**Maximizing Profitability of North-Point's Mailing Campaign through Predictive Customer Selection**

Rajesh Dasari Timmisetty

**Executive Summary**

North-Point is a software production company that sells games and educational products. Recently, it has updated its list of items and is preparing to introduce these new items to customers through mailing campaigns. Along with the products customers are its main other assets. To increase its customer base north-point company joined a consortium pool of listing firms. In this consortium, companies can exchange customers details. North-point company shared 200,000 customers details with the pool and received the 200,000 customers randomly. Its known fact that all customers that randomly selected from the pool won’t make a purchase. So north-point company run a test that select 20,000 customers and mailed its products to all the 20,000 customers. But only 1065 customers responded to the mail and made purchase. So, with the collected data north-point want to use the predictive analysis to select potential customers from the remaining 180,000 customer. As the available data is un balanced the company decided to work on a sample set of only 2000 customers details by oversampling the data. The oversampled data is divided into train, validation and test splits. After trying with different models, it is found that forward selection logistic model is best model for predicting the probability of the customer becoming purchaser and forward selection model is the best model for predicting the spending amount of the customer for a purchase. By using the models, the spending amount and probability is predicted, as the data is oversampled probability and predicted value is adjusted back to original state for the test partition and plotted the decile lift chart. From the decile chart it is clear that selecting top 10% high spending customers from model will gain 4.17 times profit. So North-point company can use the model to select only 10% customer and mail the products list to those selected customer which will gain more profit with less investing amount.

Table Of Contents

1. Introduction.………………………………………………………………………...5
   1. Business Background.…………………………………………………………..5
   2. Business Goal.…………………………………………………………………..5
   3. Objectives.……………………………………………………………………...6
   4. About Dataset.………………………………………………………………….6
   5. Process To Achieve Goal.………………………………………………………8
2. Data Preprocessing and Exploratory Data Analysis………………………………...8
   1. Exploratory Data Analysis.……………………………………………………..8
   2. Data Preprocessing.……………………………………………………………..43
3. Dimension Reduction.………………………………………………………………45
4. Partition the Data.…………………………………………………………………...45
5. Modelling for Classification Task.………………………………………………….46
   1. Logistic Regression.……………………………………………………………47
      1. Measuring Performance.………………………………………………..49
      2. Improving Model Performance.………………………………………...50
6. Classifier Model Selection...…………………………………………………….….55
7. Modelling for Regression Task.……………………………………………………..57
   1. Linear Regression Model.……………………………………………………...57
      1. Measuring Performing.………………………………………………….59
      2. Improving Model Performance.………………………………………...59
   2. Decision Tree for Regression (Regression Tree).………………………………60
      1. Measuring Performance.………………………………………………..62
      2. Improving Model Performance.………………………………………...62
8. Regression Model Selection.………………………………………………………..63
9. Predicting New Data.……………………………………………………………….64
10. Conclusion.………………………………………………………………………….65

1. **Introduction**
   1. **Business Background**

North-Point is a software production company that sells games and educational products. Over time, the company has expanded its offerings to include third-party titles. Recently, it has updated its list of items and is preparing to introduce these new items to customers through mailing campaigns. Apart from its products, North-Point considers its customer base as another big asset. In an attempt to expand its customer base company recently joined in consortium of listing firms that specialized in computer hardware and software products. In this consortium, companies can exchange customers details. For instance, if a company shares 1000 customers with the pool, it can receive details for 1000 customers in return.

* 1. **Business Goal**

North-Point Company shared 200,000 customer details with a pool totaling 5,000,000 names. From this pool, the company randomly selected 200,000 names. Mailing the company's product list to a customer costs $2. However, recognizing that not all customers will make a purchase, the company decided to conduct a test. They randomly selected 20,000 customers and mailed the product list, leading to an investment of $40,000 (i.e., 20,000 \* $2). The test yielded only 1,065 customers, indicating that the invested money on the remaining 18,935 customers is wasted.

The gross profit obtained from the initial 20,000 customers is the total amount spent by customers minus the amount invested by the company, equal to $205,250 - $40,000 = $165,250. If the company were to mail the remaining 180,000 customers, the investment would be $180,000 \* $2 = $360,000. The estimated spending would be the average spending of the initial 20,000 customers multiplied by 180,000, i.e., ($205,250 / 20,000) \* 180,000 = $1,847,250. The estimated gross profit for these 180,000 customers is $1,847,250 - $360,000 = $1,487,250.

To make a more profitable investment, the company decided to use predictive analysis on the data of 20,000 customers to identify potential purchasers among the remaining 180,000 customers. Due to data imbalance, with fewer purchasers compared to non-purchasers, the company recognizes the challenge in building accurate models. Understanding that accurate models can still build with a smaller dataset, the company decided to work with a sample of 2,000 customers. This sample is divided into 1,000 purchasers and 1,000 non-purchasers, achieved through oversampling.

* 1. **Objectives**

The primary objective is to utilize the sampled data to build a model that can classify or predict customers more likely to make a purchase and understand the spending patterns of those customers. The model can be used by the company to identify customers with high chances of making a purchase and mail the product list exclusively to those customers. This way, the company can avoid wasting investment on all 180,000 customers and gain more profit.

* 1. **About Dataset**

The Sample dataset created by oversampling the 20,000 customers from the pool is used for the predictive analysis. Let’s try to understand the what each column represents in the dataset.

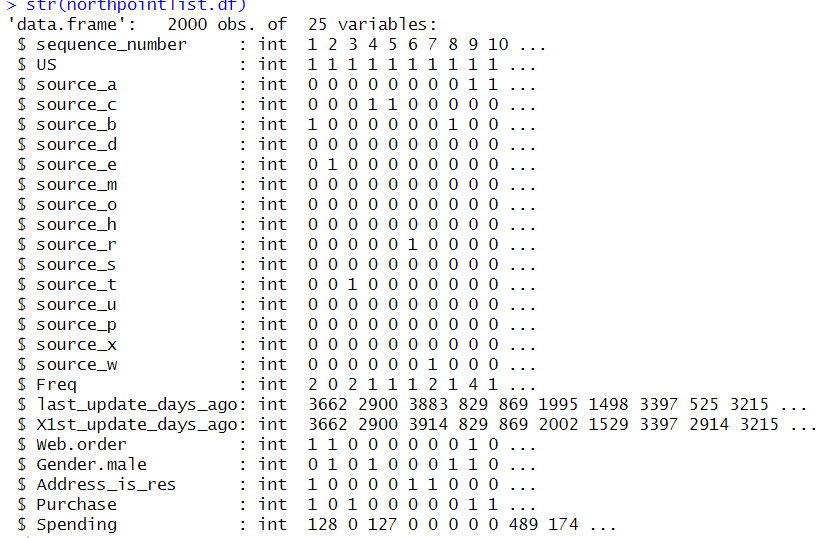
* Sequence\_number*:* A uniquely generated number that uniquely represents each row in the dataset
* US: This column represents weather the customer is located in USA or not. It is a binary column 1 represents customer in USA and 0 represents customer is not in USA
* Source\_a, source\_c, source\_b, 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 tells from which source the customer came from. All are the binary columns in which 1 represents the customer from that source and 0 represents that customer from any other sources.
* Freq: This column represents the number of purchases/transactions made by the customer. It is a numeric column.
* last\_update\_days\_ago: It indicates how many days ago the last update was made to the customer data. It is a numeric column.
* X1st\_update\_days\_ago: This column indicates how many days ago the first update was made to the customer record. It is a numeric column.
* Web order: This column represents whether the purchase was made through a web order or not. It is a binary column where 1 indicates a web order, and 0 indicates a non-web order.
* Gender = male: This column indicates whether the customer is male or not. It is a binary column where 1 represents a male customer, and 0 represents a non-male customer.
* Address\_is\_res: This column indicates whether the customer's address is residential or not. It is a binary column where 1 represents a residential address, and 0 represents a non-residential address.
* Purchase: This column indicates whether the customer made a purchase or not. It is a binary column where 1 represents a customer who made a purchase, and 0 represents a customer who did not make any purchases.
* Spending: This column represents the amount the customer spent in a purchase. It is a numeric column.
  1. **Process to Achieve Goal**

There are two outcome columns in the dataset purchase and spending. First, build a classifier machine learning model to predict the probability of customers becoming purchaser. After that, regression models are used to predict Spending amount. Based on the probability and spending value, the company can decide to which customers the products information should be mailed in order to maximize profits.

1. **Data Preprocessing and Exploratory Data Analysis**
   1. **Exploratory Data Analysis**

In the section, each predictor will be analyzed, and its impact with the target column is also analyzed.

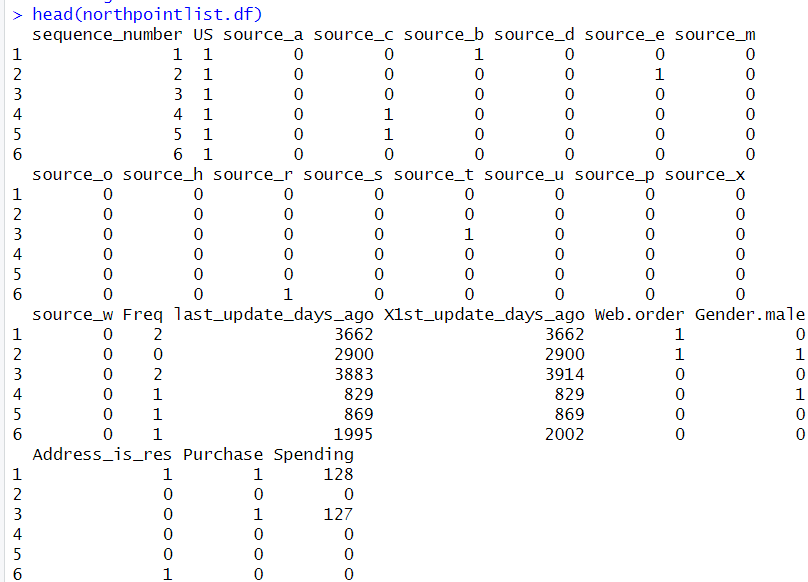
In the first step of EDA let’s try to explore the structure of the data.



**Figure 2.1 Columns structure in the dataset**

From the figure 2.1, it is evident that all columns are of integer datatype. Null values in the dataset can be identified using the function sum(is.na ()), and found that there are no null values present in the dataset.

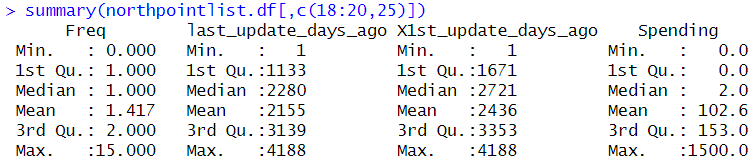
Below figure shows the first 6 rows of the dataset.



**Figure 2.2 First 6 rows of the dataset**

**Summary Statistics Analysis**

Exploring the summary statistics for all numerical columns

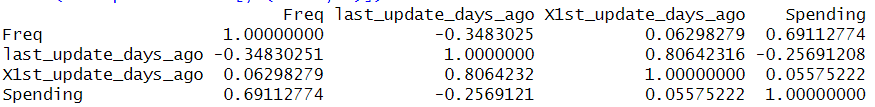


**Figure 2.3 Summary statistics for numerical columns**

From the figure 2.3 it is clear that the different variables have very different range of values.

**Correlation Anlysis**

Using the cor() function in R, correlation of all numerical columns can be obtained.

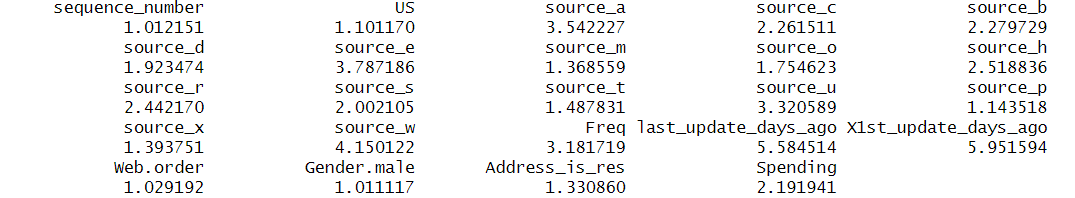


**Figure 2.4 Correlation for all numerical columns**

Figure 2.4 presents the correlation for all numerical columns. From the correlation matrix, it can be concluded that columns last\_updated\_days\_ago and x1st\_updated\_days\_ago are highly positively correlated, indicating that these two columns contain a lot of overlapping information. Only frequency column is strongly correlated with the target column spending, suggesting that freq might be important for predicting spending for a customer.

**Multi-collinearity Analysis**

By using the variance inflation factor (VIF), multicollinearity of the columns can be identified.



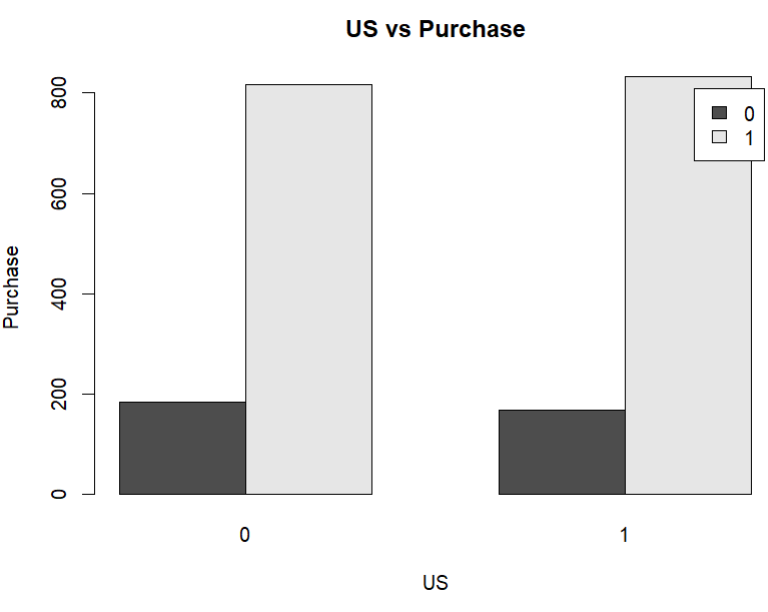
**Figure 2.5 variance inflation factor**

From Figure 2.5, it can be observed that last\_update\_days\_ago and x1st\_update\_days\_ago have potential multicollinearity, as the VIF value is > 5.

From now on let’s analyze how each predictors impact target columns.

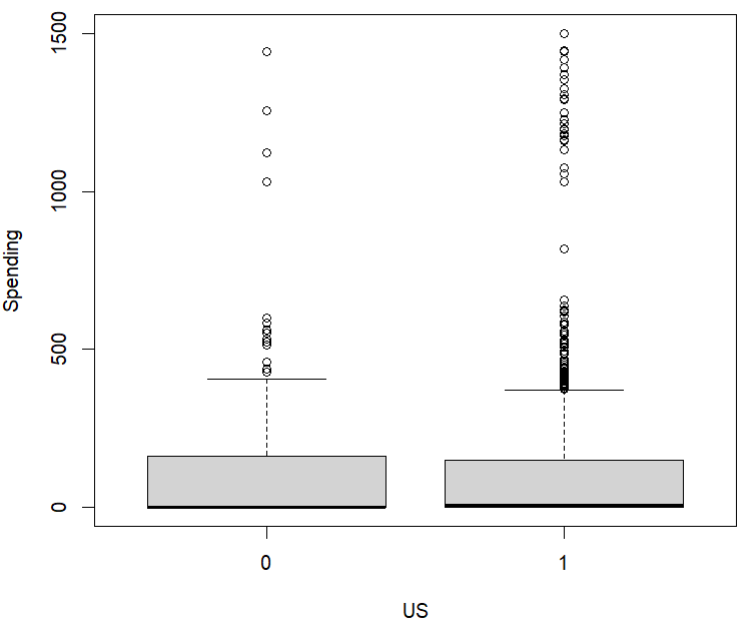
**US Column**

US Column represents weather the customer is located in USA or not. Now let’s analyze how the US column impacting the target columns purchase and spending.



**Figure 2. 6 US vs Purchase**

From Figure 2.6, it is evident that the count of purchasers from the U.S. and non-U.S. is the same.

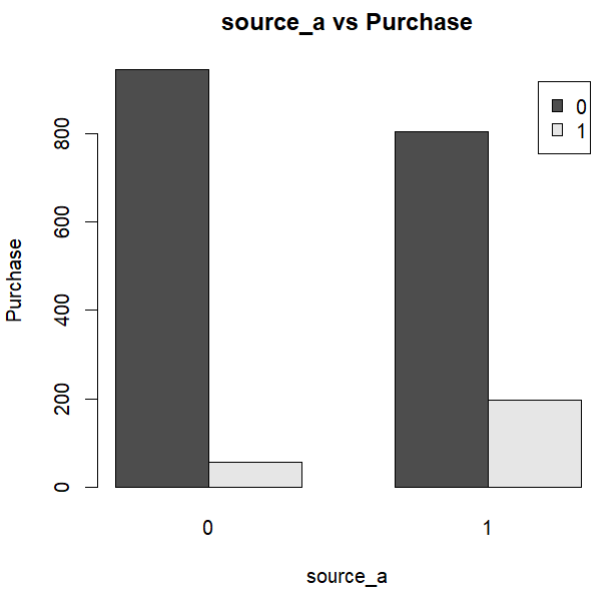


**Figure 2.7 US vs Spending**

From Figure 2.7, it can be inferred that US customers are spending more than non-US customers. The maximum amount spent by a US customer is $1500. The average amount spent by non US customers is more than that by US customers. Most of the customers from US spent between 0$ to 500$.

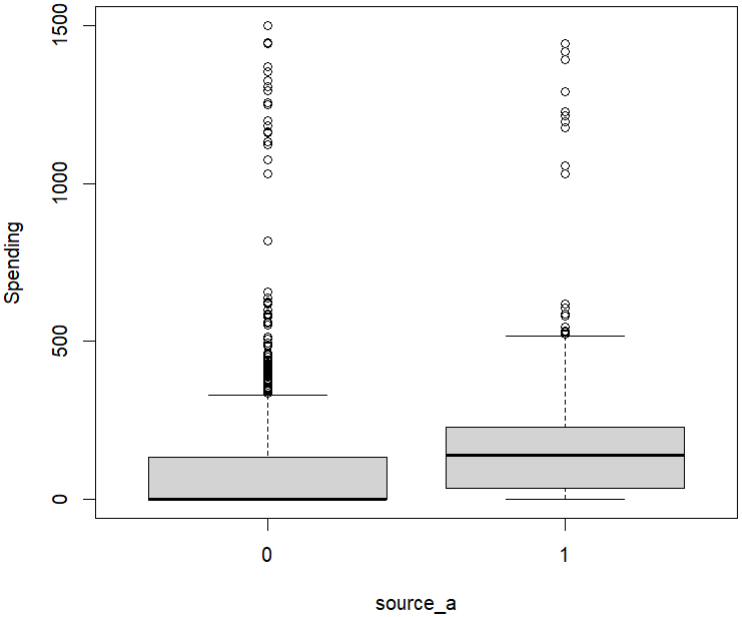
**Source\_a Column**

Source\_a Column tells if the customer is from source a or not. Lets analyze how the source a customers impacting the target columns.



**Figure 2.8 Source\_a vs Purchase**

From Figure 2.8, it can be observed that approximately 200 customers from source A are purchasers.

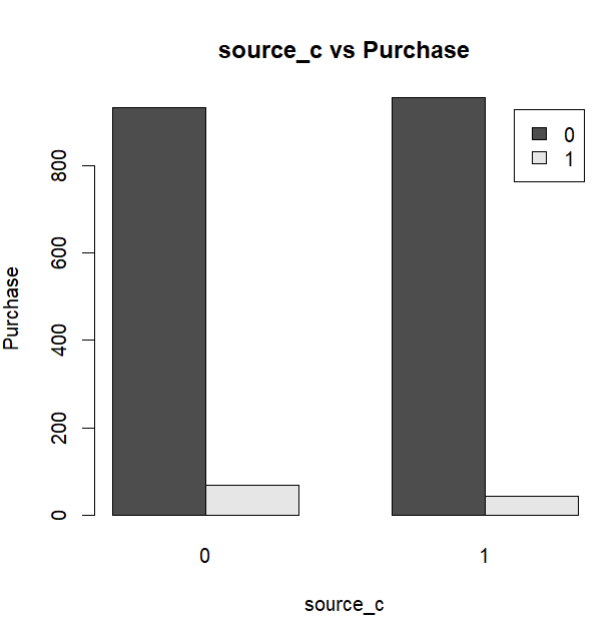


**Figure 2.9 Source\_a vs Spending**

from Figure 2.9, it's evident that the most of the customers from source\_a spent between 0$ to 600$

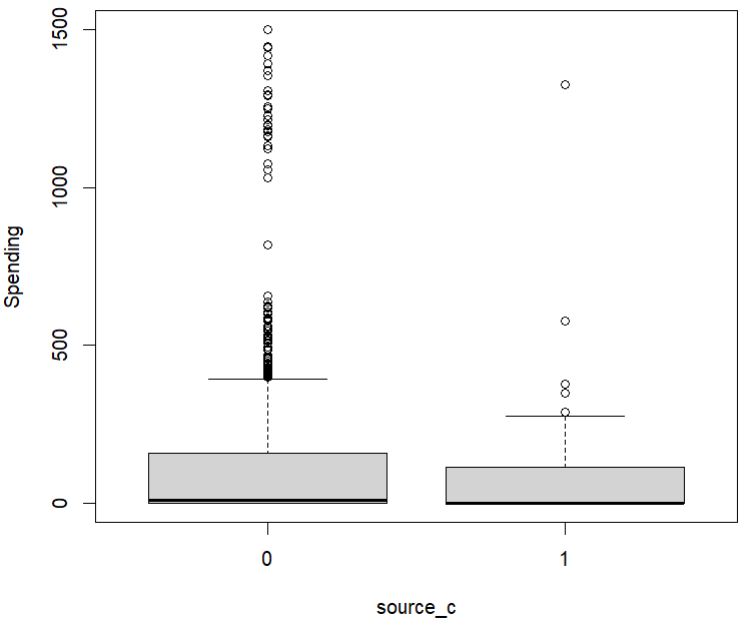
**Source\_c**

Source\_c Column tells if the customer is from source c or not. Lets analyze how the source c customers impacting the target columns.

****

**Figure 2.10 Source\_c vs Purchase**

From figure 2.10, it can be observed that approximately 44 customers from source c are purchasers

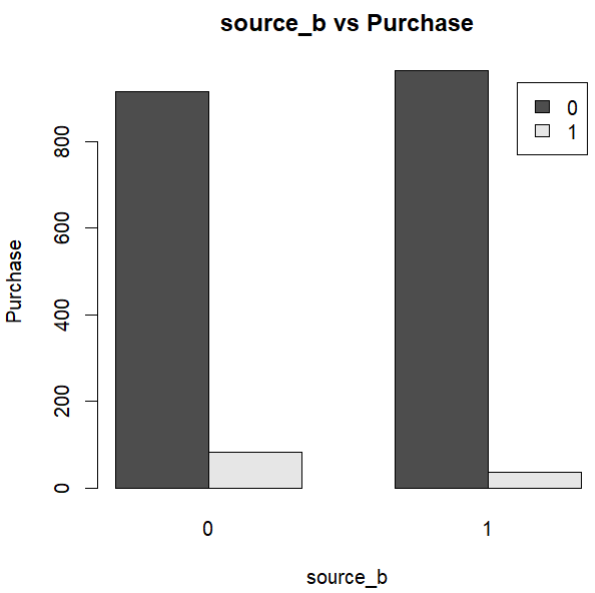


**Figure 2.11 Source\_c vs Spending**

From Figure 2.11, the maximum amount spent by Source\_C customers is approximately $1300. Additionally, the average amount spent by customers from Source\_C is approximately $200. Most of the customers from source\_c spent between o$ to 300$.

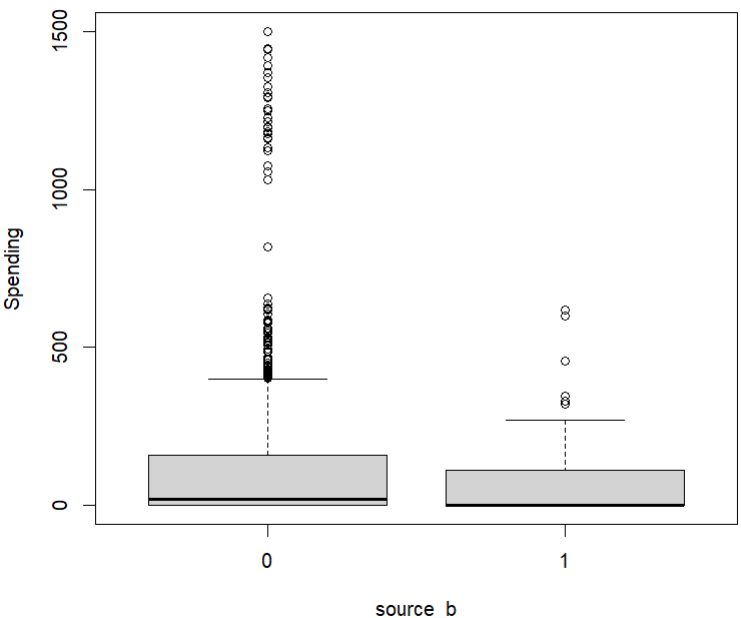
**Source\_b column**

Source\_b Column tells if the customer is from source b or not. Lets analyze how the source b customers impacting the target columns.

****

**Figure 2.12 Source\_b vs Purchase**

From Figure 2.12, it can be observed that approximately 36 customers from source b are purchasers

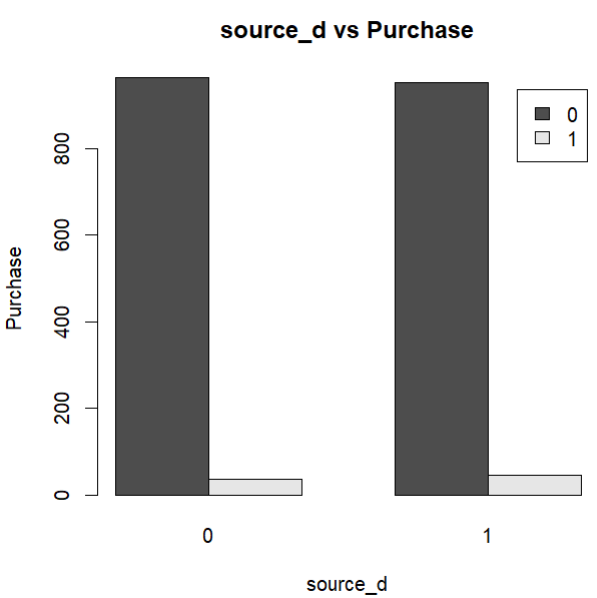


**Figure 2.13 Spending vs source\_b**

From Figure 2.13, it is evident that the maximum spending by customers originating from Source\_b is less than $1000. Most of the customers from source\_b spent between 0$ to 200$

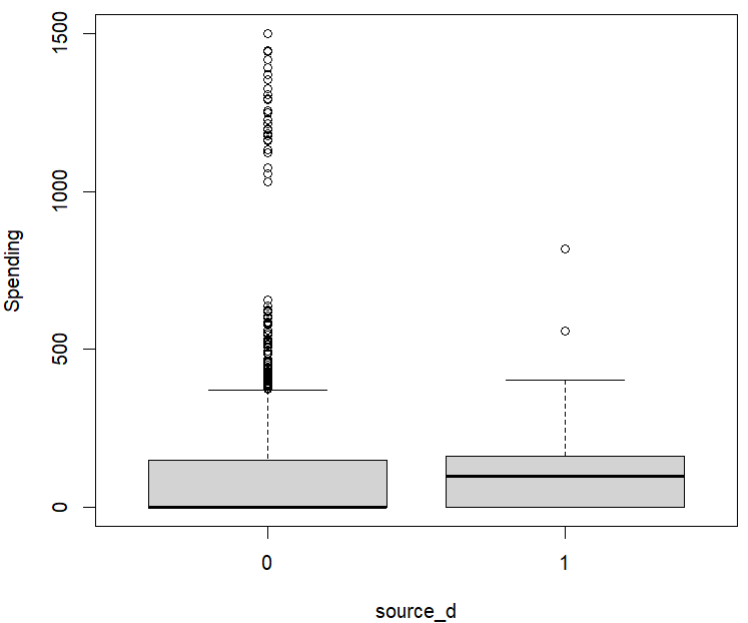
**Source\_d**

Source\_d Column tells if the customer is from source d or not. Lets analyze how the source d customers impacting the target columns.

****

**Figure 2.14 Source\_d vs Purchase**

From figure 2.14, it can be observed that approximately 47 customers from source d are purchasers

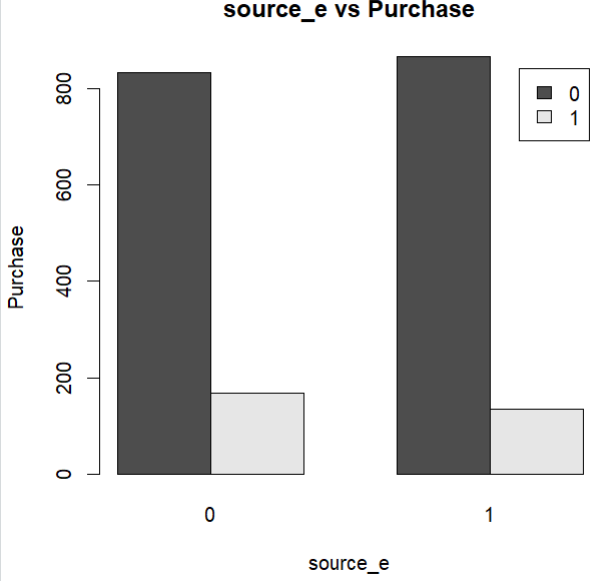
****

**Figure 2.15 source\_d vs Spending**

From the figure 2.15 the maximum spending by customers originating from source\_d is less than $500. The average amount spends by the customers from source\_d is approximately 300$ which is more than all other customers came from different sources.

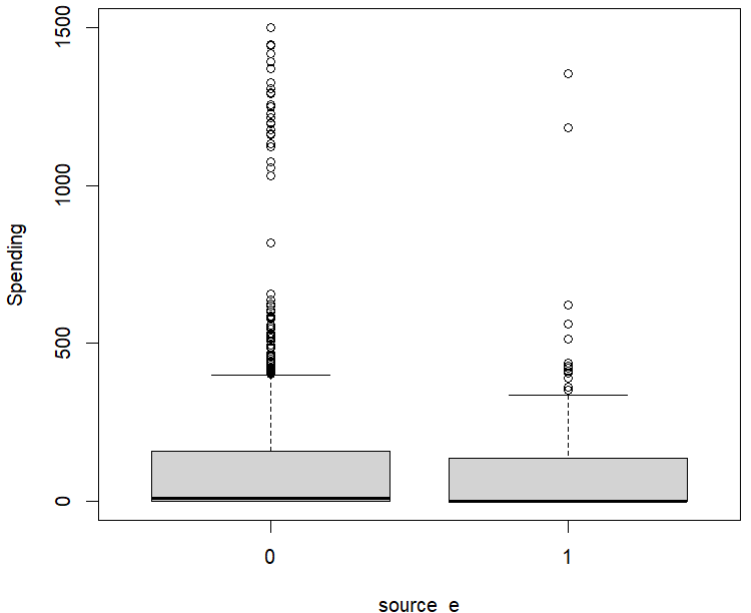
**Source\_e column**

Source\_e Column tells if the customer is from source e or not. Lets analyze how the source e customers impacting the target columns.

****

**Figure 2.16 Source\_e vs Purchase**

From figure 2.16, it can be observed that approximately 134 customers from source e are purchasers

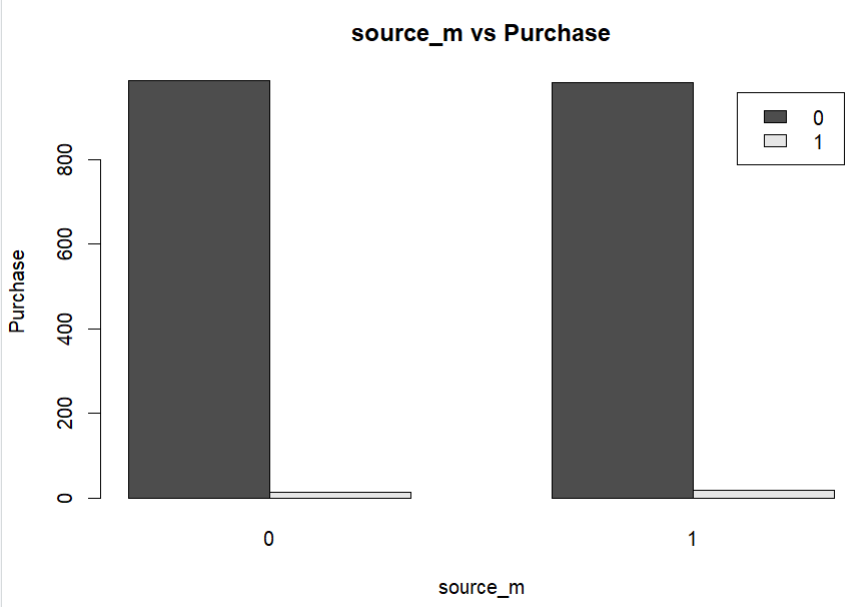


**Figure 2.17 Source\_e vs Spending**

From the figure 2.17 the maximum spending by customers originating from source\_e is less than 1500$ and the average amount spent by the customers is approximately 200$. Most of the customers from source\_e spent between 0 to 200$.

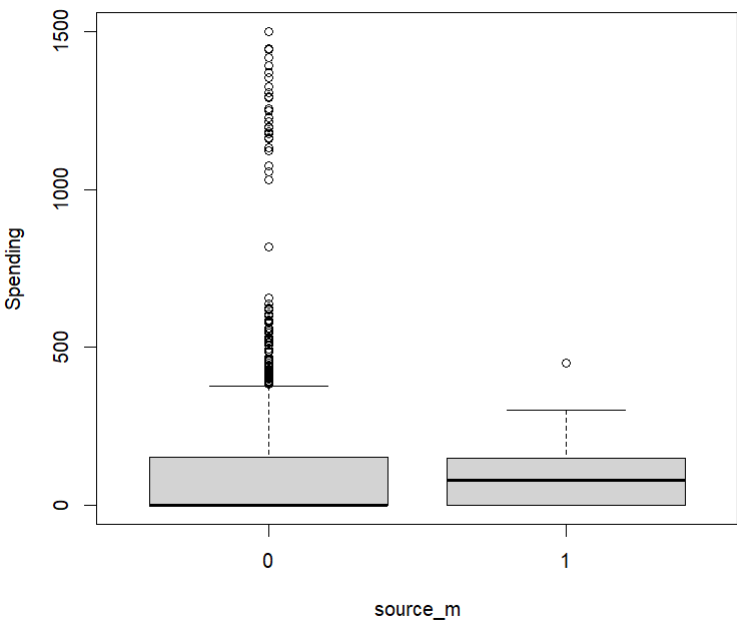
**Source\_m**

Source\_m Column tells if the customer is from source m or not. Lets analyze how the source m customers impacting the target columns**.**

****

**Figure 2.18 Source\_m vs Purchase**

From figure 2.18, it can be observed that approximately 19 customers from source m are purchasers.

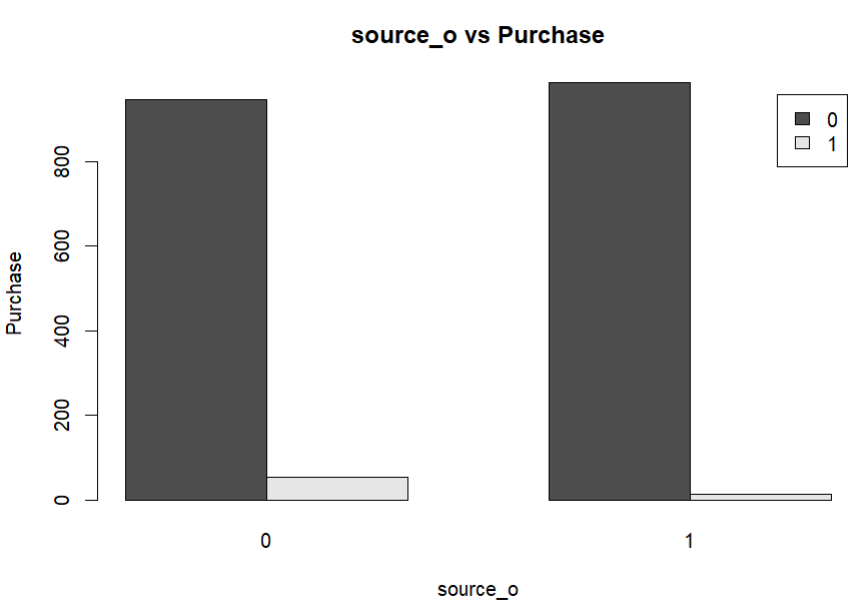
****

**Figure 2.19 Source\_m vs spending**

From figure 2.19 it is clearly that the maximum spending by customers originating from source\_m is under $500, with an average spending of around $300.Most of the source\_m customers spent between 0 to 400$.

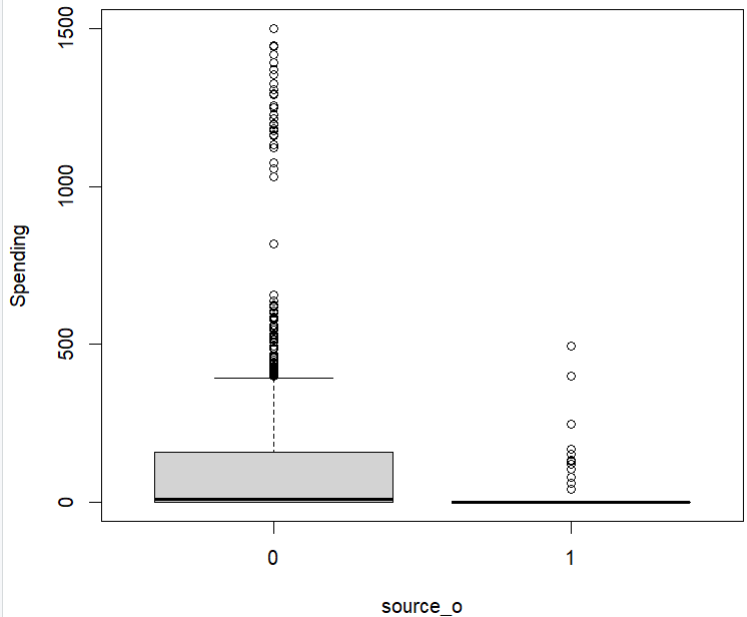
**Source\_o**

Source\_o Column tells if the customer is from source o or not. Lets analyze how the source o customers impacting the target columns.

****

**Figure 2.20 Source\_o vs Purchase**

From figure 2.20, it can be observed that approximately 30 customers from source o are purchasers.

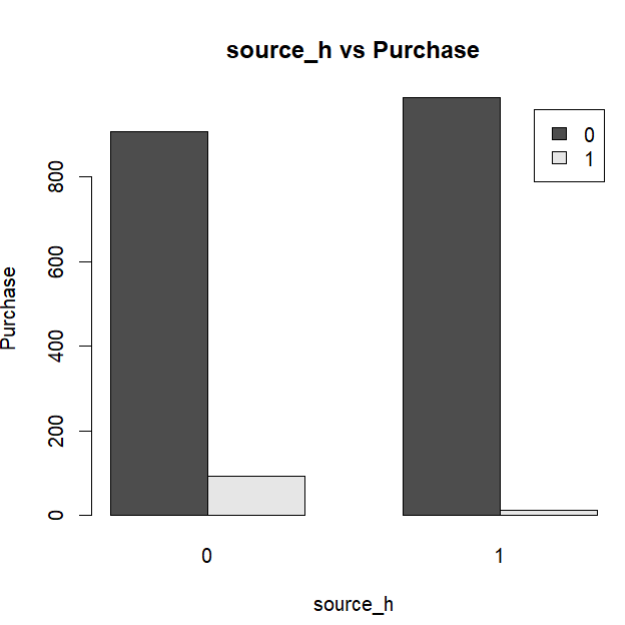
****

**Figure 2.21 source\_o vs Spending**

From the figure 2.21 it is clear that the maximum amount spent by source\_o customers is less than 1000$. The average amount spend from source\_o customers is less than 100$. Most of the customers from source\_o spend less than 100$.

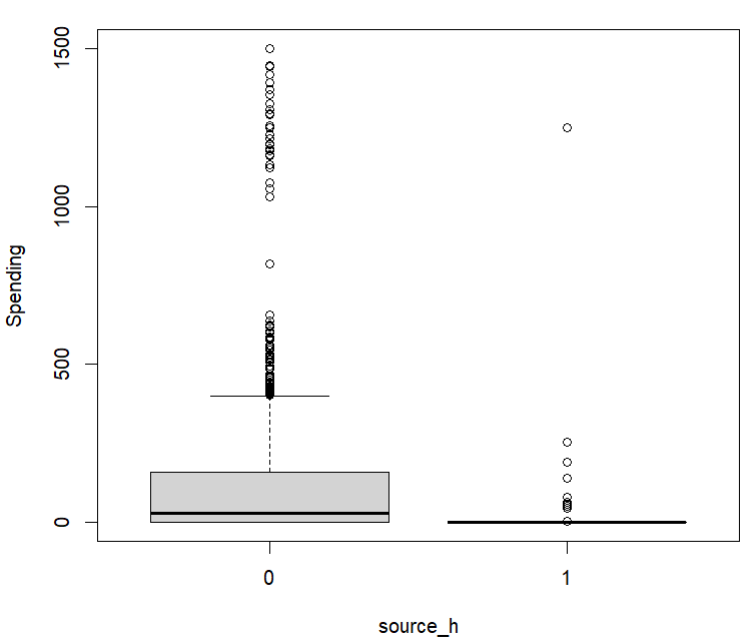
**Source\_h**

Source\_h Column tells if the customer is from source h or not. Lets analyze how the source h customers impacting the target columns.

****

**Figure 2.22 Source\_h vs Purchase**

From figure 2.22, it can be observed that approximately 12 customers from source h are purchasers

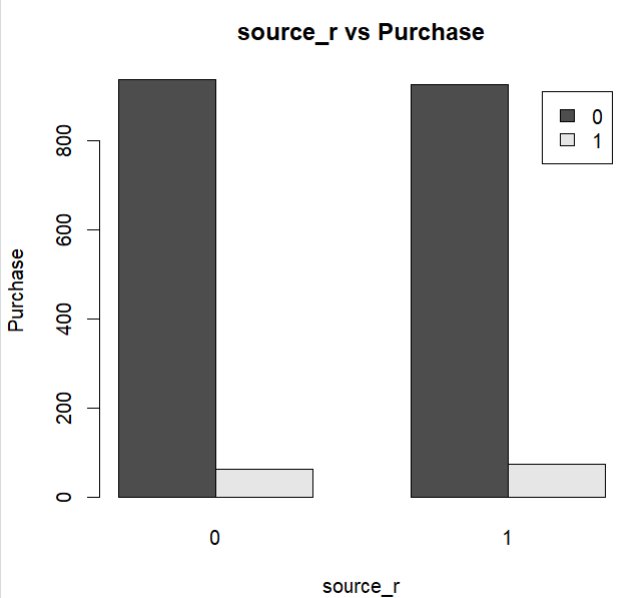


**Figure 2.23 Source\_h vs Purchase**

From figure 2.23 The maximum amount spent by source\_h customers is less than 1500$. The average amount spent by the source\_h customers are less than 50$. Most of the customers from source\_h spent between 0$ to 100$.

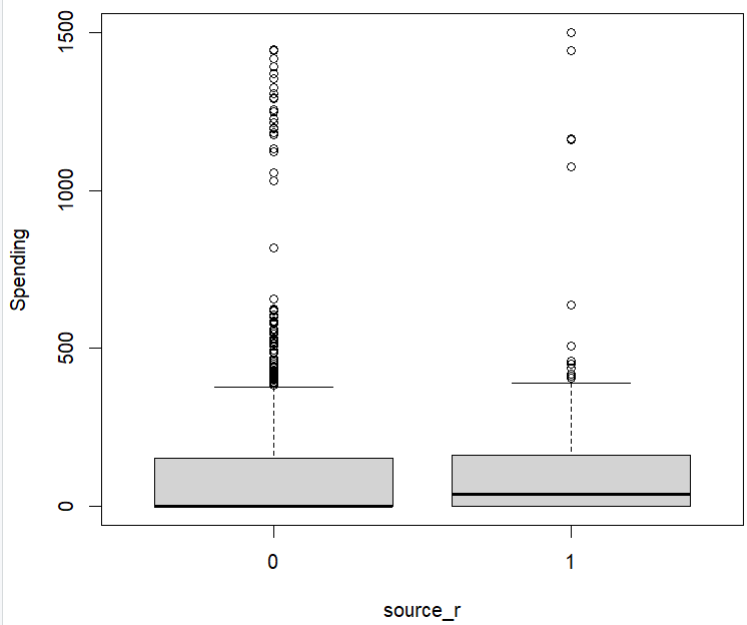
**Source\_r**

Source\_r Column tells if the customer is from source r or not. Lets analyze how the source r customers impacting the target columns.

****

**Figure 2.24 Source\_r vs Purchase**

From figure 2.24, it can be observed that approximately 90 customers from source r are purchasers

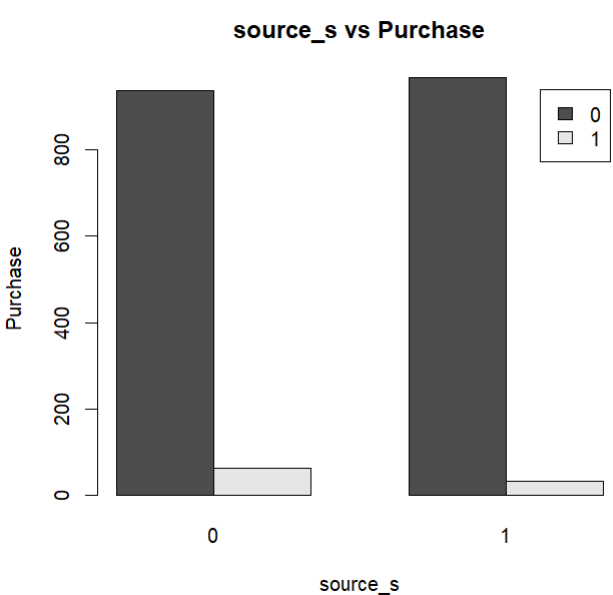
****

**Figure 2.25 Source\_r vs Spending**

From the figure 2.25 it is clear that the maximum amount spent by the source\_r customers is 1500$ and Average amount spent by source\_r customers are 200$. Most of the source\_r customers spent between 0 to 400$.

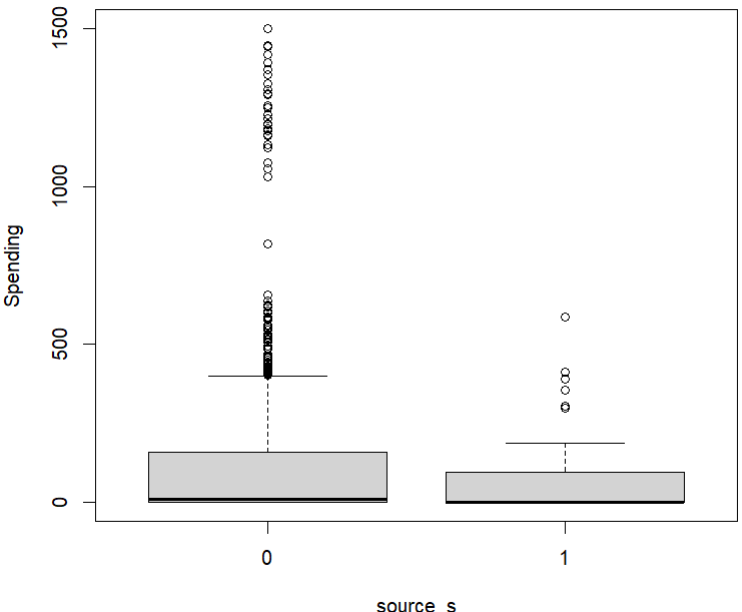
**Source\_s**

Source\_s Column tells if the customer is from source s or not. Lets analyze how the source s customers impacting the target columns.

****

**Figure 2.26 Source\_s vs Purchase**

From figure 2.26, it can be observed that approximately 32 customers from source s are purchasers.

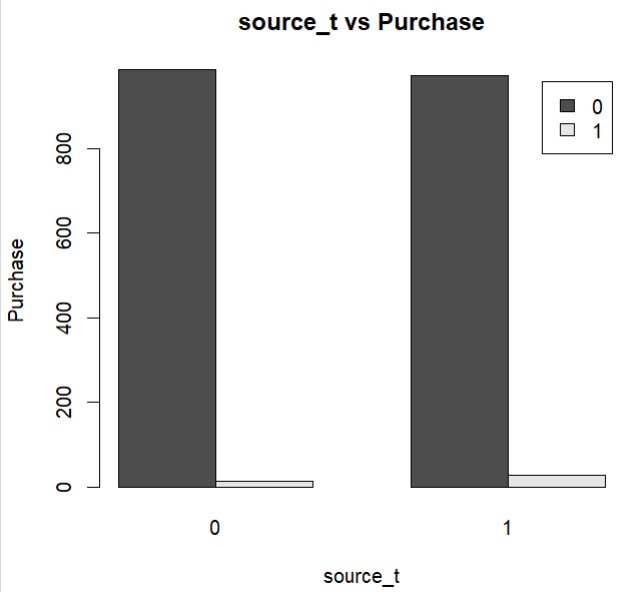
****

**Figure 2.27 source\_s vs Spending**

From the figure 2.27 it is clear that the maximum amount spent by the source\_s customers is less than 1000$. The average amount spent by source\_s customers approximately 200$. Most of the source\_s customers spent between 0$ to 300$

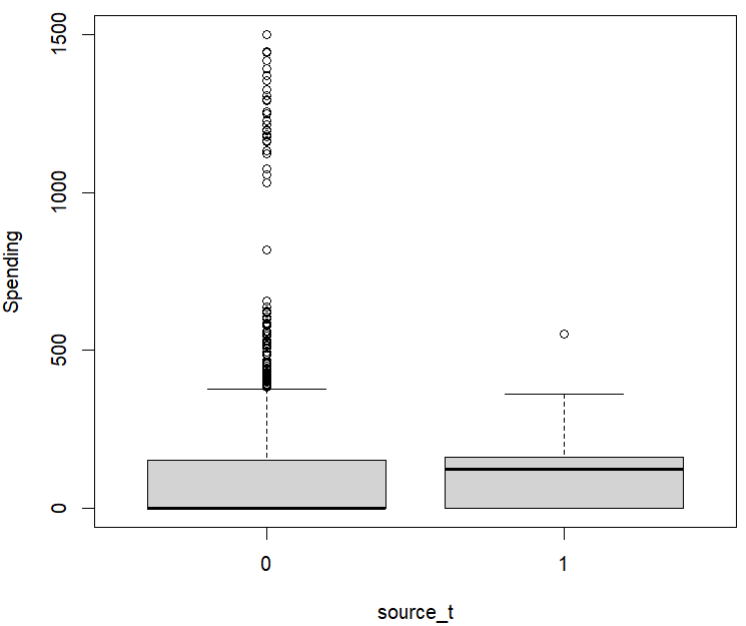
**Source\_t**

Source\_t Column tells if the customer is from source t or not. Lets analyze how the source t customers impacting the target columns.



**Figure 2.28 Source\_t vs Purchase**

From figure 2.28, it can be observed that approximately 29 customers from source t are purchasers.

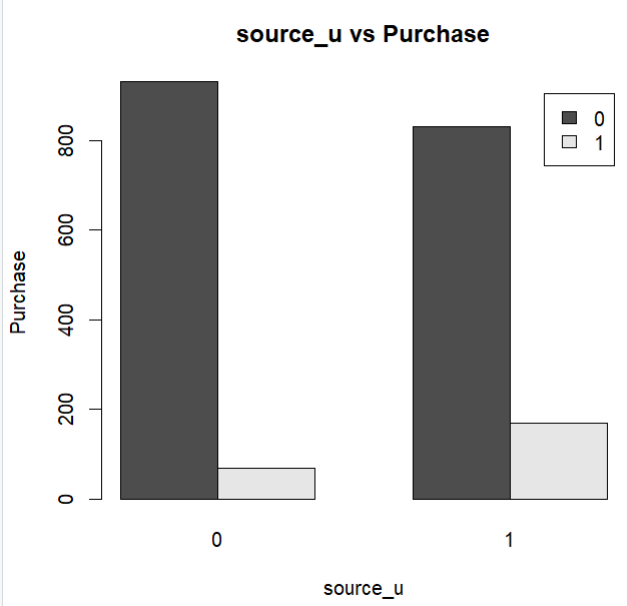
****

**Figure 2.29 source\_t vs Spending**

From the figure 2.29 it is clear that the maximum amount spent by customers from source\_t is less than 800$. The average amount spent by the source\_t customers is approximately 300$. Most of the source\_t customers spent between 0$ to 600$

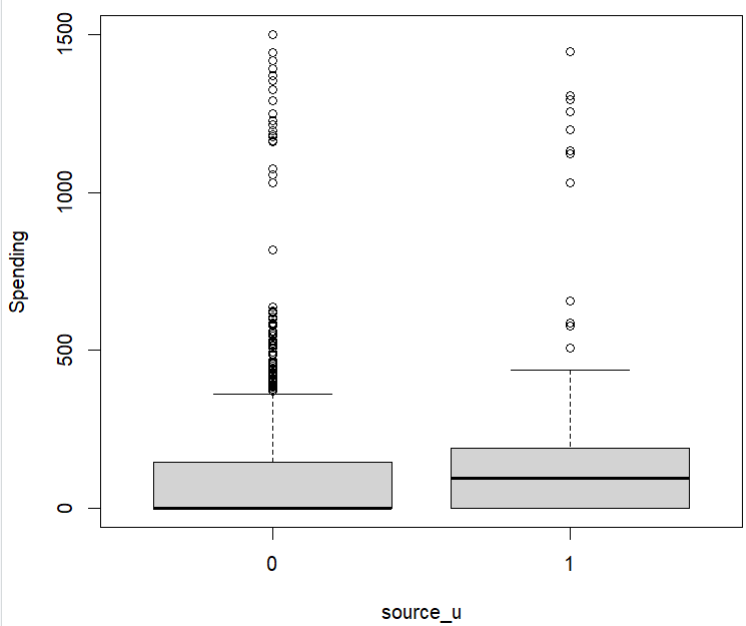
**Source\_u**

Source\_u Column tells if the customer is from source u or not. Lets analyze how the source u customers impacting the target columns.

****

**Figure 2.30 Source\_u vs Purchas**

From figure 2.30, it can be observed that approximately 169 customers from source u are purchasers

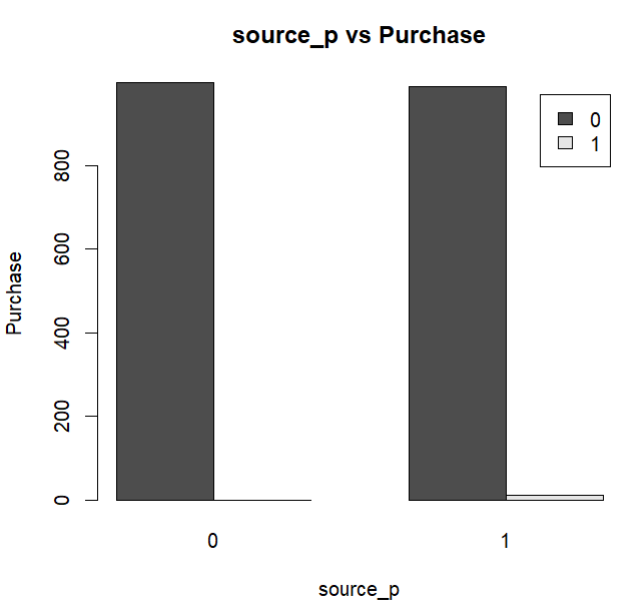
****

**Figure 2.31 Source\_u vs Spending**

From the figure 2.31 it is clear that the maximum amount spent by source\_u customers is approximately 1400$. The average amount spent is $300. Most of the source\_u customers spent between 0$ to 400$.

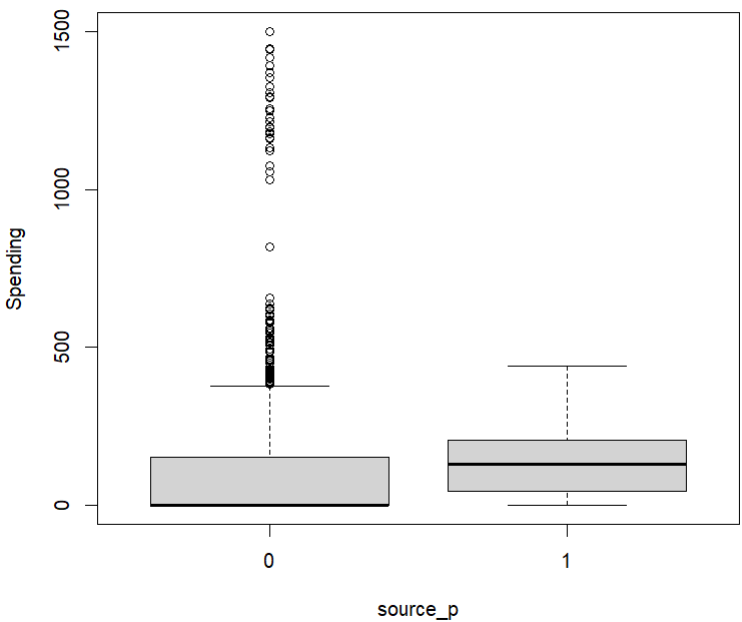
**Source\_p**

Source\_p Column tells if the customer is from source p or not. Lets analyze how the source p customers impacting the target columns.



**Figure 2.32 Source\_p vs Purchase**

From figure 2.23, it can be observed that approximately 11 customers from source p are purchasers

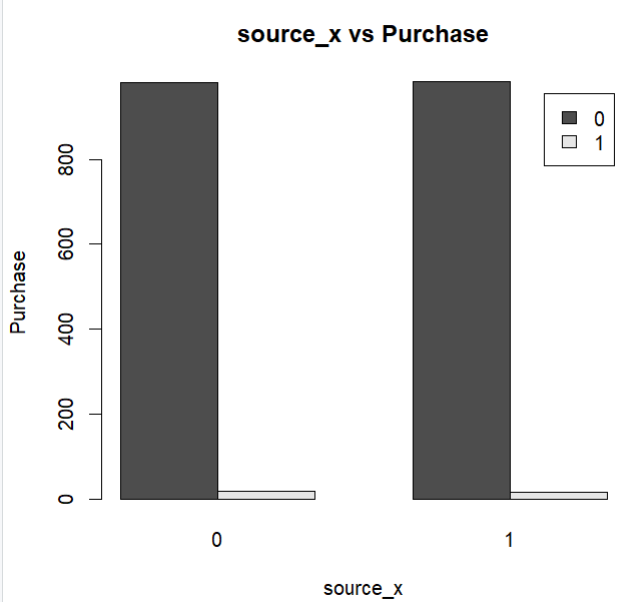
****

**Figure 2.24 source\_p vs Spending**

From the figure 2.24 it is clear that the maximum amount spent by the source\_p customer is below 400$ and the average amount spent is approximately 300$. Most of the source\_p customers spent between 0$ to 400$.

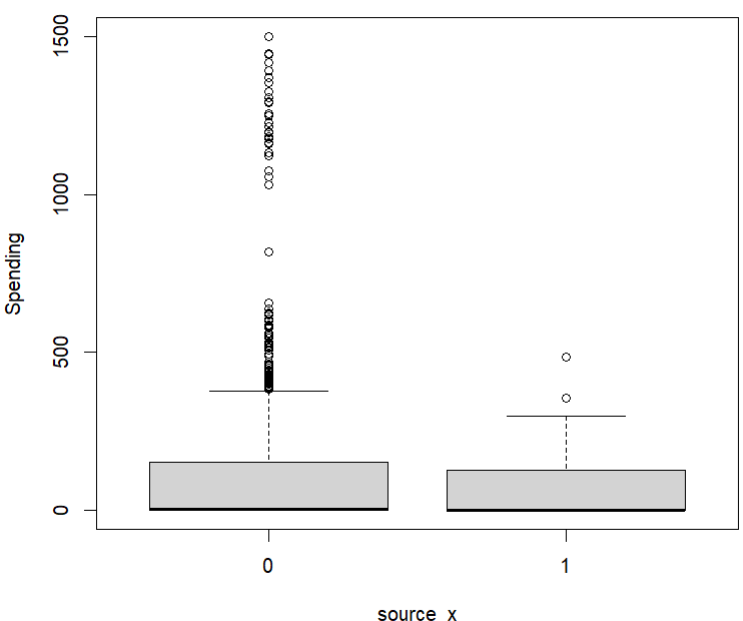
**Source\_x**

Source\_x Column tells if the customer is from source x or not. Lets analyze how the source x customers impacting the target columns.



**Figure 2.25 Source\_x vs Purchas**

From figure 2.25, it can be observed that approximately 17 customers from source x are purchasers

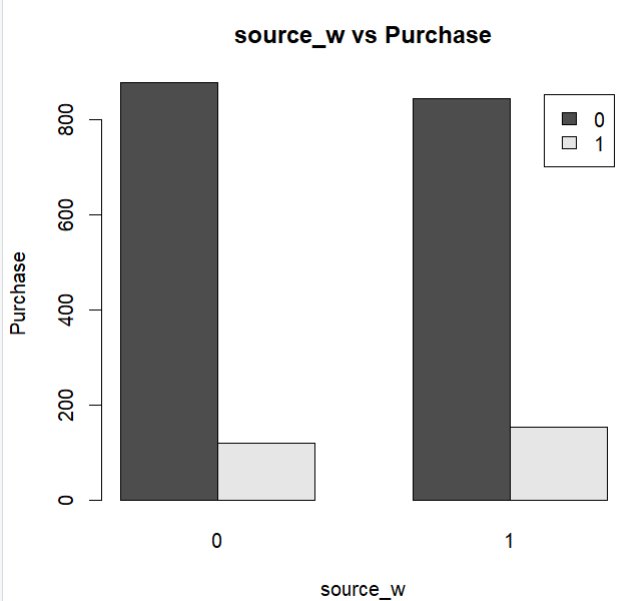
****

**Figure 2.26 source\_x vs Spending**

From the figure 2.26 it is clear that the maximum amount spent by source\_x customers is 600$. The average spending is approximately 300$ and most of the customers spent between 0 to 400$.

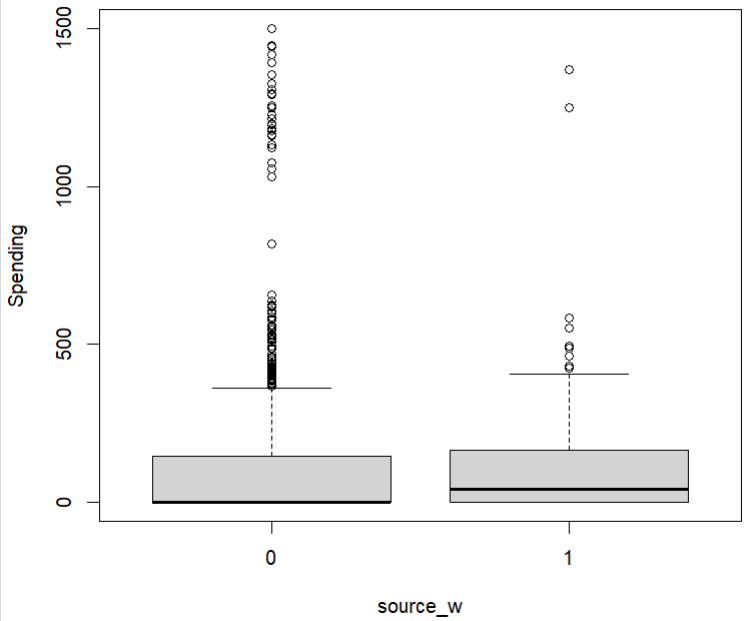
**Source\_w**

Source\_w Column tells if the customer is from source w or not. Let’s analyze how the source w customers impacting the target columns.



**Figure 2.27 Source\_w vs Purchase**

From figure 2.27, it can be observed that approximately 154 customers from source w are purchasers

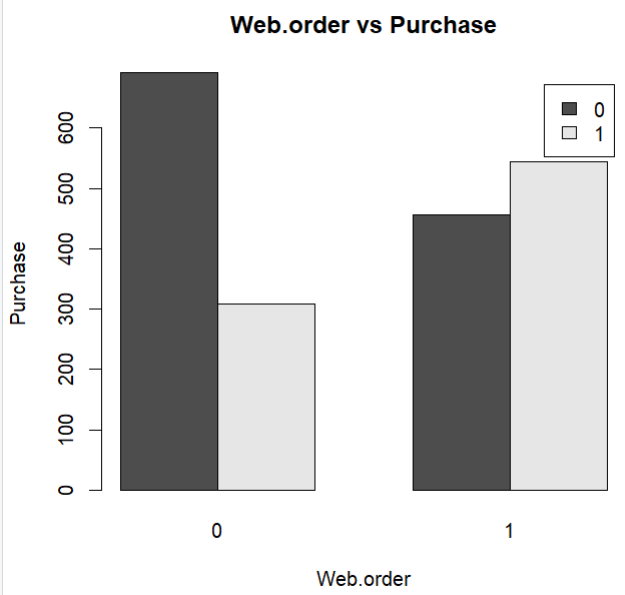
****

**Figure 2.28 source\_w vs Spending**

From the figure 2.28 it is clear that the maximum amount spent by source\_w customers is less than 1500$ and average amount spent is approximately 300$. Most of the customers spent between 0 to 499$.

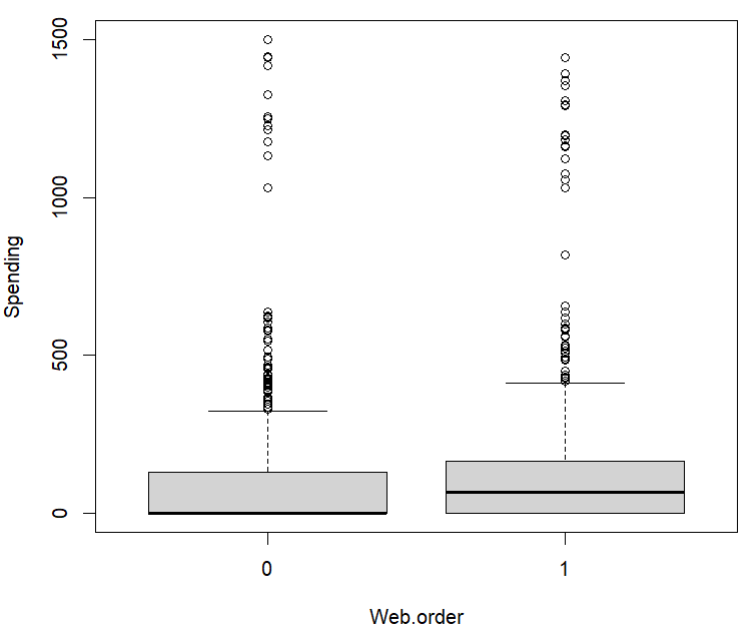
**Web.Order**

Web.Order Column tells whether the purchase was made through a web order or not. Now lets analyze how the column is impacting the target columns



**Figure 2.29 web.order vs Purchase**

From the figure 2.29 it is clear that there are 852 customers who made purchase from online and 1148 customers who made purchase on other ways. Out of 1148 customers 692 customers are non-purchasers and 456 customers are purchaser. Out of 852 customers 308 customers are non-purchasers and 544 customers are purchasers.

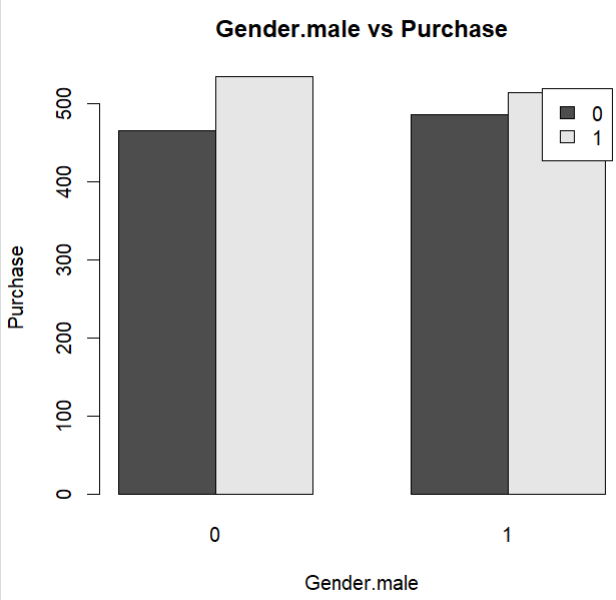
****

**Figure 2.30 web.order vs Spending**

The maximum spending by customers who made web orders is less than $1500, while for customers who made non-web orders, it's exactly $1500. Additionally, the average spending by customers who made web orders is lower than the average spending by customers who made non-web orders. Most customers who made web orders spent between $0 and $400, while customers who made purchases through non-web orders typically spent between $0 and $300.

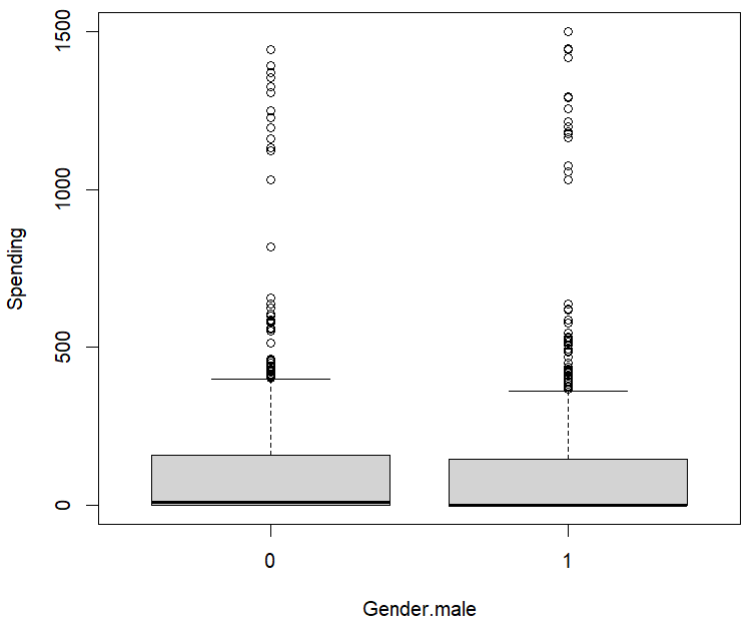
**Gener.male column**

Gener.male Column tells whether the customer is male or not.



**Figure 2.31 gener.male vs Purchase**

From the figure 2.31 it is clear that there are 1049 male customers and 951 non male customers. Out of 951 customers 465 are non-purchasers and 486 are purchasers. Out of 1049 customers 535 are non-purchasers and 514 are purchasers.

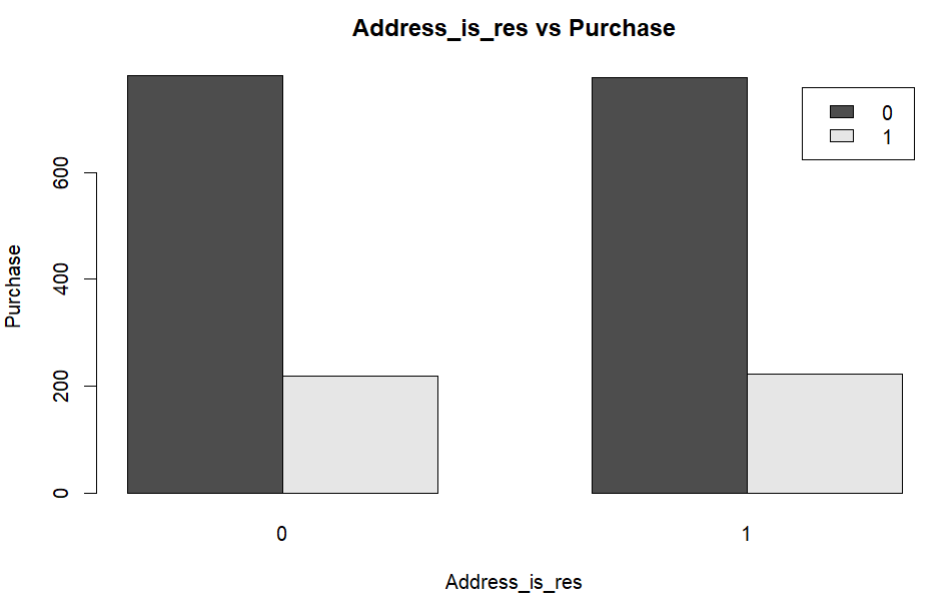
****

**Figure 2.32 gender.male vs spending**

The maximum spending by male customers is $1500, while the maximum spending by non-male customers is less than $1500. Additionally, the average spending by male customers is lower than the average spending by non-male customers. Most male and non-male customers spent less than $500.

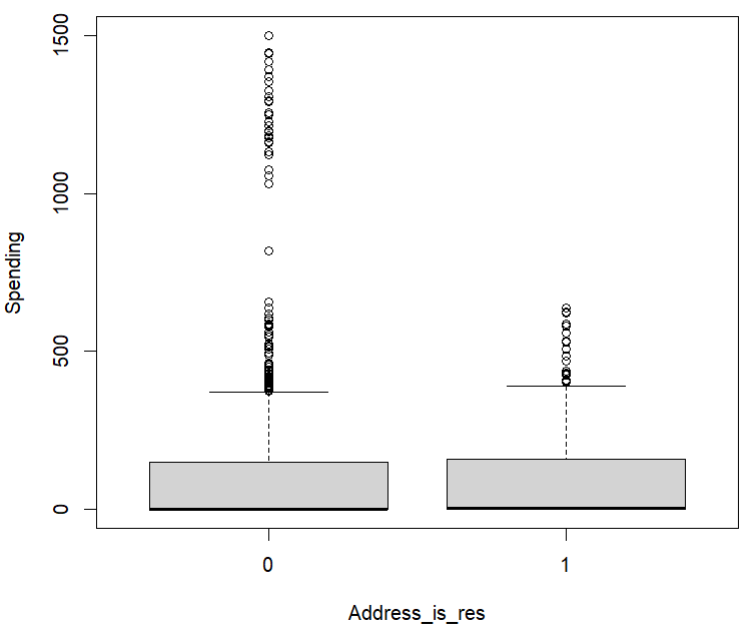
**Address\_is\_res column**

Address\_is\_res column tellswhether the customer's address is residential or not

****

**Figure 2.33 Address\_is\_res vs Purchase**

From the figure 2.33 it is clear that there are 1558 customers whose address is not residential and 442 customers address is residential. Out of 1558 customers 781 customers are non-purchaser and 777 customers are purchaser. Out of 442 residential customers 219 are non-purchasers and 223 are purchasers.

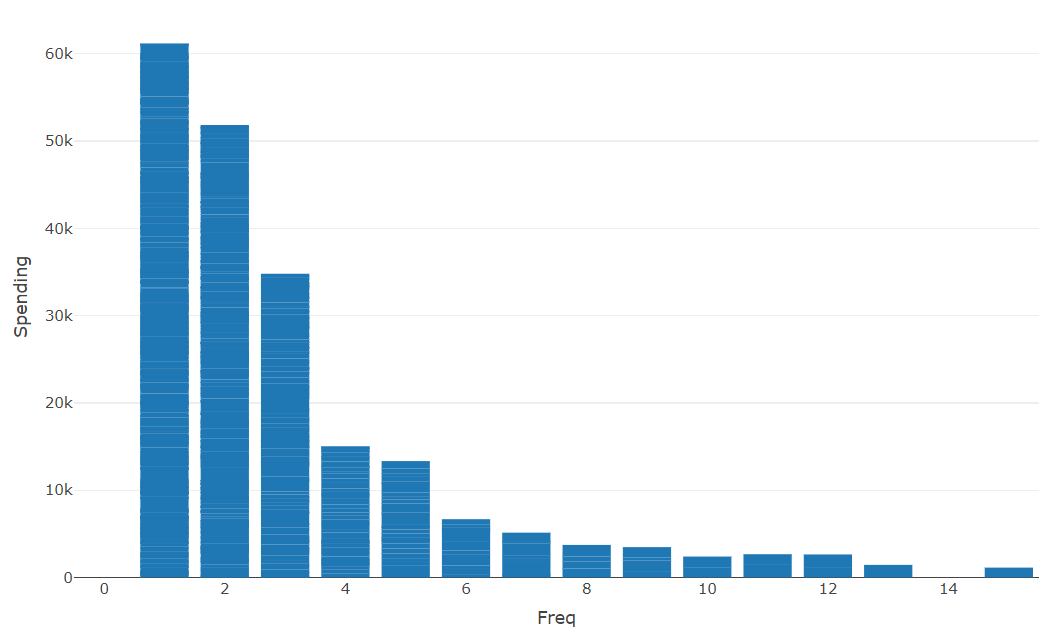
****

**Figure 2.34 address\_is\_res vs Spending**

From the figure 2.34 it is clear that for non-residential customers, the maximum spending is $1500, while for residential customers, it's $700. Additionally, the average spending by residential customers is less than that of non-residential customers. Most residential customers spent between $0 to $300, and the maximum number of residential customers spent between $0 to $400.

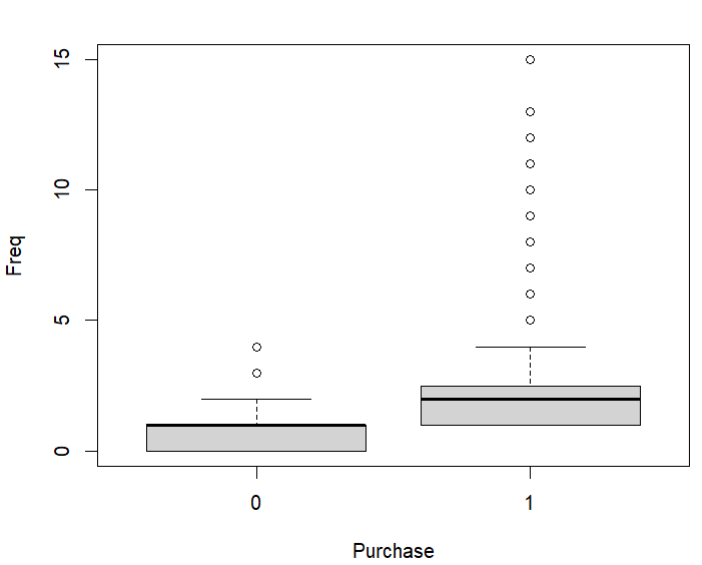
**Freq Column**

Freq column tells what is the number of purchases/transactions made by the customer. Lets analyze how the freq column impacting the target columns.

****

**Figure 2.35 freq vs Spending**

From the figure 2.35 it is clear that as the frequency of transactions increase the amount spending on purchasers is decreasing.

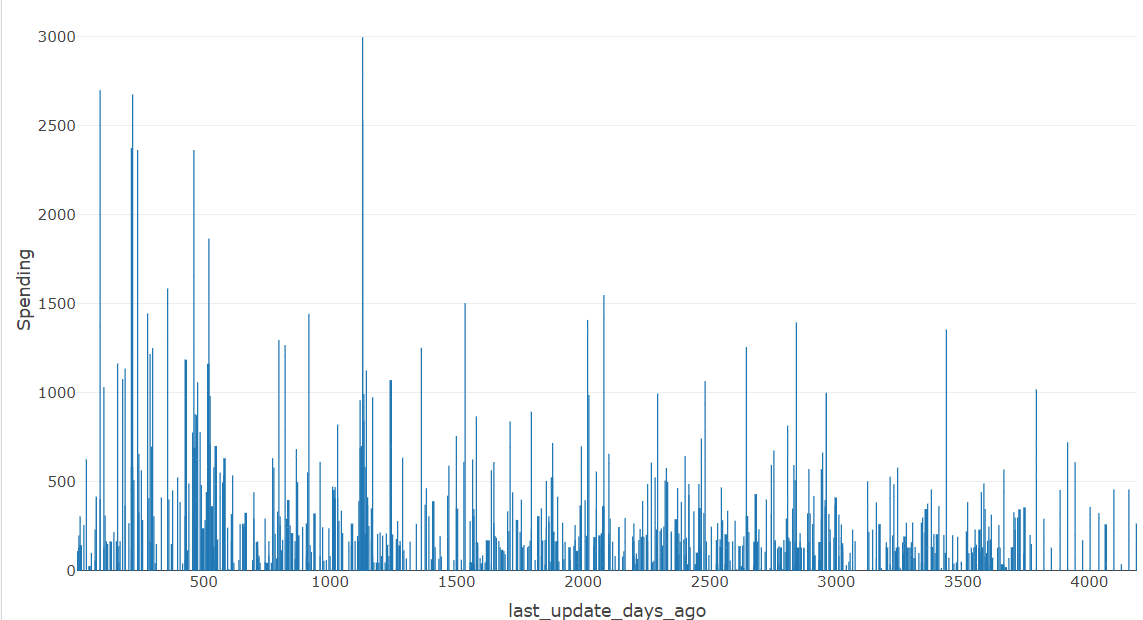


**Figure 2.36 freq vs Purchase**

From the figure 2.36 it is clear that number of transactions made by purchasers is more than the non purchaser customers.

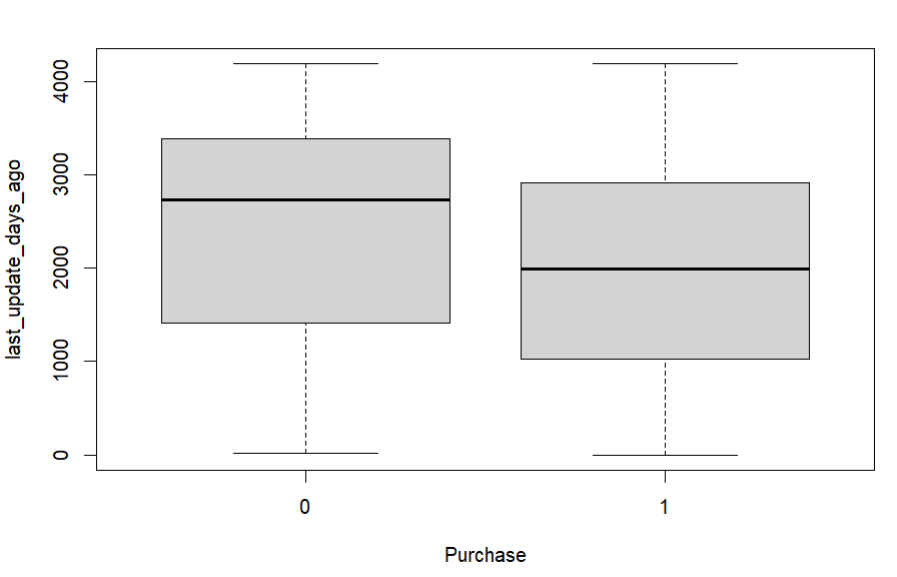
**Last\_update\_days\_ago**

Last\_update\_days\_ago column tells how many days ago the last update was made to the customer data. Lets analyze how this column is impacting the target columns.



**Figure 2.37 last\_update\_days\_ago vs spending**

From the figure 2.37 it is clear that customer who lastly updated their data less than 1500 days ago has spent more on purchases.

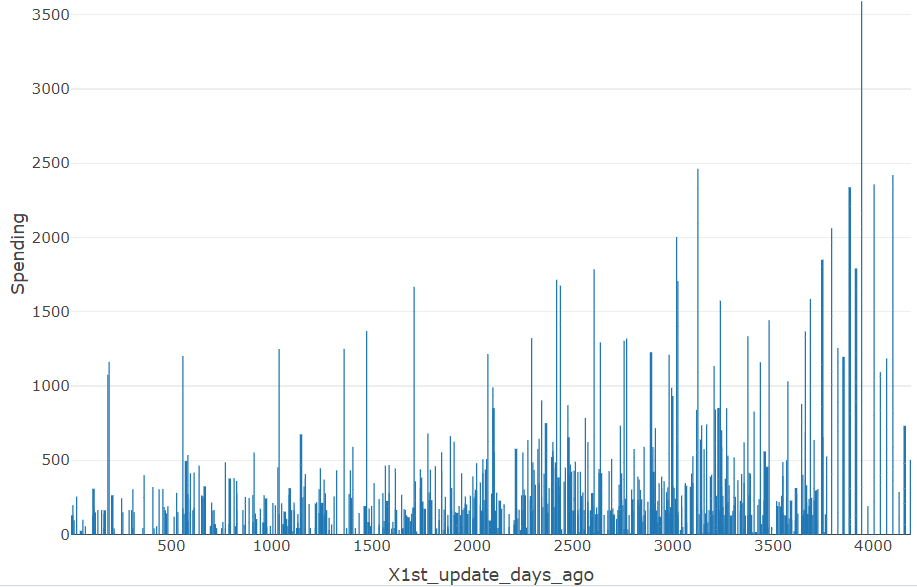


**Figure 2.38 last\_update\_days\_ago vs Purchase**

From the figure 2.38 it is clear that most of the customer who are purchasers last updated their data between 1000 and 3000 days ago.

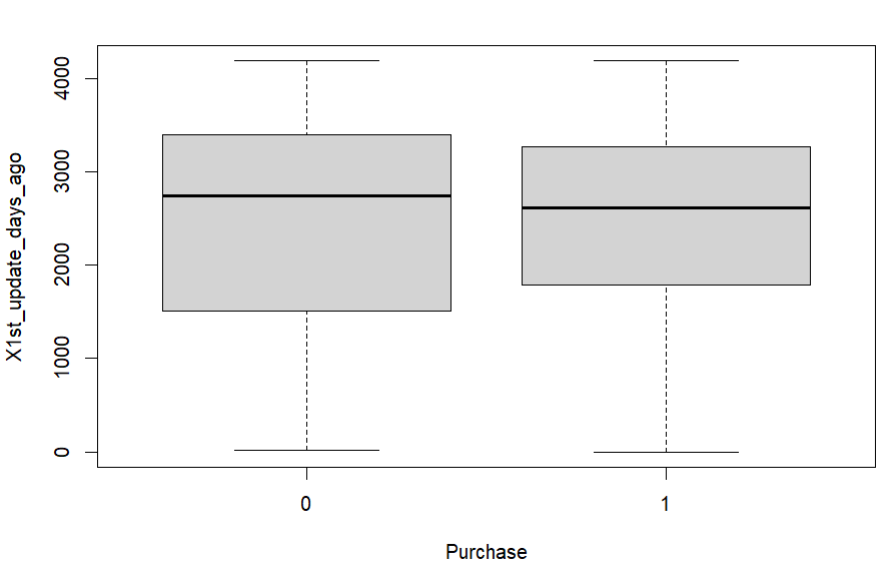
**X1st\_update\_days\_ago Column**

X1st\_update\_days\_ago Column tellshow many days ago the first update was made to the customer record.

****

From the figure 2.39 it is clear that as the customer first updated days ago increases the spending also increasing.

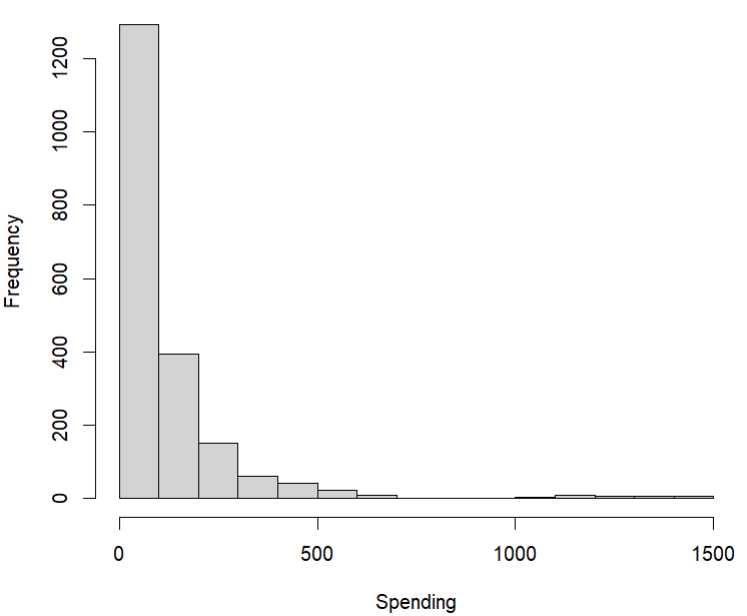
**Figure 2.39 x1st\_upadate\_days\_ago vs spending**

****

**Figure 2.40 x1st\_upadate\_days\_ago vs spending**

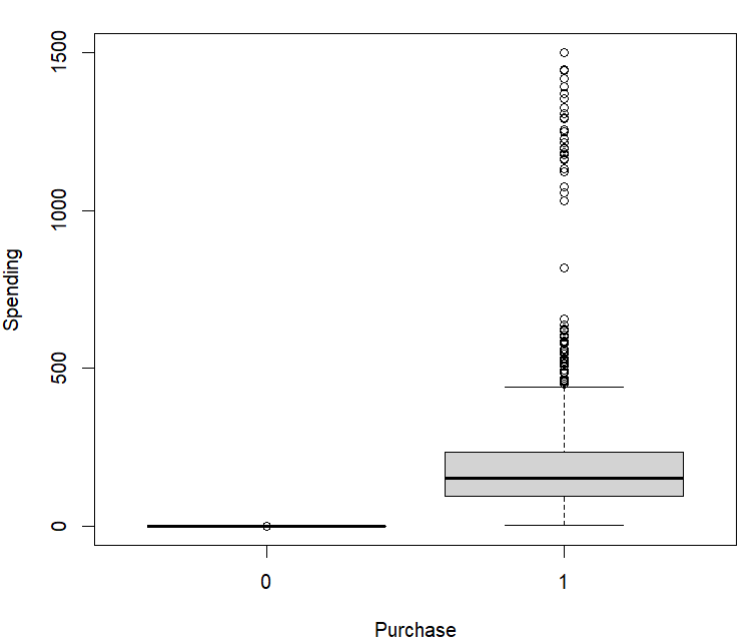
From the figure 2.40 it is clear that most of the customer who are purchasers first updates their data between 3200 and 1800.

Next, will try to understand the distribution of the target column Spending and how purchase impacting spending

****

**Figure 2.41 histogram of spending**

Spending column is left skewed data and most of the customers have spending amount less than 500$. Very few customers spent 1500$.

****

**Figure 2.42 Purchase vs Spending**

From the figure 2.42 it is clear that there is only one customer who is not a purchaser but has the spending value 1$.

* 1. **Data Preprocessing**

**Handling Categorical Columns**

Based on the analysis above, it's noted that the purchase column has an integer data type. Given that it serves as the target column and is categorical in nature, it's essential to convert the data type of the purchase column to a factor to work well for model like decision tree.

**Missing Values**

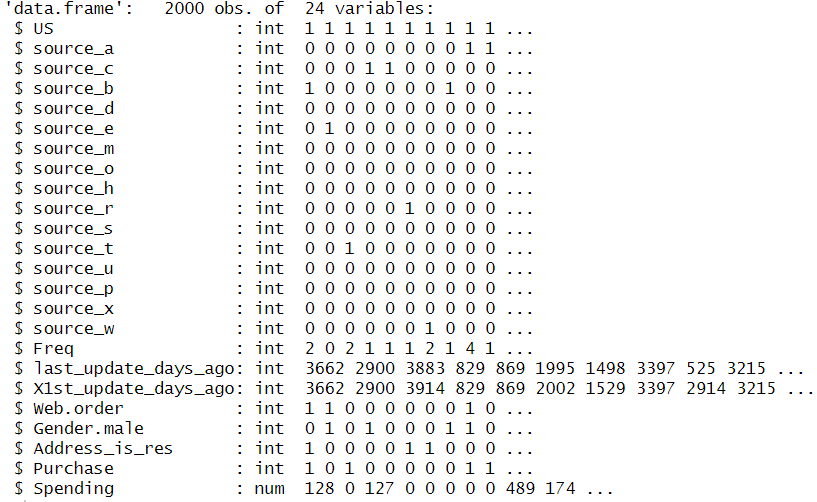
From the EDA analysis there is no missing values to handle.

**Outliers**

For customers who are not purchasers, spending should be zero. However, Figure 2.42 shows records with a spending amount of $1 for non-purchasers, which is an outlier. Therefore, the spending amount for those records needs to be adjusted to $0.

**Normalizing and Rescaling Data**

From Figure 2.43, it is noted that the columns in the data are in different scales. However, some models like KNN will perform well if the data is scaled. Therefore, if KNN is used to predict the outcome column, the data need to be scaled using standardization.



**figure 2.43 structured of data**

1. **Dimension Reduction**

Dimension reduction is important because it's challenging for north-point company to obtain error-free data when dealing with a large number of predictor columns.

**Practical Considerations**

According to the EDA analysis, all the columns are important for the task of classifying and predicting whether customers will make a purchase or not.

**Dimension Reduction using correlation analysis**

From the figure 2.4there is a correlation exists between last\_update\_days\_ago and X1st\_update\_days\_ago but both columns tell different information about the customer these columns are not duplicated column so those columns are not removed.

**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. To overcome these issues and prevent model overfitting, 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.

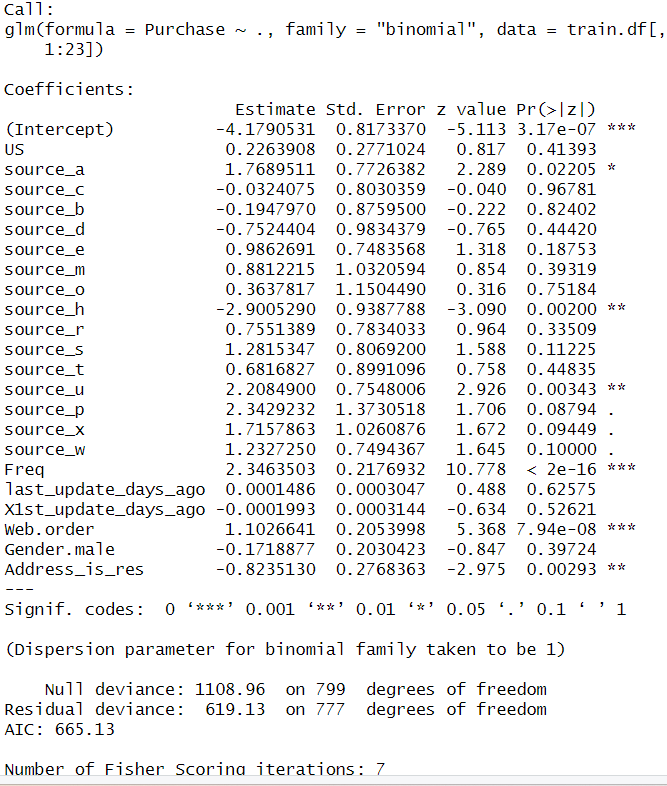
Partitioning the data into train, validation, and test sets is done randomly according to the proportions of 40%, 35%, and 25%, respectively. The customer records in the train, validation, and test partitions should be distinct. After partitioning the dataset, the train partition contains 800 records, the validation partition contains 700 records, and the test partition contains 500 records.

1. **Modelling for Classification Task**

In this phase, the focus shifts to fit classification machine learning models to predict the probability of customer becoming purchaser. Based on the probability estimation company can decide which customer to contact for Malling the products. The target variable for classification model is Purchase column, while the predictors include all columns except Sending. The performance of the models is measured using a confusion matrix. Here class 1(purchaser) considered as the important class because it’s important to classify purchasers properly than classifying non-purchasers. To predict the probability only logistic regression type model is used because these are the only model that can accurately predict the probabilities when compared to all other models.

* 1. **Logistic Regression**

First, will use the logistic regression model which is highly popular and powerful classification model. 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 the target column Purchase is used as it is without factoring.



**Figure 5.1 Summary of the logistic regression for train data**

The model estimated equation is :

Logit(purchase = 1) = -4.1790531+ (US \* 0.2263908)+ (source\_a\*1.7689511) + (source\_c \* -0.0324075) + (source\_b \* -0.1947970) + (source\_d \* -0.7524404) + (source\_e \* 0.9862691) + (source\_m \* 0.8812215) + (source\_o \* 0.3637817) + (source\_h \* -2.9005290) + (source\_r \* 0.7551389) + (source\_s \* 1.2815347) + (source\_t \* 0.6816827) + (source\_u \* 2.2084900) + (source\_p \* 2.3429232) + (source\_x \* 1.7157863) + (source\_w \* 1.2327250) + (Freq \* 2.3463503) + (last\_update\_days\_ago \* 0.0001486) + (X1st\_update\_days\_ago \* -0.0001993) + (Web.order \* 1.1026641) + (Gender.male \* -0.1718877) + (Address\_is\_res \* -0.8235130)

**Model Interpretation Based on coefficients**

Customer from source\_a, source\_e, source\_m, source\_o, source\_h, source\_r, source\_s, source\_t, source\_u, source\_p, source\_x, source\_w has positive coefficients which indicates that customers associated with these specific sources are more likely to be classified as purchasers.

Customers from source\_c, source\_b, source\_d, has negative coefficients which indicates that customers associated with these specific sources are less likely to be classified as purchasers.

Customers who made purchases through web orders has the positive correlation which indicates that those customers are more likely to get classified as purchasers

X1st\_update\_days\_ago has negative coefficient indicates that customers with higher values of X1st\_update\_days\_ago are more likely to be classified as non-purchasers

Gender.male has the negative coefficients which indicates that male customers are less likely to be classified as purchaser

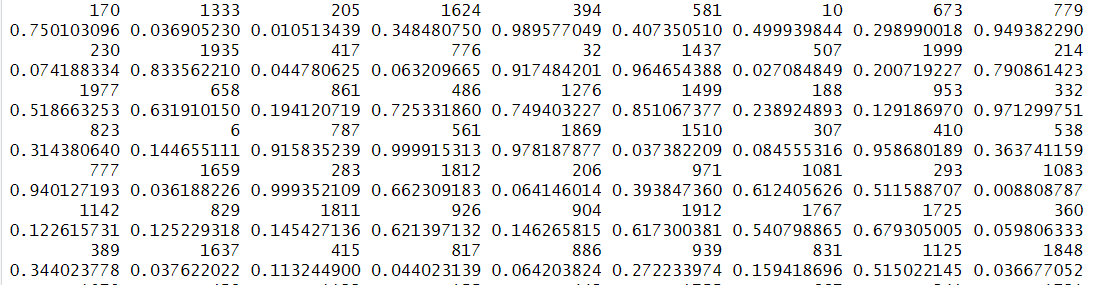
Freq has the positive coefficients which indicates that higher the customers freq value are more likely to get classified as purchasers.

last\_update\_days\_ago has the positive coefficients which indicates that customers having the higher values of last\_update\_days\_ago are more likely to classified as purchasers.

Address\_is\_res has the negative coefficient which indicates that customers who are residential are less likely to get classified as purchasers.

Suppose logistic regression is the final model. North-Point Company can use the logistic model to predict the probability of a new customer becoming a purchaser. If the probability is above the company's threshold value, for example, 0.5, is considered a threshold value. If the predicted probability for a new customer is 0.5 or higher, then the new customer is assigned to the purchaser class (purchase =1).

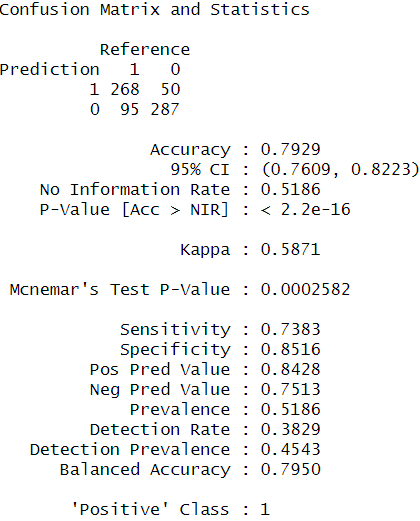
Now will predict the validation data using the model



**figure 5.2 predicted propensities for validation data**

* + 1. **Measuring Performance**

The predictive performance of the model is assessed on the validation data using a confusion matrix. This matrix summarizes the correct and incorrect classifications that classifier produced for certain dataset. In practical most of other accuracy measures are derived from the confusion matrix. The chosen threshold value is 0.5, meaning that propensities greater than 0.5 are classified as class1.



**Figure 5.3 Confusion Matrix for validation data**

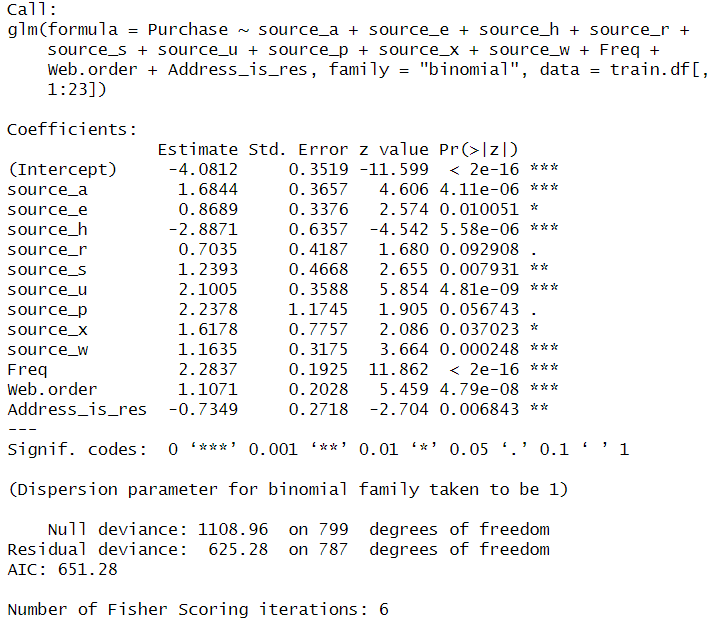
The accuracy of the validation set is 0.79. Rather than focusing only on overall accuracy, it is more crucial to evaluate the classifier's ability to correctly identify important class members, specifically in Class 1. Therefore, sensitivity is chosen as the measuring metric. The sensitivity value of 0.7383 indicates that 73.83% of Class 1(purchasers) members are correctly classified.

* + 1. **Improving the model performance**

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

**Backward selection algorithm**

Now will try to improve the model performance by using the Backward selection algorithm which will start with all predictors and, in each step, it will eliminate the least useful predictor.



**figure 5.4 Summary of Backward selection model for train data**

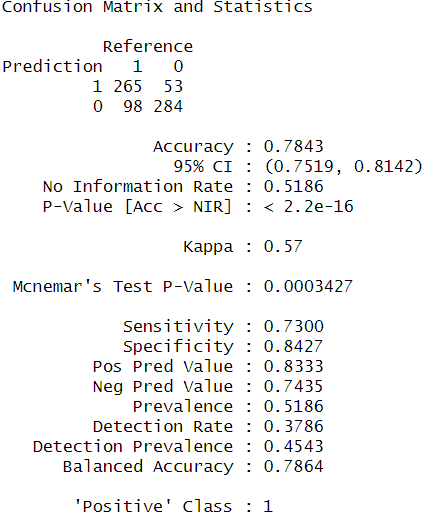
The model estimated equation is

Logit(purchase =1) = -4.0812 + (source\_a \* 1.6844) + (source\_e \* 0.8689) + (source\_h \* -2.8871) + (source\_r \* 0.7035) + (source\_s \* 1.2393) + (source\_u \* 2.1005) + (source\_p \* 2.2378) + (source\_x \* 1.6178) + (source\_w \* 1.1635)+ (Freq \* 2.2837)+ (Web.order \* 1.1071)+ (Address\_is\_res \* -0.7349)

**Model Interpretation Based on coefficients**

customers from sources source\_a, source\_e, source\_r, source\_s, source\_u, source\_p, source\_x, and source\_w are more likely to be predicted as purchasers. Additionally, customers with a higher frequency of transactions are more likely to get classified as purchasers. Moreover, customers who made purchases through web orders are more likely to be classified as purchasers. Conversely, customers whose addresses are residential are more likely to get classified as non-purchasers and customers from source\_h are more likely to classified as purchaser.

**Measuring Performance on validation data**

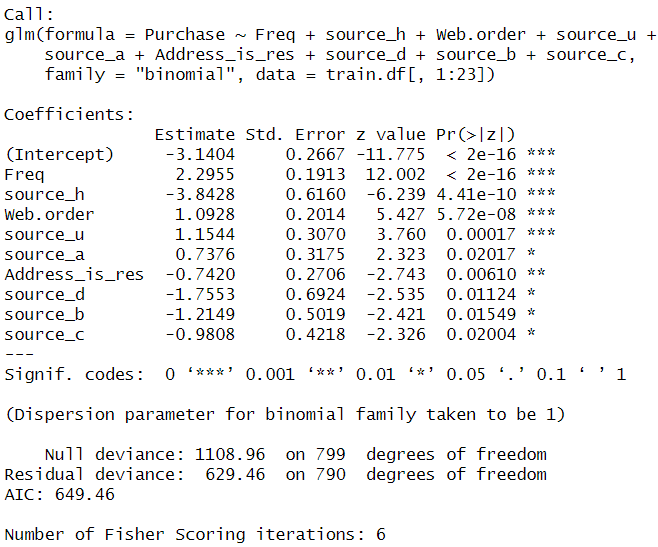
****

**Figure 5.5 Confusion Matrix of backward on validation data**

The sensitivity is 0.7300 which indicates that 73% of Class 1(purchasers) members are correctly classified.

**Forward selection model**

Forward selection algorithm will start with zero predictors and in each step will add predictors one by one.



**figure 5.6 Summary of Forward Selection Model for train**

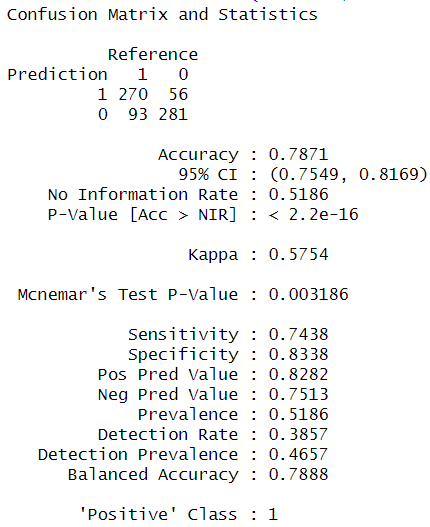
The Model estimated equation is

Logit(purchase =1) = -3.1404 + (Freq \* 2.2955) + (source\_h \* -3.8428) + (Web.order \* 1.0928) + (source\_u \* 1.1544) + (source\_a \* 0.7376) + (Address\_is\_res \* -0.7420) + (source\_d \* -1.7553) + (source\_b \* -1.2149) + (source\_c \* -0.9808)

**Model Interpretation Based on coefficients**

Customers from sources source\_h, source\_d, source\_b, and source\_c are less likely to be classified as purchasers. Additionally, customers with residential addresses are less likely to get classified as purchasers. Conversely, customers with a higher frequency of transactions are less likely to be classified as purchasers. Moreover, customers who made purchases through web orders are more likely to be classified as purchasers. Customers from source\_u and source\_a are more likely to get classified as purchasers.

**Measuring Performance on validation data**

****

**Figure 5.7 Confusion Matrix of forward selection model on validation data**

The sensitivity is 0.7438 which indicates that 74.38% of Class 1(purchasers) members are correctly classified.

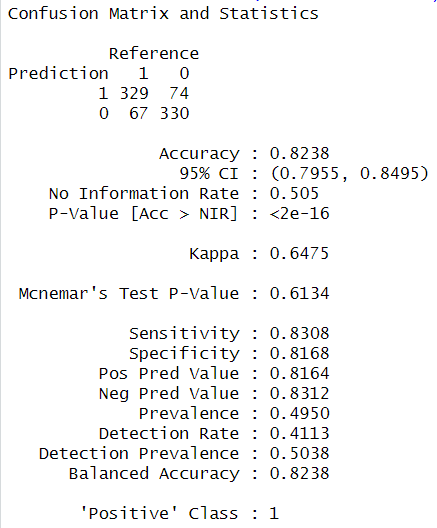
1. **Classifier Model Selection**

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

**Table 6.1 Accuracy Comparison of models**

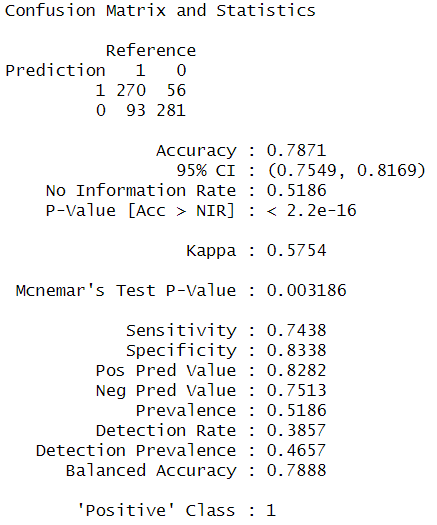
|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Sensitivity |
| Logistic Regression | 79.29% | 73.83% |
| Backward selection | 78.43% | 73% |
| Forward Selection | 78.71% | 74.38% |

Based on the sensitivity percentage of all the models, the forward selection logistic regression model correctly classified 74.38% of purchasers in the sample, which is higher compared to all other models. Therefore, forward selection model is considered as the best classifier model. Now let’s check if the forward selection model is overfitting or not by comparing the training accuracy with validation accuracy



**Figure 6.1 Confusion matrix of forward selection model on train data**

As the training data accuracy more than the validation data accuracy the forward selection model is not overfitted.



**Figure 6.2 Confusion Matrix of the validation data for the best classifier**

Now will try to understand how the best classifier model is performing from various aspects

Accuracy: 78.71%

This suggests that among the 700 customers in validation partition, 78.71% are accurately classified into their respective classes.

Error rate: (56+93)/700 = 0.21 = 21%

The error rate of 21% indicates that out of the 700 customers in validation partition, only 21% are misclassified.

Sensitivity: 74.38%

This indicates that 74.38% of the purchasers (Class1) are correctly classified.

Specificity: 83.38%

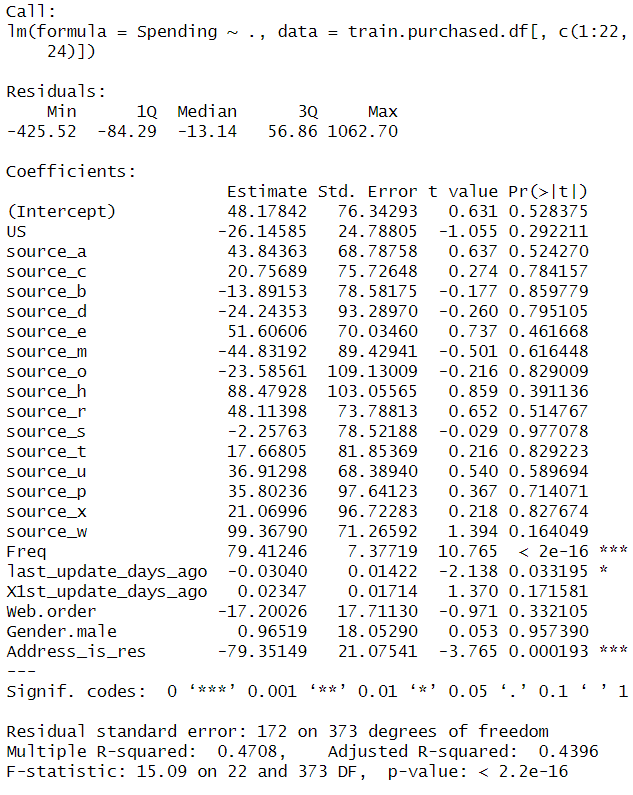
The specificity of 83.38% suggests that 83.38% of the non-purchasers (Class 0) are correctly classified.

1. **Modelling for Regression Task**

From now on, the prediction will focus on determining the amount customers spend. In this process, only purchased customer data will be filtered for both the training and validation datasets. After filtering, there are 396 records in the training data and 363 records in the validation data. Spending column is considered as the target column and remaining all columns except purchase is 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 multiple linear regression model**

The model estimated equation is

Spending= 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)

**Interpretation of the Model**

The regression coefficients generated by the model are used to predict the individual spending amount of the new purchaser customer.

* + 1. **Measuring Performance**

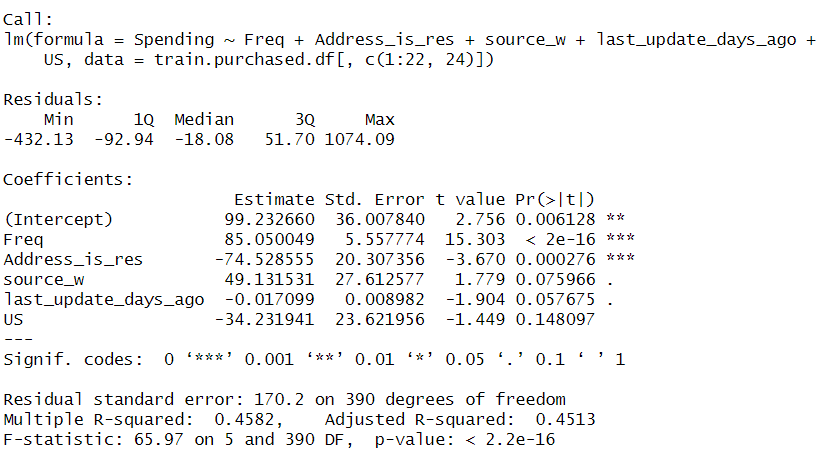
To measure performance, RMSE is used. The model with lowest value of RMSE value is the best model. The RMSE value for the validation data is 167.2498.

* + 1. **Improving Model Performance**

In this section, the aim is to construct Forward selection models to assess whether there is an improvement in predictive performance compared to multiple linear regression model.

**Forward selection model**

Forward selection algorithm will start with zero predictors and in each step will add predictors one by one.



**Figure 7.3 Summary of forward selection Model**

The model estimated equation is

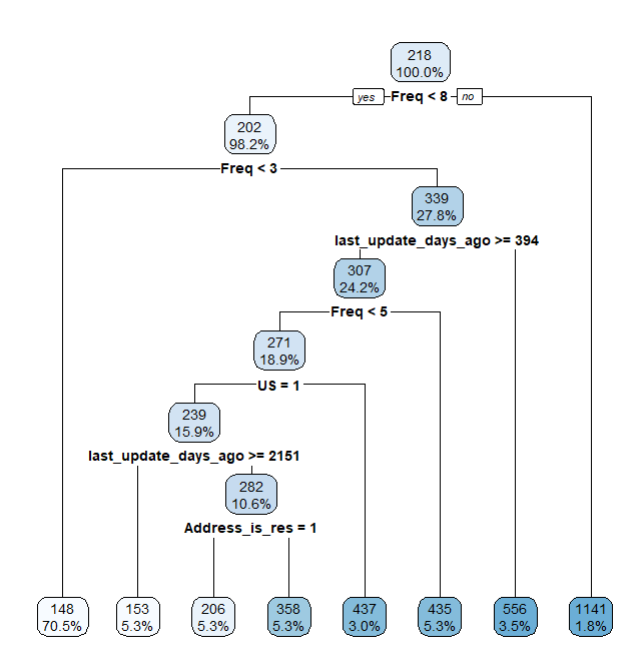
Spending = 99.232660 + (Freq \* 85.050049) + (Address\_is\_res \* -74.528555) + (source\_w \* 49.131531) + (last\_update\_days\_ago \* -0.017099) + (US \* -34.231941)

**Performance measure**

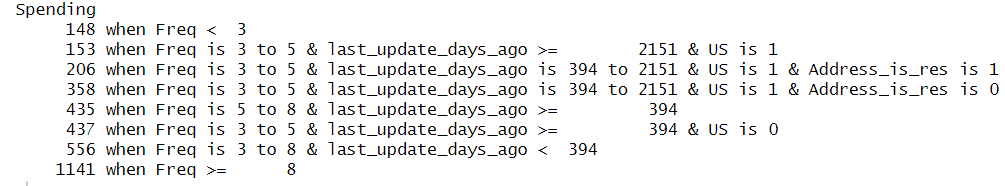
The RMSE of the forward selection model is 165.6397. When compared with the multiple linear regression model, the forward selection model is the best.

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

Regression tree is non-parametric so there is no assumption for the model. So, train data that filtered only purchased customer data can be directly use the train data that filtered only purchased customer data. Regression tree can be applied in R by using RPART.

****

**Figure 7.4 Decision tree**

****

**Figure 7.5 Rules Generated by the Decision Tree**

**Interpretation**

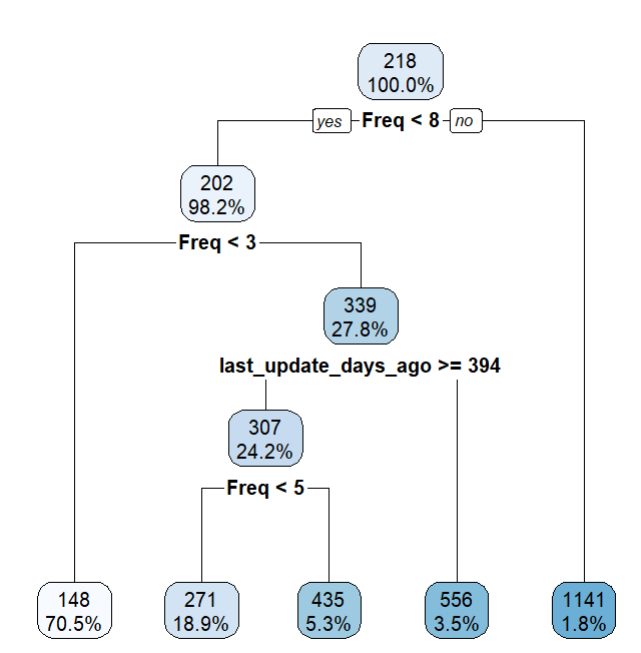
When predicting a new purchaser customer spending amount, the record will drop down to the terminal node based on the predictors' information, and the prediction will be the average spending value in the terminal node.

* + 1. **Measuring Performance**

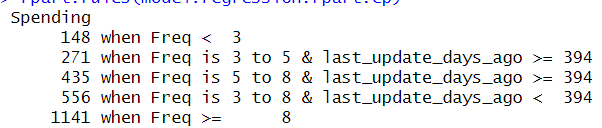
The RMSE value of the purchaser’s validation data is 181.0081.

* + 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 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10. The Best Cp value is 0.02.



**Figure 7.6 Decision tree for CP value 0.02**



**Figure 7.7 Rules for cp value 0.02**

**Performance measure**

The RMSE value for the validation purchaser’s data is 173.6269

1. **Regression Model Selection**

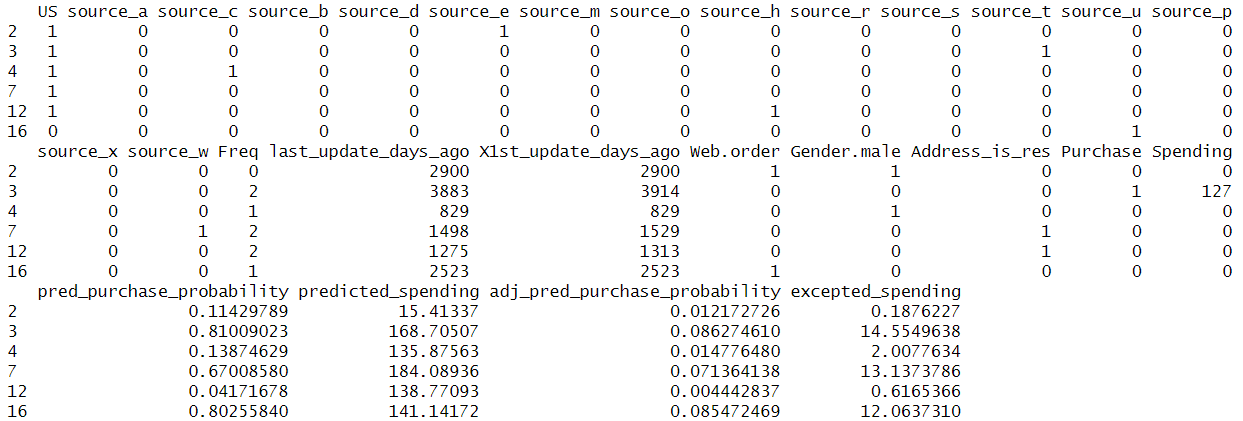
**Table 8.1 Performance Comparison**

|  |  |
| --- | --- |
| Model | RMSE on Validation Data |
| Multiple Linear Regression | 167.2498 |
| Forward Linear Regression | 165.6397 |
| Decision Tree (Regression Tree) | 181.0081 |
| Decision Tree (Regression Tree cp 0.02) | 173.6269 |

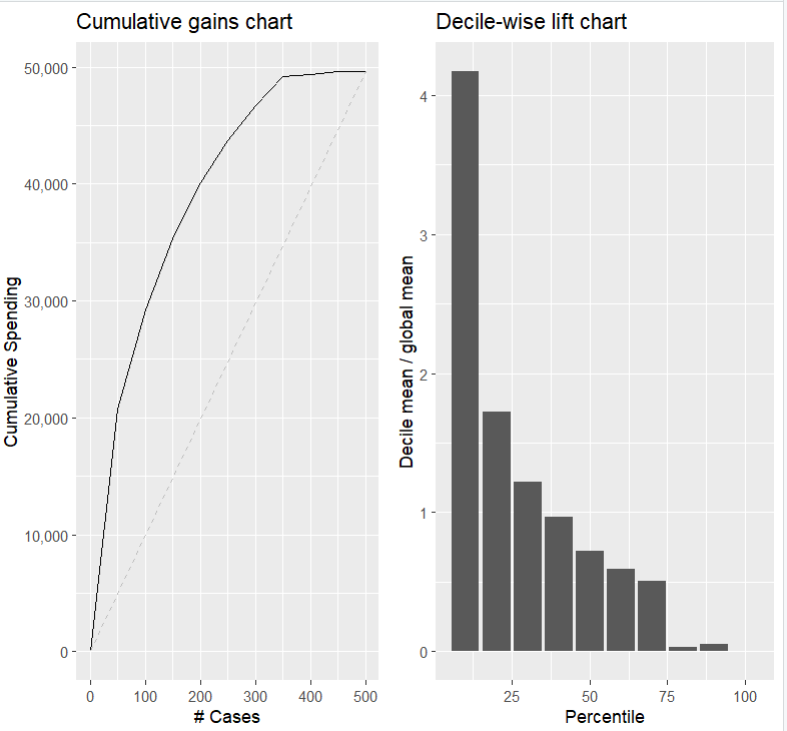
From table 8.1, it is evident that the Forward selection model is the optimal choice for predicting the spending value of a customer who is a purchaser. Because of its low RMSE value compared to all other models. Now Validation RMSE value from forward model is compared with the Train RMSE value to find if the model is overfitting or not. The RMSE value for the training data is 168.9053, which is higher than the validation RMSE. Since there is only a minimal difference, the model is considered non-overfitting. If the difference were too large, then it would be considered overfitting.

1. **Predicting New data**

In this section, the selected classifier forward selection logistic regression model is utilized to classify the customers in the test partition. Additionally, the best-selected regression model forward selection model is applied to predict the spending value of the customer in the test data. Since the data used to build the model is oversampled, the predicted probability needs to be adjusted by multiplying it with the purchase rate obtained from the test mailing, which is 0.1065. Moreover, the spending value predicted from the model also needs to be adjusted by dividing the predicted value by the adjusted probability.



**Figure 9.1 first 6 rows of the test data after adding all columns**

****

**Figure 9.2 Cumulative gain chart and decile-wise lift chart of spending vs expected spending**

The figure 9.2 shows the lift chart and decile chart based on fitting the selected model on data. Form the gain chart it is clear that the predictive performance of the models in the terms of lift is better than the baseline model. From the decile chart it is clear that by choosing only top 10% of highest spending customers predicted by model can gain 4.17 times the profit. So, instead of investing in mailing the product list to all 180,000 customers and risking lower profits, North-Point Company can use the model to target specific customers, maximizing profit with less investment.

1. **Conclusion**

In conclusion, North-point can utilize predictive analysis to make best decisions regarding customer selection to mail the products list. Rather than investing $360,000 to mail the product list to all 180,000 customers, the company can focus its investment solely on the top 10% of high-spending customers predicted by the model. This targeted investment, amounting to $36,000, results in a 4.17 times higher profit compared to investing in all 180,000 customers.