**North Point Software Listing Company**

**Mahendrakumar D Chaudhari**

**Data Analytics, Webster University**

**CSDA 6010: Data Analytics Practicum**

**Prof: Dr. Jiangping Wang**

**Table of Contents**

[Introduction 6](#_Toc166084884)

[1.0 Business Problems and Goals 6](#_Toc166084885)

[1.1Business Problems: 6](#_Toc166084886)

[1.2 Analytics Goals: 7](#_Toc166084887)

[2.0 Data Exploration and Preprocessing 7](#_Toc166084888)

[2.1 Attributes Definition: 7](#_Toc166084889)

[2.2 Data Exploration: 9](#_Toc166084890)

[2.3 Distribution of Variables 11](#_Toc166084891)

[3.0 Predictors Analysis and Relevancy 14](#_Toc166084892)

[4.0 Dimension Reduction 19](#_Toc166084893)

[PCA analysis: 20](#_Toc166084894)

[5.0 Data partitioning method 21](#_Toc166084895)

[6.0 Classifier Models Selection and Model Performance: 22](#_Toc166084896)

[6.1 Logistic regression model: 23](#_Toc166084897)

[6.2 Evaluating the logistic regression model Performance: 24](#_Toc166084898)

[6.3 Stepwise Regression Model: 25](#_Toc166084899)

[6.4 Evaluating the Stepwise Backward model Performance. 27](#_Toc166084900)

[6.5 Evaluating the Stepwise Forward model Performance. 28](#_Toc166084901)

[6.6 K-NN Model 30](#_Toc166084902)

[6.7 Evaluating K-NN model performance. 30](#_Toc166084903)

[6.8 Tree Model 31](#_Toc166084904)

[6.9 Evaluating Tree model Performance: 32](#_Toc166084905)

[6.10 Naïve Bayes Model 34](#_Toc166084906)

[6.11 Evaluating Naïve Model Performance. 35](#_Toc166084907)

[7.0 Best Classifier Model Selection 37](#_Toc166084908)

[8.0 Regression Model Selection 37](#_Toc166084909)

[8.1 Linear Regression Model 38](#_Toc166084910)

[8.2 Evaluating Regression Model Performance 39](#_Toc166084911)

[8.3 Stepwise- backward Regression 39](#_Toc166084912)

[8.4 Evaluating Stepwise Backward Regression Model Performance 40](#_Toc166084913)

[8.5 Regression tree 40](#_Toc166084914)

[8.6 Evaluating Regression tree model Performance. 41](#_Toc166084915)

[9.0 Best Regression Model Selection 41](#_Toc166084916)

[10.0 Profit analysis 41](#_Toc166084917)

[10.1 Gross profit 43](#_Toc166084918)

[11.0 Conclusion 44](#_Toc166084919)

**List of Figures**

[Figure 2.2.1: First Six Rows and dimension of the data 9](#_Toc166083488)

[Figure 2.2.2: Columns names of the dataset 9](#_Toc166083489)

[Figure 2.2.3: Summary of the data 10](#_Toc166083490)

[Figure 2.2.4: Data Types of the Variables 10](#_Toc166083491)

[Figure 2.2.5: Checking the missing Values 11](#_Toc166083492)

[Figure 2.2.6: Check for the Zero’s 11](#_Toc166083493)

[Figure 2.3.1: Distribution of Purchase and Gender=Male 12](#_Toc166083494)

[Figure 2.3.2: Distribution of the No of transactions 12](#_Toc166083495)

[Figure 2.3.3: Distribution of Spending amount 13](#_Toc166083496)

[Figure 2.3.4: Distribution of Web Order 14](#_Toc166083497)

[Figure 2.3.5: Distribution of US addresses and US Residence 14](#_Toc166083498)

[Figure 3.0.1: Correlation matrix plot 15](#_Toc166083499)

[Figure 3.0.2: Pairwise Plot 16](#_Toc166083500)

[Figure 3.0.3: Last Update Days Ago & X1st Update Days Ago Vs Spending 17](#_Toc166083501)

[Figure 3.0.4: Frequency Vs Spending 17](#_Toc166083502)

[Figure 3.0.5: Spending By Web Order 18](#_Toc166083503)

[Figure 3.0.6: Spending By Purchaser & Gender= Male 18](#_Toc166083504)

[Figure 3.0.7: Spending By US Address and US residence 19](#_Toc166083505)

[Figure 4.0.1: Removing the Sequence Column 19](#_Toc166083506)

[Figure 4.0.2: Variance Captured by Dimensions 20](#_Toc166083507)

[Figure 4.0.3: Summary of PCA result 21](#_Toc166083508)

[Figure 4.0.4: Variability of the predictors 21](#_Toc166083509)

[Figure 5.0.1: Data Partition 22](#_Toc166083510)

[Figure 6.1.1: Summary of the model 24](#_Toc166083511)

[Figure 6.2.1: Confusion Matrix of Logistic Regression Model 25](#_Toc166083512)

[Figure 6.3.1: Summary of Stepwise Backward Regression Model 26](#_Toc166083513)

[Figure 6.3.2: Summary of Stepwise Forward Regression Model 27](#_Toc166083514)

[Figure 6.4.1: Confusion matrix for Training data 28](#_Toc166083515)

[Figure 6.4.2: Confusion Matrix for Validation Data 28](#_Toc166083516)

[Figure 6.5.1: Confusion Matrix for Training data 29](#_Toc166083517)

[Figure 6.5.2: Confusion Matrix for Validation data 29](#_Toc166083518)

[Figure 6.6.1: Best K Values in K-NN model 30](#_Toc166083519)

[figure 6.7.1: Confusion Matrix for K-NN model on training and Validation data 31](#_Toc166083520)

[Figure 6.8.1: Default Classification Tree 32](#_Toc166083521)

[Figure 6.9.1: Confusion Matrix On training data 33](#_Toc166083522)

[Figure 6.9.2: Confusion Matrix On Validation Data 33](#_Toc166083523)

[Figure 6.10.1: Summary of Naïve Model 35](#_Toc166083524)

[Figure 6.11.1: Confusion Matrix on Training Data 36](#_Toc166083525)

[Figure 6.11.2: Confusion Matrix On Validation Data 36](#_Toc166083526)

[Figure 7.0.1: Comparison of Classification Models 37](#_Toc166083527)

[Figure 8.1.1: Summary of the Regression Model 38](#_Toc166083528)

[Figure 8.2.1: Accuracy for Validation 39](#_Toc166083529)

[figure 8.3.1: Summary of Stepwise Backward Regression Model 39](#_Toc166083530)

[Figure 8.4.1: Accuracy for Validation 40](#_Toc166083531)

[Figure 8.5.1: Regression tree plot 40](#_Toc166083532)

[Figure 8.6.1: Accuracy for Validation 41](#_Toc166083533)

[Figure 9.0.1: Comparison of regression Models 41](#_Toc166083534)

[Figure 10.0.1: Model Performance on Holdout data 42](#_Toc166083535)

[Figure 10.0.2: Cumulative Gains Chart and Decile wise lift chart 43](#_Toc166083536)

[Figure 10.1.1: Gross Profit Calculation 44](#_Toc166083537)

# Introduction

The North Point Software Company is a firm that sells games and educational software. The company wants to expand its customer base, so they join the group name consortium. This group specializes in computer hardware and software products. Every member of the group shares their customer list in a pool, and they receive the same number of customers from the pool. The North Point Software Company shared 200,000 names in the pool, for a total of 5,000,000 names in the pool. The company picked 20,000 names and did a test mailing. Out of the 20,000 customers listed, 1065 purchased after receiving mail, with a response rate of 0.053, or 5.3%. So, the company made a list of 1000 purchasers and 1000 non-purchasers (a response rate of 0.5 or 50%) to build the best prediction model. While using the prediction model in the pool, the purchase rate needs to be adjusted back down by multiplying each “case’s probability of purchase” by 5.3/50, or 0.106. The company is allowed to use a prediction model in the pool (5,000,000), so they can select the top 180,000 customers from the pool. The study will use the models to identify the purchasers and predict their spending behaviours to maximize gross profit and the customer base.

# 1.0 Business Problems and Goals

## 1.1Business Problems:

The company's primary objectives are to expand its customer base and enhance profit. After paying $2 for mail, the company wants to know if the customer will make a purchase or not. The business also wants to know if the customers are purchasers and how much they are going to spend on software or games. The company believes that they can apply the models to the pool of 5,000,000 customers. The company needs to rank those customers from the pool based on the original purchase rate (adjusted probability: 0.106) and adjusted probabilities.

The aim of maximizing gross profit while maintaining cost-effectiveness led to the decision to focus on the top customers rather than all customers. This target is based on the knowledge that the greatest returns on investment will come from concentrating on the top customers who are most likely to make purchases.

The business can maximize profit and expand its customer base by focusing its efforts on the most promising customers.

## 1.2 Analytics Goals:

The main goal of the North Point Software Company is to select all the potential customers from the pool to expand their business. The company aims to build prediction models that classify the customer as a purchaser or non-purchaser after receiving mail and their spending behaviors. Based on what the remaining 180,000 customers might spend on games and software, the company hopes to estimate the gross profit from them. By ranking all the customers from the pool, the company wants to use higher probabilities of spending customers to maximize their gross profit and the remaining average or lower probabilities of spending customers to make more cost-effective strategies.

# 2.0 Data Exploration and Preprocessing

## 2.1 Attributes Definition:

* **Sequence\_number**: It shows the distinct number or index associated with every record in the dataset.
* **US**: This column indicates whether the customer is located in the United States or not. It is likely a binary variable, with 1 representing customers from the US and 0 representing customers from other countries.
* **Source\_a, source\_b, source\_c, 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 likely represent different sources or channels through which customers were acquired. Each column may contain binary values (0 or 1) indicating whether a particular source was used to acquire the customer.
* **Freq**: This column represents the frequency or number of purchases made by the customer. It is likely a numeric variable that indicates the count of purchases.
* **last\_update\_days\_ago**: This column may indicate the number of days that have passed since the last update or interaction with the customer's data.
* **1st\_update\_days\_ago**: It may represent the number of days that have passed since the first update or interaction with the customer's data.
* **Web order**: This column may indicate whether the purchase was made through a web order or not. It could be a binary variable with 1 representing a web order and 0 representing other types of orders.
* **Gender=male**: This column may indicate the gender of the customer, with a binary value of 1 representing male and 0 representing female or other genders.
* **Address\_is\_res**: This column may indicate whether the customer's address is residential or not. It could be a binary variable with 1 representing a residential address and 0 representing a non-residential address.
* **Purchase**: This variable indicates whether a prospect responded to the test mailing and made a purchase. It is likely a binary variable with values like 0 or 1, where 0 represents no purchase and 1 represents a purchase.
* **Spending**: For those prospects who made a purchase (Purchase = 1), this variable represents the amount they spent. It is likely a numeric variable that indicates the monetary value of the purchase.

## 2.2 Data Exploration:

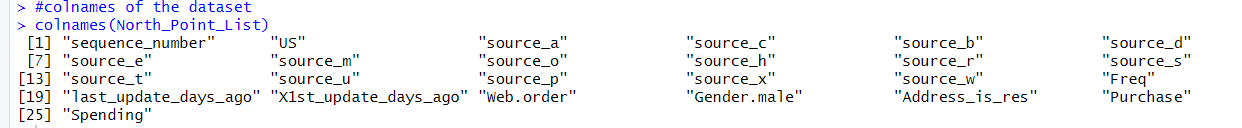
Data exploration helps to get the value inside of the dataset. First, the dataset is typically loaded using the **'read\_csv'** function. With the “**head ()”** function, which shows the first six rows and dimension of the dataset (Figure 2.2.1). In the dataset, there are 2000 rows (records) and 25 columns (attributes) stored as **‘North\_Point\_List’**. The target variables are “Purchase” and “Spending” which show whether a customer made a purchase or not after receiving mail and how much amount (in dollars) a customer spent.

A number grid with numbers

Description automatically generated with medium confidence

Figure 2.2.1: First Six Rows and dimension of the data

Figure 2.2.2 displays the column names of the dataset which includes sequence number, US, different sources, freq (frequency), last update days ago and 1st update days ago, web order, Gender=male, Address is res (US), Purchase, and Spending.

****

*<Figure 2.2.2: Columns names of the dataset>*

Summary statistics provide an overview of the dataset including mean, median, minimum, maximum,1st Quartile, and 3rd Quartile. In the summary of the dataset (figure: 2.2.3), the company got customers from various sources and the majority of the customers are from the US. However, there are very few customers whose addresses are residential (mean: 0.221). From all the sources, the company got the highest number of customers from source\_e (mean: 0.151) and the lowest number of customers from source\_p (mean:0.006). The average spending amount by the customers is $102.6 with minimum spending $0 and maximum spending $1500. Also, some variables are related to customer behaviors like frequency (the number of purchases), last\_update\_days\_ago, and spending.

A close up of a number

Description automatically generated

<Figure 2.2.3: Summary of the data >

Figure 2.2.4 shows that the data types for all the variables are numeric. In the dataset, the Sources, US, Web Order, Gender=male, Address\_is\_res, and Purchase columns have categorical data where 0 represents ‘no’ and 1 represents ‘yes’.

A screenshot of a computer code

Description automatically generated

<Figure 2.2.4: Data Types of the Variables>

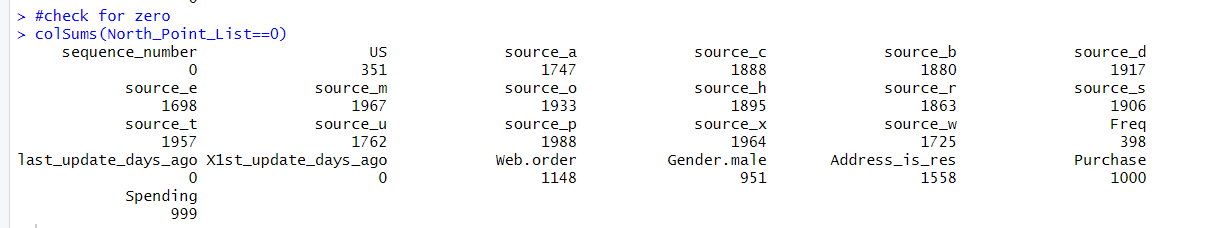
Missing values in the dataset give information about fake transactions or other information. However, checking the missing values in the dataset, figure 2.2.5 shows that there is no missing value in the dataset. This means all the transaction data and other information (e.g., sources, gender=male, address) are correctly collected.

A close-up of a list

Description automatically generated

<Figure 2.2.5: Checking the missing Values>

When it comes to checking the zeros in the dataset (Figure 2.2.6), there are 999 zeros in spending, which means there are 1001 customers who spent money, but there are only 1000 customers who purchased after receiving a mail. This might be happening due to a web order or a customer making an offline transaction. Also, 951 zeros in gender columns show that 951 customers are females and 1049 are males.



<Figure 2.2.6: Check for the Zero’s>

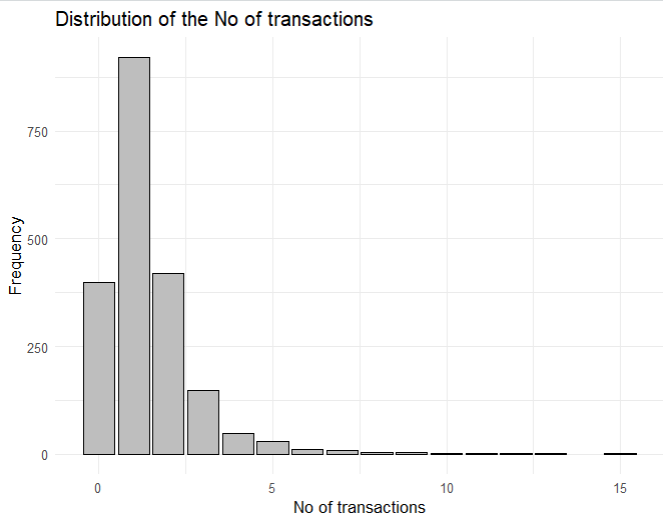
## 2.3 Distribution of Variables

The Distribution of Purchase graph (Figure 2.3.1) shows that 1000 customers made purchases and 1000 customers didn't make any purchases after receiving a mail from the company. The distribution of gender (Figure 2.3.1) shows that male customers are more than female customers.

|  |  |
| --- | --- |
| A graph of a distribution of purchase  Description automatically generated |  |

<Figure 2.3.1: Distribution of Purchase and Gender=Male>

The distribution of the number of transactions graph (Figure 2.3.2) shows that approximately 375 customers didn’t make any transactions, and around 875 customers made one transaction. There are very few customers who have made more than five transactions.



<Figure 2.3.2: Distribution of the No of transactions >

Figure 2.3.3 shows many of the customers who made a purchase through the mail and spent between 0 and 500 dollars. There are very few customers who have purchased more than $500. So, the company should know that many new customers would spend between 0 and 500 dollars. Some of the customers who spend $0 show that the company might give them free games or software.

A graph of a number of columns

Description automatically generated

<Figure 2.3.3: Distribution of Spending amount>

The distribution of the web order (Figure 2.3.4) shows that around 800 customers made a purchase through the web order, and around 1200 customers didn’t make a purchase through the web order. This shows that the majority of the company’s purchases come through other types (e.g. offline).

**A graph of a distribution of web order

Description automatically generated**

<Figure 2.3.4: Distribution of Web Order>

Figure 2.3.5 shows the distribution of US addresses and residences. Most of the customers are from the US, but not their residence. So, if a few of the customers are from outside of the US, they might purchase through the web order.



<Figure 2.3.5: Distribution of US addresses and US Residence >

# 3.0 Predictors Analysis and Relevancy

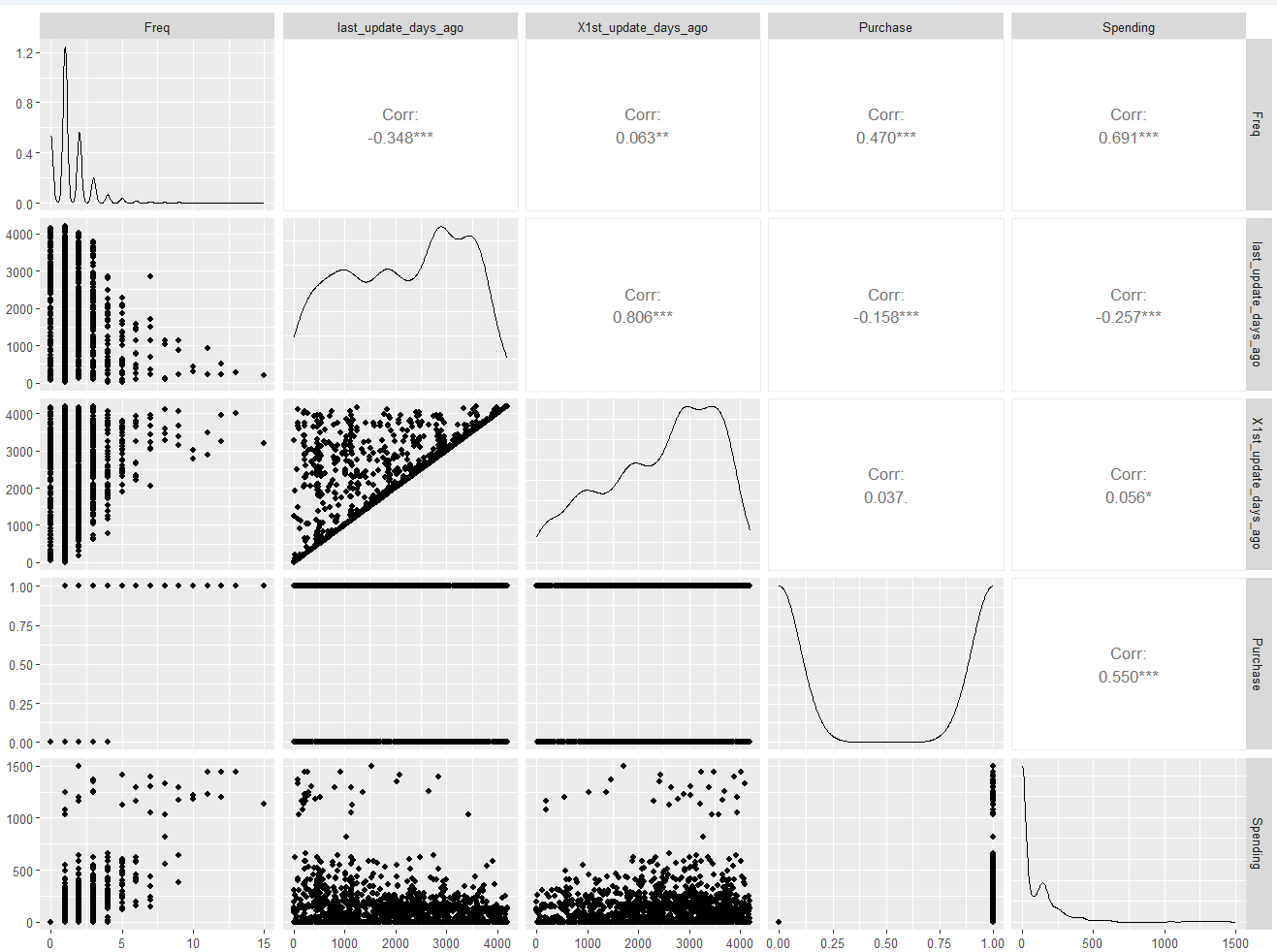
The correlation matrix gives an analysis of the relationship between predictors and target variables. From the correlation plot (Figure 3.0.1), spending and frequency have a strong positive relationship, which suggests a customer who makes more transactions spends more. The target variable spending and the predictors last\_update\_days\_ago and Xst\_update\_days\_ago have a weak relationship, which means they are not strongly correlated. The target variables ‘Purchase’ and ‘Spending’ have a negative relationship with ‘last\_update\_days\_ago’, which shows that the higher number in 'last\_update\_days\_ago' has less chance or no purchase and spending. Last\_update\_days\_ago and X1st\_Update\_days\_ago have highly strong relationship with each other which may affect on modelling.

A screenshot of a computer screen

Description automatically generated

<Figure 3.0.1: Correlation matrix plot >

From the matrix and pairwise plots (Figure 3.0.2), for the predictors, more than 10 frequencies (10-time transactions) show more than $1,000 in spending, which shows a strong correlation between spending and frequency. Also, more than 5 frequencies show that customers purchased after receiving mail, and frequencies between 0 and 5 show that some customers made transactions but did not purchase through mail. These transactions might be offline.



<Figure 3.0.2: Pairwise Plot >

The scatter plot of 1st update days ago and last update days ago vs. spending (figure 3.0.3) shows that the majority of customers who spent more than $1000 made the first update or interaction between 3000 and 4000 days ago, and the last interaction or update happened between 0 and 1000 days ago, which means they might like the game or software and again make purchases.



<Figure 3.0.3: Last Update Days Ago & X1st Update Days Ago Vs Spending >

The scatter plot of spending vs. frequency (Figure 3.0.4) shows that less frequency means less spending, except for some customers. Customers who purchase more than 10 times spend more than $1000 on games and software.

A graph with black dots and white text

Description automatically generated

<Figure 3.0.4: Frequency Vs Spending >

The box plot of spending by web order (figure 3.0.5) shows less spending by the customers through the web order than the amount of spending by the customers through the other type (e.g., offline).

**A graph with numbers and lines

Description automatically generated with medium confidence**

<Figure 3.0.5: Spending By Web Order >

The boxplot of spending by the purchaser (Figure 3.0.6) shows that there is one customer who didn’t make any purchases but spent an amount, which might be anomalies or offline transaction. The boxplot of spending by gender (male) shows that males spend more money than females.

|  |  |
| --- | --- |
| A diagram of a purchase  Description automatically generated | A graph with numbers and symbols  Description automatically generated |

<Figure 3.0.6: Spending By Purchaser & Gender= Male >

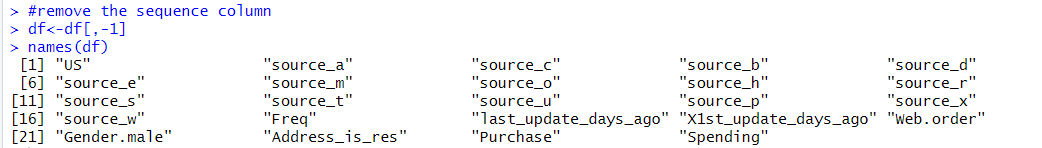
Figure 3.0.7 shows the boxplots of spending by US address and US residence, which display the highest amount of money spent by customers who live in the US and who are not residents of the US.

|  |  |
| --- | --- |
| A graph of a number of people  Description automatically generated with medium confidence | A graph of a number of people |

<Figure 3.0.7: Spending By US Address and US residence>

# 4.0 Dimension Reduction

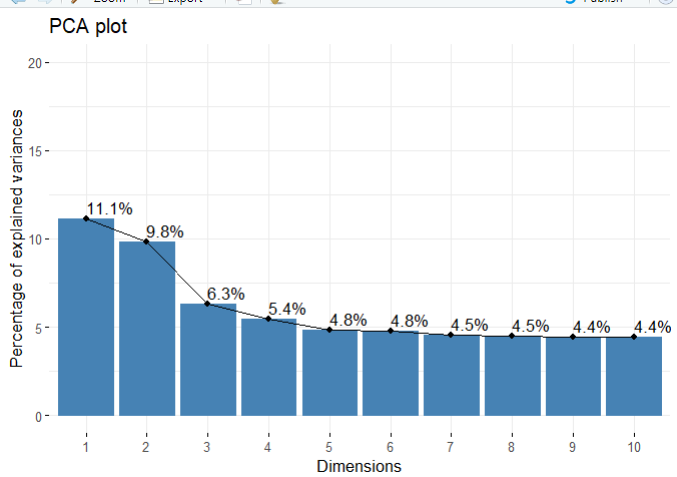
In the dataset, all the predictors are important for the classification and regression models. However, there is a predictor named sequence\_number, which is removed from the data frame using **df<-df[,-1] ( Figure: 4.0.1)** because this is a unique identifier, which is not useful for the prediction model. All the sources are relevant to the customer’s purchases. This dataset is a balanced dataset with 1000 purchasers and 1000 non-purchasers.



<Figure 4.0.1: Removing the Sequence Column >

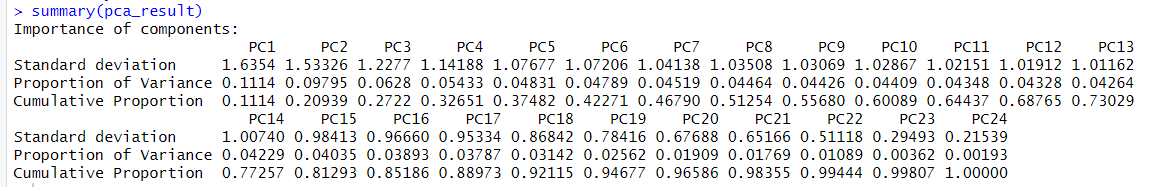
## PCA analysis:

Figure 4.0.2 shows the PCA plot, which shows how much variance is captured by the dimension. Dimension 1 shows the highest variance captured, which accounts for 11.1% of the cumulative variance.



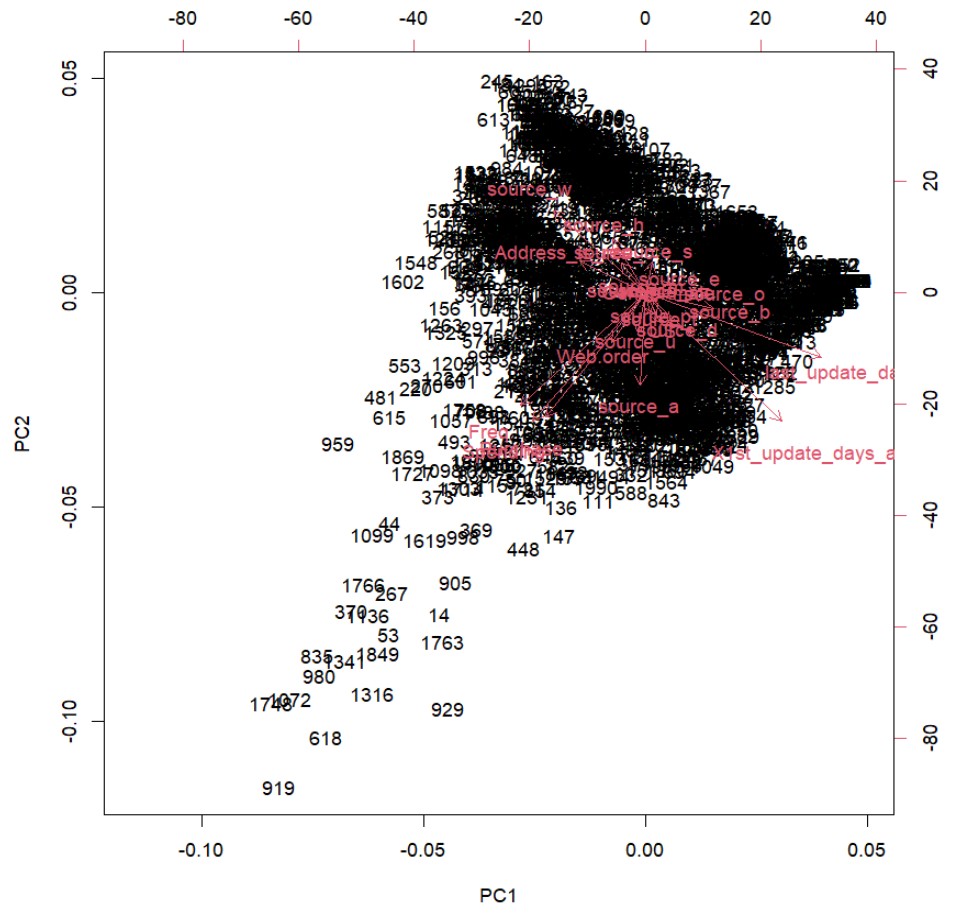
<Figure 4.0.2: Variance Captured by Dimensions>

Figure 4.0.3 shows the standard deviation, proportion of variance, and cumulative proportion. The first two PCs (PC1 and PC2) together account for 20.94% of the variance, the first three PCs account for 27.22%, and so on, according to the cumulative proportion. Principal component (PC23 and PC24) doesn't do much to explain the variance in the data, based on its incredibly small standard deviation and proportion of variance.



<Figure 4.0.3: Summary of PCA result>

The biplot (Figure: 4.0.4) shows the variability of the predictors, all the variables look in a similar space which shows that there are no clusters in the data set.



<Figure 4.0.4: Variability of the predictors>

# 5.0 Data partitioning method

In the data partitioning, the data is divided into training (40%), validation (35%), and holdout (25%). In the training data, there are 800 random observations, which are going to be used to train the model. In the validation data, 700 random observations are not included in the training data used to evaluate model performance. In holdout data, 500 observations are not in training data, and validation data is used to evaluate the performance of new unseen data. Holdout data is fresh data or new data for the model and never seen or used by the model, which will use as how the model will works on the pool data.

So, the company can target customers who are classified as “purchaser” to send a mail.

Data Partition:

First, the target variable “Purchase” is converted into a factor, with 0 as non-purchaser and 1 as purchaser. For the classification, the spending column is removed as per business purpose to select a customer who is a “purchaser” or “non-purchaser”.

First six raw of training, validation, and holdout data:

|  |
| --- |
|  |
|  |

<Figure 5.0.1: Data Partition>

# 6.0 Classifier Models Selection and Model Performance:

For the model selection, the below model will be used to predict class, whether customers make a purchase or not.

* Logistic regression model
* Stepwise regression model
* K-NN model
* Tree model (default tree and deep tree)
* Naïve Bayes model

## 6.1 Logistic regression model:

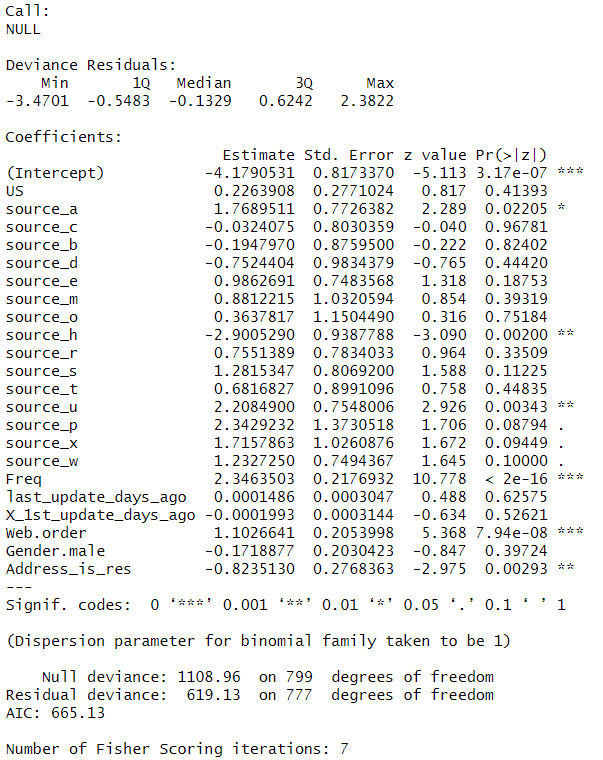
The logistic regression model is a supervised algorithm used for classification. In the dataset, target variable is “Purchase”, the logistic regression model gives the probability of customer purchase or not. First, the model is build using the ‘glm’ function on the training dataset.

From the below model (Figure : 6.1.1) , logistic regression equations:

++…+

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

If the logit value is greater than 0 then success is more likely to happen. The positive coefficient shows the higher predictor value is related to a high chance of getting a purchase. In the dataset, higher frequency means a high chance of purchase. The model output in figure 6.1.1 shows the intercept and coefficient values for the predictors in the model. Also, star marked variable shows the importance of predictors in the model.



<Figure 6.1.1: Summary of the model>

## 6.2 Evaluating the logistic regression model Performance:

When it comes to model performance, the logistic regression model gives an 82.5% accuracy rate with a sensitivity of 0.8207 and a specificity of 0.8292 for the training model. While testing this model on new unseen data (validation data), it correctly identified the class with an accuracy of 79.29%. For business purposes, a logistic regression model can classify around **73.83%** of the positive class (Purchaser) through this predictor information and select top customers among them who are more likely to purchase and send them mail. Figure 6.2.1 shows the logistic regression model performance on training and validation data.

|  |  |
| --- | --- |
|  |  |

<Figure 6.2.1: Confusion Matrix of Logistic Regression Model>

## 6.3 Stepwise Regression Model:

The figure 6.3.1 shows the summary of the stepwise backward and figure 6.3.2 shows the summary of forward models, the AIC (Akaike Information Criterion) of the backward model is 651.28 and the forward model is 649.46. The summary also shows the coefficient values for all the predictors, which shows the importance of the predictors in the model. Also, it gives information about the changes in the model; for example, a high frequency shows a high chance of getting a positive class (purchaser). Stars indicate significant predictors in the model. Here, the low AIC shows the better performance of the model.

**A screenshot of a computer

Description automatically generated**

<Figure 6.3.1: Summary of Stepwise Backward Regression Model>

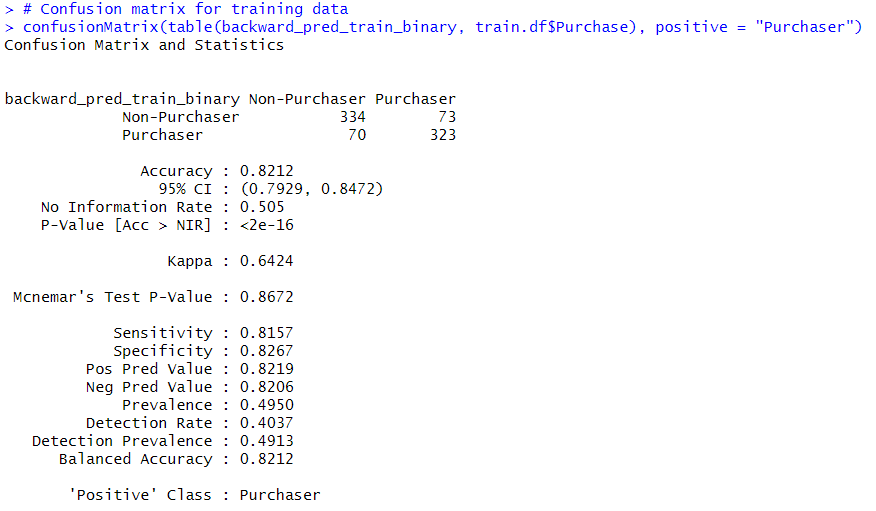
A screenshot of a computer

Description automatically generated

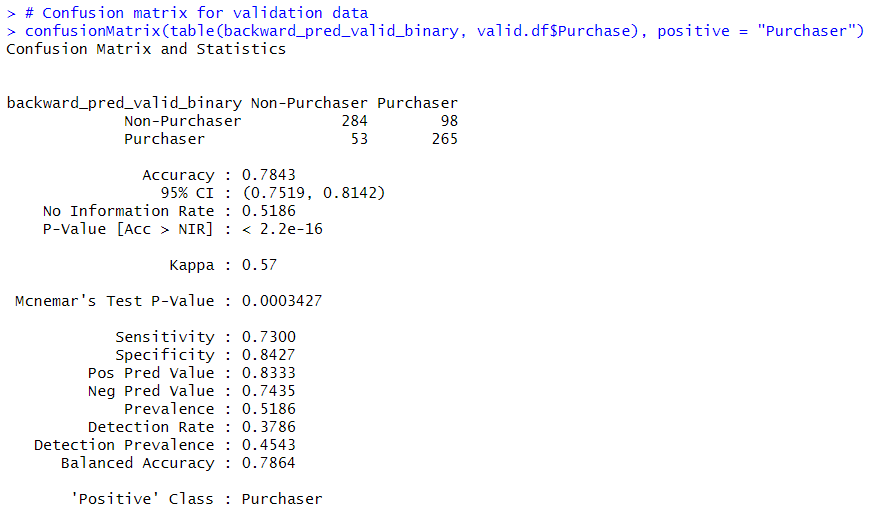
<Figure 6.3.2: Summary of Stepwise Forward Regression Model>

## 6.4 Evaluating the Stepwise Backward model Performance.

With regards to backward regression, the confusion matrix (Figure: 6.4.1) shows that the model performs 82.12% accurately on training data and 78.43% on validation data (Figure: 6.4.2). For validation data, 284 non-purchasers and 265 purchasers were correctly classified, whereas 98 non-purchasers and 53 purchasers were misclassified. As for business purposes, the backward regression model correctly identifies around 73% of the positive class from pool data, which helps the business owner decide which top customers from the pool to send mail to.



<Figure 6.4.1: Confusion matrix for Training data>



<Figure 6.4.2: Confusion Matrix for Validation Data>

## 6.5 Evaluating the Stepwise Forward model Performance.

With regards to forward regression, the below confusion matrix gives the accuracy of training data (Figure: 6.5.1) as 82.38% and validation data (Figure: 6.5.2) as 78.71%. Looking on the validation data, the model identified 281 non-purchasers and 270 purchasers correctly classified. However, 56 purchasers and 93 non-purchasers were misclassified. Businesses can use this model to classify the purchaser, which performs around 74.38% accurately classify positive class on new data.

A screenshot of a computer

Description automatically generated

<Figure 6.5.1: Confusion Matrix for Training data>

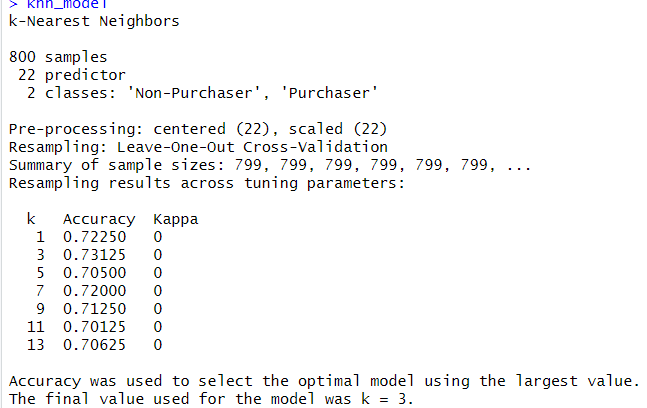
A screenshot of a computer

Description automatically generated

<Figure 6.5.2: Confusion Matrix for Validation data>

## 6.6 K-NN Model

The K-NN model is used for both classification and regression. In classification, the K-NN model predicts the new class based on the majority of its neighbour’s class. First, check the best k value for the model using the cross-validation method. From figure 29, k = 3 gives a more accurate model than others.



<Figure 6.6.1: Best K Values in K-NN model>

## 6.7 Evaluating K-NN model performance.

Using k = 3, the k-nn model classified accuracy of 85.38 on training data (Figure 6.7.1) and 72.29% on validation data (Figure 6.7.1). For validation data, 267 non-purchasers and 239 purchasers were correctly classified, whereas 124 non-purchasers and 70 purchasers were misclassified. A sensitivity for validation shows that 65.84% of the positive class ‘Purchaser’ were correctly identified, and a specificity for validation data shows that 79.23% of the negative class “Non-Purchaser” were correctly identified.

|  |  |
| --- | --- |
|  |  |

<Figure 6.7.1: Confusion Matrix for K-NN model on training and Validation data>

## 6.8 Tree Model

From the default classification tree, if the frequency is less than 1, then it will be classified as a non-purchaser. Also, there are other predictors from tree shows about the class chances.

****

<Figure 6.8.1: Default Classification Tree>

## 6.9 Evaluating Tree model Performance:

About the default tree model, it is 84.38% accurate on training data (Figure: 6.9.1) and 77% accurate on validation data (Figure: 6.9.2). To classify new data, the default tree classified 75.48% correctly identifies the new class.

**A screenshot of a computer

Description automatically generated**

<Figure 6.9.1: Confusion Matrix On training data>

**A screenshot of a computer

Description automatically generated**

<Figure 6.9.2: Confusion Matrix On Validation Data>

## 6.10 Naïve Bayes Model

A Naïve Bayes model is used for classification when there are more predictors. For the dataset, the Naïve Bayes model is built by loading the library library (e1071), which is used for statistical analysis. Then train the model using a training dataset, and it gives an output as a conditional probability for each predictor. Also, Figure 6.10.1 shows the output of a priori probabilities: there is a 50.5% chance of getting a non-purchaser and a 49.5% chance of getting a purchaser. The conditional probabilities for each predictor show the probability that the predictor belongs to a particular class. Then the model predicts the results using these probabilities and gives an output for the chance of the target variable.

.**A white background with black text

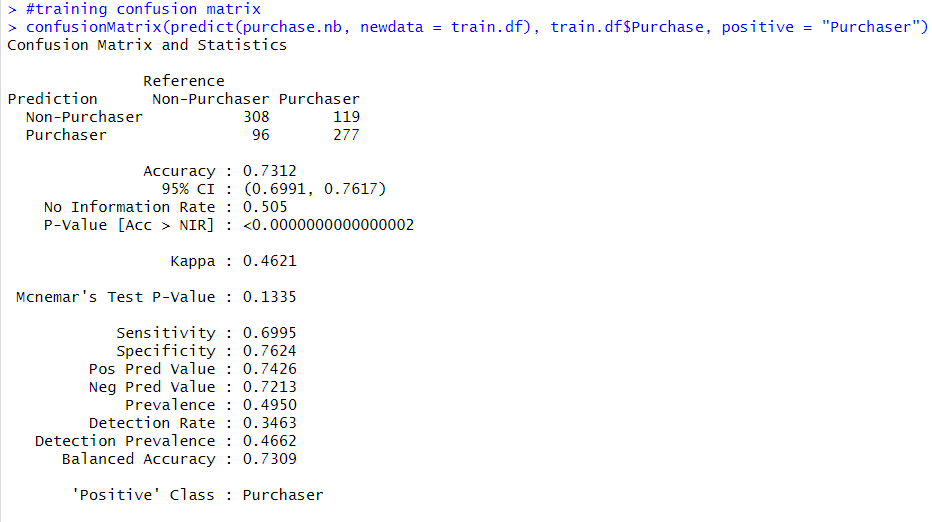
Description automatically generated**



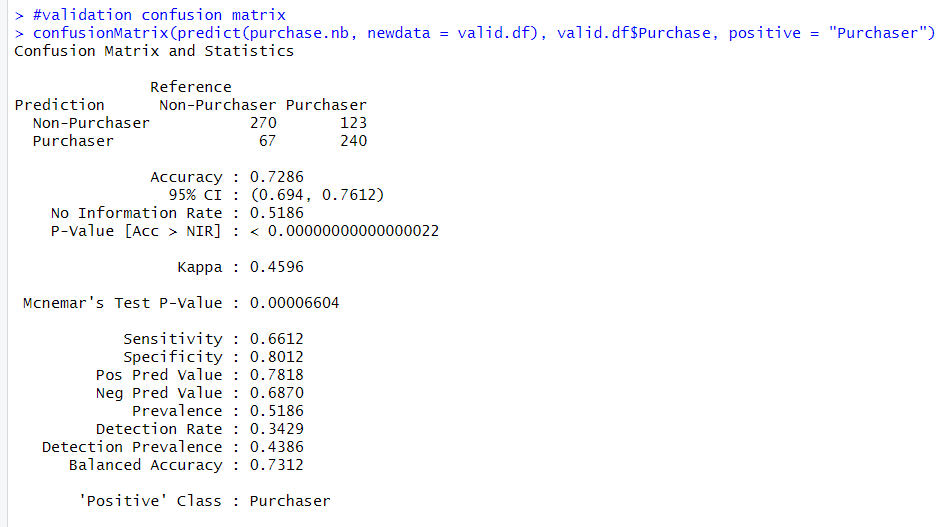
<Figure 6.10.1: Summary of Naïve Model>

## 6.11 Evaluating Naïve Model Performance.

The model gives 73.12% accuracy on training data (Figure: 6.11.1) and 72.86% accuracy on new validation data (Figure: 6.11.2). 270 non-purchasers were correctly identified as non-purchasers, and 240 purchasers were correctly identified as purchasers in the validation data.

.****

<Figure 6.11.1: Confusion Matrix on Training Data>

****

<Figure 6.11.2: Confusion Matrix On Validation Data>

# 7.0 Best Classifier Model Selection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Training Accuracy** | **Validation Accuracy** | **Validation**  **Sensitivity** | **Validation**  **Specificity** |
| **Logistic Regression** | 0.825 | 0.7929 | 0.7383 | 0.8516 |
| **Stepwise Backward** | 0.8212 | 0.7843 | 0.73 | 0.8427 |
| **Stepwise Forward** | 0.8238 | 0.7871 | 0.7438 | 0.8338 |
| **K-NN** | 0.8538 | 0.7229 | 0.6584 | 0.7923 |
| **Default Tree** | 0.8438 | 0.77 | 0.7548 | 0.7864 |
| **Naïve Bayes** | 0.7312 | 0.7286 | 0.6612 | 0.8012 |

<Figure 7.0.1: Comparison of Classification Models >

Among all the classification models, logistic regression performance is better than that of other models on the validation dataset with an accuracy of 79.29% and Sensitivity of 0.7383. Using this model, the probability of purchaser will be calculated, which gives the probability of customer likely to be a purchaser after receiving mail.

# 8.0 Regression Model Selection

For the model selection, the below model will be used to predict the spending amount for the customer who made a purchase after receiving a mail.

* Linear Regression model
* Stepwise- Backward model
* Regression tree

Before the modelling process, only purchaser-based partitions will be used for the training and validation datasets. The “sequence” and “purchase” will be removed for modelling. So, there are 396 records and 23 columns in the training dataset, and the validation data set has 363 records and 23 columns.

A close up of numbers

Description automatically generated

## 8.1 Linear Regression Model

Using the regression function lm, the model is trained using a training dataset. The targeted variable, spending, can be predicted using the intercept and coefficient values for the predictors. The model output (Figure: 8.1.1) marked the stars as the important predictors. The minimum residual is -425.52, and the maximum residual is 1062.70. The positive coefficient of the predictors shows the more amount of spending on games and software. The negative coefficient shows the less amount of spending on games and software.

**A screenshot of a computer

Description automatically generated**

<Figure 8.1.1: Summary of the Regression Model>

## 8.2 Evaluating Regression Model Performance

From figure 8.2.1, the model gave the RMSE (Root Mean Square Error) of 167.2498. which shows the predictor deviates from the actual value is around 167 units.

**A close up of numbers

Description automatically generated**

<Figure 8.2.1: Accuracy for Validation>

## 8.3 Stepwise- backward Regression

In the stepwise-backward regression model, check for all the important predictors by removing less effective predictors and training the model using the training dataset. The stepwise-backward model summary (Figure: 8.3.1) shows that five variables give better performance on the training dataset. The model marked stars as the important predictors.**A screenshot of a computer program

Description automatically generated**

<Figure 8.3.1: Summary of Stepwise Backward Regression Model>

## 8.4 Evaluating Stepwise Backward Regression Model Performance

With regards to accuracy for validation data (Figure 8.4.1), the model gave 165.6397 RMSE on new unseen data. Also, accuracy parameters ME (mean error) shows the mean average value (-1.618712) for the validation dataset.

****

<Figure 8.4.1: Accuracy for Validation>

## 8.5 Regression tree

The regression tree model is built by using rpart() function on the training dataset. The tree plot (Figure: 8.5.1) shows tree rules for the model. For example, if frequency is less than 3 than the average spending is 148 which 70% accurate.

****

<Figure 8.5.1: Regression tree plot>

## 8.6 Evaluating Regression tree model Performance.

RMSE vale of regression tree model for validation data (Figure: 8.6.1) is 181.0081, which shows the predictor deviates from the actual value is around 181 units.

**A close up of a number

Description automatically generated**

<Figure 8.6.1: Accuracy for Validation>

# 9.0 Best Regression Model Selection

|  |  |
| --- | --- |
|  | **RMSE** |
| **Linear Regression Model** | 167.2498 |
| **Stepwise Backward Regression** | 165.6397 |
| **Regression tree** | 181.0081 |

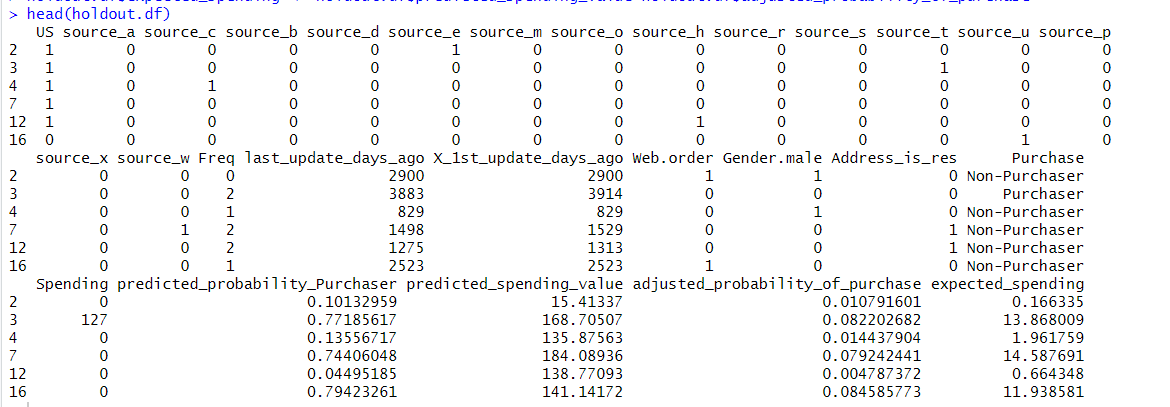
<Figure 9.0.1: Comparison of regression Models >

Among all the linear models, stepwise backward regression performs better on the validation dataset, with an RMSE of 165.6397. Using this model, the predicted spending value will be by the customers. It provides the estimated spending amount that customers are going to spend on software and games.

# 10.0 Profit analysis

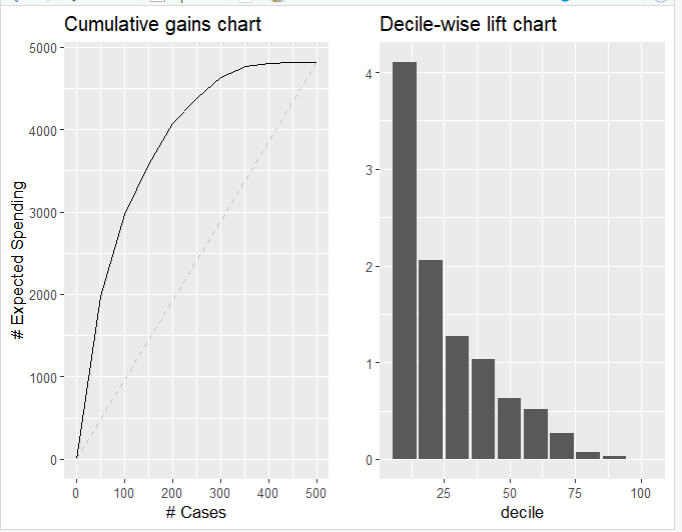
For modeling and performance, logistic regression model will be used for classification and stepwise backward regression for prediction. These models will be used for the holdout data, which was not used before by these models. For the modelling, the dataset has 2000 customer details, which includes 1000 purchasers and 1000 non-purchasers, which means the response rate is 50%. However, the actual response rate is 0.1065. So, by adjusting the probability of the purchaser and their expected spending amount with an actual response rate of 0.1065.

First, the column of the probability of the purchaser is added using a linear regression model, which shows the chance that the customer is likely to make purchases after receiving the mail. Also, another column with the predicted spending value of the customer is added using a stepwise backward model. The probability of a purchaser is adjusted by multiplying it by the original purchase rate of 0.1065. For the adjusted spending value, multiply the predicted spending values by the adjusted probabilities of purchase. Figure 10.0.1 shows the output of these processes, which includes the first six rows of the holdout data.



<Figure 10.0.1: Model Performance on Holdout data >

A cumulative gain chart and a decile-wise lift chart are used to see the model's performance on new, unseen data. These models are compared with the random model, which showed a baseline. The cumulative gain chart (Figure 10.0.2) shows the expected spending by the top customers. For example, here it shows that the North Point company could expect around $2000 in spending from the top 50 customers out of 500 customers. Also, a decile-wise lift chart shows how much a predictive model performs as compared to a random model. From figure 10.0.2, a lift of around 4 shows that the predictive model can identify the top 10% of 500 customers four times better than randomly selected customers.



<Figure 10.0.2: Cumulative Gains Chart and Decile wise lift chart>

## 10.1 Gross profit

So, Holdout data is used to calculate the expected spending value for new customers. The sum of the expected spending for 500 customers is **$4815.239**. Each list booklet costs approximately $2 to mail, which includes printing, postage, and mailing other mailing costs. The company need to spend $360,000 to send the mail to 180,000 customers. The company could expect the spending amount of $1,733,486 from remaining180,000 customers. So, the gross profit is **$1,373,486** that firm could expect from the remaining 180,000 customers.

Figure 10.1.1 shows the calculation of this process.

A screenshot of a computer program

Description automatically generated

<Figure 10.1.1: Gross Profit Calculation>

# 11.0 Conclusion

In conclusion, the company’s main customers come from the US, and male customers are more frequent than female customers. When customers buy more software or games again, the company generates more revenue from those customers. For the classification, logistic regression performs better on new data with an accuracy of 0.7929 and sensitivity of 0.7383, while the stepwise backward model performs well on new data with an RMSE of 165.6397. Using these models, the company could expect **$1,373,486** in gross profit from the remaining 180,000 customers. In the recommendation, if the company can select the customers from the 5,000,000, then they should select or ask for contact details of the top-ranked customers by the expected spending, then use them partially, like first using 20,000 customers from the top 180,000 customers, and so on. Also, the company can use first the high probability of purchasers like top 10% and the high expected spending of customers to send the mail. Then, the company can use other marketing strategies for those customers with a lower probability and a lower spending amount. For example, the company can provide some additional discounts on games or some free games or software.