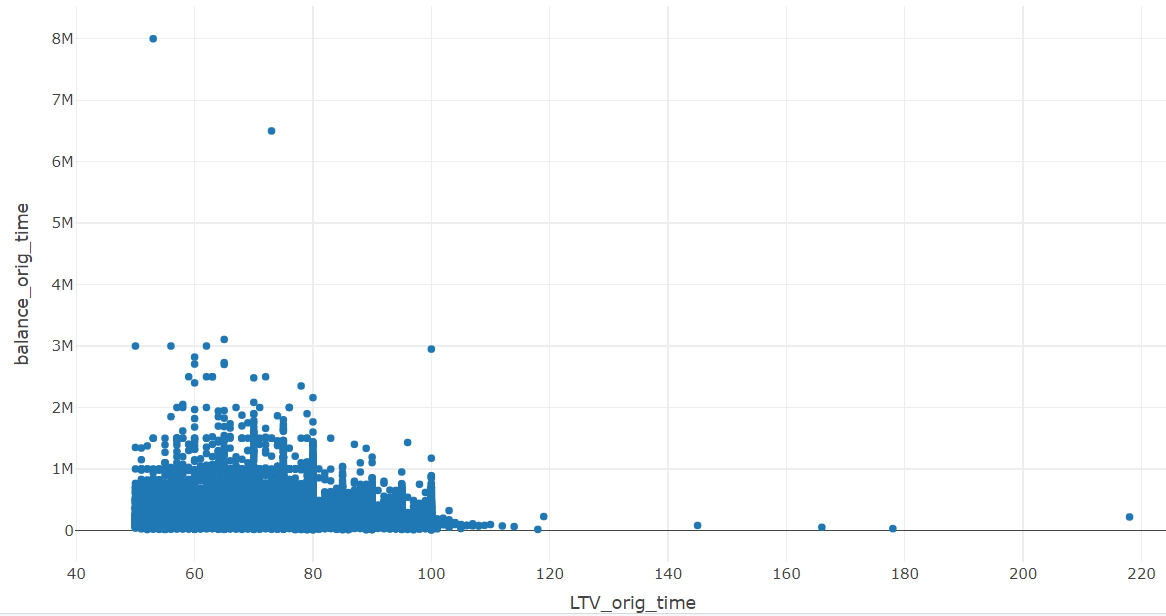
**LTV\_orig\_time**

This column represents loan to value ratio of the property at the loan origination time. Now lets try to understand how the column is impacting the target columns

****

**Figure 2.26 ltv\_orig\_time vs status\_time**

Based on Figure 2.26, it is evident that most customers whose property has an LTV origination ratio between 60% and 100% are classified as payoff.

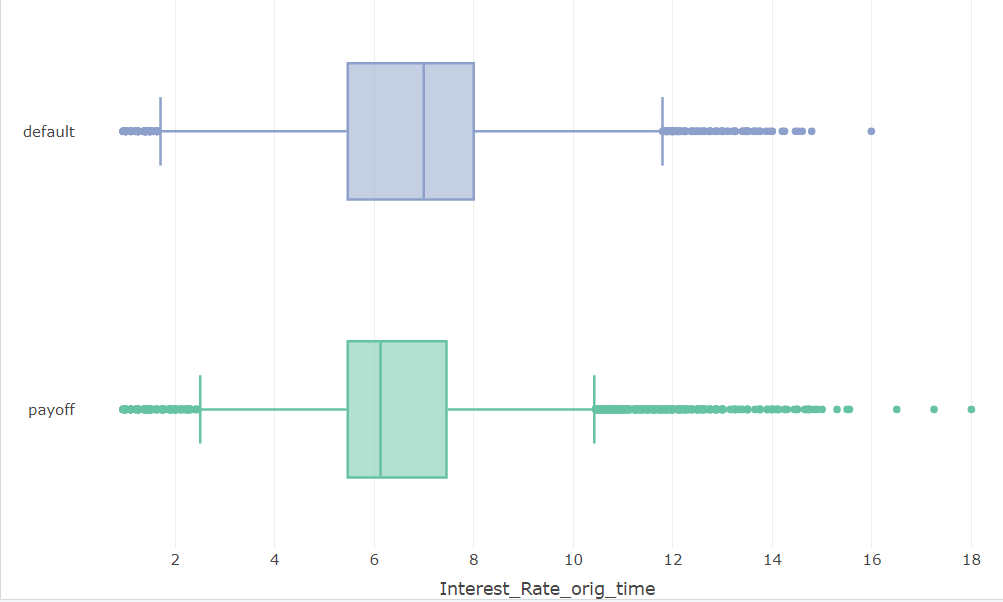
****

**Figure 2.27 LTV\_orig\_time vs balance\_orig\_time**

Based on Figure 2.27, it is evident that customers whose property has an LTV ratio greater than 100% at the loan origination time are offered less loan amount. Therefore, the bank may consider offering lesser loan amounts for customers whose property LTV ratio at origination is greater than 100%.

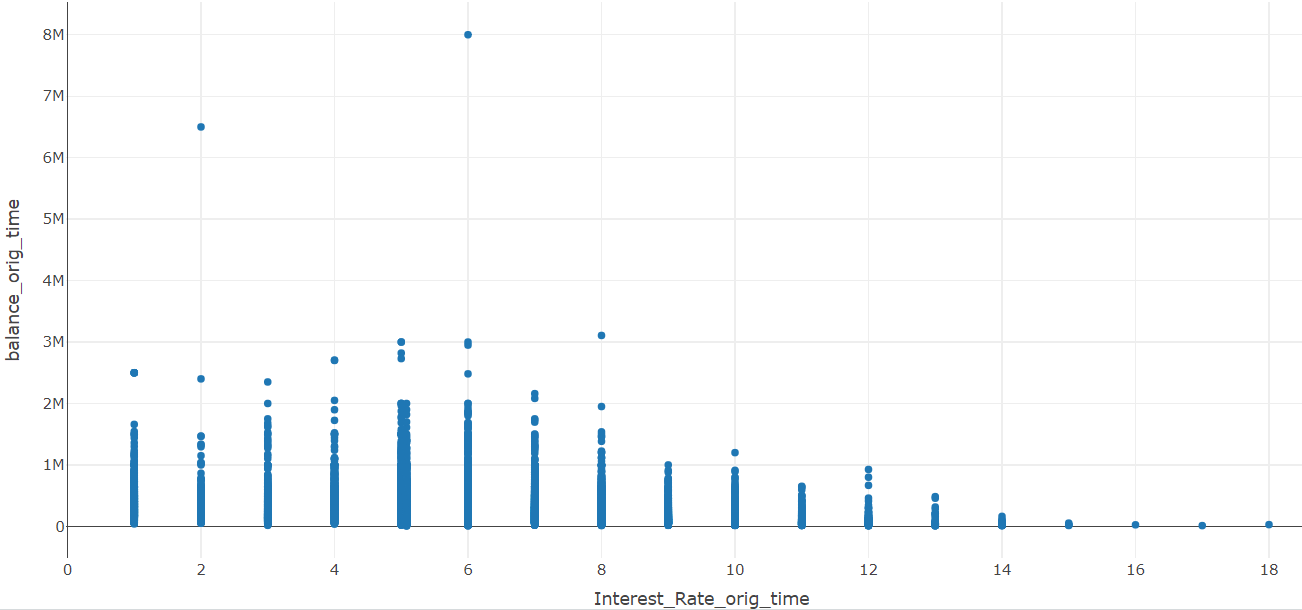
**interest\_rate\_orig\_time**

This column represents the interest rate at the time of loan orgination time. Let’s explore the interest\_rate\_orig\_time is impacting the target column.

****

**Figure 2.28 Interest\_Rate\_orig\_time vs offer\_top\_up**

Based on Figure 2.28, it is evident that most customers who have an interest rate at loan origination time of more than 7% are classified as default.

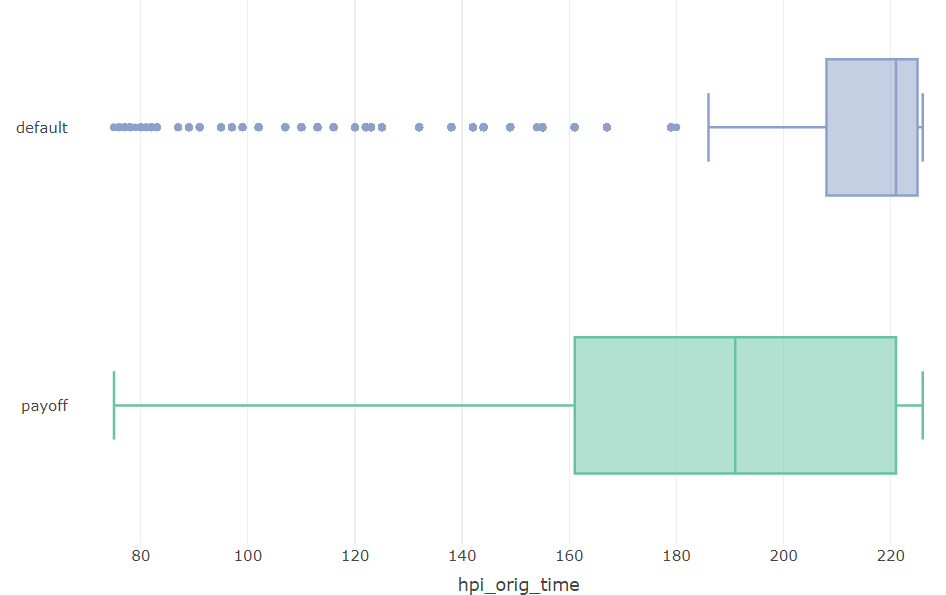
****

**Figure 2.29 Interest\_Rate\_orig\_time vs balance\_orig\_time**

Based on Figure 2.29, it appears that customers paying an interest rate at origination time greater than 15% have lesser loan amounts. Therefore, the bank may consider offering a higher interest rate at origination time for customers who borrow lesser amounts.

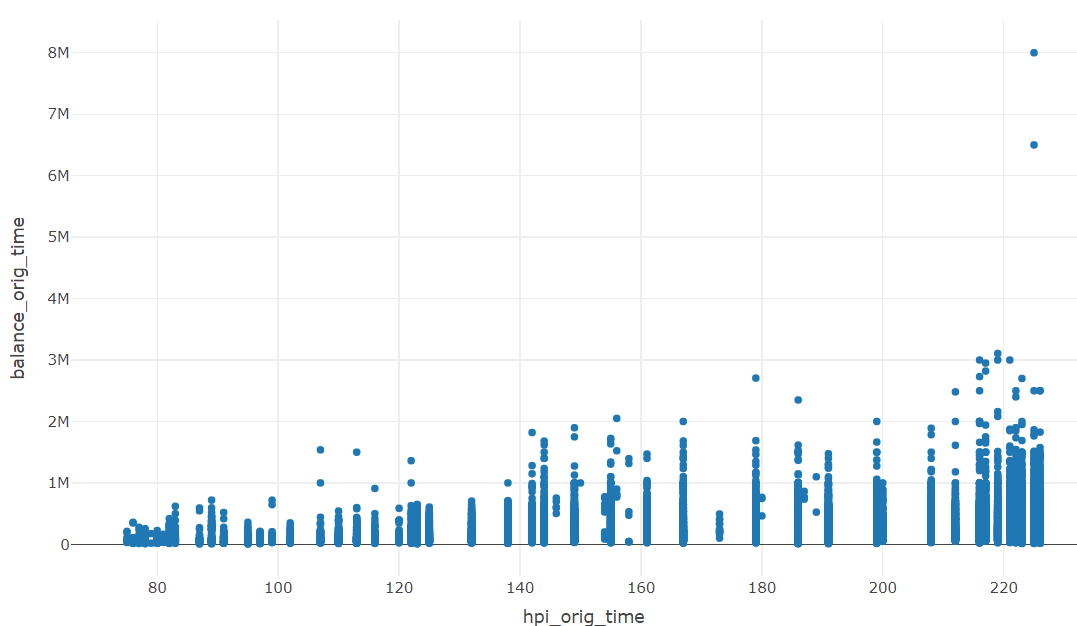
**hpi\_orig\_time**

This column indicates hpi value for the property at the origination of the loan. Now lets see how the hpi origination time is impacting the target columns.

****

**Figure 2.30 hpi\_orig\_time vs status\_time**

Based on Figure 2.30, it is evident that customers whose property has an HPI value at loan origination less than 210 are classified as payoff.

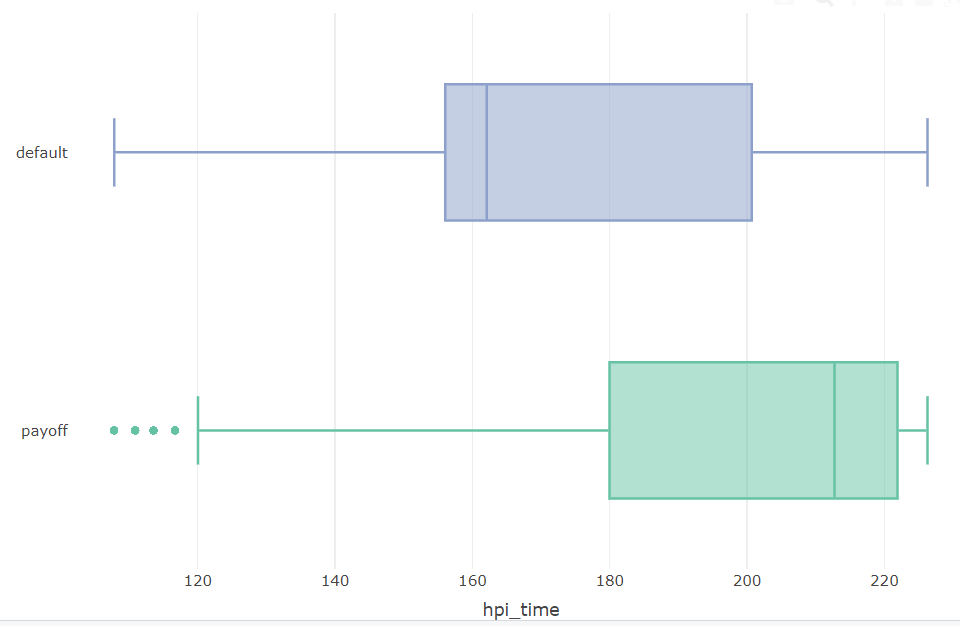
****

**Figure 2.31 hpi\_orig\_time vs balance\_orig\_time**

Based on Figure 2.31, it is evident that customers whose property has an HPI value at origination time greater than 140 have higher loan amounts. Therefore, the bank may consider offering more loan amount to customers whose HPI value at origination time is greater than 140.

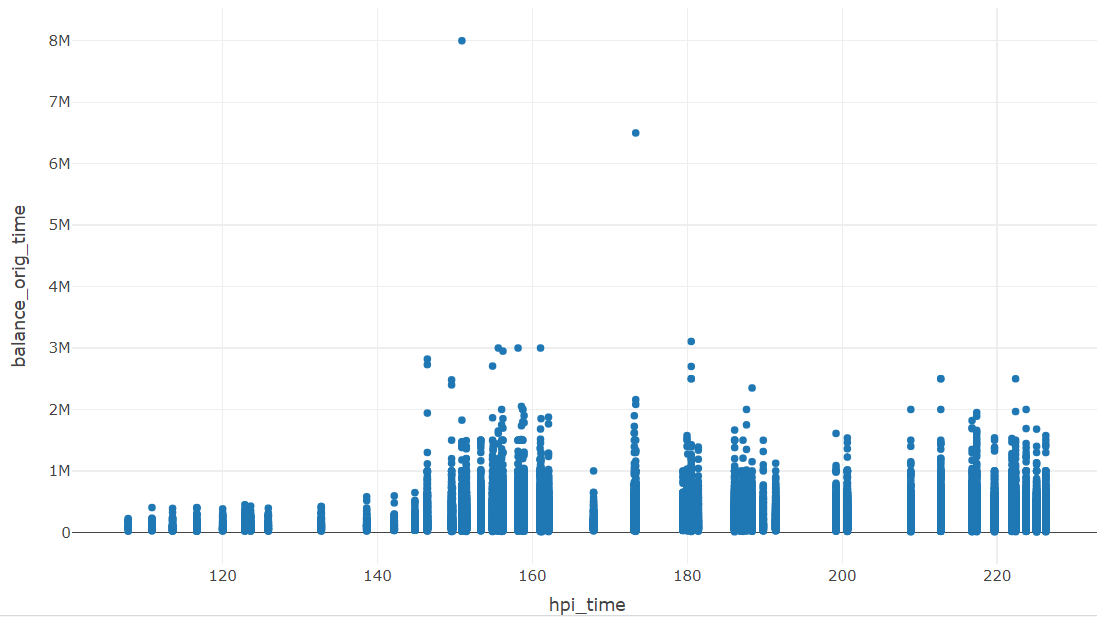
**hpi\_time**

This column represents the hpi index value for the customer property. Now lets see how the hpi origination time is impacting the target columns.

****

**Figure 2.32 hpi\_time vs status\_time**

Based on Figure 2.32, it is evident that most of the customers whose property has an HPI value greater than 180 are classified as payoff

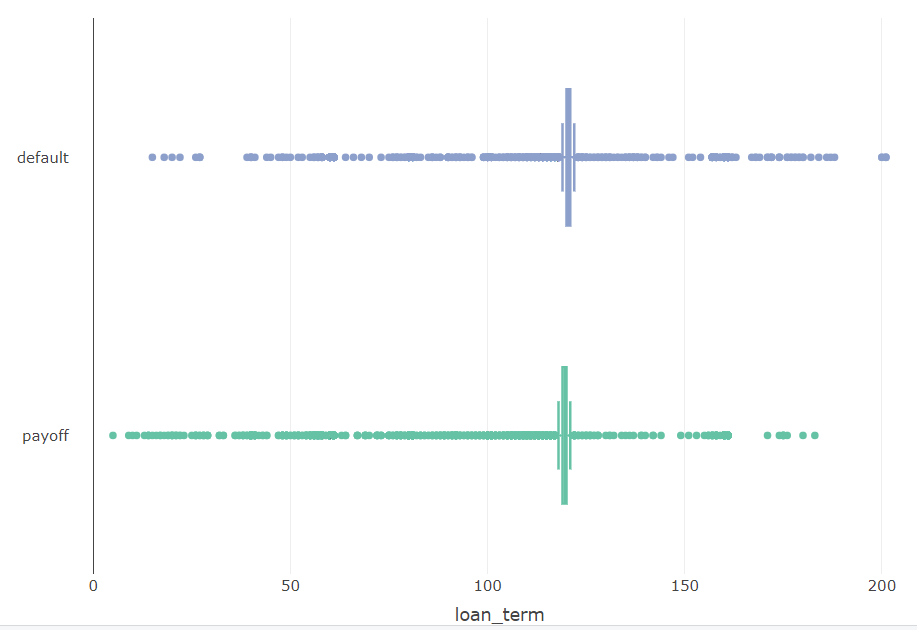
****

**Figure 2.33 hpi\_time vs balance\_orig\_time**

Based on Figure 2.33, it is evident that customers whose property has an HPI value greater than 145 have higher loan amounts. Therefore, the bank may consider offering more loan amount to customers whose HPI value is greater than 145.

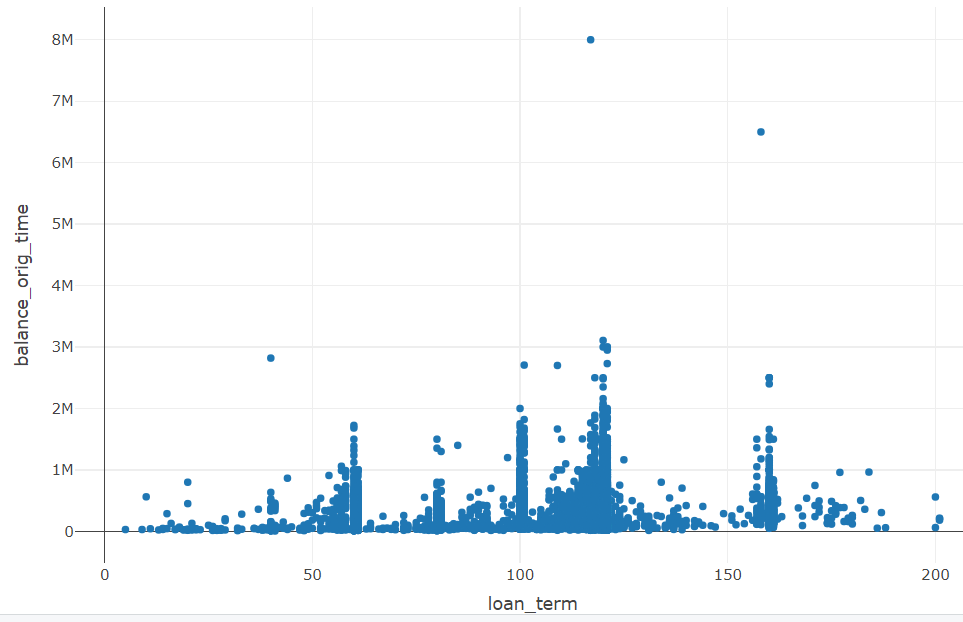
**Loan\_term**

This column indicates the actual loan term of the customer now let’s see how the column is impacting the target columns.

****

**Figure 2.34 loan\_term vs status\_time**

From the figure 2.34 it is clear that there is almost equal distribution for both default and payoff. So this column might not be useful for classifying the customer.

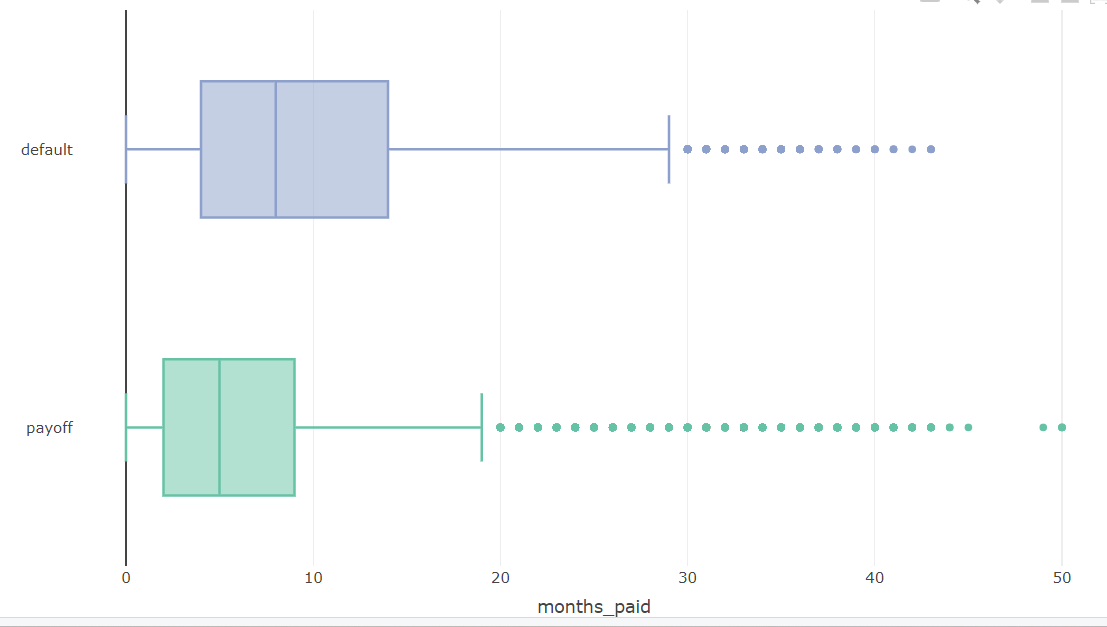


**Figure 2.35 loan\_term vs balance\_orig\_time**

From the figure 2.35 it is clear that customer who has the loan term between 100 to 150 have higher loan amount. So bank can provide higher loan amount for the customer whose loan term is in between 100 to 150.

**Months\_paid**

This column indicates the how many months customer paid the loan. Now lets find how the column is impacting the target columns

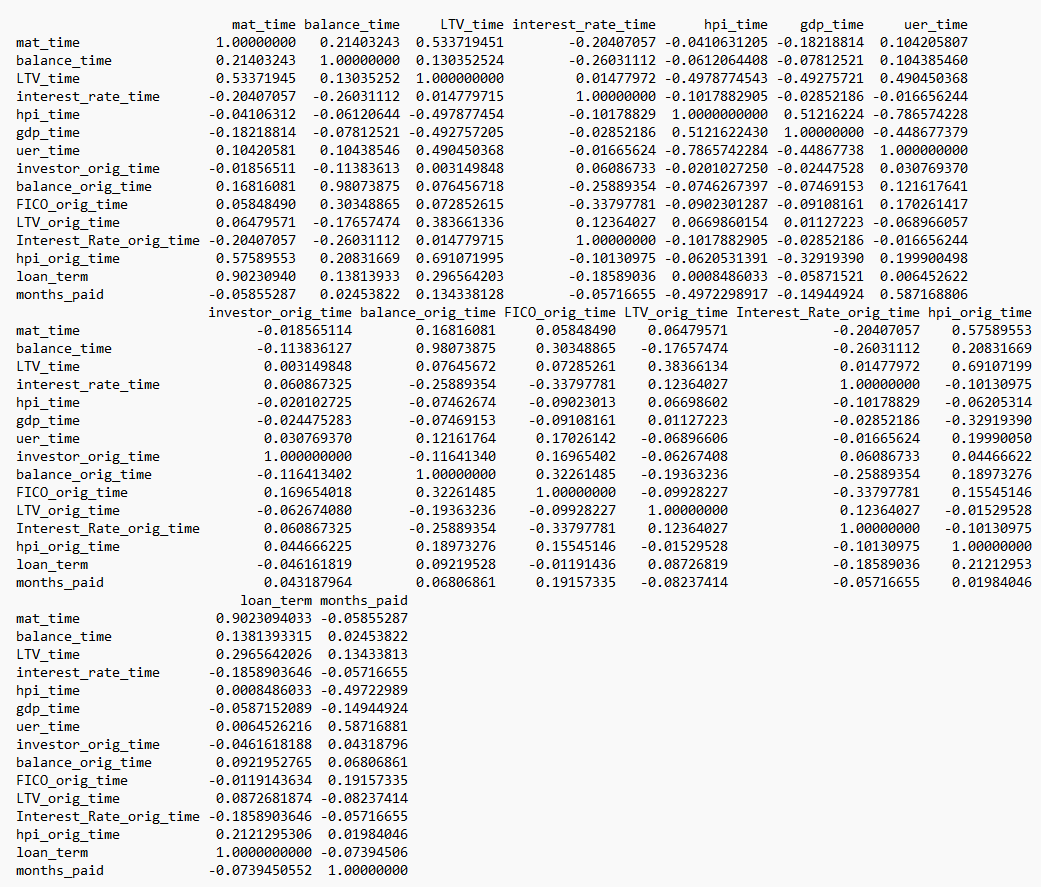
****

**Figure 2.36 months\_paid vs status\_time**

From the figure 2.36 it is clear that customers who paid less than 10 months are classified as payoff.

**Correlation Analysis**

Now let’s try to understand the correlation between all the numerical predictors.

****

**Figure 2.37 correlation analysis**

Findings of the correlations from figure 2.37 are

Mat\_time has the high positive correlation with ltv\_time and hpi\_orig\_time, loan\_term.

Balance\_time has the high positive correlation with balance\_orig\_time.

Ltv-time has the high positive correlation with mat\_time column

Interest\_rate\_time has the high positive correlation with interest\_orig\_time.

Hpi\_time has the high negative correlation with uer\_time.

Gdp\_time does not have high correlation with any other column.

Uer\_time has high negative correlation with hpi\_time and high positive correlation with months\_paid.

Investor\_orig\_time does not have high correlation with any other column.

Fico\_orig\_time dose not have high correlation with any other column

Ltv\_orig\_time does not have high correlation with any other column.

Interest\_rate\_orig\_time has high positive correlation with interest\_rate\_time.

Hpi\_orig\_time has high positive correlation with ltv\_time and mat\_time.

Loan\_term has high correlation with mat\_time.

Months\_paid has high positive correlation with uer\_time.

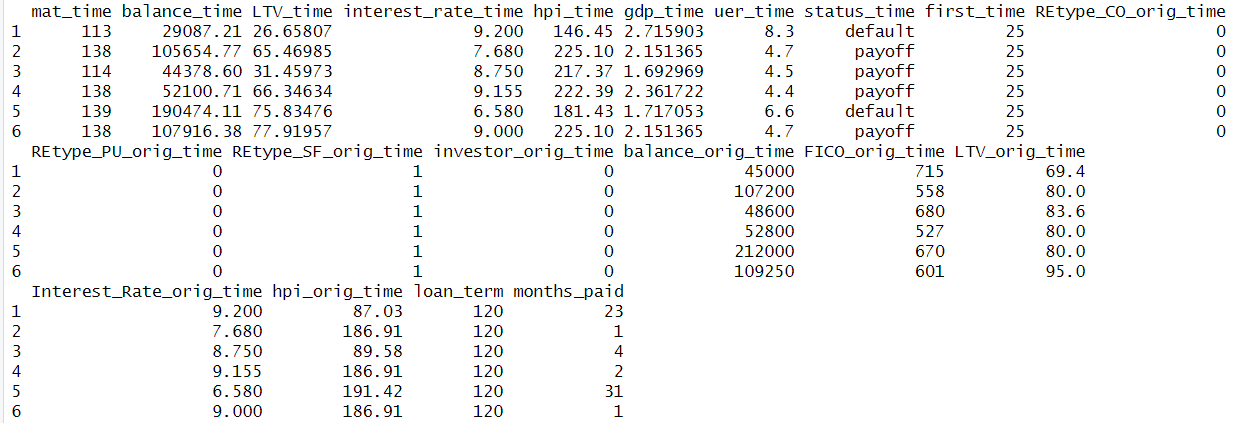
Columns with high positive or negative correlations indicate duplicative information between them.

Only balance\_time has a high correlation with the target column balance\_orig\_time. However, according to domain knowledge, balance\_time may not be a useful predictor for estimating balance\_orig\_time, as it simply represents the balance amount that the customer needs to pay to clear the loan.

**Outliers**

During the exploratory data analysis (EDA), outliers were detected in all numerical columns. However, these outliers are considered expected values based on domain knowledge. Therefore, they are retained in the dataset without any treatment or removal.

**Normalizing and Rescaling Data**

****

**Figure 2.38 First 6 rows of the data**

Figure 2.38 clearly shows that the columns have values on different scales. When utilizing KNN models or neural networks, standardization or normalization of the data is necessary.

1. **Dimension Reduction**

**Practical considerations**

Based on domain knowledge and insights acquired from the Exploratory Data Analysis (EDA), different columns are utilized as predictors for specific tasks. For the classification task of identifying new customers as either default or payoff, the predictors include: gdp\_time, uer\_time, REtype\_CO\_orig\_time, REtype\_PU\_orig\_time, REtype\_SF\_orig\_time, investor\_ orig\_time, FICO\_orig\_time, LTV\_orig\_time and hpi\_orig\_time.

For the regression task aimed at estimating the balance\_orig\_time, the predictors are the same: gdp\_time, uer\_time, REtype\_CO\_orig\_time, REtype\_PU\_orig\_time, REtype\_SF\_orig\_time, investor\_orig\_time, FICO\_orig\_time, LTV\_orig\_time, and hpi\_orig\_time.

For the classification task of determining whether ongoing customers will default or payoff, the predictors used are: mat\_time,balance\_time, LTV\_time, hpi\_time, gdp\_time, uer\_time, status\_time, first\_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, loan\_term, and months\_paid.

**Dimension Reduction using correlation analysis**

All the columns that are high positive or negative correlations have the duplicate information. So, one of the columns can be removed to reduce the multicollinearity. Out of all columns mat\_time and loan\_term has highest correlation of 0.90 so mat\_time column is removed in modelling because from the domain knowledge loan\_term is important predictors in determining whether an ongoing customer will become default or payoff.

**Dimension Reduction using regression Models.**

Forward selection and backward selection algorithms utilize predictors that are statistically significant for predicting the outcome columns. Predictors that are not included in the forward or backward selection model can be removed.

**Dimension reduction using Decision Tree**

The resulting tree from the decision tree is used to identify important features. Predictors that are not included in the decision tree can be removed.

1. **Partition the Data**

As it is a supervised task, utilizing the complete data to build and test the model's performance may introduce optimism bias. This bias arises because the model may perform well with the current dataset, but it might not generalize effectively to real-world scenarios due to factors such as slight variations in the data. So, the data needs to be partitioned into train, validation, and test sets.

The train partition is the largest partition containing the data to train various models. The same training data is used across different models. The validation partition is utilized to assess the predictive performance of each model and select the best model from the options available. The test partition is used to evaluate the performance of the chosen model with new data.

Partitioning the data into train, validation, and test sets is done randomly according to the proportions of 60%, 20%, and 20%, respectively. The records in the train, validation, and test partitions should be distinct. After partitioning the dataset, the train partition contains 25038 records, the validation partition contains 8346 records, and the test partition contains 8347 records.

1. **Modelling For Classification Task of New Customers**

In this phase, the focus shifts to fitting classification machine learning models to classify new customer as default or payoff. Based on the classification, Bank can selectively offer loan to new customers who are classified as payoff thereby reducing the incidence of loan become defaults. The target variable for the classification model is the status\_time column, while the predictors include gdp\_time, uer\_time, REtype\_CO\_orig\_time, REtype\_PU\_orig\_time, REtype\_SF\_orig\_time, investor\_orig\_time, FICO\_orig\_time, LTV\_orig\_time and hpi\_orig\_time.. The performance of the models is measured using a confusion matrix. Here, the payoff class category is considered the important class because it’s crucial to correctly classify payoff customers rather than default customers.

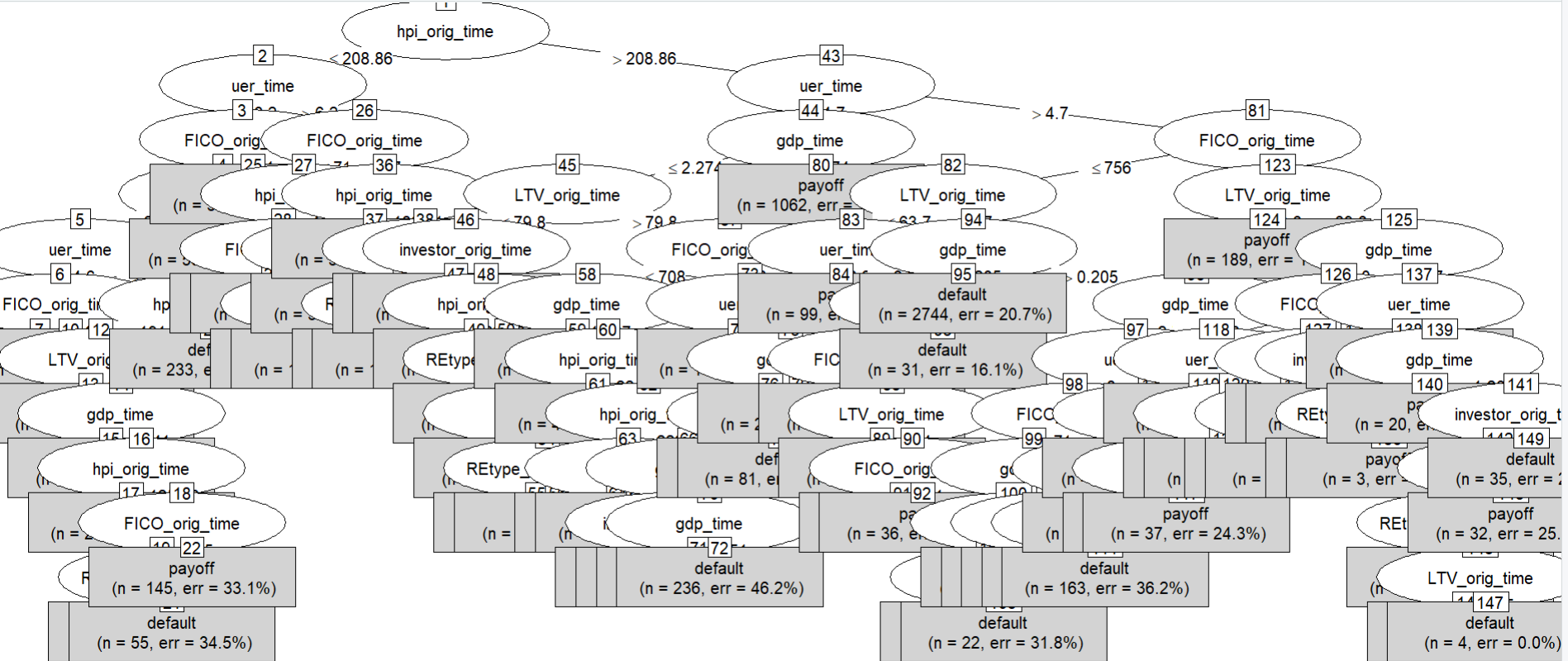
* 1. **Decision Tree**

Decision tree is based on separating records into subgroups by creating split on predictors. These splits create logical if then prediction rules which are simple and easy to interpret. Decision trees can be applied for both classification and regression. There are multiple ways to fit the classification decision tree in one of the ways is using C5.0 library.

**Model Interpretation**

When classifying new records, the record will dropdown to the terminal node and based on the rules created by the decision tree the new record will be classified.

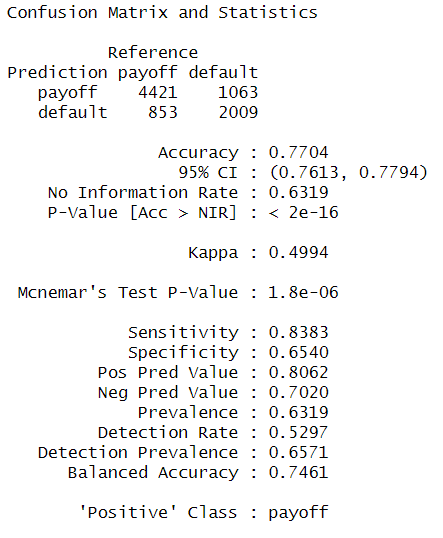
Using the plot function in R, one can visualize the tree structure.

****

**Figure 5.1 Tree structure**

Predictors used in the decision tree are considered as the important predictors for classifying if a new customer will be payoff the loan or default.

* + 1. **Measuring Performance**

****

**Figure 5.2 confusion matrix on valid data**

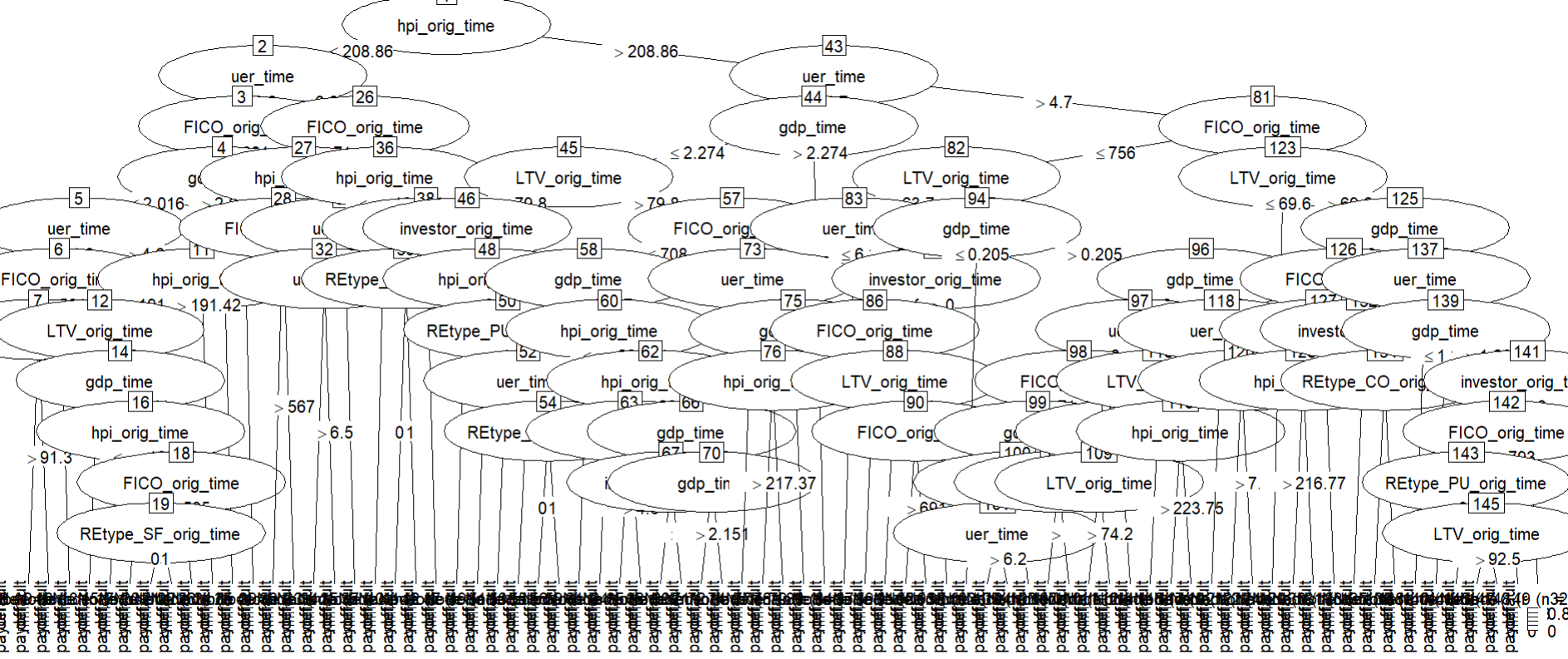
From Figure 5.2, the sensitivity rate is 83.83%, which indicates that model correctly classified 83.83% of class payoff members.

* + 1. **Improving Model Performance**

The performance of the decision tree model can be improved by using an extra parameter trail while building the model.

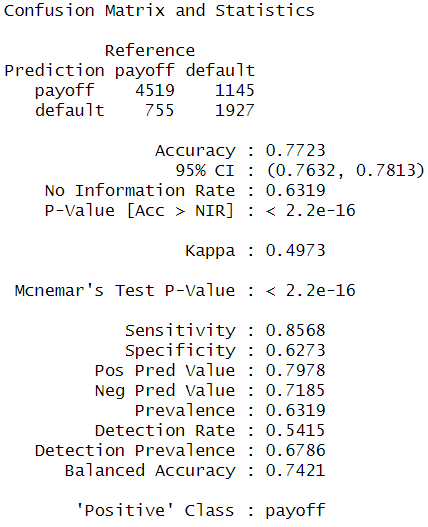
**Trails option 6**

After trying various trail options such as 6, 10, 15, and 20, it was found that trail option 6 has the highest sensitivity rate. Therefore, trail option 6 is considered as the best model among all other trail models.

****

**Figure 5.3 Tree Structure for trail 6**

**Measuring Performance**

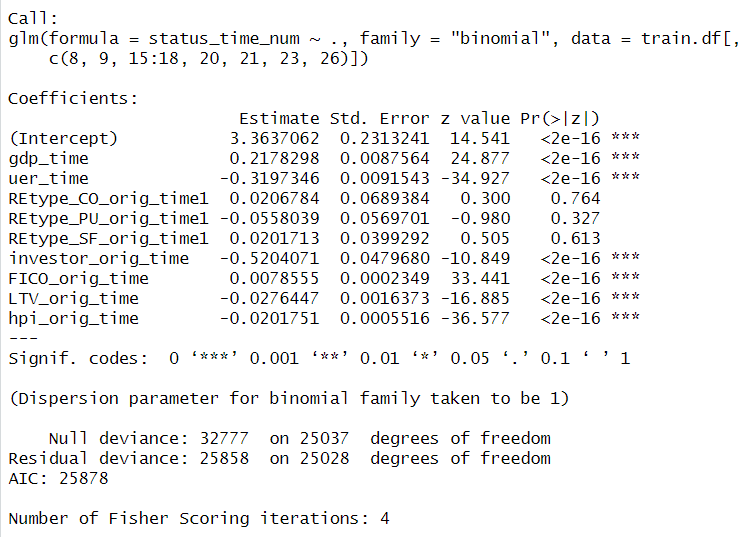
****

**Figure 5.4 confusion matrix on validation set for trail option 6**

From the figure 5.4 the sensitivity rate is 85.68% which indicates that 85.68% of the class payoff members in the validation set is correctly classified.

* 1. **Logistic Regression**

Logistic regression provides estimates of propensities indicating the likelihood of each record belonging to each class. Subsequently, Use the threshold values to classify each case into one of the classes. In R GLM is used to fit the logistic regression. The GLM model works well if the target column has 0s and 1s, so before applying logistic regression, the target column status\_time is converted to 1s and 0s by assigning payoff category to 1 and default to 0 in the train data.

****

**Figure 5.5 Summary of the logistic regression for train data**

**Estimated Model Equation**

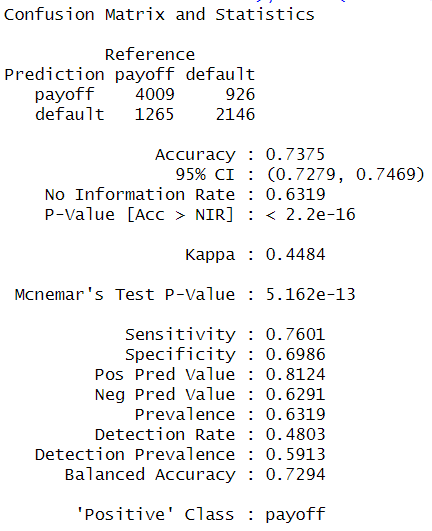
Logit(status\_time\_num = yes) = 3.3637062+(gdp\_time \* 0.2178298)+ (uer\_time \*-0.3197346)+ (REtype\_CO\_orig\_time1\*0.0206784)+ (REtype\_PU\_orig\_time1\*-0.0558039) + (REtype\_SF\_orig\_time1 \*0.0201713)+ (investor\_orig\_time \* -0.5204071)+ (FICO\_orig\_time \*0.0078555) + (LTV\_orig\_time\*-0.0276447)+(hpi\_orig\_time\*-0.0201751)

**Model Interpretation**

Based on the model estimated equation the new records are classified as class 1(payoff) or 0(default).Columns gdp\_time, uer\_time,investor\_orig\_time, FICO\_orig\_time, LTV\_orig\_time and hpi\_orig\_time are most significant columns for classifying if a customer belongs to class 1(payoff) or 0(default).

* + 1. **Measuring Performance**

For easy to understand the confusion matrix class 1 and 0’s are converted as payoff and default.

****

**Figure 5.6 confusion matrix for logistic regression on valid data**

The sensitivity value of 76.01 indicates that 76.01% of payoff members are correctly classified.

* + 1. **Improving Model Performance**

In this section, the aim is to construct Backward selection models to assess whether there is an improvement in predictive performance compared to logistic regression.



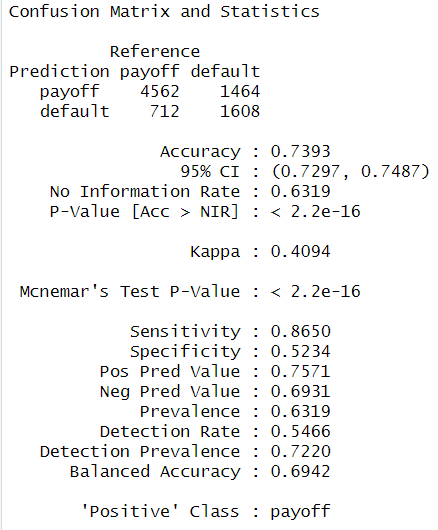
**Figure 5.7 Summary of the backward selection model**

**Model Interpretation**

Based on the model estimated equation the new records are classified as class 1(payoff) or 0(default).Columns gdp\_time, uer\_time,investor\_orig\_time, FICO\_orig\_time, LTV\_orig\_time and hpi\_orig\_time are most significant columns for classifying if a customer belongs to class 1(payoff) or 0(default).

**Measuring Performance**

For easy to understand the confusion matrix class 1 and 0’s are converted as payoff and default.

****

**Figure 5.8 Confusion matrix for backward selection model**

The sensitivity is 86.50 which indicates that 86.50% of payoff members are correctly classified.

1. **Classifier Model Selection**

The model which is giving high sensitivity is considered as the best model. So, all models sensitivities are compared

**Table 6.1 Accuracy Comparison of models**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Sensitivity |
| Decision Tree | 77.04% | 83.83% |
| Decision Tree with trail 6 | 77.23% | 85.68% |
| Logistic regression | 73.75% | 76.01% |
| Backward selection model | 73.93% | 86.50% |

From table 6.1 the sensitivity rate of backward selection model is higher than all other models. So, the backward selection model is considered as the best model to classify the customer new customer as payoff or default.

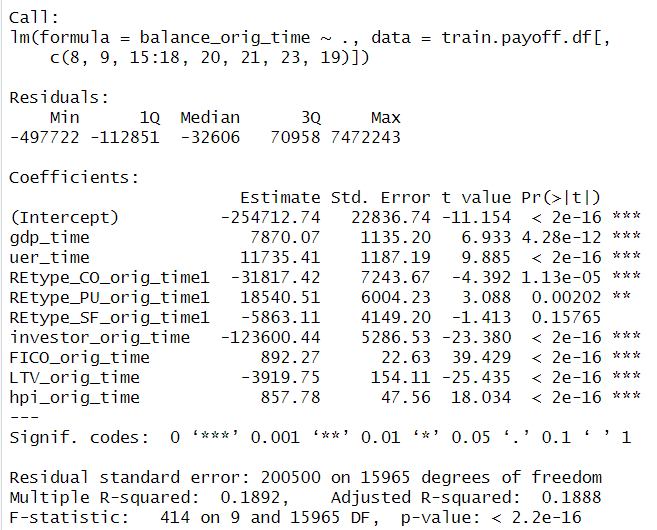
So bank can use the best classifier model to classify if the new customer who applied new loan will payoff or default. Based on the model output bank can selectively offers loan to customer and reduce the risk of loan going defaults.

1. **Modelling for regression Task**

From now on, the prediction will focus on determining the appropriate loan amount for customers who are classified as payoff customers. In this process, only payoff customer data will be filtered for both the training and validation datasets. After filtering, there are 15975 records in the training data and 5274 records in the validation data. Balance\_orig\_time column is considered as the target column and gdp\_time, uer\_time, REtype\_CO\_orig\_time, REtype\_PU\_orig\_time, REtype\_SF\_orig\_time, investor\_orig\_time, FICO\_orig\_time, LTV\_orig\_time and hpi\_orig\_time columns are considered as the predictors.

* 1. **Linear Regression Model**

The lm function is used to fit a multiple linear regression model. The multiple linear regression tries to form linear relationship between predictors and outcome column.

****

**Figure 7.1. summary of linear regression model**

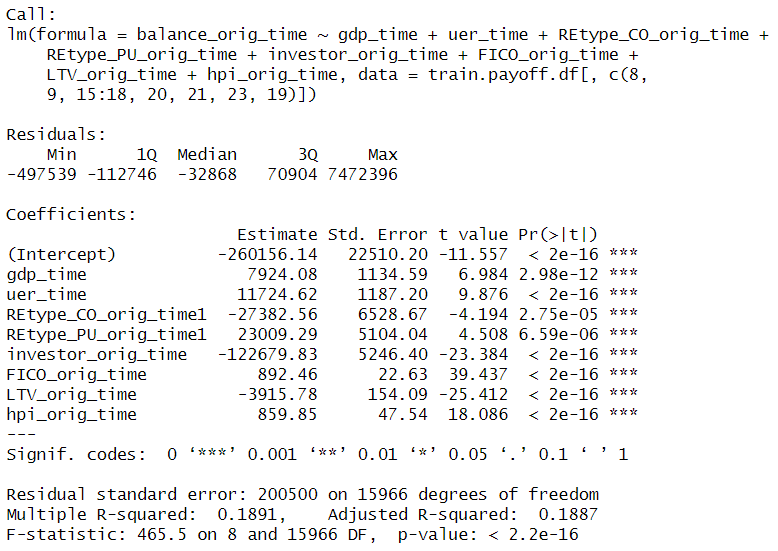
The columns gdp\_time, uer\_time,investor\_orig\_time, FICO\_orig\_time, LTV\_orig\_time and hpi\_orig\_time, REtype\_CO\_orig\_time, REtype\_PU\_orig\_time are important predictors for estimating the appropriate loan amount.

* + 1. **Measuring Performance**

To measure performance, RMSE is used. The model with lowest value of RMSE value is the best model. The RMSE value for the validation data is 180139.7.

* + 1. **Improving Model Performance**

In this section, the aim is to construct Backward selection models to assess whether there is an improvement in predictive performance compared to multiple linear regression model. Backward selection algorithm will start with all predictors and in each step will remove predictors one by one that are not important.



**Figure 7.2. summary of backward linear regression model**

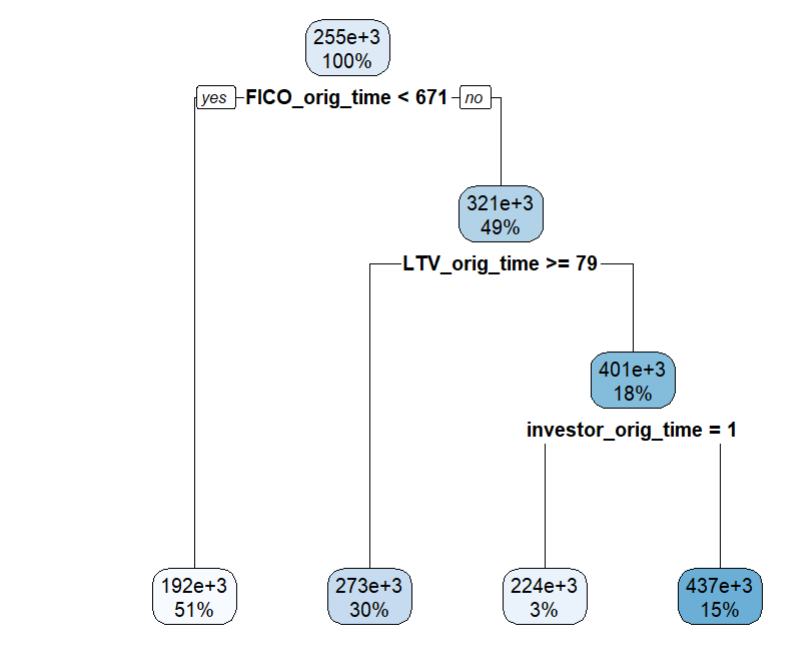
The columns gdp\_time, uer\_time,investor\_orig\_time, FICO\_orig\_time, LTV\_orig\_time and hpi\_orig\_time, REtype\_CO\_orig\_time, REtype\_PU\_orig\_time are important predictors for estimating the appropriate loan amount.

**Performance measure**

The RMSE of the backward selection model is 180177.4

* 1. **Decision Tree For regression (Regression Tree)**

Regression tree is non-parametric so there is no assumption for the model.Regression tree can be applied in R by using RPART.



**Figure 7.3 Decision tree**

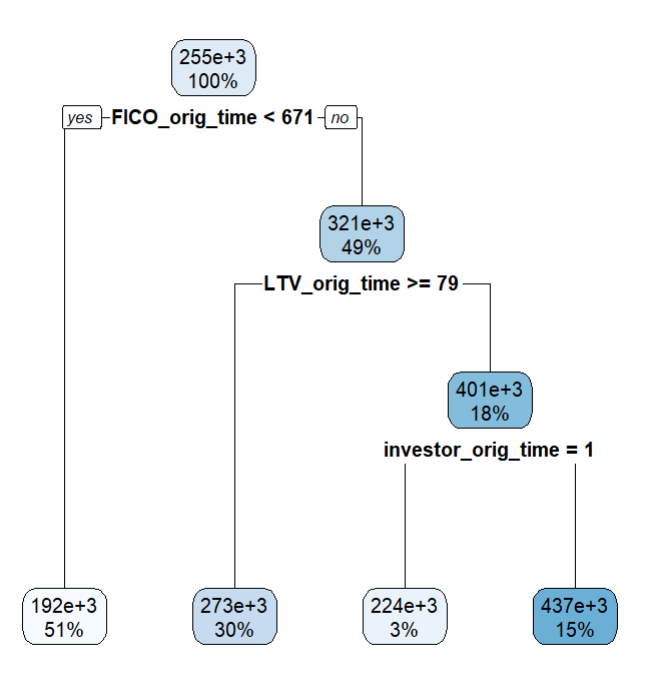
When predicting a new customer balance\_orig\_time, the record will drop down to the terminal node based on the predictors' information, and the prediction will be the average balance\_orig\_time value in the terminal node. Columns fico\_orig\_time, ltv\_orig\_time and investor\_orig\_time are considered as the important predictors for estimating the balance\_orig\_time of the customer.

* + 1. **Measuring Performance**

The RMSE value of the purchaser’s validation data is 188239.1.

* + 1. **Improving Model performance**

The regression tree model can be improved by adjusting the CP values and selecting the one that got less RMSE value. After trying all different CPs like 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10. The Best Cp value is 0.02.



**Figure 7.4 Decision tree for CP value 0.02**

Columns fico\_orig\_time, ltv\_orig\_time and investor\_orig\_time are considered as the important predictors for estimating the balance\_orig\_time of the customer.

**Performance measure**

The RMSE value for the validation purchaser’s data is 188239.1

1. **Regression Model Selection**

**Table 8.1 Performance Comparison**

|  |  |
| --- | --- |
| Model | RMSE on Validation Data |
| Multiple Linear Regression | 180139.7 |
| Backward Linear Regression | 180177.4 |
| Decision Tree (Regression Tree) | 188239.1 |
| Decision Tree (Regression Tree cp 0.02) | 188239.1 |

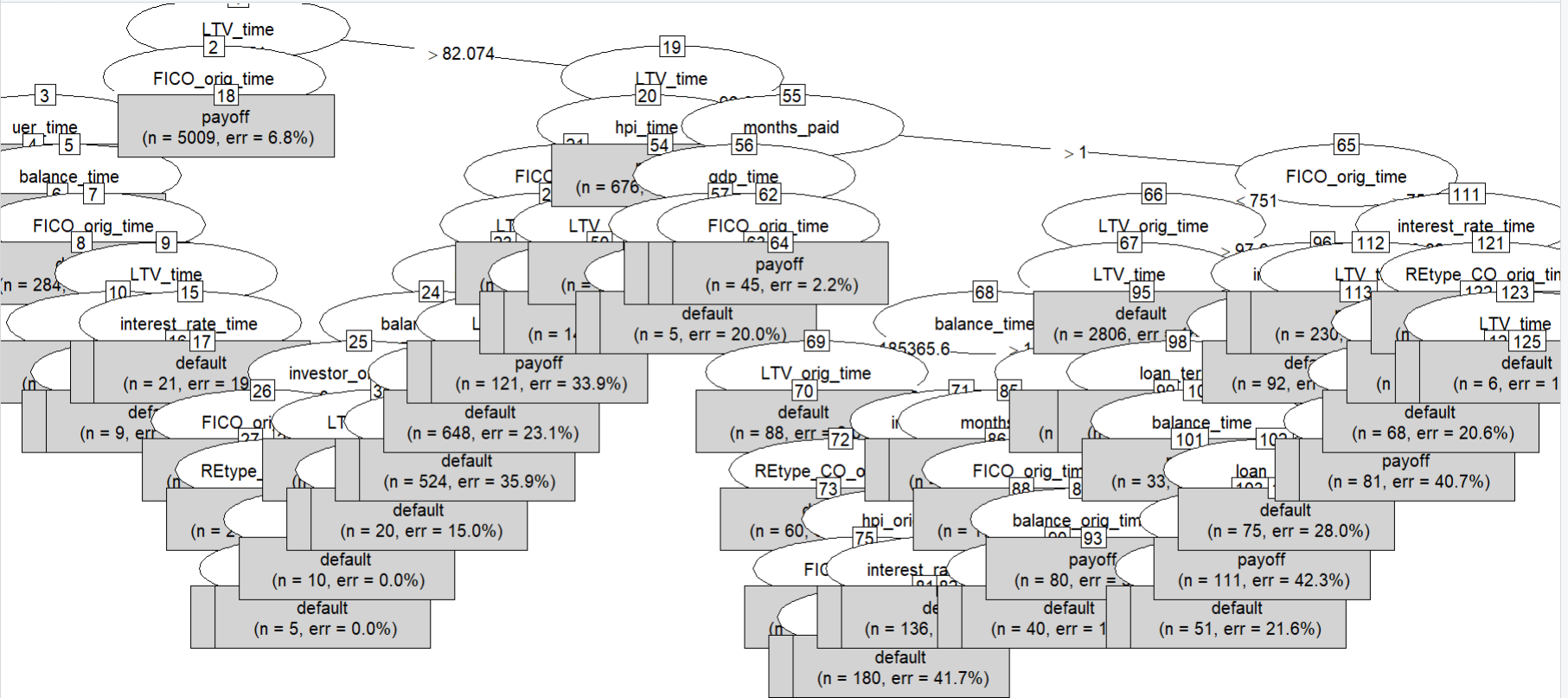
From Table 8.1, it is evident that both the multiple linear regression and the backward selection model are optimal choices for predicting the balance\_orig\_time of a customer. This is because both models have the same RMSE value, which is lower when compared to other models. Among these, the backward linear regression model is considered the best model because it uses only the most important predictors while predicting.

bank can use the backward selection model to estimate the appropriate loan amount that bank can offer to the customer who are classified as payoff which reduces the loan becoming default.

1. **Classification Modelling for classifying on-going customers**

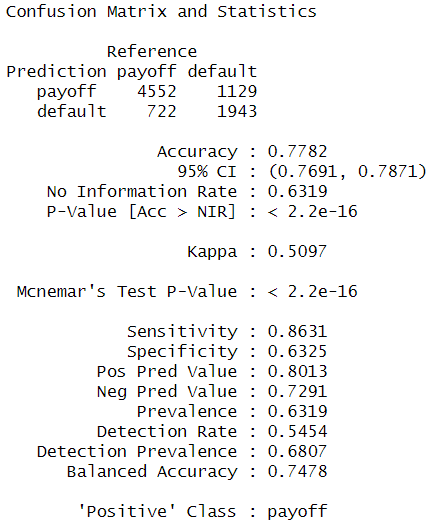
From now the classification models are build to classify the ongoing customers as payoff or default.By using these models bank can identify risk early in ongoing customers and can take action required to improve the financial stability. Status\_time is used as the target column and columns balance\_time, LTV\_time, interest\_rate\_time, hpi\_time, gdp\_time, uer\_time, first\_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, loan\_term, and months\_paid are used as predictors. The performance of the models is measured using a confusion matrix. The payoff class category is considered the important class because it is crucial to accurately classify payoff customers rather than default customers. If a payoff customer is misclassified as default and the bank takes appropriate measures, it could lead to the loss for the bank and customer.

* 1. **Decision Tree**

****

**Figure 9.1 Tree structure**

* + 1. **Measuring Performance**

****

**Figure 9.2 Confusion matrix for decision tree**

The sensitivity rate is 86.31 which indicates that 86.31% of payoff class members are correctly classified.

* + 1. **Improving Model Performance**

The performance of the decision tree model can be improved by using an extra parameter trail while building the model.

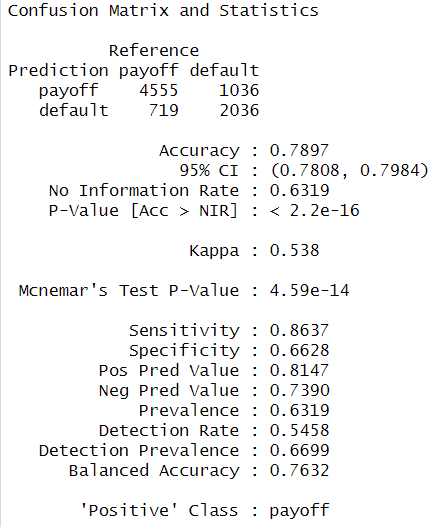
**Trails option 6**

After trying various trail options such as 6, 10, 15, and 20 it was found that trail option 6 has the highest sensitivity rate. Therefore, trail option 6 is considered as the best model among all other trail models.

****

**Figure 9.3 tree structure for decision tree with trail 6**

**Measuring Performance**

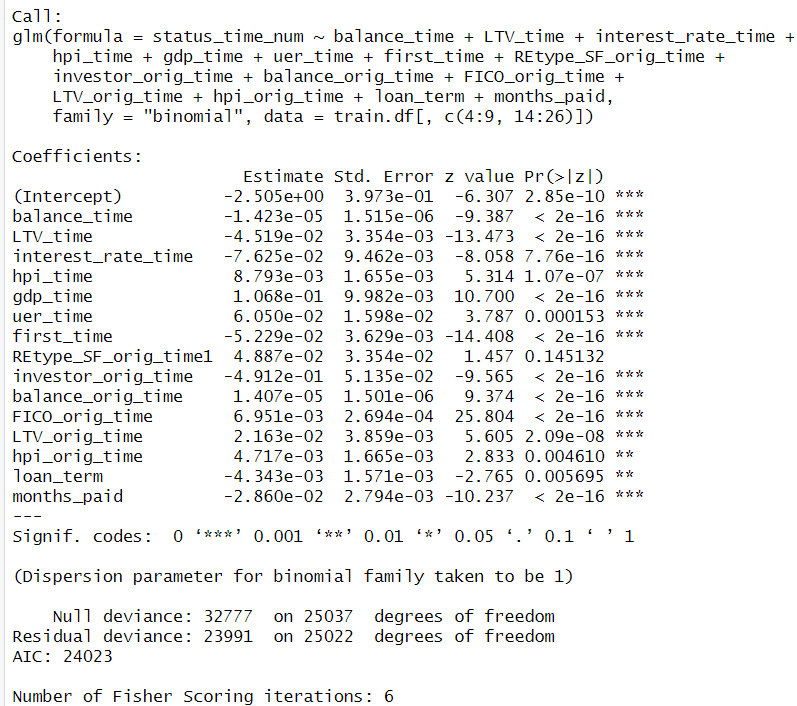
****

**Figure 9.4 confusion matrix for decision tree with trail 6**

The sensitivity rate is 86.37 which indicates that 86.37% payoff customer in valid data is classified correctly.

* 1. **Backward selection Logistic Regression Model**

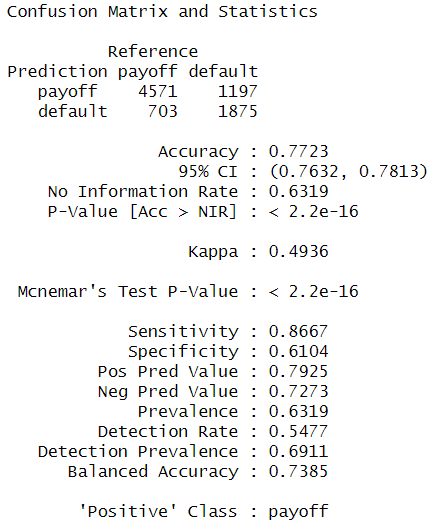
The Backward selection model works well if the target column has 0s and 1s, so before applying logistic regression, the target column status\_time is converted to 1s and 0s by assigning payoff category to 1 and default to 0 in the train data.

****

**Figure 9.5 Summary of backward selection model**

From the figure 9.5 it is clear that columns balance\_time ,LTV\_time interest \_rate\_ time, hpi\_time ,gdp\_time ,uer\_time ,orig\_time ,first\_time.investor\_orig\_time ,balance\_ orig\_time ,FICO\_orig\_time ,LTV\_orig\_time ,months\_paid.

* + 1. **Measuring Model Performance**

****

**Figure 9.6 Confusion matrix of backward selection model**

From the figure 9.6 the sensitivity rate is 86.67 which indicates that 86.67% of payoff members are correctly classified.

1. **Ongoing Customers Classification Model Selection**

The model which is giving high sensitivity is the best model. So, let’s compare the all models sensitivities.

**Table 10.1 Performance Comparison**

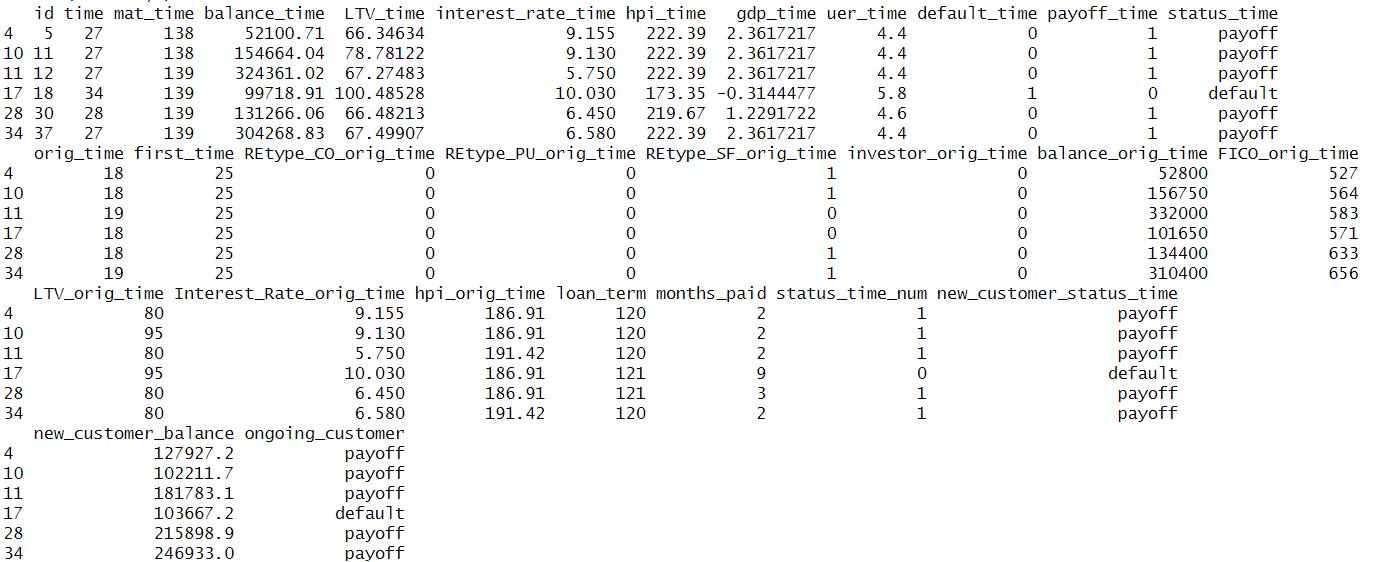
|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Sensitivity |
| Decision Tree | 77.82% | 86.31% |
| Decision Tree with trail 6 | 78.97% | 86.37% |
| Backward selection model | 77.23% | 86.67% |

From the table 10.1 it is clear that backward selection model is considered as the best classifier for classifying ongoing customer as default or payoff.

bank can use the backward selection logistic regression model to classify if a ongoing customer is going to payoff the loan or default it. With this information bank can take required action on the ongoing customer who are classified as default to increase their financial stability.

1. **Predicting New Data**

In this section, all selected models are applied to the test data. Let's consider that the test data contains information on new customers who have applied for a mortgage loan. Bank first applies the best classifier model, the backward selection, to determine whether a new customer will payoff the loan or default. If the new customer is classified as payoff, then the best regression model, the backward linear regression model, is used to estimate the appropriate loan amount. Now, consider that the test data also includes information on ongoing customers. Suppose the bank wants to determine whether an ongoing customer will default or pay off their loan; then the best-selected classification model, the backward selection logistic regression, is used.

****

**Figure 11.1 Sample records in test data with all predicted values**

1. **Conclusion**

Bank is currently experiencing low financial stability for two primary reasons: first, a significant number of loans are defaulting because the bank has been approving loans for customers who are unable to repay them. The approved loan amounts are not sustainable for these customers. Second, the bank is unable to estimate which actively paying customers are going to default. To overcome these challenges and improve financial stability, Bank has decided to utilize predictive analytics on historical data containing the loan progression data of 50,000 customers. Among all the models applied, the backward selection logistic regression model has been chosen as the best model for classifying whether a new customer will pay off the loan or not. The backward selection regression model is considered the best for estimating the loan amount. Finally, the backward selection logistic regression model has been selected as the best model to classify whether an ongoing customer will pay off the loan or not.