Hema Lakshmi Prasanna Sanaka

Webster University

CSDA 6010 DATA ANALYTICS PRACTICUM

SPRING 2024

FINAL PROJECT (Case-1)

JP WANG

NORTH-POINT SOFTWARE PRODUCTION COMPANY

TABLE OF CONTENTS

INTRODUCTION ……………………………………………………….6

BUSINESS GOAL……………………………………………………….6

ANALYTICS APPROACH………………………………………………7

1.DATA PREPROCESSING …………………………………………….7

* 1. ATTRIBUTE DEFINITION………………………………………….7
  2. DATA EXPLORATION………………………………………………8

1.2.1 CHECK FOR NULL VALUES……………………………………….9

1.2.2 DESCRIPTIVE STATISTICS OF DATA…………………………….9

1.2.3 CHECKING FOR OUTLIERS……………………………………….10

1.2.4 CORRELATION FOR NUMERICAL COLUMNS………………….14

2. DATA ANALYSIS………………………………………………….…14

3. DIMENSIONALITY REDUCTION……………………………….….24

4. DATA PARTITIONING METHODS …………………………….….25

4.1 TRAINING DATA ……………………………………………….….25

4.2 VALIDATION DATA …………………………………………….….25

4.3 HOLDOUT DATA ………………………………………………….26

5. CLASSIFIER MODEL SELECTION ……………………………….26

5.1 LOGISTIC REGRESSION MODEL.…………………………………26

5.1.1 FORWARD STEPWISE LOGISTIC REGRESSION ………………30

5.1.2 BACKWARD STEPWISE LOGISTIC REGRESSION ………………………………………………………………………………31

5.2 DECISION TREE MODEL………………………………………32

5.3 DECISION TREE BY USING THE C5.0 ……………………………….36

5.4 COMPARISION OF CLASSIFIER MODELS………………………….42

6. REGRESSION MODELS ………………………………………………….43

6.1.1 MULTILINEAR REGRESSION MODEL ……………………………43

6.1.2 FORWARD STEPWISE REGRESSION MODEL …………………46

6.2 REGRESSION TREE MODEL ………………………………………….48

6.3 COMPARISION OF REGRESSION MODELS…………………………51

7. REPORT ON MODEL PERFORMANCE ……………………………….52

8. HOLDOUT MODEL ACCURACY …………………………………….54

8.1 CUMULATIVE GAIN CHART FOR EXPECTED SPENDING ……….57

9. CONCLUSION …………………………………………………………….59

**Executive Summary:**

* Northpoint Software, specializing in games and educational software, faced challenges in expanding its customer base through mailing campaigns. To address a significant 5.3% response rate from a test sample, so the company decided to building predictive models.
* In this need to build the 2 predictive models one is for classification model to classify the potential customers and other one is for regression model to estimate the spending value.
* Choose forward step wise logistic regression is a best classifier model among all other models because it will provide the accurate results of probabilities which will help to figure out the potential purchasers based on the probability. The model with accuracy 78.71% and the sensitivity is 74.39%.
* Choose forward step wise linear regression is a best regression model with lowest RMSE value 165 among all other models because it will provide best accurate results and able to predict the spending value.
* Evaluation metrics such as predicted probabilities, expected spending, adjusted probability, and estimated spending values were incorporated. Cumulative gain and decile charts demonstrated that targeting the top 10% of customers yields 4.17 times more profit than sending mails to all customers.
* Focus on the top 10% of customers identified by the models for mailing campaigns, optimizing returns on investment.

**INTRODUCTION:**

North-Point is a software firm which sells the games and educational software. Later it’s started the third-party titles to its offerings. In the part of expanding their customer base they put their redesigning items with a new collection that will be informed to public through direct mail. Here customer list is an asset to the north-point. In the strategy of expanding their customer base they joined in one pool drive that is specialized in software and hardware products. In this pool drive North-Point get a good opportunity that is, members can use a shared list of customers to pick names for their mailings. They share their own customer lists with the group and can take back a similar number of names every quarter. This collaboration helps them reach more people and make their marketing efforts more effective.

**BUSINESS GOAL:**

The main goal is to enhance the customer base with their strategic move by mailing directly to the customer about their new offerings. But each mail costs $2 if they sent mail to each 200,000 customers it will charge more so, they did a test run on first 20,000 customers they got only 5.3% response rate which will give more lose to the business so they thought to build the predictive model to predict the purchasers who ever interested to purchase, and they will send mails to those customers and make more profit instead of losing. Let’s calculate the actual gross profit for the 20,000 customers which is the amount of total spending - investment of each customer is 205250 – 40,000 is $1,65,250. Now, let’s calculate the estimated gross profit for remaining 1,80,000 customers if company wants to mail which is estimated spending - investment of each customer is 18,47,250 – 3,60,000 is 14,87,250. The estimated gross profit is $14,87,250 for remaining 1,80,000 customers. Now, let’s build the predictive model to predict the potential customers and amount of spending by the customer. The first goal of this is to build a predictive model based upon the response of the customer to the testing mails and purchasing something. From that data we can classify weather the customer will purchase or not. The second goal is to build a predictive model to predict the amount of spending by the customer for purchases.

**Analytical Approach:**

The analytical approach towards the enhance the customer base by building the predictive models to overcome the impact of huge lose. We will have two approaches those are classifying the potential purchasers by using the classification model and predicting the amount of spending value by each customer then we use these models to predict the probabilities and estimate the expected spending values. Then we will make decision by using the ranking performance to select the potential customers based upon their spending and we can make the profit on investment and can enhance the customer targeting

**1.DATA PRE-PROCESSING:**

**1.1 ATTRIBUTE DEFINITIONS:**

**Sequence Number**: It indicates the unique number of the records present in the dataset.

**US**: It indicates weather the customer present in US or not. This column is in binary format, 1 represents the customer present in US and 0 represent the customer does not present in US.

**Source\_a, source\_c, source\_b, source\_d, source\_e, source\_m, source\_o, source\_h, source\_r, source\_s, source\_t, source\_u, source\_p, source\_x, source\_w:** These columns indicate the different channels obtained by the customers. These columns also in binary format, 1 represents the obtained channel by the customer and 0 represents the does not obtained by the customer.

**Freq:** This column indicates the number of purchases made by the customer. This is in numeric format and represents the count of purchases by the customer.

**Last\_updated\_days\_ago:** This column represents the number of days since the last updated.

**1st\_updated\_days\_ago:** This column represents thenumber of days since the first updated.

**WEB ORDER:** This column represents whether the purchase was made via web order or not. This column also in the binary format, 1 represent the purchase was made via web order and 0 represents the purchase was made by other type of orders.

**Gender=Male:** This column represents gender of the purchaser; it is also in the binary format 1 represents the male and 0 represents the female and others.

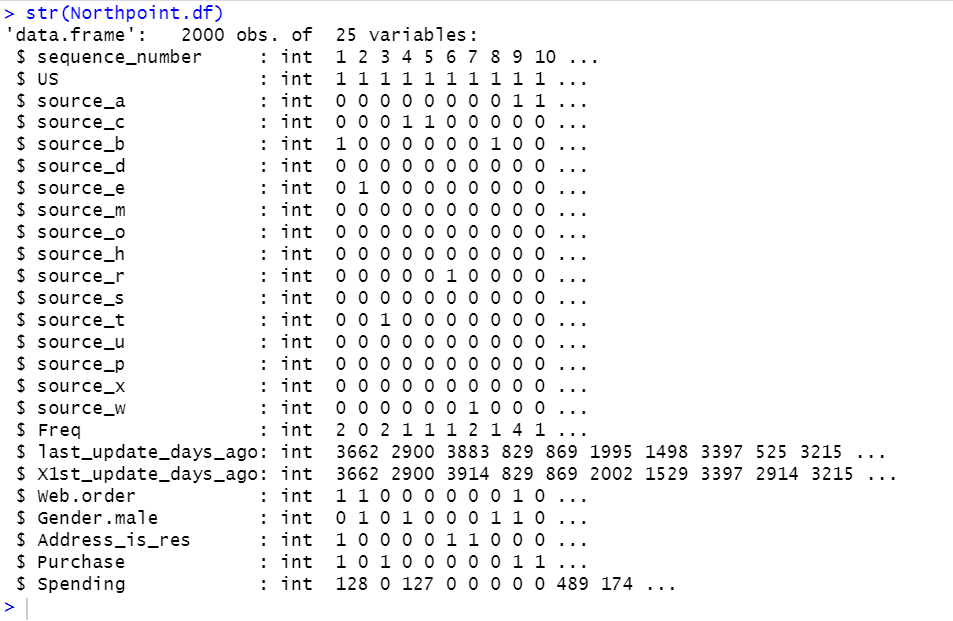
**Address\_is\_res:** This column represents the whether the purchaser address is a residential or not. This column is also in the binary format 1 represents the address is residential and 0 represents the address is not residential.

**Purchase:** This column represents the whether the purchaser respond to the email or not and made a purchase. It is also in the binary format 1 represents the made purchase and 0 represents the not purchased.

**Spending:** This column represents individuals who made a purchase, along with the corresponding amount they spent.

**1.2 DATA EXPLORATION:**

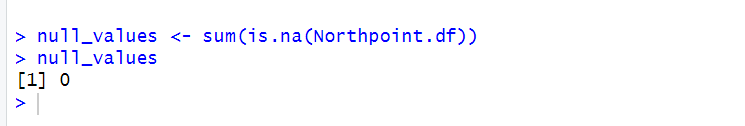
In this dataset have total 2000 records and 25 columns. In those 25 columns one is unique columns that is sequence number, 20 columns are in binary format which is represented as 0’s & 1’s and the remaining 4 columns are in numeric. The structure of the dataset is



**Figure 1.2 Structure of the data**

**1.2.1 CHECK FOR NULL VALUES:**

Here checked for the null values

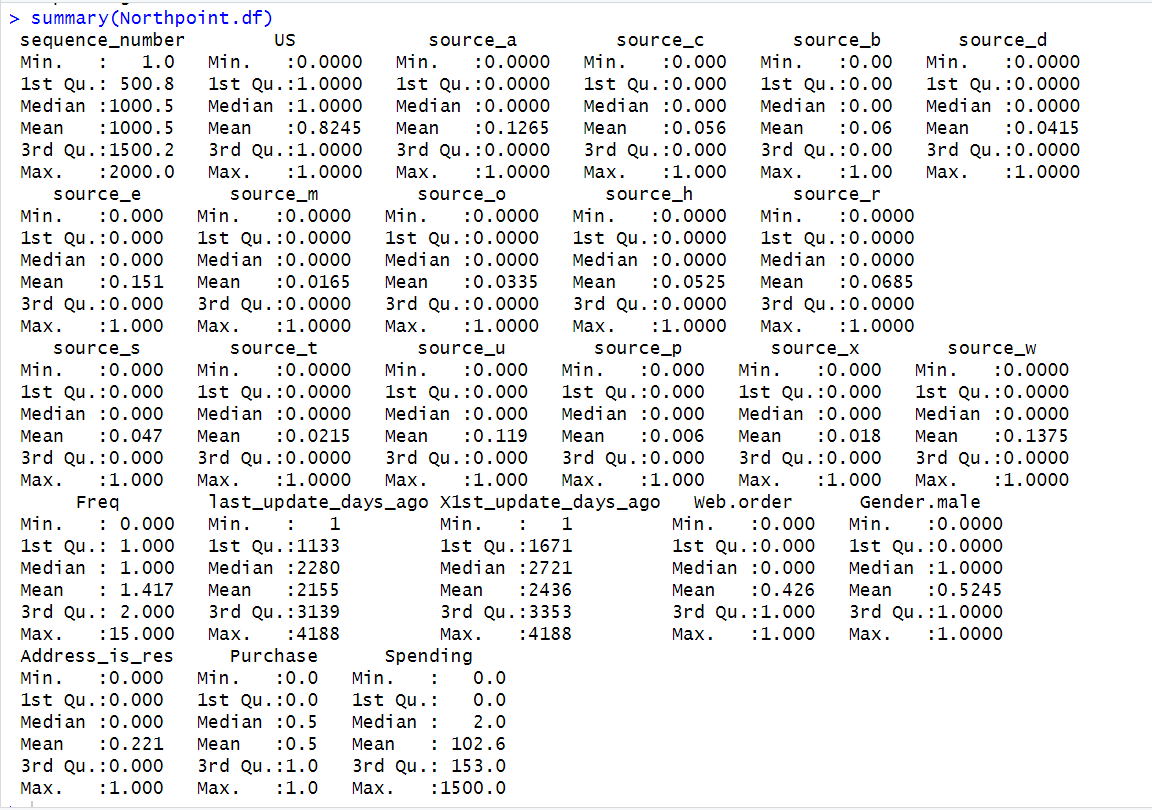


**Figure 1.3 Null Values**

There are no null values present in the data so, there is no further action required on this.

**1.2.2 DESCRIPTIVE STATISTICS**

Now let’s have a look on the descriptive statistics of each attribute in the dataset, from the below screenshot observed that maximum and minimum values of each attribute.

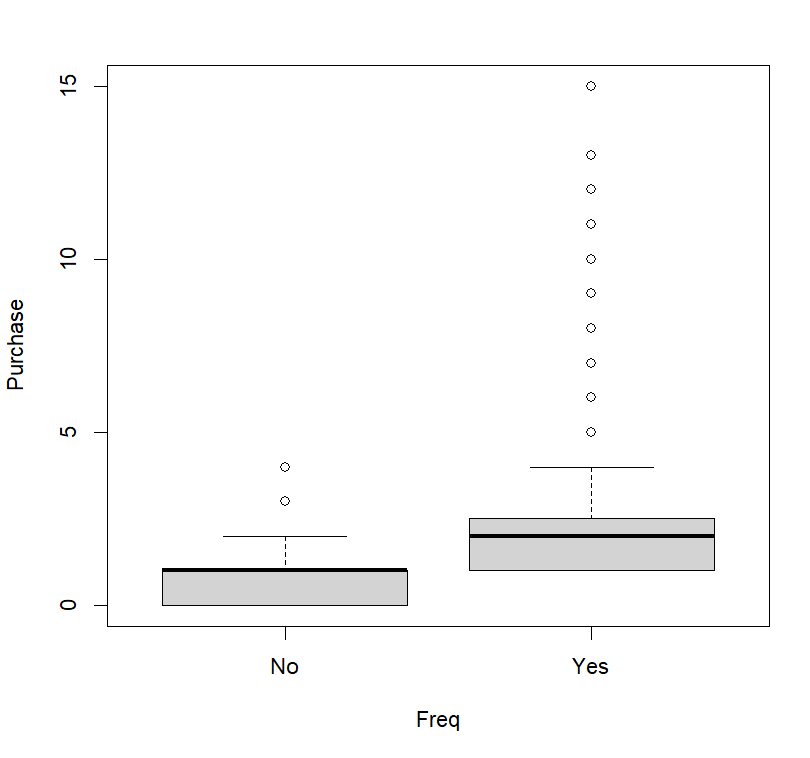


**Figure 1.4 Descriptive Statistics of the data**

By looking at the above figure 1.4 can be seen that descriptive statistics of numerical columns can check for the mean, median, 1st and 3rd quartiles. The min&max value of sequence number is 0 &2000, min&max value of US is 0&1, from source\_a to source\_w min&max value is 0&1, min&max value of Freq is 0&15, min&max of last\_updated\_days\_ago is 1&4188, min&max value of X1st\_updated\_days\_ago is 1&4188, min&max value of web.order is 0&1, min&max value of Gender.male is 0&1, min&max value of Address\_is\_res is 0&1, min&max value of purchase is 0&1, min&max value of spending is 0&1500, mean of the freq is 1.417, mean of last\_updated\_days\_ago is 2155, mean of X1st\_updated\_days\_ago is 2436.

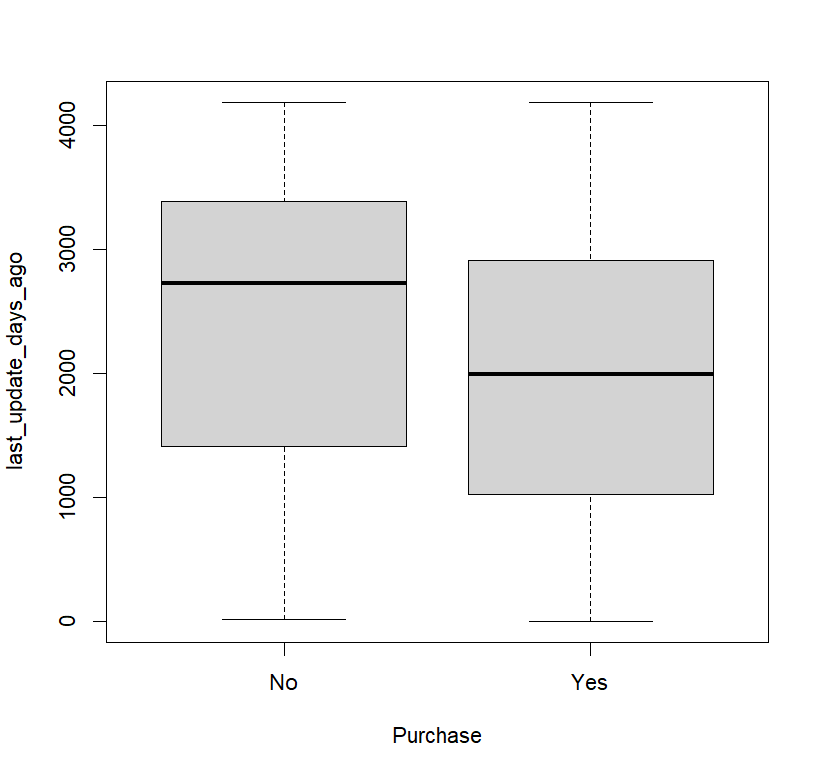
**1.2.3 CHECKING FOR OUTLIERS:**

To check the outliers used box plot for the numerical variables that is frequency, last\_updated\_days\_ago, 1st\_updated\_days\_ago with purchase column and spending is the outcome variable.



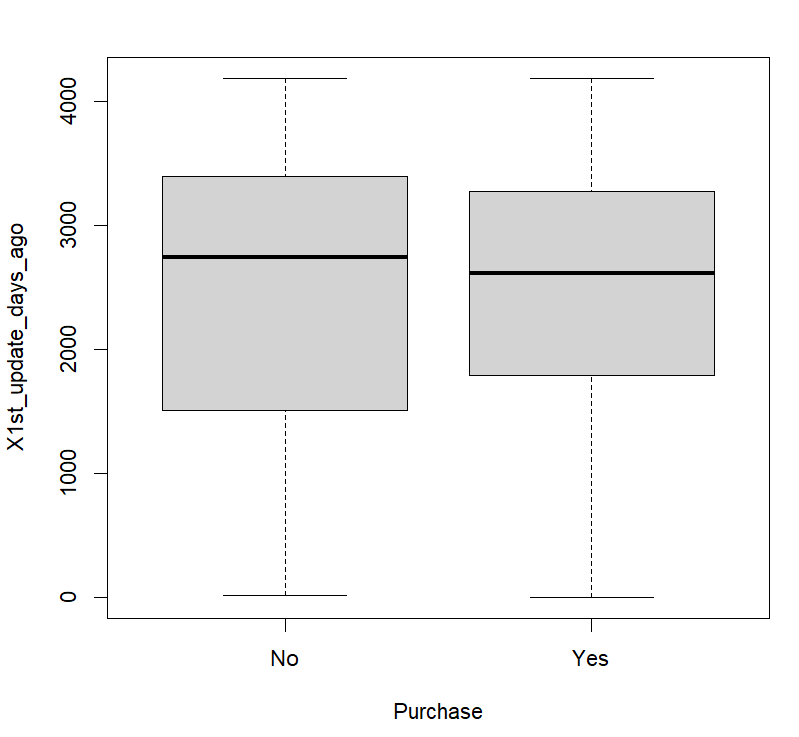
**Figure 1.5 Box plot for Freq vs Purchase**

By looking at the above figure 1.5 can be seen that in frequency those are not outliers because those are actual count of purchase made by the individual customers so, we are not considered as outliers.



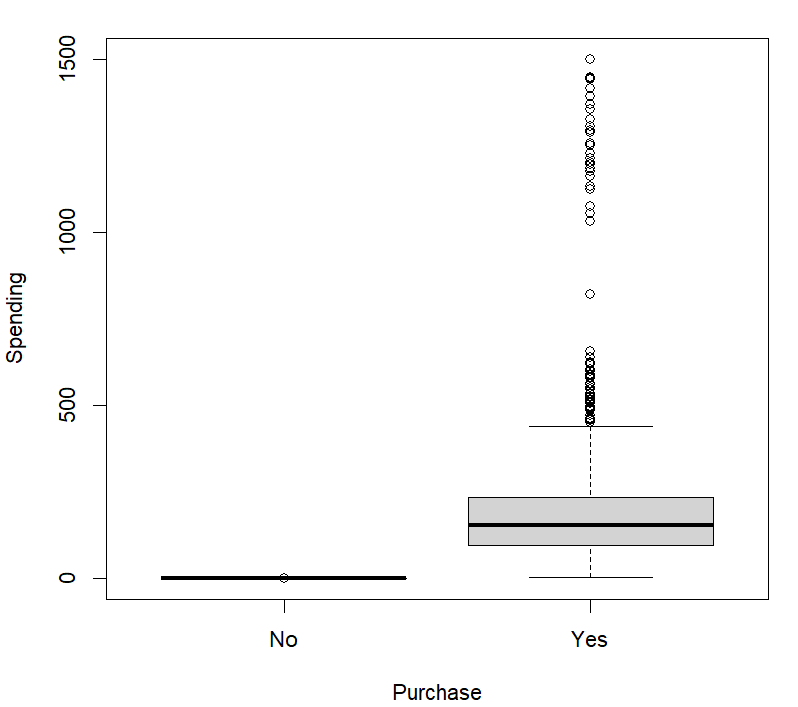
**Figure 1.6 Boxplot for Purchase vs last\_updated\_days-ago**

By looking at figure 1.6 can be seen that there are no outliers in last\_updated\_days\_ago and the first quartile of the last\_updated\_days\_ago for no is around 1300 days ago and the third quartile of the last\_updatded\_days\_ago for no is around 3300 if we have data points below or above the quartiles then only, we consider as outliers. The first quartile of the last\_updated\_days\_ago for yes is around 1133 and the third quartile of the last\_updated-days-ago for yes is around 2900.



**Figure 1.7 Box plot for Purchase vs X1st\_updated\_days\_ago**

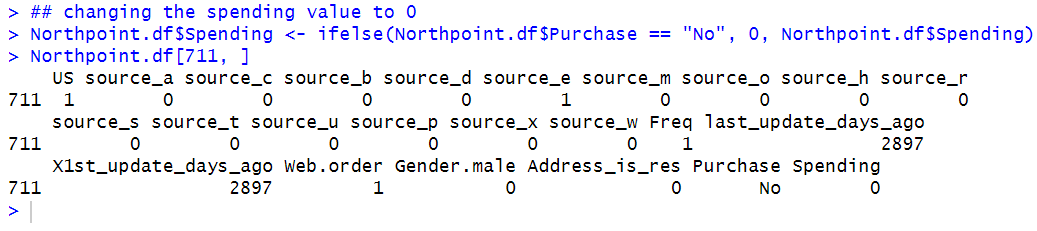
By looking at the above figure 1.7 can be seen that there are no outliers in X1st\_updated\_days\_ago. The first quartile of the X1st\_udated\_days\_ago for no is around 1300 and the third quartile of the X1st\_updated\_day\_ago for no is around 3300 and the first quartile of the X1st\_updated\_days\_ago for yes is around 1800 and third quartile of the X1st\_updated\_days\_ago for yes is around 3200 If we have data points below or above the quartiles then only, we consider as outliers.



**Figure 1.8 Box plot for Purchase vs Spending**

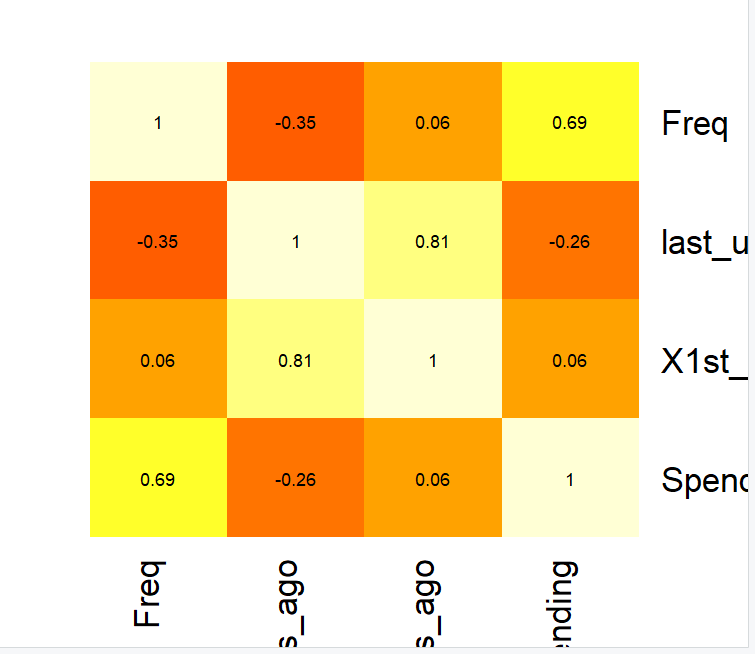
By looking at the above figure 1.8 can be seen that it is apparent that only one outlier is present, indicating no purchase. However, the spending value is available, and it is considered an outlier. It is necessary to adjust the spending value to 0.

Let’s have a look on the adjusted value of spending to 0.



**Figure 1.9 adjusting the spending value to “0”**

**1.2.4 Correlation For numerical Columns:**

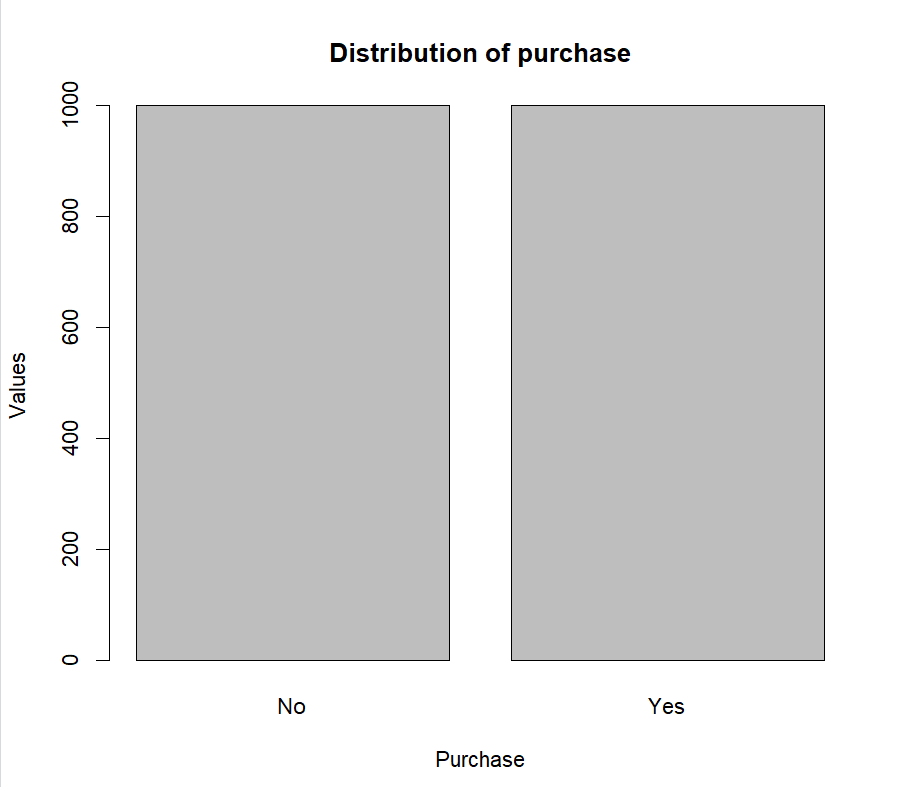


**Figure 1.10 Correlation heat for numerical columns**

Figure 1.10 reveals that last\_updated\_days\_ago and 1st\_updated\_days\_ago are highly positively correlated, with a correlation coefficient of 81%. This high correlation suggests redundancy between these two variables, indicating that can choose either one for modelling to avoid duplication and multicollinearity. Additionally, the variable last\_updated\_days\_ago shows a low negative correlation with Freq and spending, suggesting a slight inverse relationship between these variables. The other variables in the dataset appear to have little to no significant correlation, indicating that they are relatively independent of each other.

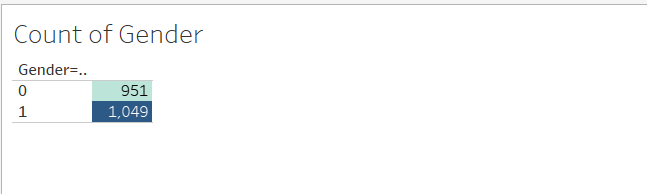
**2. DATA ANALYSIS**

First, checking for the outcome variable whether it is not uniformly distributed or not. let’s have a look on the graph.



**Figure 2.1 Distribution plot of Purchase variable**

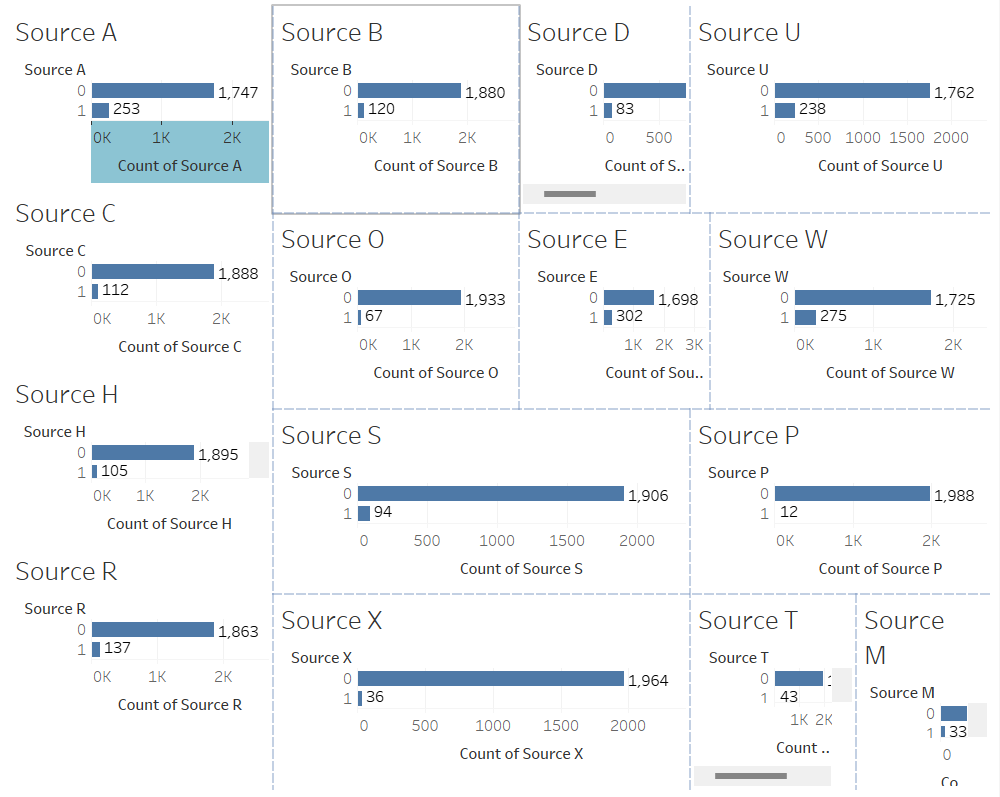
By looking at above figure 2.1 can be able to say that distribution plot of purchase variable is distributed uniformly.



**Figure 2.2 Count of Gender**

By looking at the above figure 2.2 can say that the count of male purchasers is more when compared with the count of purchasers of females and others.

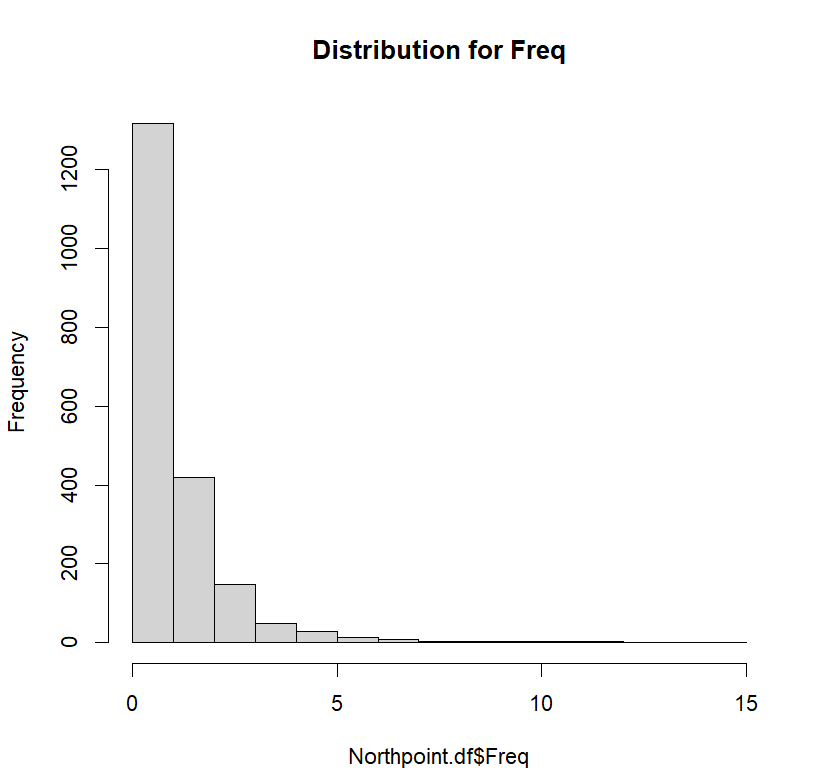
**Count Of All Sources:**

****

**Figure 2.3 Count of all sources**

By looking at the above figure 2.3 can say that the count of purchasers of each source. The count of purchasers of source\_a is 253, source\_b is 120, source\_c is 112, source\_d is 83, source\_e is 302, source\_h is 105, source\_m is 33, source\_o is 67, source\_p is 12, source\_r is 137, source\_s is 94, source\_t is 43, source\_u is 238, source\_w is 275, source\_x is 36. If we observe source\_e has the highest count.

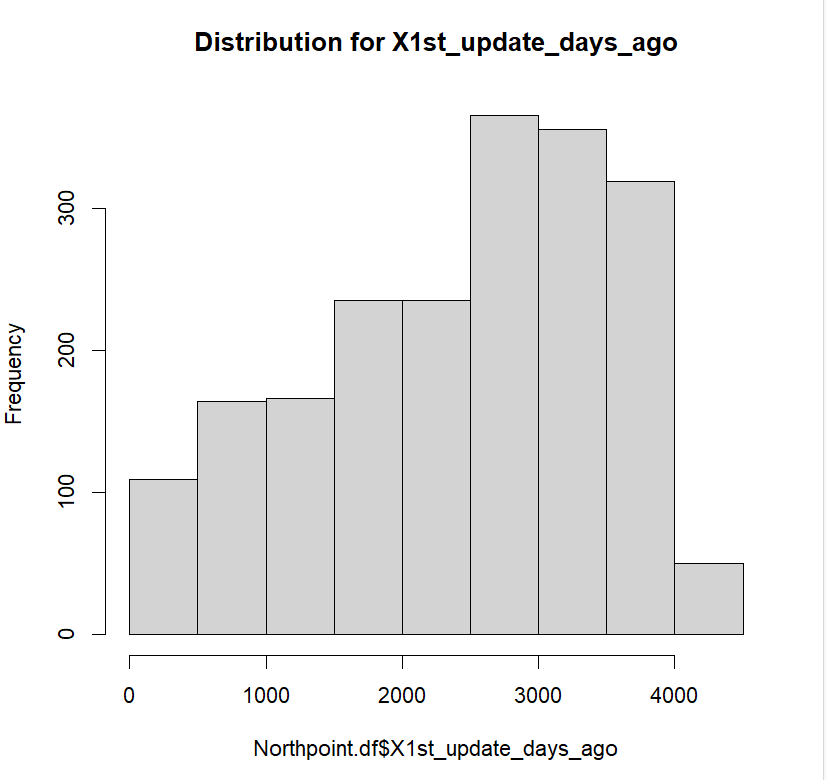
**Distribution of Frequency:**

****

**Figure 2.4 Distribution of Frequency**

By looking at above figure 2.4 can be seen that frequency has the right skewed data. Most of the Purchasers between the 0 to 5 transactions.

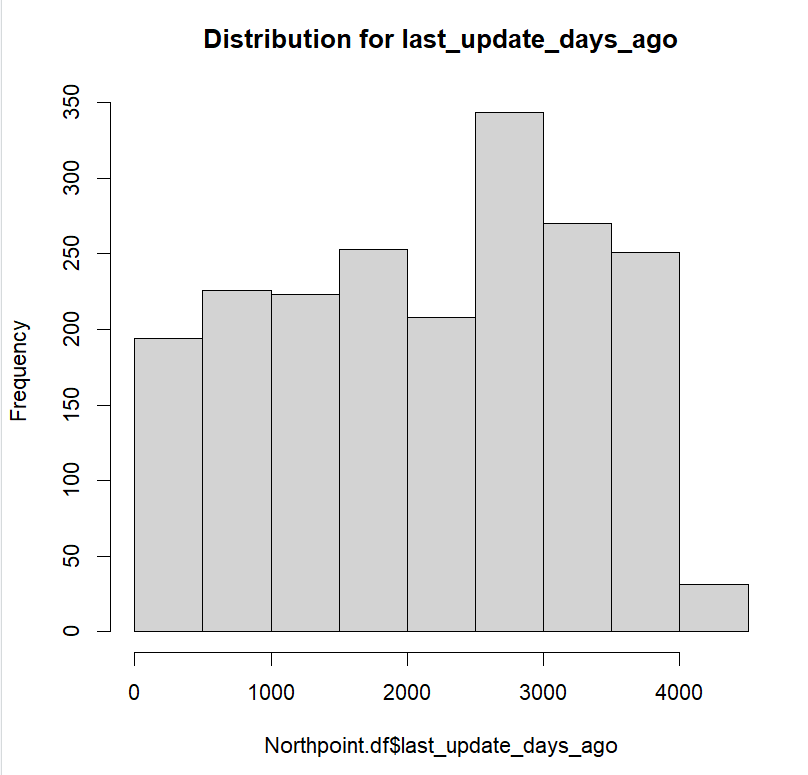
**Distribution Of X1st\_updated\_days\_ago:**

****

**Figure 2.5 Distribution of X1st\_updated\_days\_ago**

By looking at the above figure 2.5 can be seen that Ist\_updated\_days\_ago is skewed to the right. Observe that most of the purchaser’s data updated around 3000 to 4000 days ago. Some purchaser’s data updated around 1000 to 2000 days ago.

**Distribution of last\_updated\_days\_ago:**

****

**Figure 2.6 Distribution of last\_updated\_days\_ago**

By looking at above figure 2.6 can be seen that last updated days ago data distributed uniformly that means all the purchaser’s data were updated uniformly but few more purchasers updated around 3000 days ago.

**Spending vs 1st\_updated\_days\_ago:**

By looking at the below figure 2.7 can observe that most of the customers spending is below or equal to 500 only at the time of first updated and only few customers spending is above 1000 and 1500.

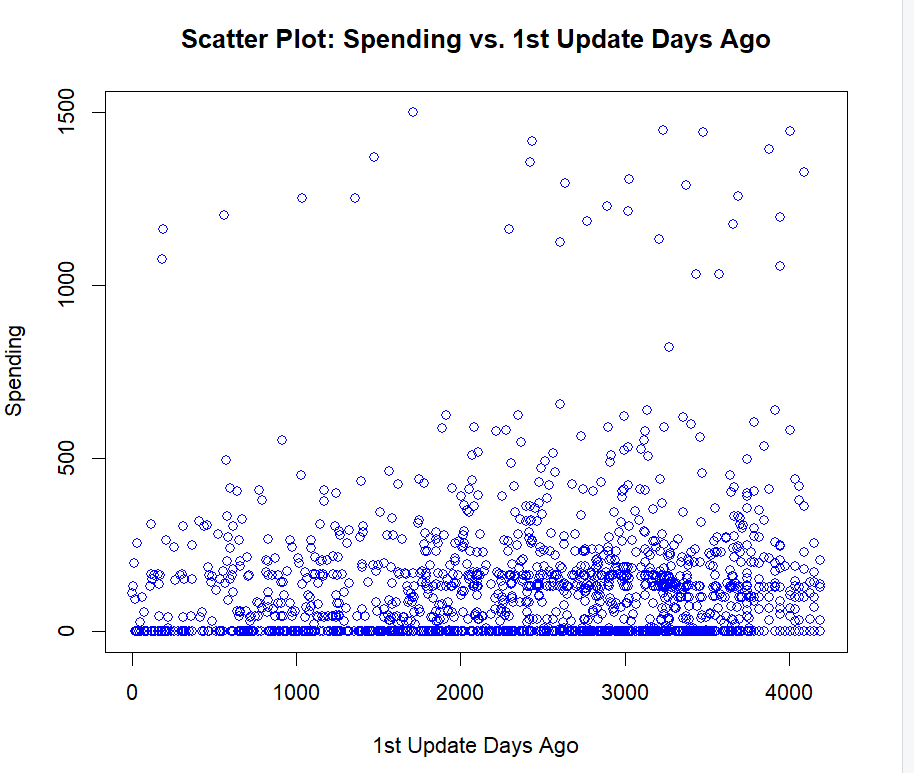


Figure 2.7 Spending Vs 1st updated days ago

**Spending Vs last\_updated\_days\_ago:**

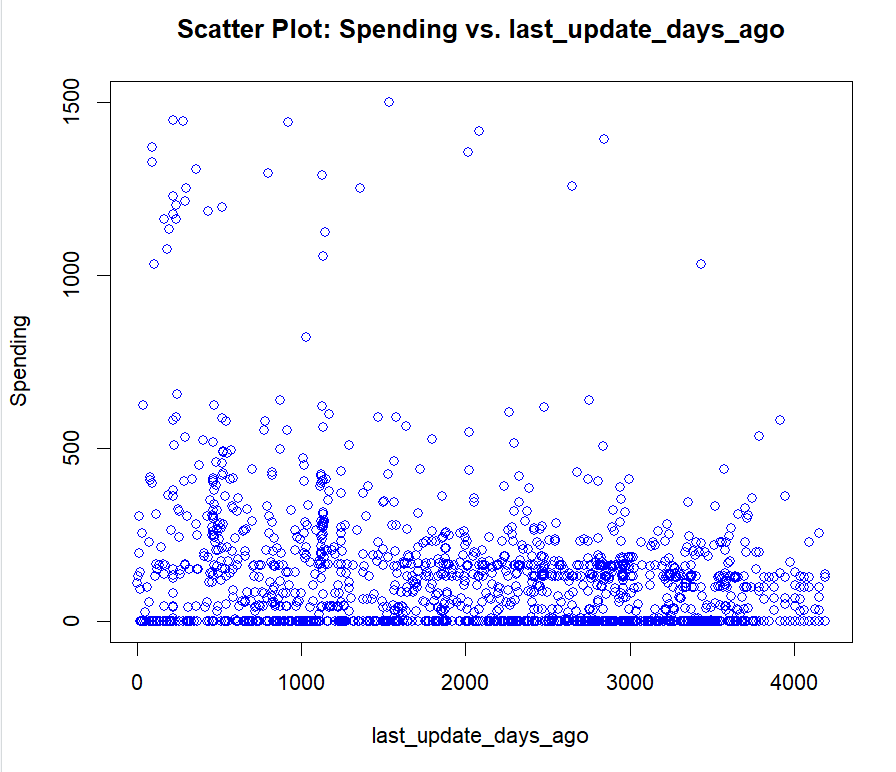


Figure 2.8 Spending Vs last updated days ago

By looking at the above figure 2.8 can say that most of the customers spent below 500 or equal to 500 in last updated days ago and only few customers spent above 1000 to 1500 in last updated 2000 days ago, very few customers spent above 1000 in last updated days ago.

**Spending VS Frequency:**

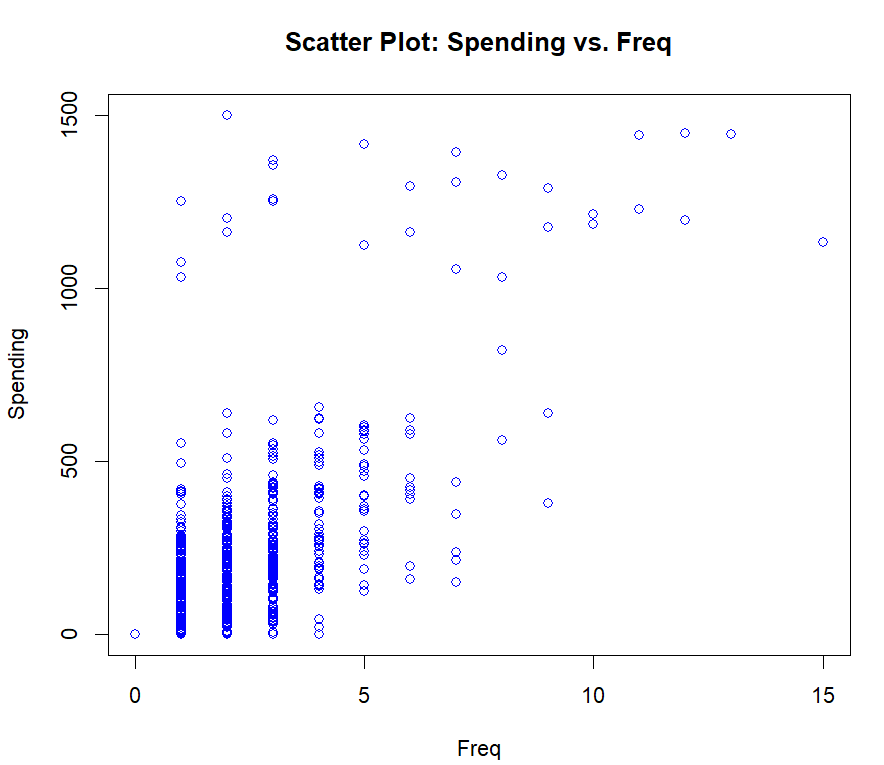
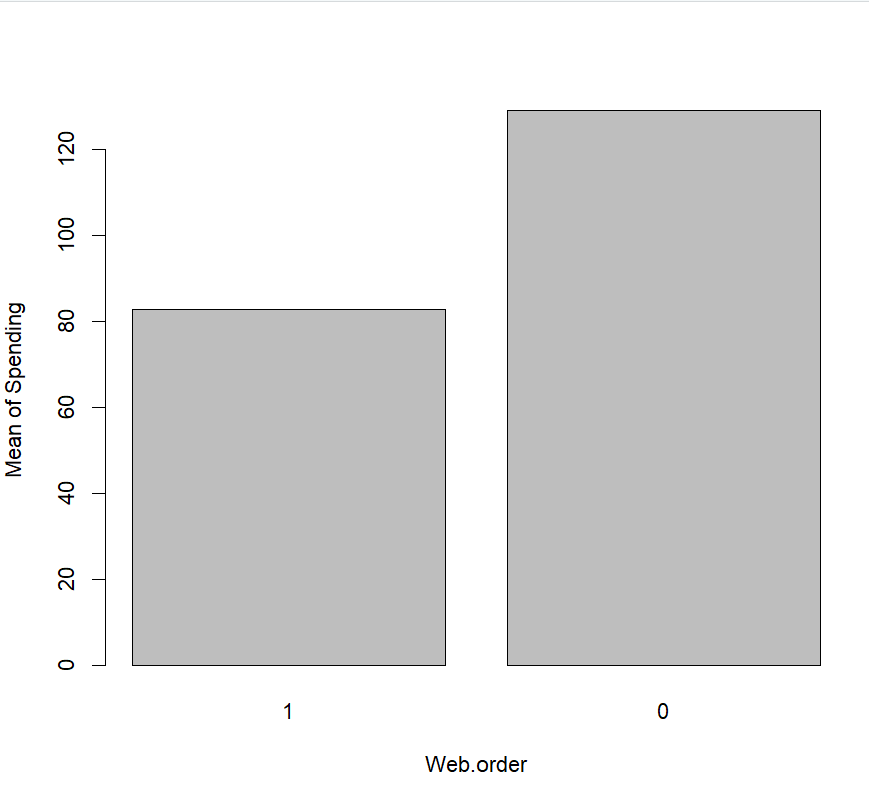
****

Figure 2.9 Spending vs Frequency

By looking at the above figure 2.9 can be seen that most of the customers transactions frequency less than 8 and they spent less than or equal to 500. Only few customers made above 1000 to 1500 spending amount.

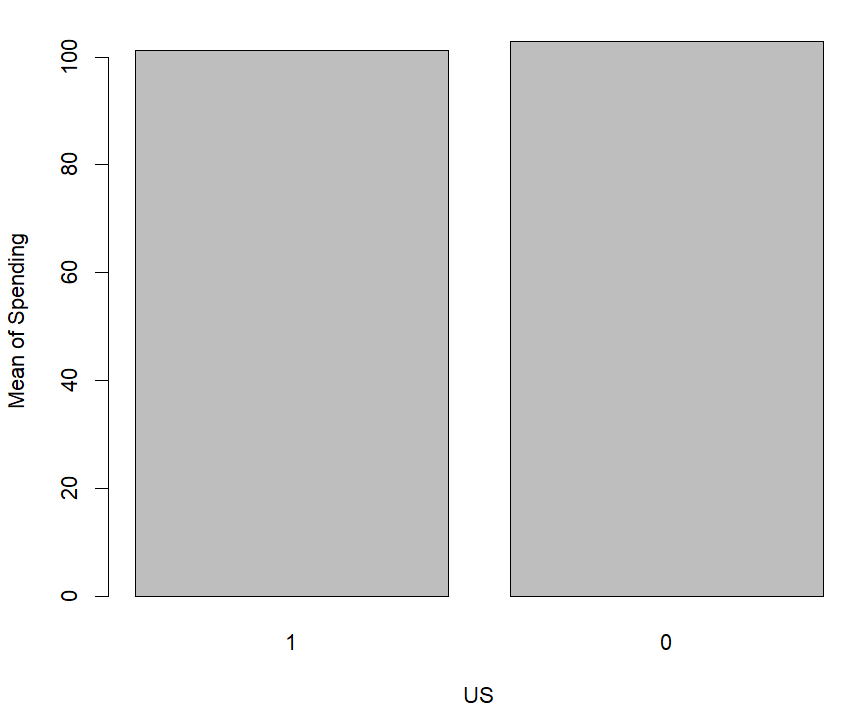
**Web Order vs Spending:**

By looking at below figure 2.10 can be seen that spending amount from web order is less than the spending amount from web orders so, can say that many people are not buying from the web orders.



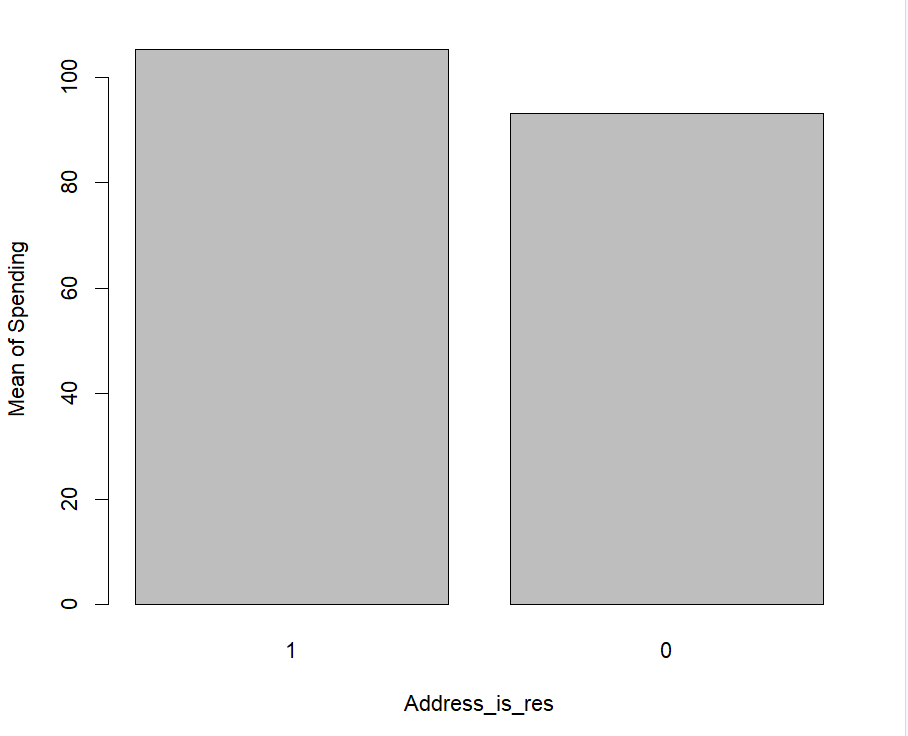
**Figure 2.10 Spending vs Web.order**

**Spending Vs US:**

****

**Figure 2.11 Spending vs US**

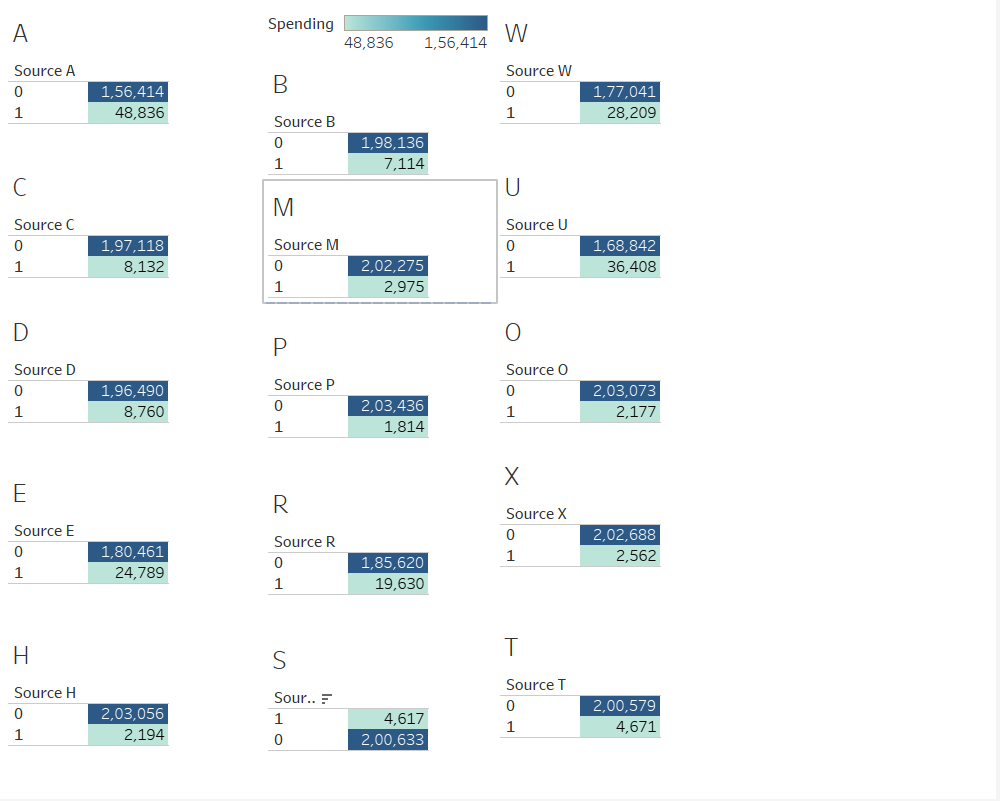
By looking at above figure 2.11 can be seen that mean spending is nearly the same for purchasers located in the US and those not present in the US.

**SpendingVSAddress\_is\_res **

**Figure 2.12 Spending vs Address\_is\_res**

By looking at above figure 2.12 can say that most the purchasers are residential, and they spent more money than the non-residentials. The spending amount of residentials is more than the non-residentials.

**Spending VS All sources:**



**Figure 2.13 Spending amount of each source**

By looking at above figure 2.13 can say that the spending amount of the source\_a is 48,836, source\_b is 7,114, source\_c is 8,132, source\_d is 8,760, source\_e is 24,789, source\_h is 2,194, source\_m is 2,975, source\_p is 1,814, source\_r is 19,630, source\_s is 4,617, source\_w is 28,209, source\_u is 36,408, source\_o is 2,177, source\_x is 2,562, source\_t is 4,671. If we observe source\_a spent high amount on purchase.

**Purchase Vs Spending among Male & Female:**

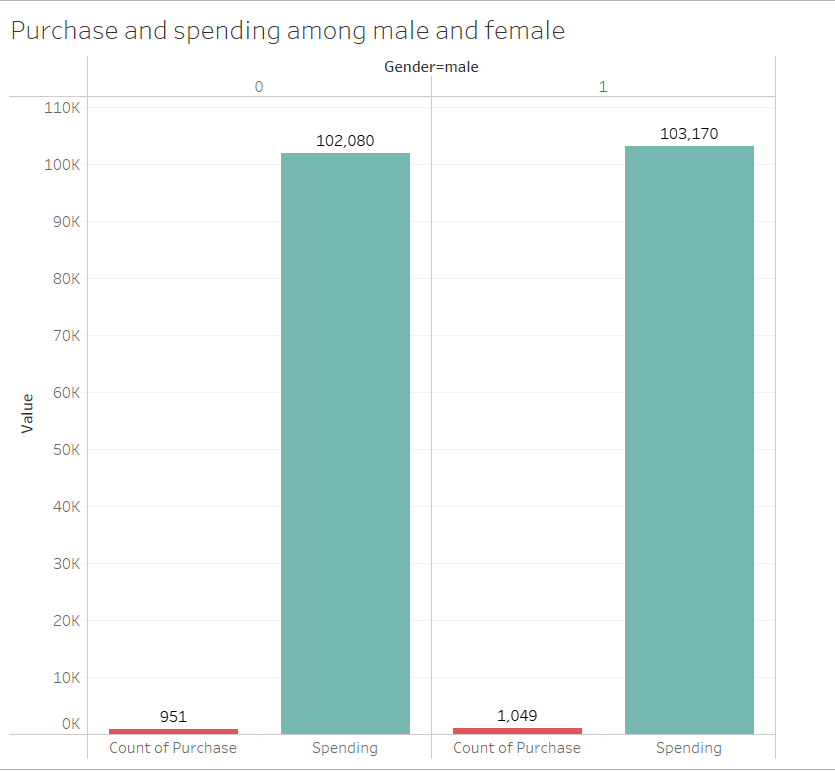
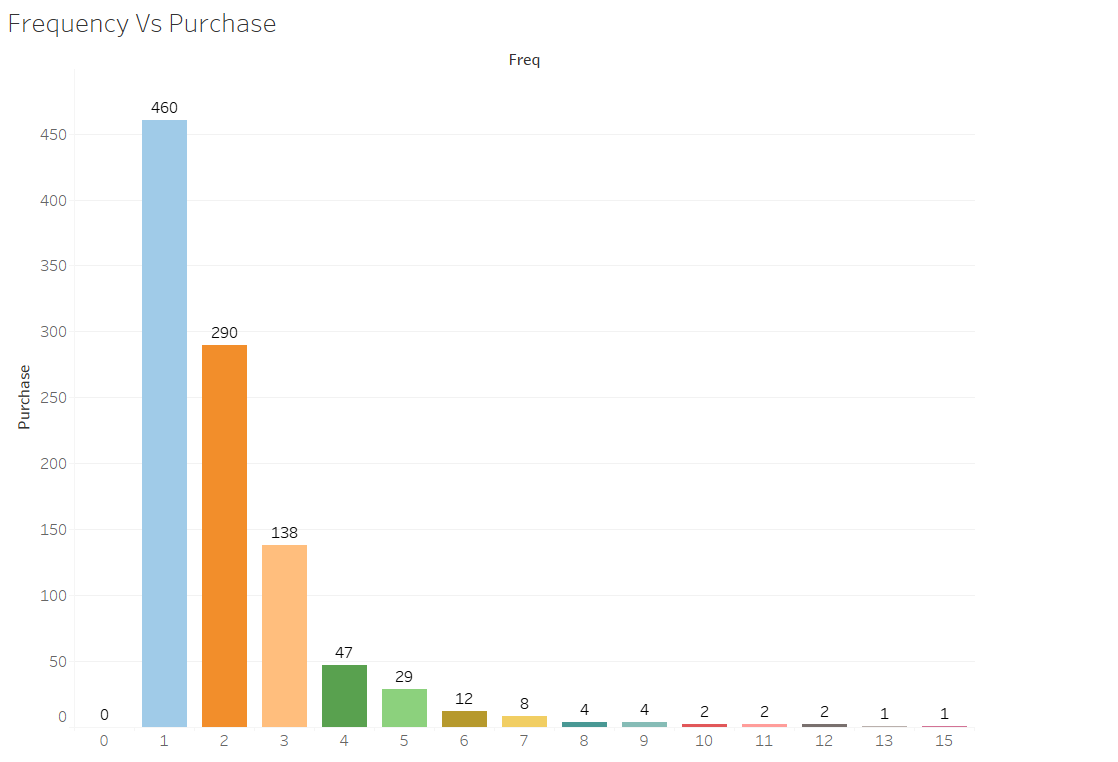


Figure 2.14 Based upon the gender amount of spending and purchase

By looking at the above figure 2.14 can say that the count of the males is greater than females and others then the male’s spending amount is more than the females and others.

**Frequency Vs purchase:**

 Figure 2.15 Frequency Vs Purchase

By looking at above figure 2.15 can be seen that 460 customers made only one time purchase, 290 customers made only two times purchase, 138 customers made three times purchase, only 47 customers made 4 times purchase, only 29 customers made 5 times purchase, only 12 customers made 6 times purchase, only 8 customers made 7 times purchase, 4 customers made 8 times purchase, 4 customers made 9 times purchase, only 2 customers made 10 times purchase, 2 customers made 11times purchase, 2 customers made 12 times purchase, 2 customers made 10 times purchase. If we observe 0

**3.Dimensionality Reduction:**

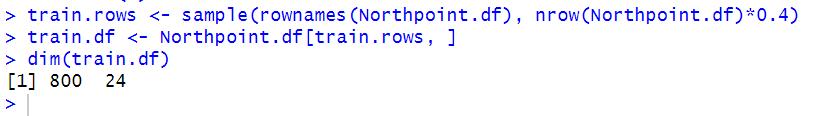
Dimensionality reduction is typically used when there's a large number of predictors, especially when some are highly correlated or redundant. This process helps reduce overfitting and improves model performance. However, in this case, since the dataset has a limited number of predictors, and most of them are relevant, dimensionality reduction might not be necessary. In this only two variables that are highly correlated and share similar variance, suggesting that keeping both would add redundancy without additional value. Can be able to remove one of these variables without compromising accuracy. Beyond this, each variable in dataset has its unique contribution, suggesting that they all play a role in providing accurate results. Interestingly, decision trees have a built-in mechanism for dimensionality reduction. When building a decision tree, the algorithm automatically selects the most important variables based on their contribution to reducing entropy or impurity. Less relevant variables are naturally excluded during tree construction, effectively reducing dimensionality. After building a decision tree, you can identify the important features, which inherently achieves dimensionality reduction. Stepwise algorithm models also used for the dimensionality reduction those are consider only important variables to get the accurate results. This is also a better approach to dimensionality reduction.

**4.DATA PARTITIONING:**

Before partitioning the data, we use the set.seed() function to ensure consistent results from random processes. This guarantees that the same sequence of numbers is generated every time the code is run, allowing reproducibility. Training a model on the entire dataset and then testing its performance on the same data does not provide a clear indication of how well the model will generalize to new or unseen data. To ensure accurate evaluation and improve the model's ability to generalize, we divide the data into three partitions: 40% for training, 35% for validation, and 25% for testing. This partitioning approach allows for effective model training and validation. The training data is used to fit the model, the validation data is used to evaluate and tune(choose) the model, and the test data is used to measure the model's performance on unseen data. This way, we can accurately gauge how well the model performs in a real-world context.

**4.1 Training Data:**

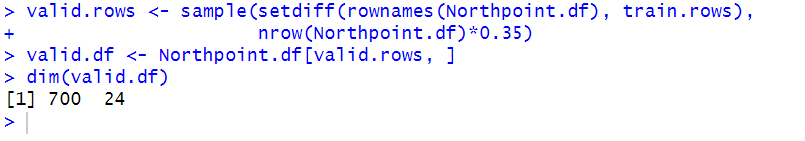
In this training data will going to use 40% of the data which is 800 records. The dimensions of the training data are 800 records and 24 columns. By using this data to train the model and fit the model.



**Figure 4.1. Training Data**

**4.2 Validation Data:**

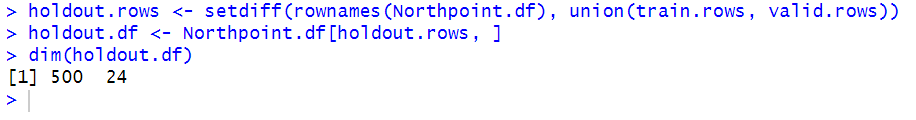
In this validation data will going to use 35% of the data which is 700 records. Dimensions of the validation data are 700 records and 24 columns. By using this data, to assess the predictive performance of each model so that we can compare the models and can tune the model to select the best model.



**Figure 4.2. Validation Data**

**4.3Holdout Data**

In this holdout data will going to use the remaining25% of the data which is 500 records. Dimensions of the holdout data are 500 records and 24 columns. By using this data, to measure the performance of the chosen model with the unseen data and able to see the accurate with that data.



**Figure 4.3 Holdout Data**

**5.Classifier Model Selection:**

Here will choose a most appropriate model for our business problem from the multiple models. In this step we will build the predictive model and train a model to make the new prediction with the unseen data. Here there are few classifier models they are logistic regression, decision tree, decision tree with c5.0 etc… Now will try to evaluate the performance on each model and choose the best model.

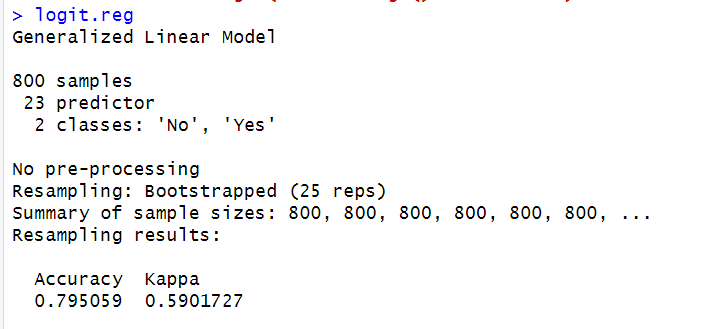
**5.1 Logistic Regression Model:**

In this model will going to use the glm(). To access this function, need to install and import the “Car” and “Caret” packages. Logistic regression extends the idea of the linear regression to the situation where the outcome variable is categorical. In this business task the outcome variable is purchase which is categorical whether the customer will purchaser or not. Here the following equation for logistic regression

logit.reg <- caret::train(Purchase ~ . - Spending, data = train.df,

method = "glm", family = "binomial")

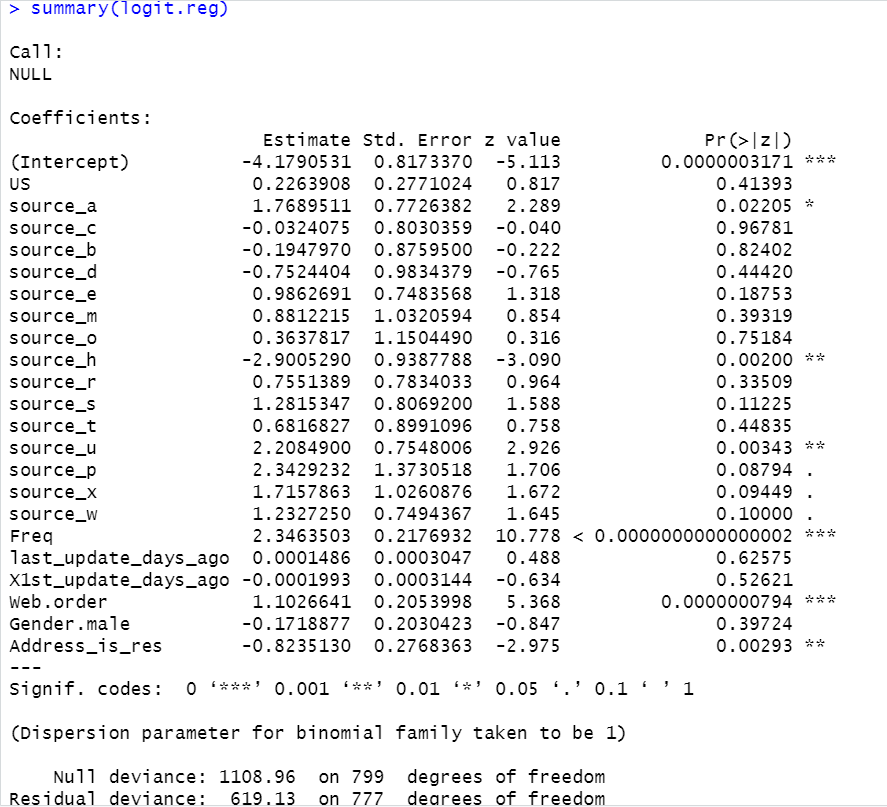
In the above equation purchase is a dependent variable, “.” It represents the all independent variables in the data, -Spending it represents the excluding that column and method glm() represents the function of logistic regression and will use training data to train the model.



**Figure 5.1 Fitting the training data in logistic regression**

By looking at the above figure 5.1 can say that the accuracy of the training data is 79.5%

Let’s interpret the summary of the logistic regression



**5.2 Summary of the logistic regression**

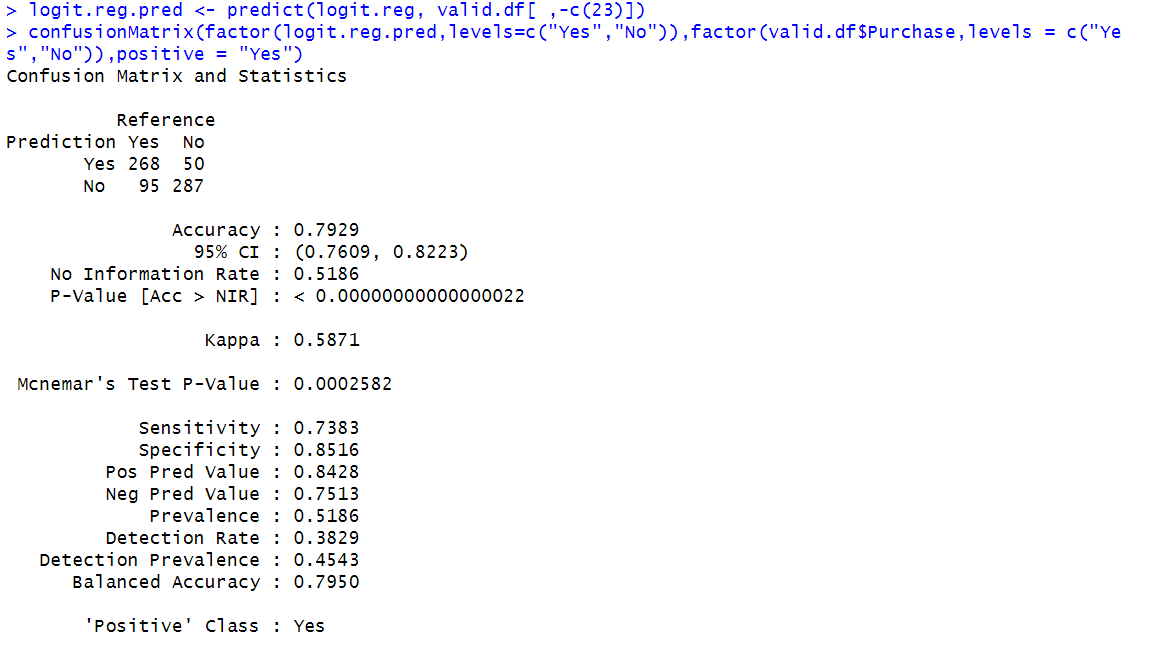
**Interpretation Of Logistic Regression:**

Here is the estimated equation for the logistic

Logit(Purchasers = Yes) = -4.1790 + 0.2263 US +1.7689 source\_a – 0.0324 source\_c -0.1947 source\_b – 0.7524 source\_d + 0.9862 source\_e + 0.8812 source\_m + 0.3637 source\_o – 2.9000 source\_h + 0.7551 source\_r + 1.2815 source\_s + 0.6816 source\_t + 2.2084 source\_u + 2.3429 source\_p +1.7157 souce\_x +1.2327 souce\_w + 2.3463 Freq + 0.00014 last\_updated\_days\_ago – 0.00019 X1st\_updated\_days\_ago + 1.1026 web.order – 0.1718 Gender.male – 0.8235 Address\_is\_res.

The positive coefficients for the categorical variables US, Source\_a, Source\_e, Source\_m, Source\_o, Source\_r, Source\_s, Source\_t, Source\_u,Source\_p, Source\_x, Source\_w, web.order are associated with higher probabilities of purchaser. The negative coefficients for the categorical variables Source\_c, Source\_b, Source\_d, Source\_h, Gender.male, Address\_is\_res are associated with the lower probablities of purchaser. For the numerical predictors, the poisitive coefficients of freq and last\_updated\_days\_ago is associated with a higher probability of purchasing the channels. Similarly, negative coefficient of X1st\_updated\_days\_ago is associated with a lower probablility of purchaser.

After fitting the model, will make the predictions with the validation data and evaluate the performance with that model equation. From this will define the common metrics for classification like accuracy, error rate, specificity, sensitivity, precision, and recall.

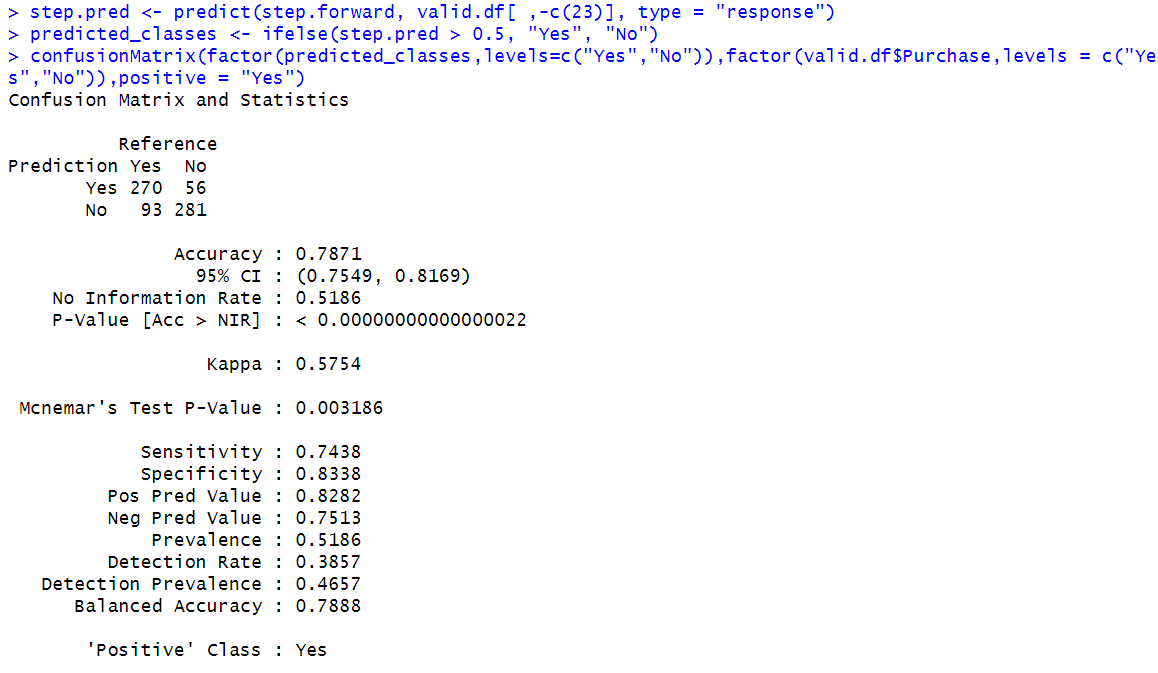


**Figure 5.3 Confusion matrix and statistics of logistic regression**

In this the important class is purchase. The model aims to predict how many customers are likely to be purchasers. Sensitivity helps to estimate how many purchasers the model can identify correctly. Based on the figure 5.3, the logistic regression model has an accuracy of 79.29%, a sensitivity of 73.83%, and a specificity of 85.16%. This means the model can correctly identify 73.83% of customers who are actual purchasers. Using this information, business owners can make informed decisions about customer targeting and marketing strategies.

**5.1.1 Forward Stepwise Logistic Regression:**

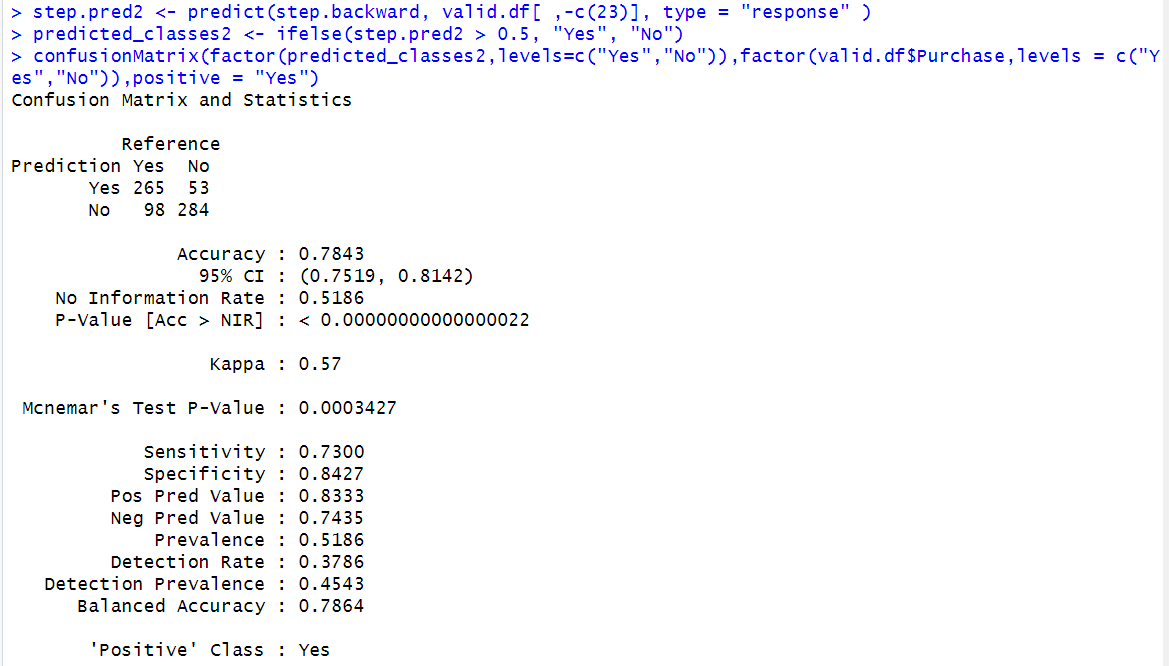
To improve the model performance did the stepwise logistic regression to perform the forward and backward selection of the individual variables. In forward selection it will starts with null, and it will add the one-by-one predictors to get the lower results of AIC. Here we have different criteria to evaluate the model in this we consider the AIC. Let’s compare the selection approach to choose the best one. First will have the look on the forward selection predictions to see the common metrics. By looking at below figure 5.4 can say that the accuracy is 78.71%, sensitivity is 0.7438 and specificity is 0.83. The sensitivity is increased when compared with the logistic regression.



**Figure 5.4 Confusion Matrix and statistics of forward stepwise logistic regression**

**5.1.2 Backward Stepwise logistic Regression:**

In this it will starts with the full model, and it will reduce the predictors. Let’s have the look on backward selection to see the common metrics.



**Figure 5.5 Confusion Matrix and statistics of backward stepwise logistic regression**

By looking at the above figure 5.5 can say that accuracy is 78.43%, sensitivity is 0.7300 and specificity 0.8427. Here the sensitivity is slightly decreased when compared with the logistic regression. That’s not a major difference. By comparing all the sensitivity values, can be able to say that forward selection logistic regression is greater than the backward selection logistic regression. The performance is increased by doing the forward selection logistic regression. A higher sensitivity indicates that the model is ability to predict purchasers more accurately which is more helpful to business owners to make the informed decisions about the customers.

Now compare all models of accuracy and sensitivity to select the best estimated model equation

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** |
| Logistic | 79.29% | 0.7383 |
| Forward Step Wise logistic | 78.71% | 0.7438 |
| Backward Step Wise Logistic | 78.43% | 0.7300 |

**Comparison table for Accuracy and sensitivity**

Comparing all values, we can say that the forward stepwise logistic model is the best one, as it has the highest sensitivity among the models.

**5.2 Decision Tree By Using rpart():**

Decision trees can be used for both classification and prediction. In this going to use for classification to classify the new record. It works by recursively splitting the dataset into rectangle based on the most important variable at each rectangle is as homogeneous or pure as possible. Will use the rpart and rpart.plot packages in R to build and visualize classification and regression trees. The class with highest vote is assigned to the new record. The equation for the decision tree is

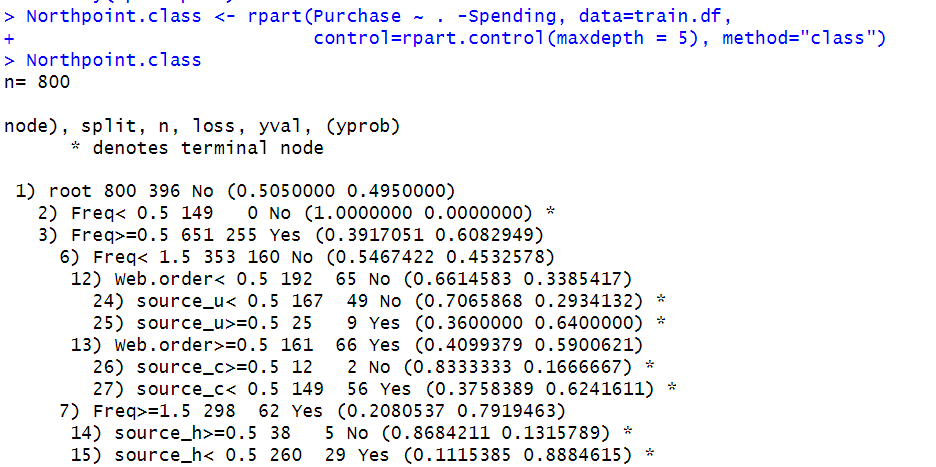
library(rpart)

library(rpart.plot)

Northpoint.class <- rpart(Purchase ~ . -Spending, data=train.df,

control=rpart.control, method="class")

we build the model with the training data. The above equation represents the rpart is a function will use from the rpart package, purchase is a dependent variable and going to use all predictor variables except spending. Let’s have look on the results of the tree structure.

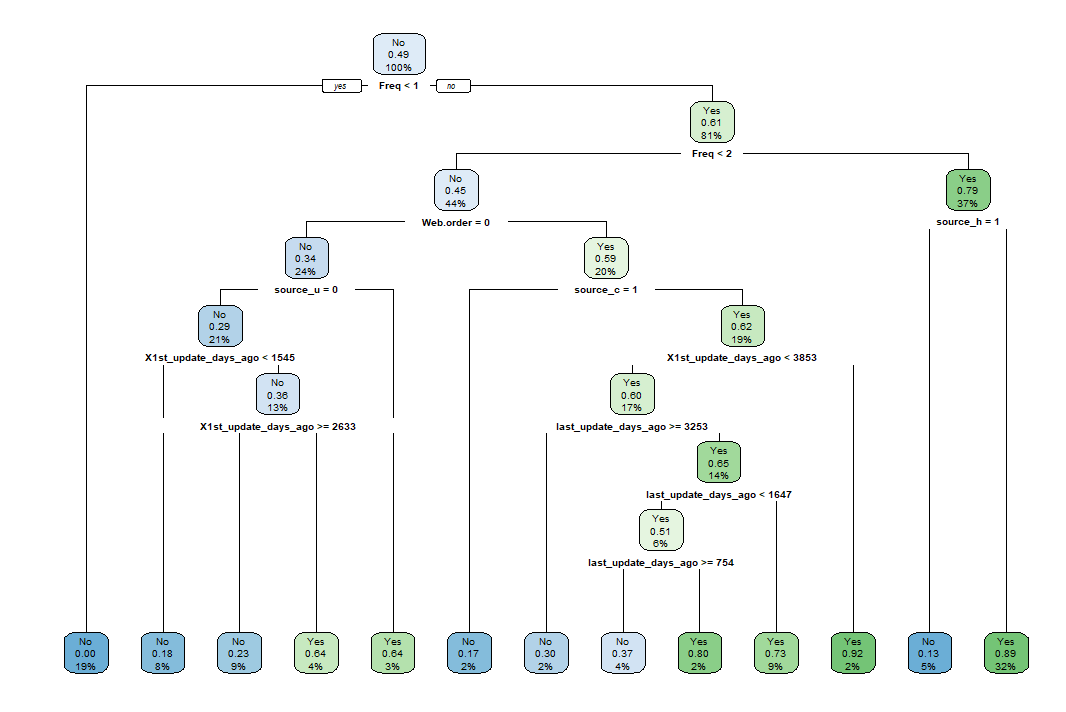


**Figure 5.2.1 Classification Rules**

Plotted the grown tree for the better understanding of the classification rules. Each leaf node is equivalent to a classification rule.

**Classification Rules:**

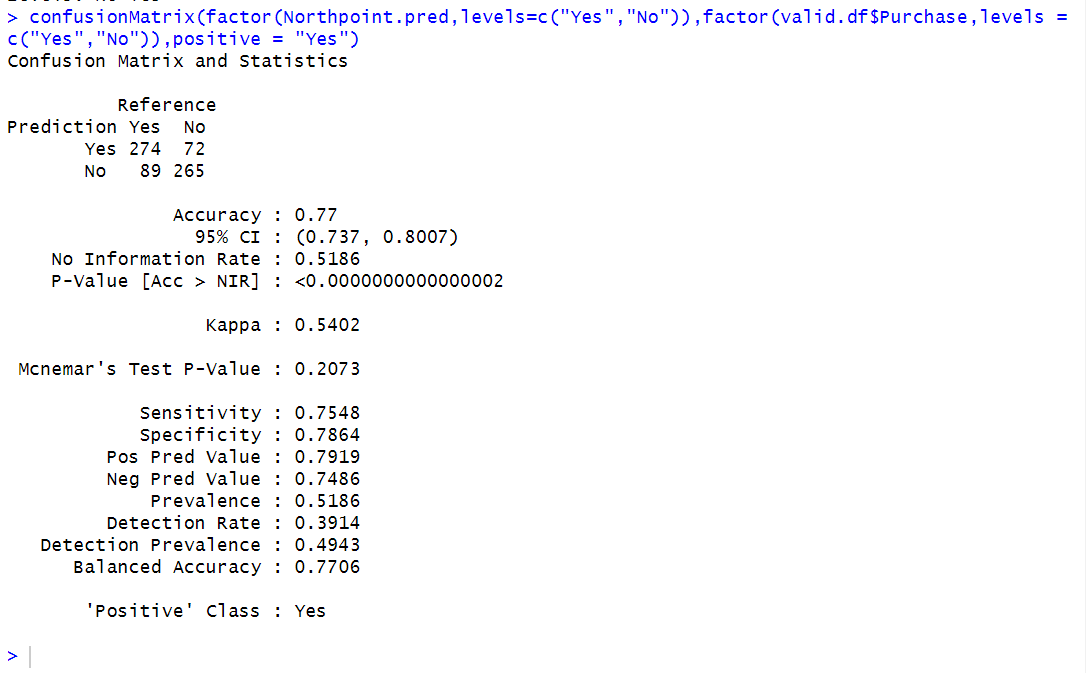
1. IF Freq<1 THEN Class is “No”.
2. IF Freq>=1 AND Freq<2 AND Web.order<0 AND Source\_u<0 AND x1st\_updated\_days\_ago<1646 THEN Class is “No”.
3. IF Freq>=1 AND Freq<2 AND Web.order<0 AND Source\_u<0 AND X1st\_updated\_days\_ago>1646 AND X1st\_updated\_days\_ago<2633 THEN Class is “No”.
4. IF Freq>1 AND Freq<2 AND Web.order<0 AND Source\_u<0 AND X1st\_updated\_days\_ago>= 2633 THEN Class is “YES”.
5. IF Freq>1 AND Freq<2 AND Web.order<0 AND Source\_u=0 THEN Class is “Yes”.
6. IF Freq>1 AND Freq<2 AND Web.order<0 AND Source\_c<1 THEN Class is “NO”.
7. IF Freq>2 AND Freq<2 AND Web.order=0 AND Source\_c=1 AND X1st\_updated\_days\_ago>3853 THEN Class is “Yes”.



**Figure 5.2.2 Plotted Tree**

1. IF Freq>1 AND Freq<2 AND Web.order=0 AND Source\_c=1 AND X1st\_updated\_days\_ago<3853 AND last\_updated\_days\_ago<3253 THEN Class is “No”
2. IF Freq>1 AND Freq<2 AND Web.order=0 AND Source\_c=1 AND X1st\_updated\_days\_ago<3853 AND last\_updated\_days\_ago>=3253 AND last\_updated\_days\_ago<754 THEN Class is “No”.
3. IF Freq>1 AND Freq<2 AND Web.order=0 AND Source\_c=1 AND X1st\_updated\_days\_ago<3853 AND last\_updated\_days\_ago>=3253 AND last\_updated\_days\_ago<1647 THEN Class is “YES”.
4. IF Freq>1 AND Freq<2 AND Web.order=0 AND Source\_c=1 AND X1st\_updated\_days\_ago>3853 THEN Class is “Yes”.
5. IF Freq >2 AND Source\_h <1 THEN Class is “NO”.
6. IF Freq >2 AND Source\_h=1 THEN Class is “Yes”.

Now evaluate the performance of the model with the validation data by using already built model. Then can able to check the common metrics from the confusion matrix like accuracy, sensitivity, and specificity.



**Figure 5.2.3 Confusion matrix and statistics of decision tree using rpart**

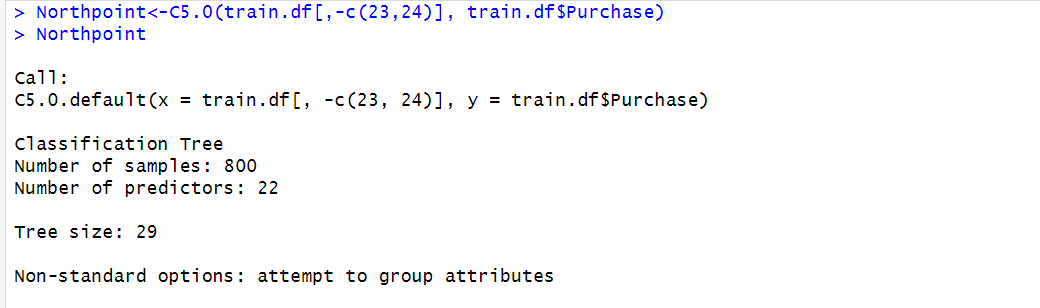
By looking at figure 5.2.3 can say that the accuracy is 77%, sensitivity is 0.7548 and specificity is 0.7864. There is a 75.48% chance of correctly classifying the purchasers by using this model. So, the business owner can make more informed decisions about the customers, and which will increase their customers and expand their business.

**5.3 Decision Tree By using the C5.0:**

This is another approach to perform the decision tree. Let’s check the results of this model and compare with the decision tree by using the rpart(). The model equation of the decision tree by using the c5.0 is

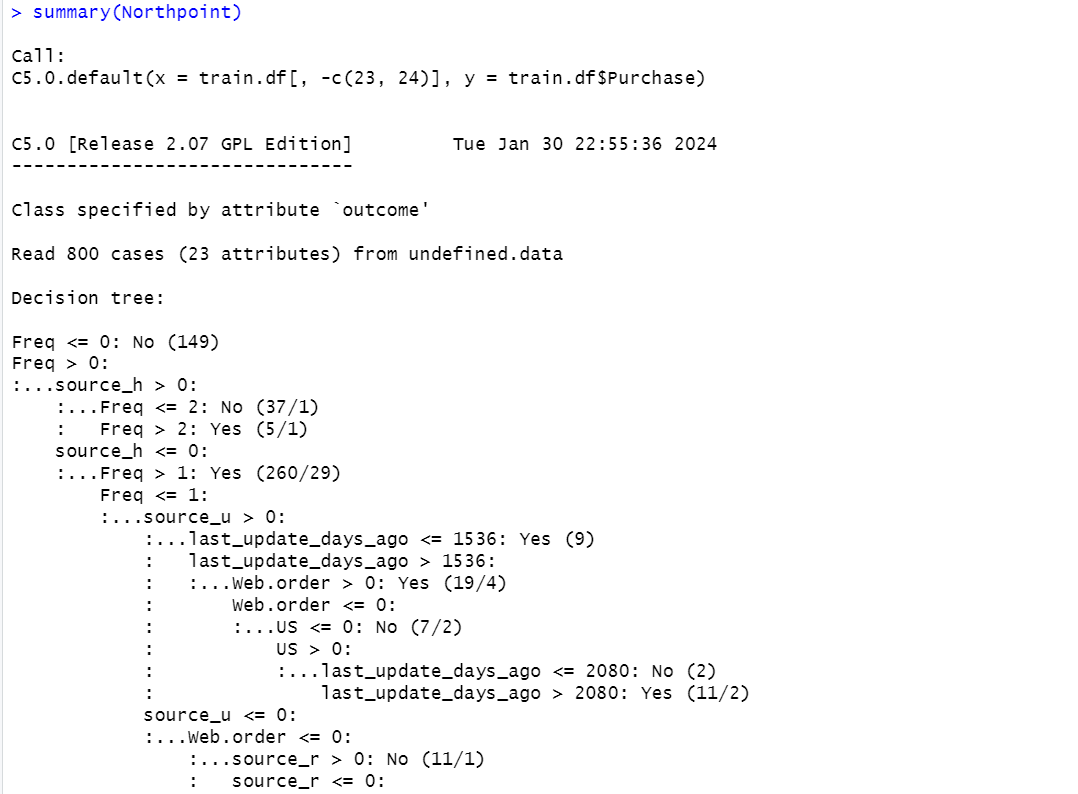
Northpoint<-C5.0(train.df[,-c(23,24)], train.df$Purchase)

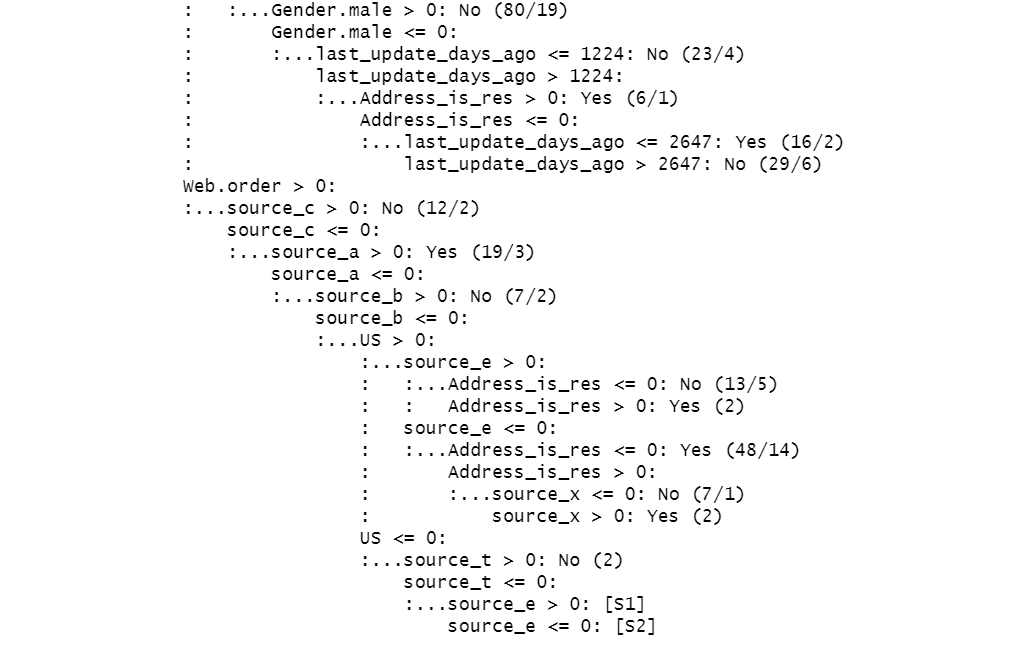
By using this equation will train the model with the training data. By looking at the below screen shot we can say that tree size is 29.

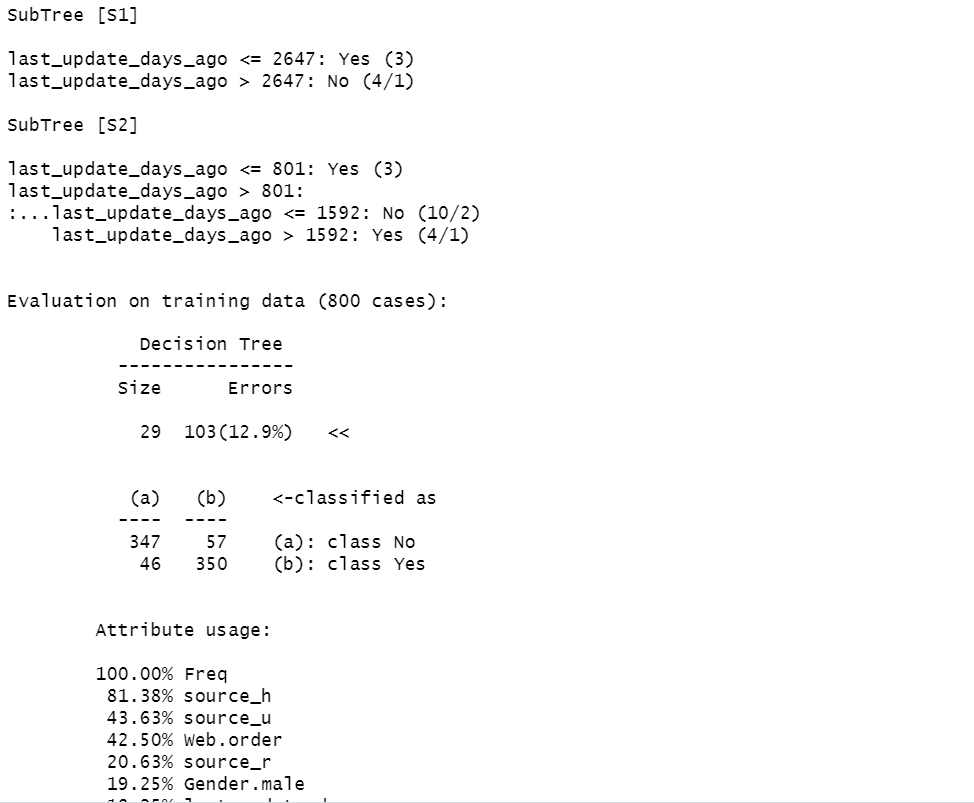


**Figure 5.3.1 Model equation of decision tree by using C5.0**

Now let’s have a look on the summary of the training data

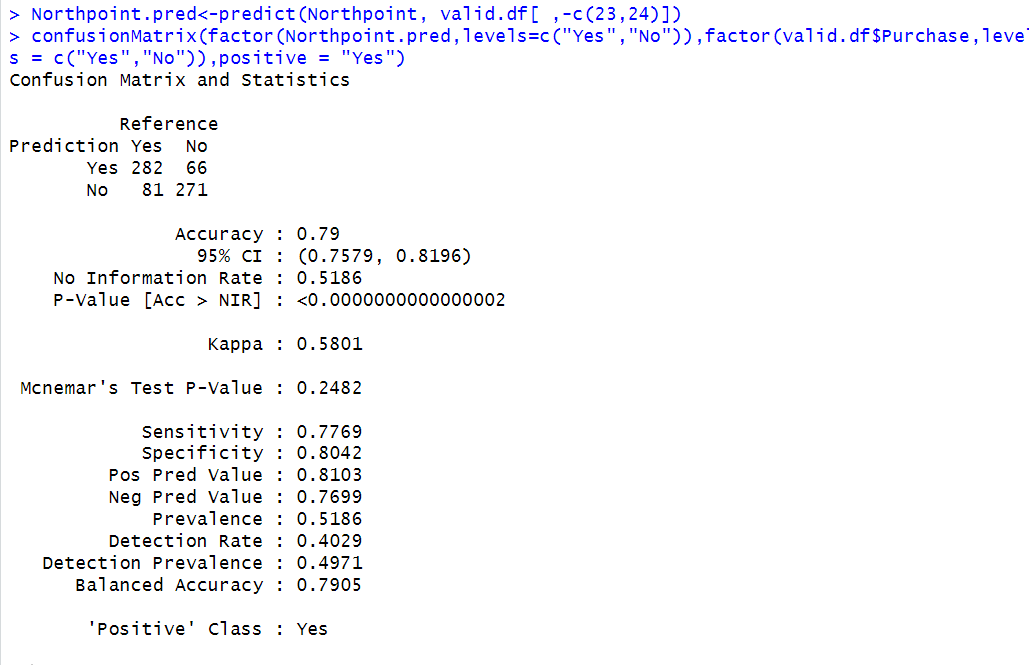






**5.3.2 Figure Summary of the model by using the training data**

Now evaluate the performance with the validation data by using the just build model with the training data and check the common metrics from the confusion matrix.



**Figure 5.3.3 Confusion matric and statistics of the decision tree using C5.0**

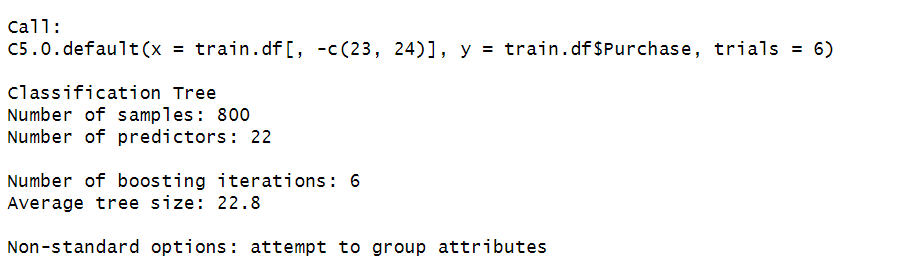
From the above figure 5.3.3 can say that accuracy is 79%, sensitivity is 0.7769, specificity is 0.8042 and the CI is (0.7579, 0.8196).

Now evaluate the performance with the trial option which specify the number of boosting iterations and we can specify the large number of iterations let’s do the trial on 6, 10 and 15.

The model equation for the trial 6 is

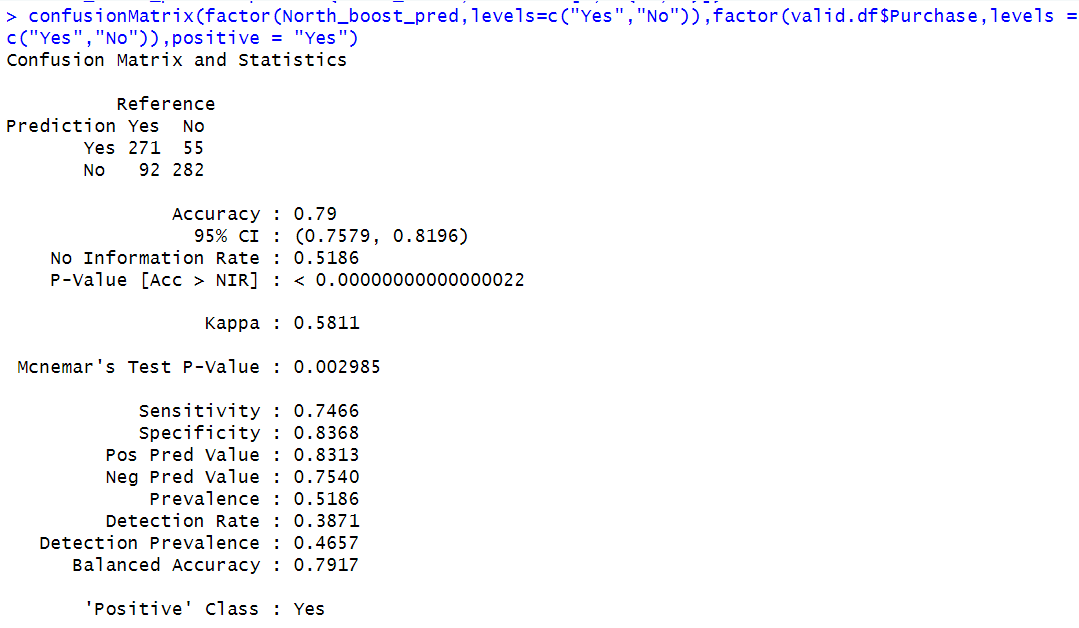
North\_boost<-C5.0(train.df[ , -c(23,24)], train.df$Purchase, trials=6)

Here trials = 6 represents the number of iterations.



**Figure 5.3.4 Trial 6 Model equation of decision tree by using C5.0**

Observing that the tree size decreased when compared with the single iteration. Let’s evaluate the model performance with the validation data by using the model equation and check the metrics.



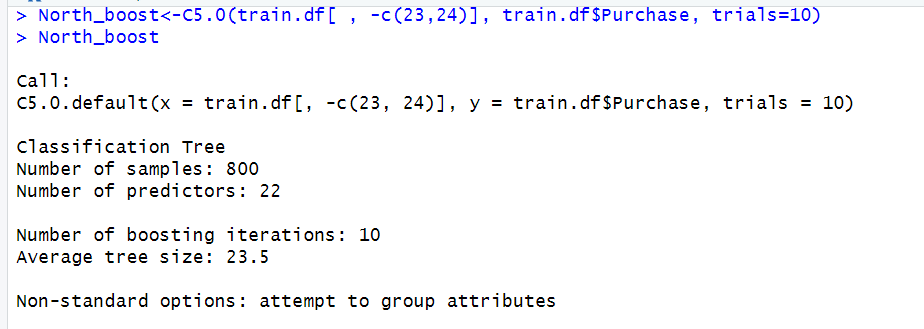
**Figure 5.3.5 Trial 6 Confusion matric and statistics of the decision tree using C5.0**

By looking at the above screenshot can say that accyracy is 79%, sensitivity is 0.7466, specificity is 0.8368 and CI is (0.7579, 0.8186).

The model equation for the trial 10 is

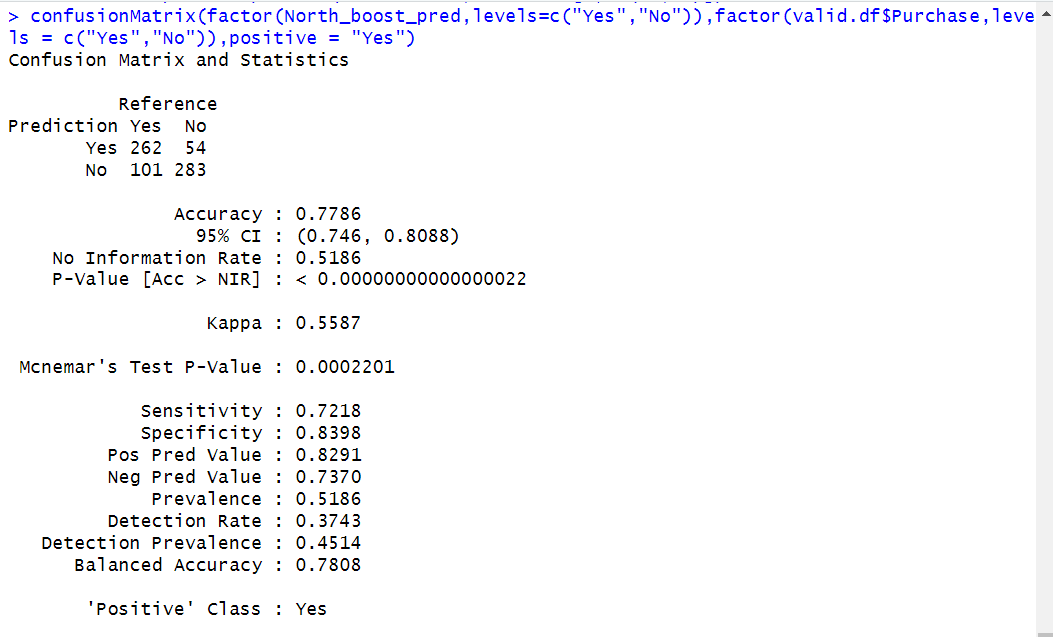
North\_boost<-C5.0(train.df[ , -c(23,24)], train.df$Purchase, trials=10)

Here the trial 10 represents the number of iterations let’s check the results for this equation



**Figure 5.3.6 Trial 10 Model equation of decision tree by using C5.0**

Observing that the tree size is decreased when compared with the single iteration and it is increased when compared with the trial 6. Now, evaluating the performance of the model with validation data and check the common metrics from the confusion matrix.



**Figure 5.3.7 Trial 10 Confusion matric and statistics of the decision tree using C5.0**

From the above figure 5.3.7 can say that the accuracy is 77.86%, sensitivity is 0.7218, specificity is 0.8398 and CI is (0.746, 0.8088).

Let’s compare the accuracy of all models which is default, trial 6 and trial 10.

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Sensitivity |
| Default C5.0 | 79% | 0.7789 |
| Trial 6 | 79% | 0.7466 |
| Trial 10 | 77.86% | 0.7218 |

**Comparison table for accuracy and sensitivity of decision tree by using C5.0**

Default model C5.0 equation have the highest accuracy when compared with the trial 6 and trial 10. So, the best one is default model equation which is single model used.

**5.4 BEST CLASSIFIER MODEL SELECTION:**

Compare the accuracy and sensitivity of all classifier models to identify the best-performing model among them.

|  |  |  |
| --- | --- | --- |
| MODEL | ACCURACY | SENSITIVITY |
| Forward Step Wise Logistic Regression | 78.71% | 0.7438 |
| Decision Tree by using rpart | 77% | 0.7548 |
| Decision Tree by using C5.0 | 79% | 0.7769 |

**Comparison table for all classifier models**

By comparing all measures with all other model measures we can conclude that Decision tree by using C5.0 is the best classifier model because it has the highest accuracy along with the highest sensitivity which means it will be the best performing model and we have the highest chance to get the more purchasers. But will choose the best model as forward step wise logistic regression because based upon our business requirement 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 purchaser and lower probability indicates that low chance of becoming a purchaser with these, we will get an accurate result. Probability of interpretation aligns well with business requirements, especially when assessing the probability of a customer becoming a purchaser.

**6.Regression Models:**

By using the regression models, we can make the predictions on the numerical output. According to our business goal we need to estimate the amount of spending on each purchase. To predict that we build a predictive model, and we can select the best performance based on the metrics.

**PREPARING THE DATA:**

According to our business requirement filtered only purchasers’ data from the whole data and created train and validation datasets. After that we train the model with the training data and evaluate the model performance with validation data

**6.1.1 Multiple Linear Regression Model:**

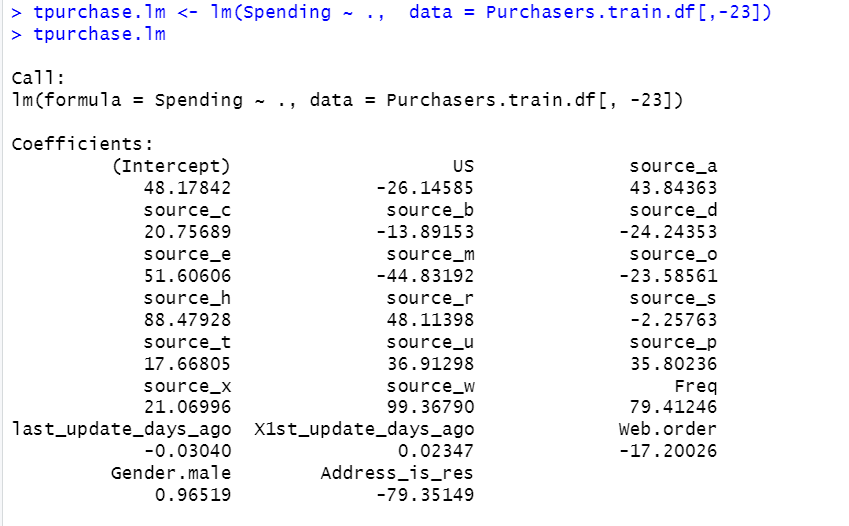
In this regression represents model of the relationship between the target variable and independent variables. We assume the linear relationship between the dependent and group of independent variables. Which may follow the multiple linear regression model. The equation for the multiple linear regression is

Y = b0+b1.x1+b2.x2+b3.x3+b4.x4+…………………….bn.xn

To access the linear regression model we use lm() function.

Later developed a regression model with the training data and evaluate the performance with the validation data. The equation of the linear model is

tpurchase.lm <- lm(Spending ~ ., data = Purchasers.train.df[,-23]).



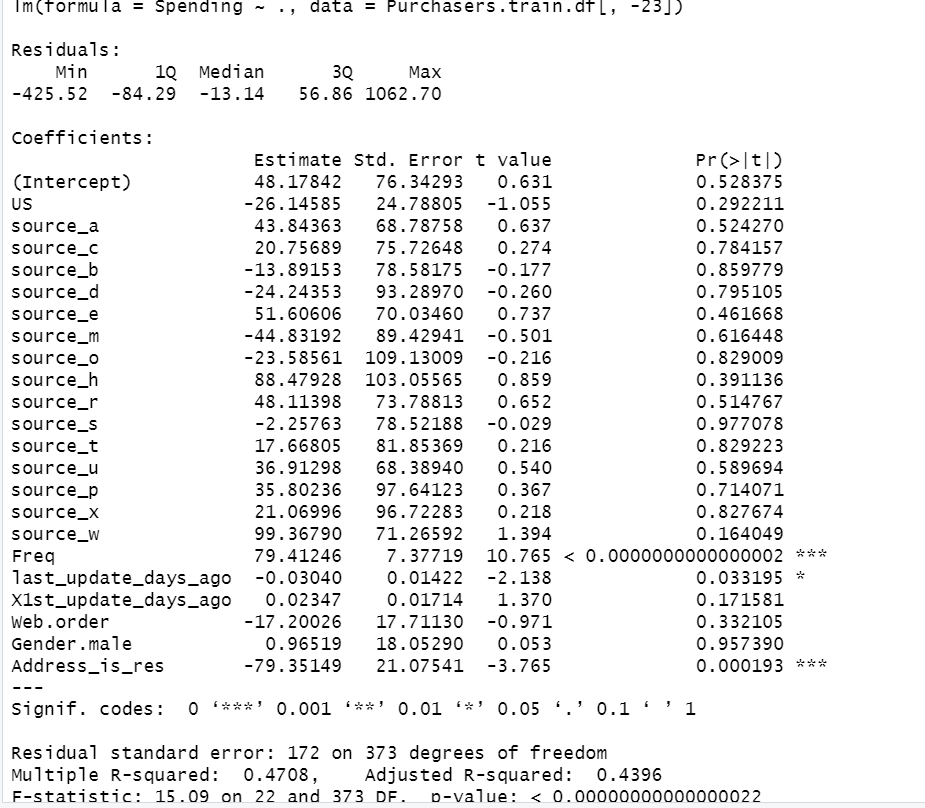
**Figure 6.1 Estimated Linear Model Equation**

The estimated equation of the linear model is

Y = 48.17842 + US\*(-26.14585) + source\_a\*43.84363 + source\_c\*20.75689 + source\_b\*(-13.89153) + source\_d\*(-24.24353 ) + source\_e\*51.60606 + source\_m\*(-44.83192) + source\_o\*(-23.58561) + source\_h\*88.47928 + source\_r\*48.11398 + source\_s\*(-2.25763) + source\_t\*17.66805 +source\_u\*36.91298 + source\_p\*35.80236 + source\_x\*21.06996 + source\_w\*99.36790 + Freq\*79.41246 + last\_update\_days\_ago\*(-0.03040) + X1st\_update\_days\_ago\*( 0.02347) + Web.order\*(-17.20026) + Gender.male\*(0.96519 ) + Address\_is\_res\*(-79.35149).

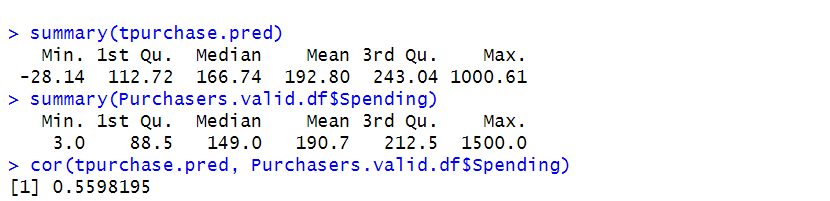
By using this equation, can predict the estimated spending amount for each purchase.

Here, the summary of the linear model is



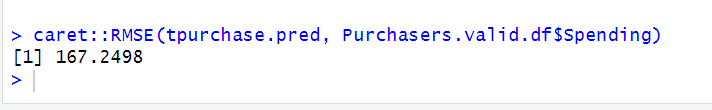
**Figure 6.2 Summary of Linear model**

By looking at the above figure 6.2 can say that the most significant variables are Freq and Address\_is\_res. R-Squared value is 0.4708 and adjusted R-squared value is 0.4396.



**Figure 6.3 Summary of the predicted and actual values.**

By looking at above figure 6.3 can be seen that there is a bigger difference between both ends predicted and actual values that means predictive power is relatively low which may relate to the skewed distribution of the data. The correlation between the actual and predicted value is potentially correlate to each other which is 0.5598.



**Figure 6.4 RMSE value of linear model**

The RMSE value of the linear model is 167.2498.

**6.1.2. FORWARD STEP-WISE LINEAR REGRESSION:**

To improve the performance of the linear regression going to perform the forward stepwise regression in this the equation starts with no predictors and it will add the one-by-one predictors.

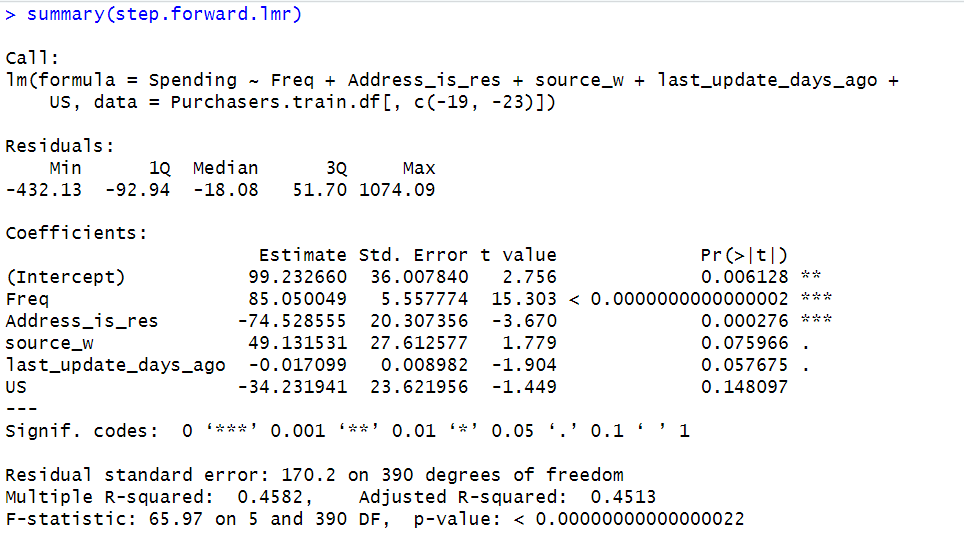
The equation of the forward step wise regression is

linear.null <- lm(Spending.log~1, data = Purchasers.train.df[,c(-19,-23,-24)])

linear.full <- lm(Spending.log ~ ., data = Purchasers.train.df[,c(-19,-23,-24)])

step.forward.lm <- step(linear.null, scope=list(lower=linear.null, upper=linear.full), direction = "forward").

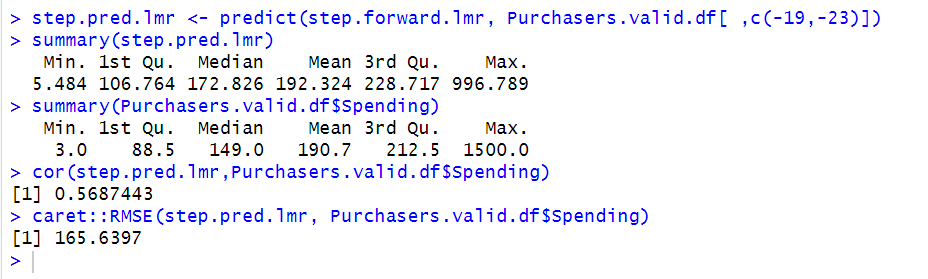
Let’s have a look on the summary of the forward stepwise regression is



**Figure 6.5 summary of forward step wise regression**

The estimated equation for the forward step wise regression is

Y = 99.232660 + Freq\*85.050049 + Address\_is\_res\*(-74.528555) + source\_w\*49.131531 + last\_update\_days\_ago\*(-0.017099) + US\* (-34.231941).



**6.6 Summary of Predicted and Observed Values**

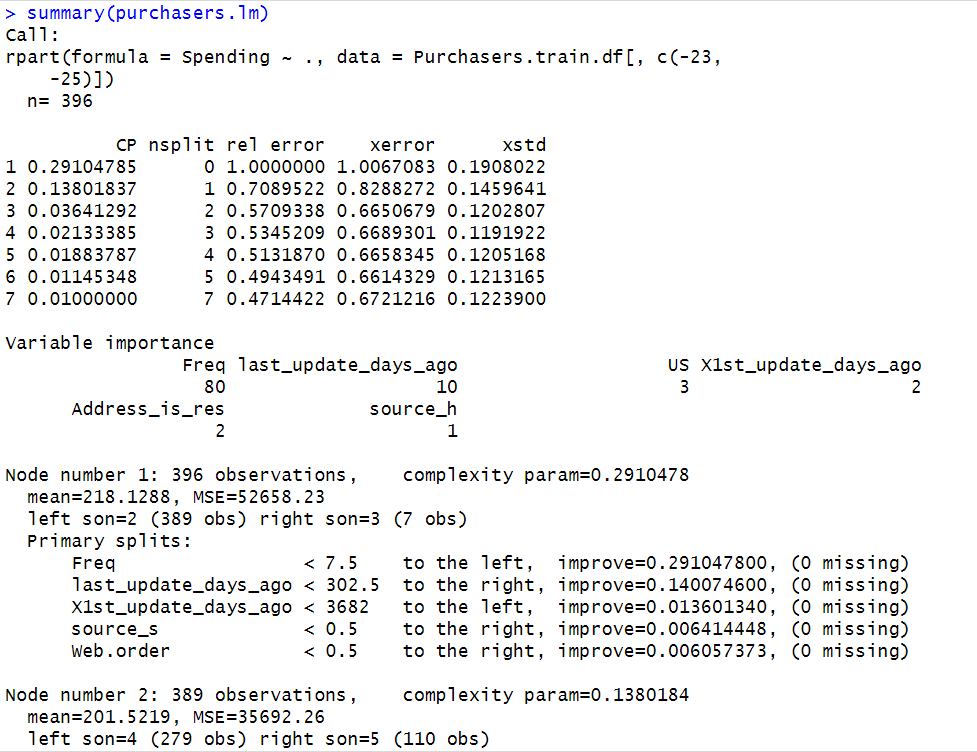
By looking at the above figure 6.6 can be observed that the difference between the actual and predicted values is more which means that the predictive power is relatively lower. This may lead to the skewed spending distribution. The correlation between the predicted and actual values is 0.568 which means that they are potentially correlated to each other. The RMSE value is 165.6397.

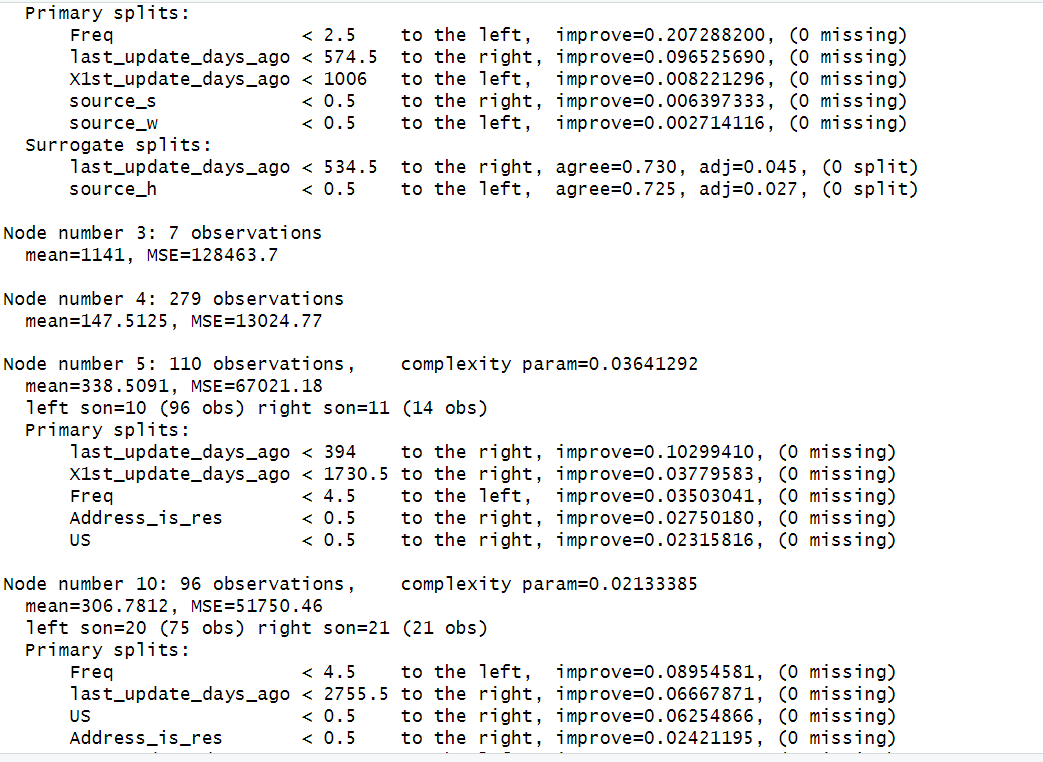
**6.2 REGRESSION TREE:**

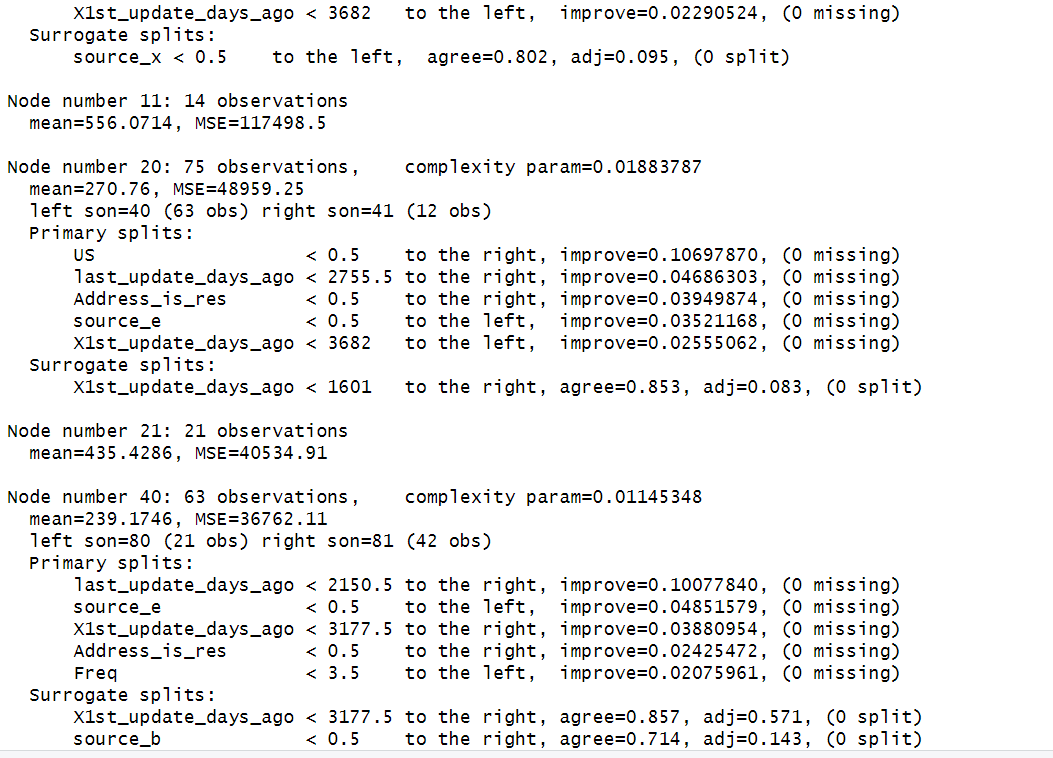
In regression tree prediction for new data are based on the rules represented on the tree in regression trees predictions are obtained by averaging the outcome values in the nodes. To access this regression tree we use r.part() function. It is based on the rules of the tree. In regression-tree prediction,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

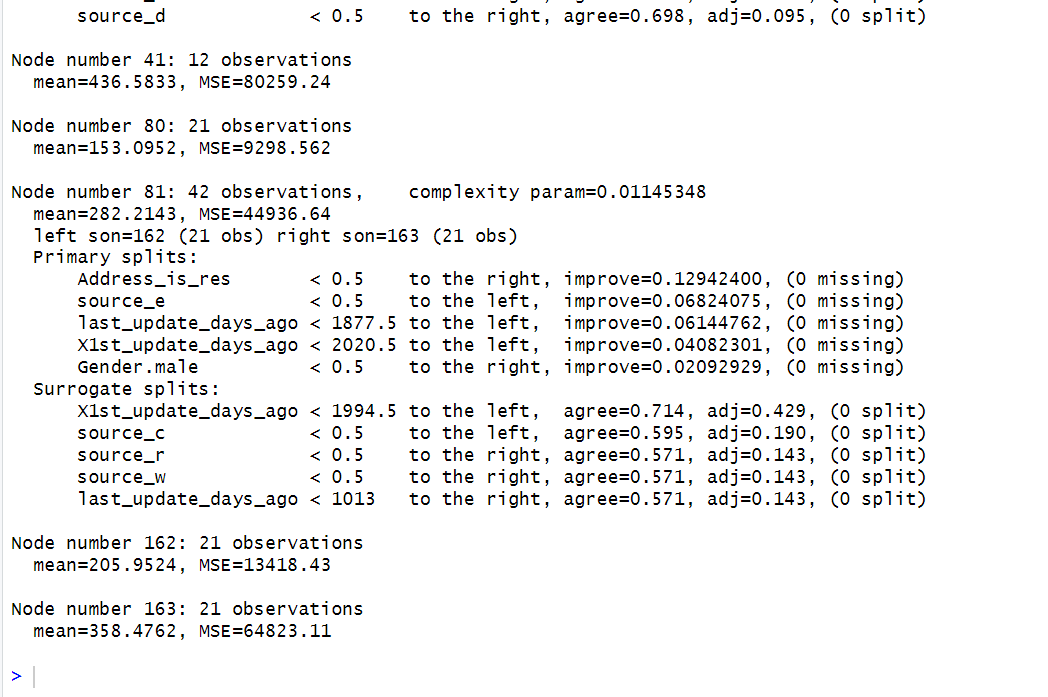
purchasers.lm <- rpart(Spending ~ ., data = Purchasers.train.df[ ,c(-23,-25)])

The summary of the regression tree is



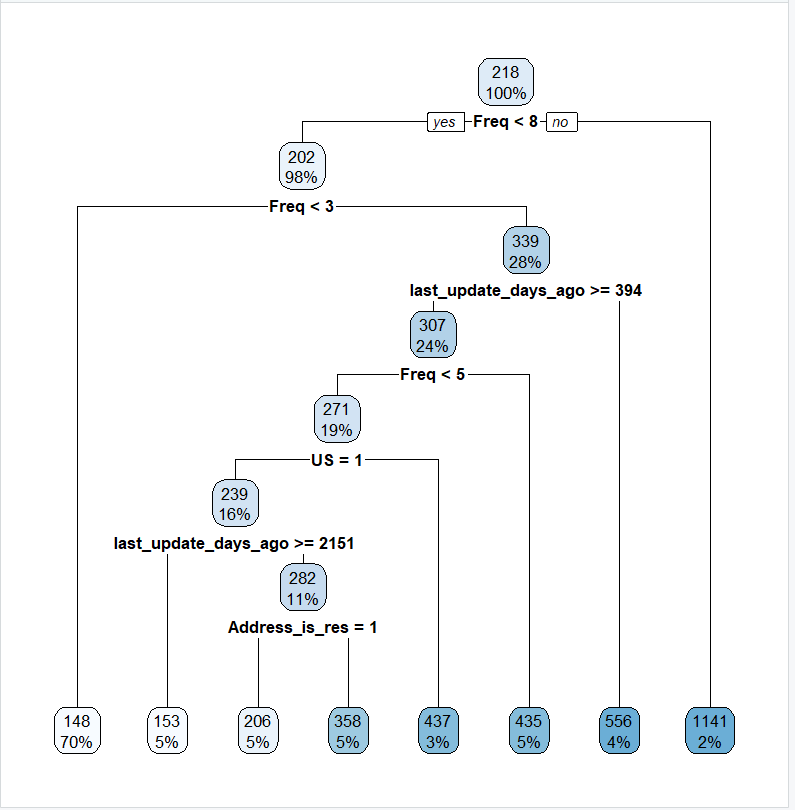






**Figure 6.7 Summary of the regression tree model**

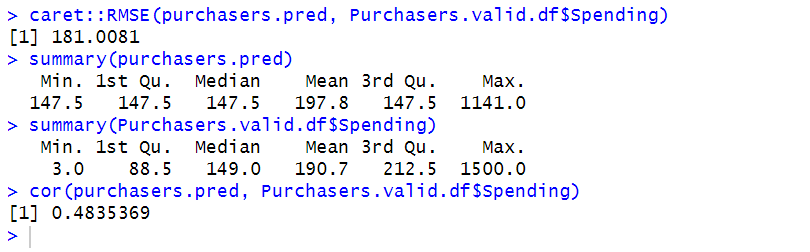
Have a look on the plotted tree



**Figure 6.8 Plotted regression tree**

Here we observed that nodes are divided based on the rule condition.

Let’s have a look on the summaries of the actual and predicted values



**Figure 6.9 Summary of Actual and predicted values of regression tree**

Observed that there is a lot of difference between the actual and predicted values which means that it has a lower predictive power and the correlation between the actual and predicted is 0.48 which means that correlated with each other and the RMSE value is 181.0081. There is no improvement in the performance when we compare with the linear equation model.

**6.3 Comparison of All Regression Models:**

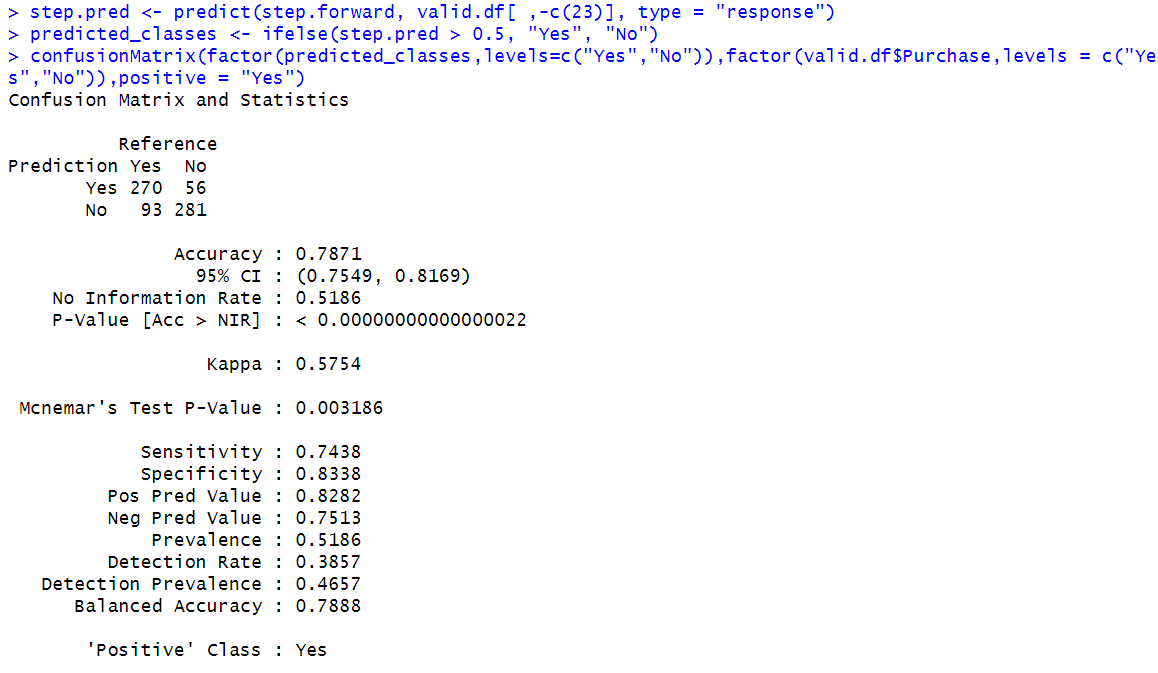
Let’s compare all RMSE values of the model and will select the best performance model based upon the lowest RMSE values.

|  |  |
| --- | --- |
| MODEL | RMSE |
| Linear Regression | 167.2498 |
| Forward Step Wise Linear Regression | 165.6397 |
| Regression Tree | 181.0081 |

By looking at the above comparison can be seen that forward stepwise linear regression has the lowest RMSE value which is 165.6397 there is a chance of providing the accurate results than all other models. By looking at the training and validation RMSE values got thought that may lead to the overfitting of the model but it’s not because based upon the seed value we may get like that, and it doesn’t make any difference with that slight difference. So, by considering all these will prefer the best model as forward step wise linear regression and will perform the evaluation with the holdout data and able to see the results.

**7.REPORT ON MODEL PERFORMANCE:**

The best classification model is forward step wise logistic regression model. The common metrics for this model is evaluated by confusion matrix. Let’s have a look on the confusion matrix



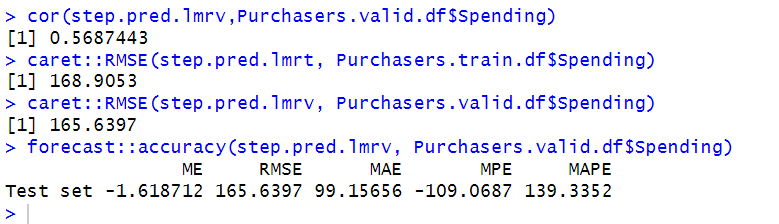
**Figure 7.1 Confusion matrix of Forward Step wise logistic regression**

Accuracy: It represents the overall measure how well the model is performing and by looking at the above screenshot can be seen that the accuracy 0.7871 which 78.71%

Sensitivity: It represent the ability of the model correctly predicting the important class according to our business the important class is purchasers in the above screenshot can be seen that sensitivity is 0.7438 which is 74.38% it means that there is a chance of getting the 74% of purchasers by using this model.

Specificity: It represent the ability of the model correctly predicting the non-important class according to our business the non-important class is not purchasers in the above screenshot can be seen that specificity is 0.8338 which is relatively high ability to correctly identify that truly belong to the non-purchaser class.

The best regression model is forward step wise linear regression among all other models with the lowest RMSE which is 165.6397 let’s interpret the results of this model

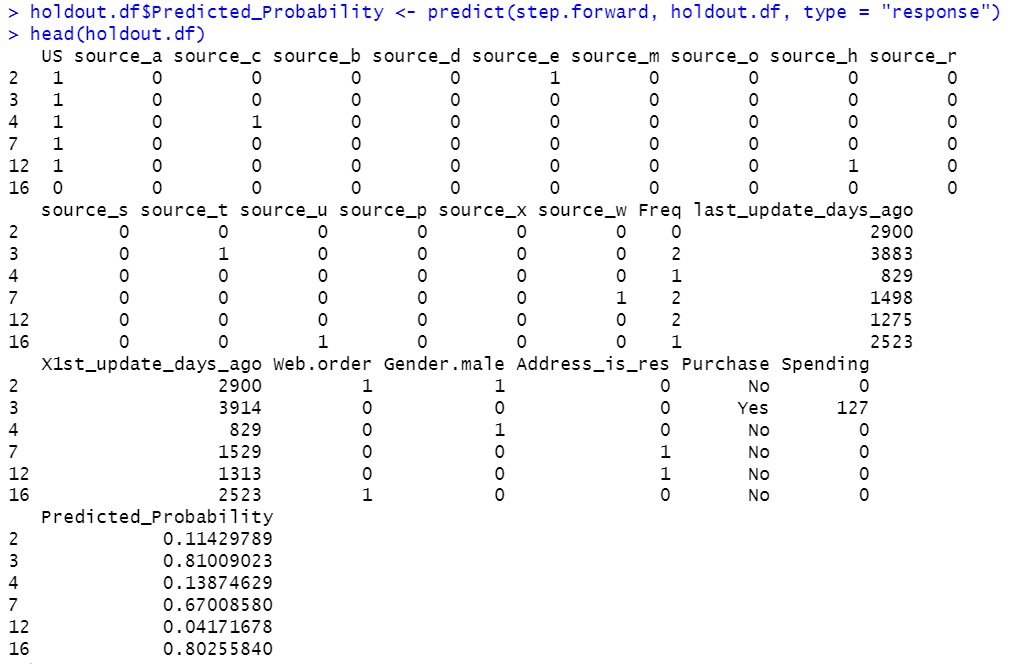


**Figure 7.2 Common Metrics of Forward step wise linear regression**

By looking at above screenshot can be seen that mean the RMSE value is 165.6397 which is lowest and will provide better accurate results while predicting the new customer spending amount.

**8.HOLDOUT MODEL ACCURACY:**

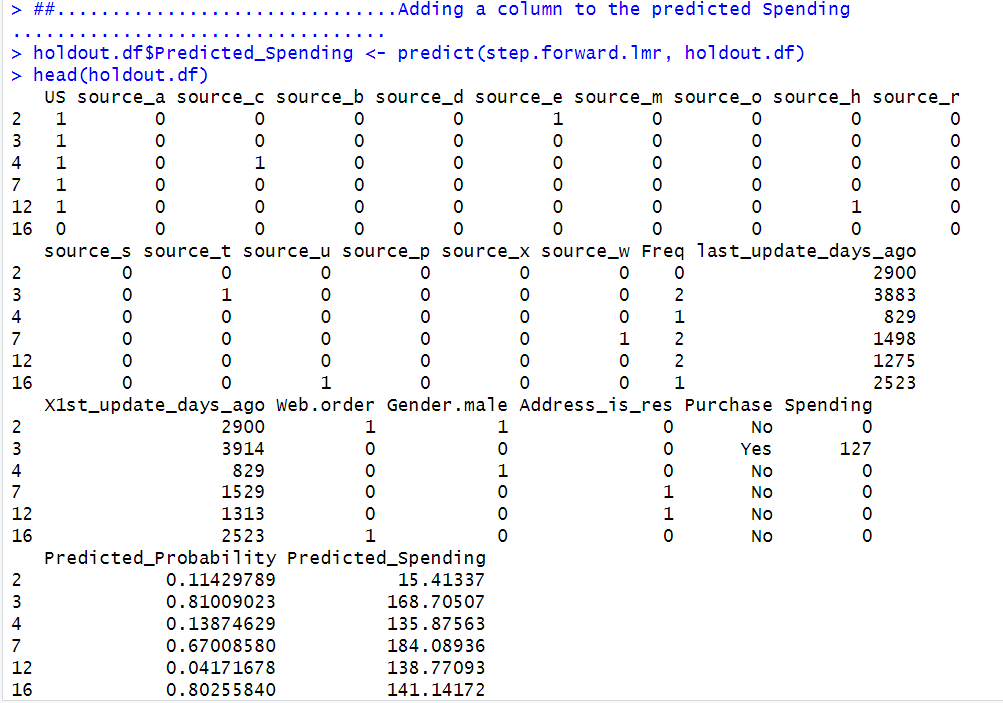
Now, considering the original holdout data for checking the model performance with the new data. First, going to add the predicted probability to the holdout data set and have look on the predicted probability.



**Figure 8.1 Predicted Probability in Holdout data**

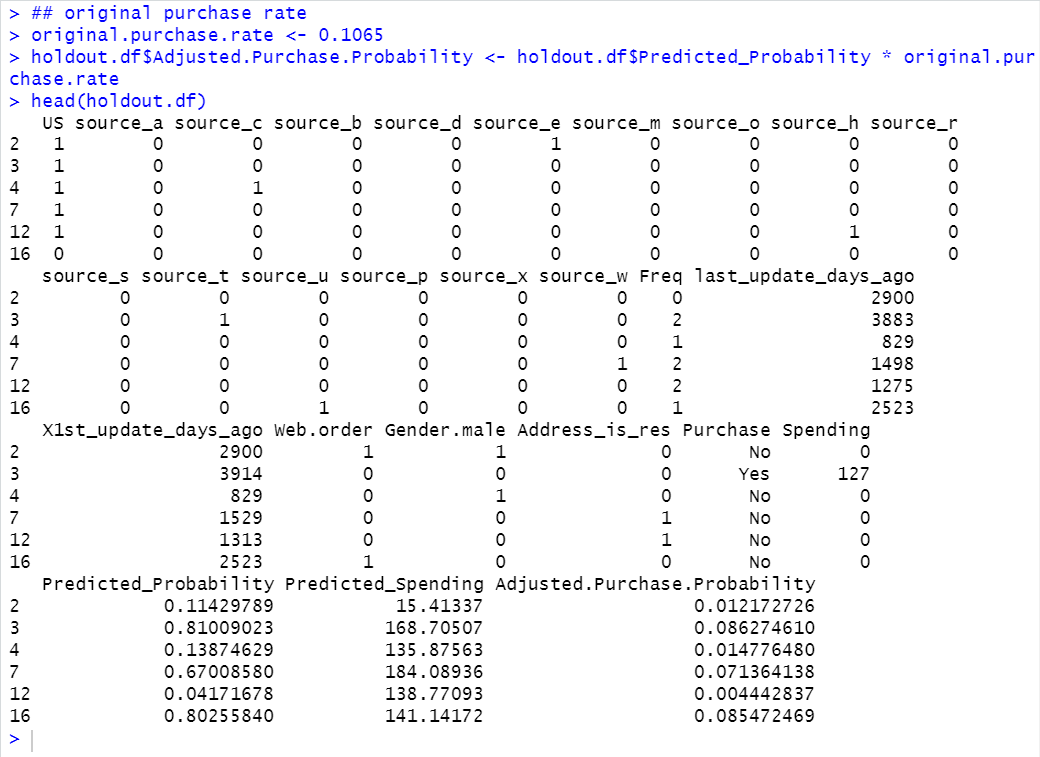
By looking at the above figure 8.1 can be observed that predicted\_probability column added to the holdout data and can see the predicted probabilities for each purchase and spending. By adding the predicted probabilities enhances the utility and interpretability of the model and providing the better understanding of the predictions of purchasers and non-purchasers.

Now, going to add the predicted spending column to the holdout data and can check for the predicted spending values. Let’s have look on the predicted spending column.

 **Figure 8.2 Predicted Spending in Holdout data**

By adding this predicted spending value promotes to make the decision making and provide the valuable estimation of each customer spending amount.

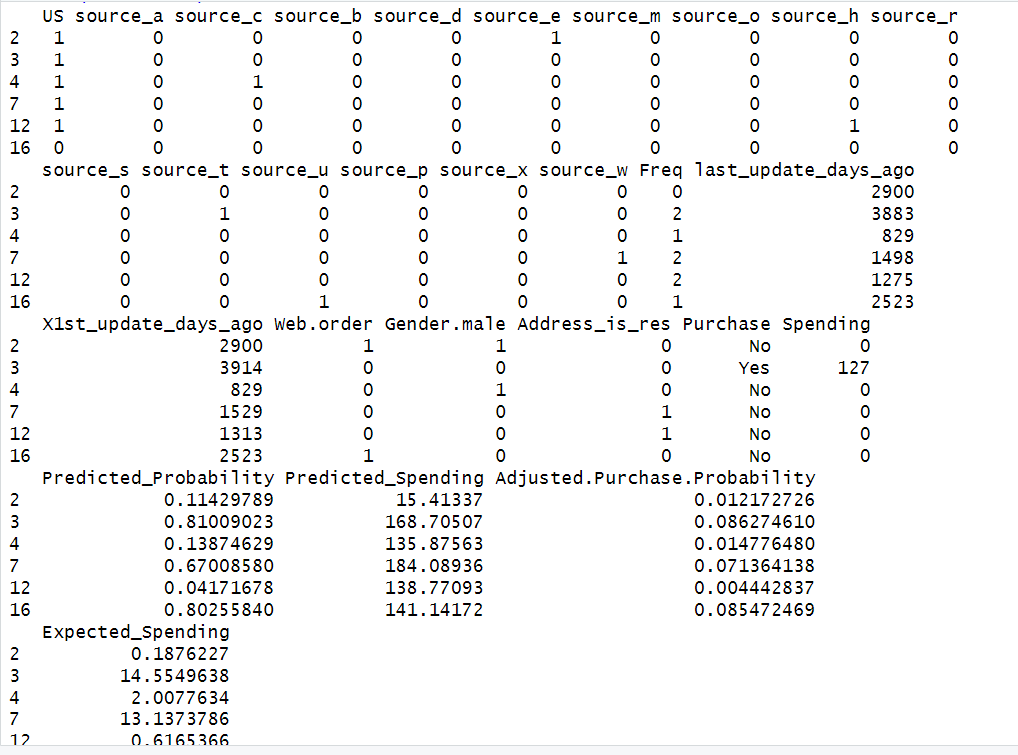
For original holdout data going to add another column that is adjusted probability by multiplying with the original purchase rate that is 0.1065. Let’s have a look on the holdout data



**Figure 8.3 Adjusted Probability in Holdout Data**

From the figure 8.3 can be able to check the adjusted probability of purchase in holdout data for each purchase values. The purpose of adding this column is correcting for oversampling and ensuring that the model's predicted probabilities are appropriately adjusted based on the original purchase rate.

Now going to add one more column to the holdout data that is Expected spending column by multiplying the adjusted purchase probability with the predicted spending value.

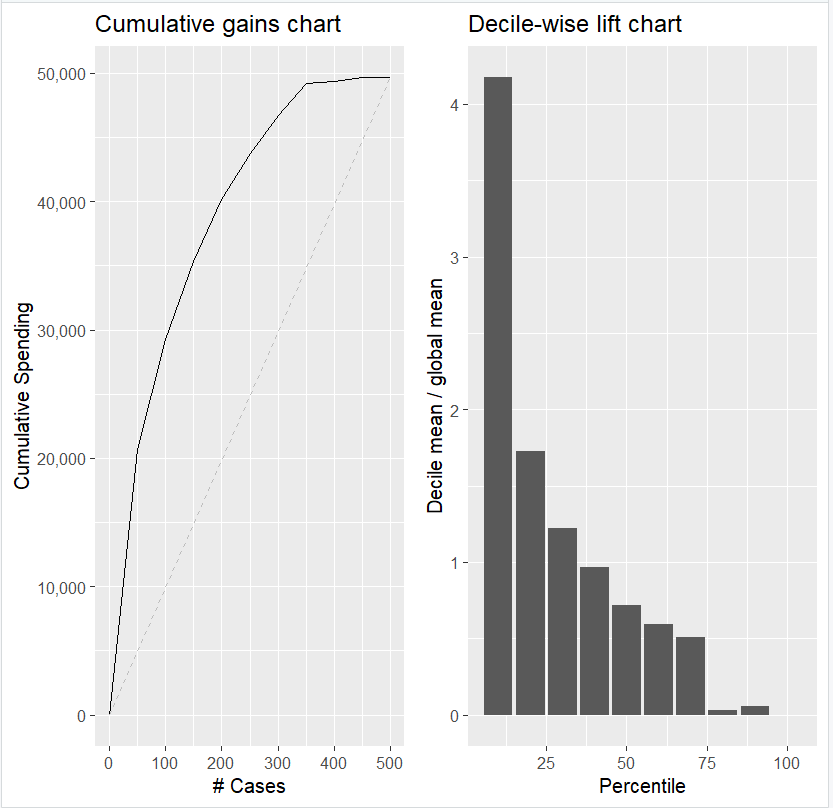


**Figure 8.4 Expected Spending in Holdout Data**

Now, by looking at above figure 8.4can be seen the expected spending values for each purchase and spending values in holdout data. By adding this we will be able to see more accurate estimation of potential spending by each customer.

**8.1 CUMULATIVE GAIN CHART FOR EXPECTED SPENDING:**

The gains and lift charts help us determine by selecting a relatively small number of records and relatively getting a larger portion of the purchasers. The cumulative chart then plots this cumulative gain expected spending column against the number of records.



**Figure 8.1.1 Cumulative gains chart and Decile-wise lift chart**

By looking at figure 8.1.1 can say that the model’s predictive performance in terms of gains is better than that of the base line model, since its cumulative gain curve is higher than that of the base line model.

The starting point of the chart represents where there are no purchasers and the there is no cumulative spending value. When considering the first 50 purchasers the cumulative spending value is $20,717. Increasing the purchasers will increases the cumulative spending value.

In decile chart the mean response is associated with each decile it calculated as the ratio of mean response to the mean spending. Here the mean response is the effectiveness of the predictive model in spending. The first decile i.e, 10% of purchasers has a mean response 4.17 times of average spending value.

Choosing the top 10% of purchasers that gave the highest spending value. We would gain 4.17 times more the amount of spending compared with the choosing all purchasers.

**9. Conclusion:**

After selecting the best classification and regression models, test the models' performance using new data to understand how well they perform and provide accurate results. By using the classification model, we can predict the potential customers based upon the probabilities and it will help us to decide towards the selecting purchasers. By using the regression model, can make a prediction on the expected spending from each individual customer, and it will help us to decide on identifying the profit on the investment. By considering the top 10% of purchasers, we observe that they exhibit 4.17 times more spending. The suggestion is that instead of selecting customers randomly and spending the amount on customers that don’t know whether they will respond or not to the email, a more effective approach is to focus on selecting the top customers by using the ranking performance for remaining customers. This approach will help to enhance the customer base, marketing strategy and maximize the profit on investment.

So, suggestion towards the business is instead of focusing on the sending the mails to all customers focus on selecting the top 10% customers will provide the better results and maximum profit on the investment.