DIMPLE KRISHNA CHINTALA

Webster University

CSDA 6010 DATA ANALYTICS PRACTICUM

SPRING 2024

FINAL PROJECT (Case1)

JP WANG

North-Point Software Production Company

BRIEF CONTENTS

Step 1: Data preprocessing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

Step 2: Predictors Analysis and relevancy . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

Step 3: Dimension Reduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

Step 4: Data partitioning methods. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .25

Step 5: Classifier models selection. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .26

Step 6: Model fitting, validation accuracy, and test accuracy. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .36

Step 7: Report models performance. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48

Step 8: Model evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .51

Step 9: Conclusion. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .59

CONTENTS IN DETAIL

1. Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
2. Objective. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
3. Data preprocessing

1.1 Attributes Definition. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .12

1.2 Data Exploration. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

1.3 Check for missing value. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .15

1. Predictors Analysis and Relevancy
   1. Exploratory Data Analysis (EDA). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .15

Data understanding . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .15

Numeric data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .16

Categorical data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .18

Predictor and outcome . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .20

* 1. Correlation Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

Correlation matrix. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

Pair plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .24

1. Dimension Reduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .24
2. Data partitioning methods. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

Training partition. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

Validation partition. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

Holdout partition. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .25

1. Classifier models selection
   1. Logistic regression model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .26
   2. Improving model performance with forward stepwise regression. . . . . . . . . . . . . . . . . .29
   3. Classification and Regression Trees. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .32
   4. Accuracy comparison. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
2. Model fitting, validation accuracy, and test accuracy
   1. linear regression model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36

6.2 Forward stepwise model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .40

6.3 Regression tree model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43

6.4 Improving Model Performance. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .46

6.5 Accuracy comparison. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47

1. Report models performance. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .48
2. Model evaluation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .51

Cumulative gains chart. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57

Decile-wise lift. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .58

1. Conclusion. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

TABLE OF FIGURE CONTENTS

Fig 1: loading data in R studio . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

Fig 2: checking missing values. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

Fig 3: Analysis of Spending count and bin . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

Fig 4: Analysis of Frequency count and bin. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .17

Fig 5: Analysis of Last\_Update\_days\_ago count and bin. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .17

Fig 6: Analysis of 1st\_update\_days\_ago count and bin. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .18

Fig 7: Purchase and nonpurchase distributions. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

Fig 8: 15 Source variables distribution. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .19

Fig 9: Binary variables. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

Fig 10: Frequency of purchase. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

Fig 11: spending range of web orders . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .21

Fig 12: Purchase with other variable analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .22

Fig 13: spending with gender and address is res . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .22

Fig 14: Correlation matrix. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .23

Fig 15: Pair plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

Fig 16: Fig 16: Partition data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .25

Fig 17: Fitting logistic regression model to training data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26

Fig 18: logistic regression model for purchase (Training data) . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . 27

Fig 19: confusion matrix. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

Fig 20: fitting a forward regression model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .29

Fig 21: Step-by-step procedure for forward stepwise model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .30

Fig 22: Summary of forward stepwise model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .31

Fig 23: confusion matrix of forward stepwise model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

Fig 24: Purchase with other variable analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33

Fig 25: Tree structure for training data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .34

Fig 26: predicting data with validation partition. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

Fig 27: calculating accuracy for the model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .34

Fig 28. Taking only purchase data out of train data and validation data . . . . . . . . . . . . . . . . . . . . . . .37

Fig 29: Histogram for spending frequency . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .38

Fig 30: Linear regression model for Train data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .38

Fig 31: Summary of linear regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .39

Fig 32: predicting value with validation data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .40

Fig 33: Accuracy of forecasting models. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .40

Fig 34: Stepwise forward regression. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

Fig 35: Summary of linear regression forward stepwise model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

Fig 36: Generating prediction and forecast measure for forward stepwise . . . . . . . . . . . . . . . . . . . . 42

Fig 37: decision tree model fitting for training data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .44

Fig 38: Tree structure for training data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44

Fig 39: Summary of predicted values. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .44

Fig 40: calculating accuracy for the model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

Fig 41: Improving model with M5P. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46

Fig 42: summary of the M5P model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46

Fig 43: correlation and forecast of the model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .46

Fig 44: Evaluating model performance for validation data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52

Fig 45: Evaluating model performance for train data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .52

Fig 46: comparing training data with validation data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .53

Fig 47: adding predicted probability column. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .55

Fig 48: Adding another column with the predicted spending. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

Fig 49: defining original purchase rate . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

Fig 50: Adding a column for adjusted probability of purchase . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .56

Fig 51: Adding a column for Expected Spending. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .56

Fig 52: creating cumulative gains chart. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .57

Fig 53: cumulative chat. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .57

Fig 54: Mean response for Decile-wise lift . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .58

Fig 55: Decile-wise lift chat. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .58

Executive summary:

North Point Software Company faces the challenge of maximizing the profitability of its mailing campaign amidst a vast pool of client names. The company aims to use predictive modeling and machine learning to improve how accurately it targets potential customers. They have randomly selected 200,000 names from a pool of 5,000,000 to work with. The primary objective is to boost the response rate of the campaign by identifying potential customers likely to purchase games and educational software. Through predictive modeling, the company aims to estimate customer spending and optimize gross profit while minimizing costs. Logistic regression, renowned for its interpretability and ability to provide insights into customer behaviour, emerges as a promising approach. By evaluating various modeling techniques, including logistic regression, North Point aims to select the most effective model to achieve its overarching business goals of enhancing campaign profitability and targeting efficiency.

1. INTRODUCTION

North Point Software Company sells education and gaming software. Initially, it started as a software manufacturer but later it got tied up with a third party and came up with a new wish list. Apart from software titles, North Point's key assets are customers. In an attempt to expand its customer base, North Point Company has recently joined a consortium listing firms that specialize in computer software and hardware products. The consortium allows the mailing list of names drawn from the pool list of customers. This company has supplied its customer list of 200,000 names to the pool, which totals over 5,000,000 names so they are allowed to pick 200,000 names out of this list for mailing. By using the names members are allowed to do a predictive model on the records in the pool so they can choose a better job of selecting customers from the pool.

1. OBJECTIVE

The North Point software business contributed the names of 200,000 of its clients to a pool of 5,000,000 names, from which 200,000 names were randomly selected. So, they must choose names that will maximize their chances of earning. So, out of the 20,000 records they selected for mail testing, 1065 sales were made, earning a response rate of 0.053 or 5.3%. For the remaining 180,000 names on the list that they wish to test using a machine learning model that provides them with higher performance in profit bases, it is more of a loss. Let us calculate the actual gross profit where their spending is $205,249. Actual investment in mailing for 20,000 customers is $40,000 where each mail costs $2. Actual gross profit is spending minus actual investment which is $205,249 - $40,000 which is $ 165,249. Estimated spending is (205,249/20,000)\*180,000 which is 1,847,241 minus 1,80,000 \* 2 which is a total of 180,000 customers and 2 dollars for mail which is 1,847,241 - 360,000 estimated spending is 1,487,241. Estimated gross profit is estimated sending minus investment which is 1,487,241 - 360,000 which is 1,127,241. The estimated gross profit is 1,127,241. The primary goal could be to maximize the profitability of their mailing campaign by efficiently targeting potential customers who are more likely to make purchases. This involves using predictive modeling and machine learning to identify high-probability purchasers from the customer pool.

*Analytics Goal:*

Predictive modeling for purchase probability involves creating a tool that can predict whether potential customers are likely to buy games and educational software based on their past behavior and personal details. This model looks at data like what products they've bought before, how often they shop, and other characteristics to determine the likelihood of them making a purchase.

On the other hand, estimating customer spending involves building another predictive model that forecasts how much money those likely buyers might spend on games and educational software. This model analyses factors such as past purchase amounts, frequency of purchases, and other relevant data to predict the potential spending of customers. It's something to predict the amount of money someone might spend on groceries based on their past shopping habits and preferences. By understanding how much customers are likely to spend, businesses can tailor their marketing strategies and product offerings to maximize revenue and meet customer needs effectively.

*Analytics Approach:*

In the process of analysing data to understand customer behaviour and predict purchasing patterns, several key steps are undertaken. Initially, data preparation and cleaning involve organizing the dataset to ensure it's suitable for analysis by addressing missing values, outliers, and formatting inconsistencies. Feature Selection and Engineering is to select relevant features (customer attributes) that are likely to impact purchase behavior and spending. Create new features if necessary to improve model performance. Evaluate the performance of the classification model using appropriate metrics such as accuracy, precision, recall, and F1 score. Validate the model using techniques like cross-validation to ensure robustness. Build a regression model to predict the spending amount of customers who are likely to make a purchase. Techniques such as linear regression, ridge regression, or ensemble methods can be used for this purpose. Interpret the results of the predictive models to understand the factors influencing purchase behavior and spending. Deploy the models into the operational environment for real-time predictions and decision-making.

1. DATA PREPROCESSING

*1.1 Attributes Definition:*

Sequence\_number: It represents the unique identifier or index for each record in the dataset.

US: This column says that if the customer is from the United States or out of the United States. This column of a binary variable. Whereas 0 indicates that the customer is not from the United States and 1 indicates that the customer is from the United States.

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: They are 15 columns with random source names and it represents source that customers are acquired. Each column has binary values (0 or 1) indicating whether a particular source was used to acquire the customer.

Freq: The number or frequency of the customer's purchases is shown in this column. It is likely a numeric variable that indicates the count of purchases.

last\_update\_days\_ago: The number of days that have passed since the last update or contact with the customer's data will be shown in this column.

1st\_update\_days\_ago: It can indicate how many days have gone by since the customer's data was updated or interacted with for the first time.

Web order: This column indicates whether the purchase was made through a web order or not. It is a binary variable with 1 representing a web order and 0 representing other types of orders.

Gender=male: This column indicates 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 indicates whether the customer's address is residential or not residential. It is a binary variable with 1 representing a residential address and 0 representing a non-residential address.

Purchase: This variable indicates whether the customer purchased or did not purchase. It is likely a binary variable with values like 0 or 1, where 0 represents no purchase and 1 represents a purchase.

Spending: This variable shows the amount that customers who made a purchase (Purchase = 1) spent. It's probably a numerical variable that represents the purchase's monetary value.

*1.2 Data Exploration:*

In this data set, we have 2 outcome variables, Purchase indicates whether the customer responded to the test mailing and purchased something. Spending indicates those who have purchased and what is their spending amount on it. Rest all other columns have binary and numerical variables. The US variable is a binary column that indicates if they are residents or not. The source has 15 different columns with different sources every individual has a different source it is defined with binary numbers 1 or 2. Similarly, web orders, gender, and address\_is\_residence are binary variables that are represented with 0 or 1 where 0 represents ‘no’ and 1 represents ‘yes’. We have 4 numeric variables which are frequency, last update days ago, 1st update days ago, and spending which has numerical values that indicate the frequency of purchase 1st, and last update of the product and spending on the purchase.

By using R code started loading data into R to explore data.

After loading data to R by using a CSV file. I have used the str () function to see data types. Checked the summary of the data file to calculate the summary of the statistics such as mean, median, standard deviation, minimum, and maximum. This helps to understand the central tendency and variability of the data. Then I checked for missing values and they are no missing values or zeros in the data file. Now data is clean enough to do analysis.

A screenshot of a computer

Description automatically generated

Fig 1: loading data in R studio

*1.3 Check for missing value:*

Typically, some records will contain missing values. If the number of records with missing values is small, those records might be omitted. However, if we have a large number of variables, even a small proportion of missing values can affect a lot of records. So, we will check missing values for the data file.

A close-up of a computer code

Description automatically generated

Fig 2: checking missing values

We don’t have any missing values in the data file. So, we don’t need any data from the data file.

2. Predictors Analysis and Relevancy

*2.1 Exploratory Data Analysis (EDA)*

Exploratory Data Analysis is an approach to analyzing data sets to summarize their main characteristics, often using visual methods. The goal of EDA is to understand the data, detect patterns, identify anomalies, and formulate hypotheses that can lead to further modeling

*Data understanding:*

The binary variable "US" helps to understand customers based on their location within or outside the United States. Meanwhile, the "Source" variables show the effectiveness of different acquisition channels, guiding marketing resource allocation. Frequency of purchases ("Freq") explains to understand customer loyalty and engagement levels, while the "Last\_update\_days\_ago" and "1st\_update\_days\_ago" variables provide customer interactions. The "Web\_order" indicator discerns between web-based and other purchase avenues, informing e-commerce optimization efforts. Gender ("Gender\_male") says weather it's male or female and residential status ("Address\_is\_res") provides demographic insights for tailored marketing approaches and logistical considerations, respectively. Finally, the "Purchase" and "Spending" variables encapsulate customer transaction behavior and their purchase expenses.

*Numeric data:*

Numeric data refers to data that consists of numerical values or measurements. In other words, it includes any data that can be represented or measured in numeric form, such as integers or real numbers. Numeric data can be continuous, like spending, which can take on any value within a range. It can also be discrete, such as counts of items sold, number of website visitors, or ratings on a scale. Numeric data is often used in statistical analysis, mathematical modeling, and data visualization to derive insights, make predictions, and support decision-making processes. We can see an analysis for numeric columns.

A graph with numbers and a bar graph

Description automatically generated

Fig 3: Analysis of Spending count and bin

In Fig 3, we can see the spending count and spending amount in that scale. For example, 1,174 records have been spent between 0 to 50 range similarly, all other records.

A graph with a bar graph

Description automatically generated

Fig 4: Analysis of Frequency count and bin

In Fig 4, we can see the frequency range and count, we can see that from range 0 to 2 have more records compared to other frequencies.

A graph of a number of bars

Description automatically generated with medium confidence

Fig 5: Analysis of Last\_Update\_days\_ago count and bin

In Fig 5, we can see the last update days range and count of it.

A graph of a number of bars

Description automatically generated with medium confidence

Fig 6: Analysis of 1st\_update\_days\_ago count and bin

In Fig 6, we can see the range and count of the 1st\_update\_days\_ago.

*Categorical data:*

Categorical data is a type of data that represents categories or groups. Unlike numerical data, which consists of numerical values that can be measured or counted, categorical data consists of labels or names that represent different groups or levels. The distribution of categorical data refers to how the different categories are spread out or distributed across the dataset. This distribution can be summarized using frequencies or proportions.

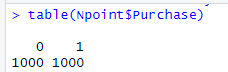


Fig 7: Purchase and nonpurchase distributions

In Fig 7, we can see that there are an equal number of records for purchases and non-purchases.

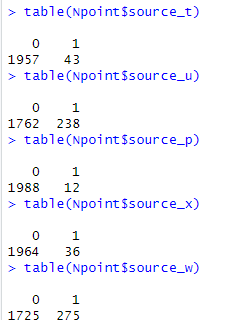
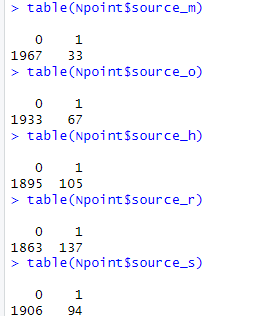
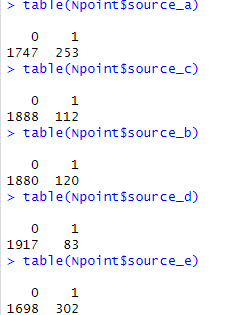


Fig 8: 15 Source variables distribution

In Fig 8, we can see the distribution of customers that are using that particular source or not using that source from that we can see that most of them are using source\_e.

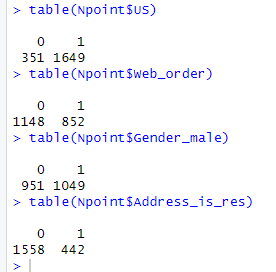


Fig 9: Binary variables

In Fig 9, we can see the binary variables and their distribution for the 1st binary variable US we can say that most of them are from the US. Similarly, most of them are not using web orders, male are more in these records and non-residents are more in these records.

*Predictor and outcome:*

So far, we have done univariant analysis, and now we will do bivariant analysis for the data with target columns. For bivariant analysis, I have done a few in R code and a few in Tableau for better understanding. Now let’s see the frequency for the purchase column. Create a Contingency table that helps to display the frequency distribution of two or more categorical variables by using the table () function.

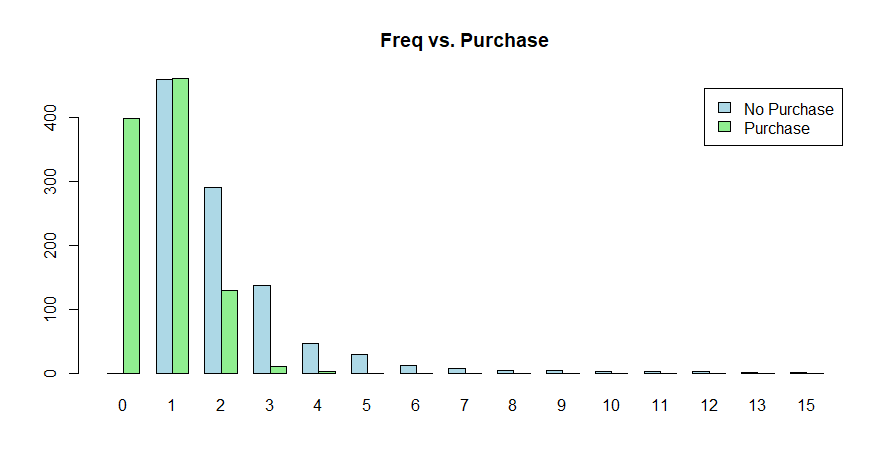


Fig 10: Frequency of purchase

In Fig 10, We can see the frequency of the people who have purchased the product. We can see there is a purchase at frequency range 0, which means there is some purchase happening from other sources Similarly with all other frequencies we can understand the purchase range and non-purchase range.

To see the spending range of web orders let us create a boxplot visualization of the spending range based on whether the purchase was made through a web order or not.

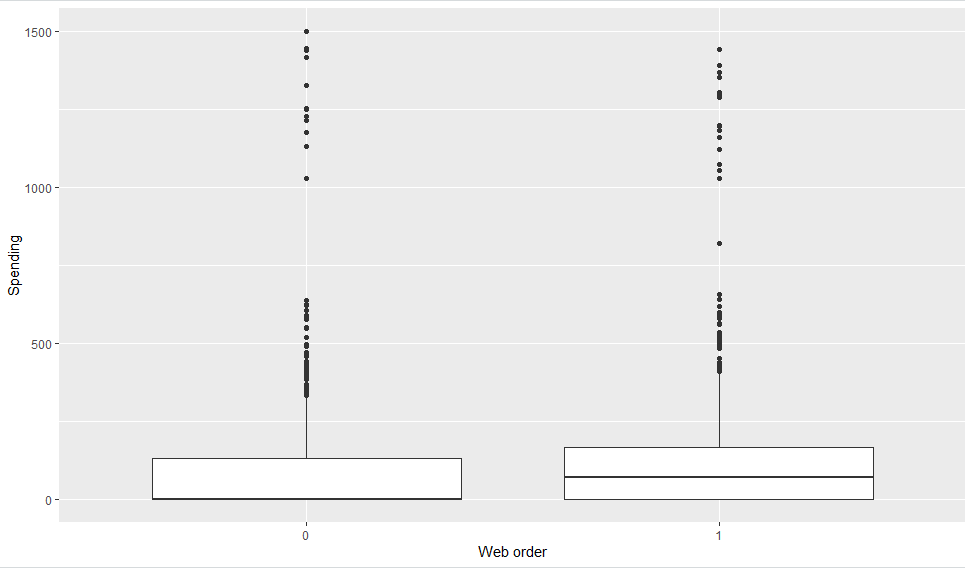


Fig 11: spending range of web orders

In Fig 11, Each boxplot represents the distribution of spending values (y) based on the categorical variable Web\_order (x). The spending values are shown in the y-axis, while the x-axis represents the categories of the Web\_order variable. The boxplot is a useful visualization for understanding the distribution and spread of a continuous variable (in this case, spending) across different categories (web order vs. non-web order).

Using Tableau, create bar chats that purchase and spending variables with other variables to analyze data.

*A screenshot of a graph

Description automatically generated*



Fig 12: Purchase with other variable analysis.

In Fig 12, we can see the purchase of sources and find out which source is very high in the purchase also, we can see how many people from the US are purchasing the products and how many are not purchasing. Source\_H, source\_P, and source\_O have the highest purchases among all other sources similarly non-purchasers are more in Source\_U, source\_w, and source\_A compared to all other sources. People who are not from the US have purchased more.

We can also check spending who is spending more and which gender is spending more and residence and non-residence.

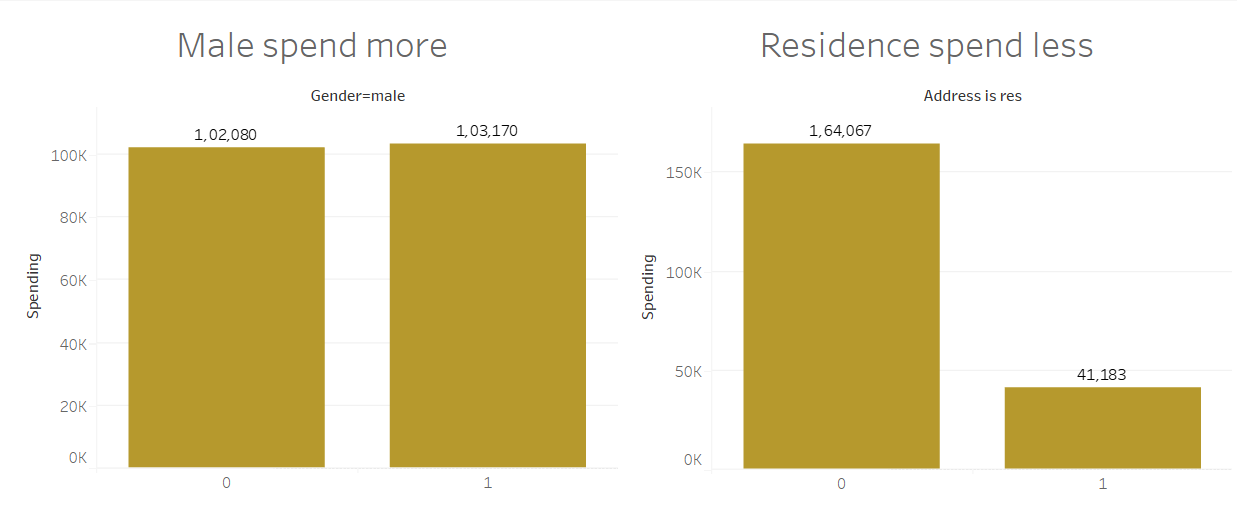


Fig 13: spending with gender and address is res

In Fig 13, we can see that gender males have spent more on their spending and non-residents spent more on their purchases.

2.2 Correlation Analysis:

*Correlation matrix:*

The correlation matrix provides insights into the relationships between pairs of numeric variables in our dataset. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to 1. The correlation matrix identifies patterns and relationships between different numeric variables in our dataset, which can be useful for exploratory data analysis and feature selection in predictive modeling tasks.

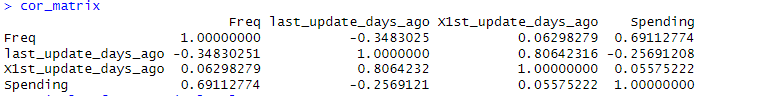


Fig 14: Correlation matrix

From Fig 14, There is a strong positive correlation between Freq and Spending (0.6911), indicating that as the frequency of some events increases, the spending tends to increase as well. There is a strong positive correlation between last\_update\_days\_ago and X1st\_update\_days\_ago (0.8064), suggesting that these two variables are closely related. The correlation between last\_update\_days\_ago and Spending (-0.2569) is negative, indicating that as the number of days since the last update increases, spending tends to decrease, although the correlation is not very strong.

*Pair plot:*

Also, let’s check data with pair plots because that allows us to visualize the relationship between pairs of numerical variables.

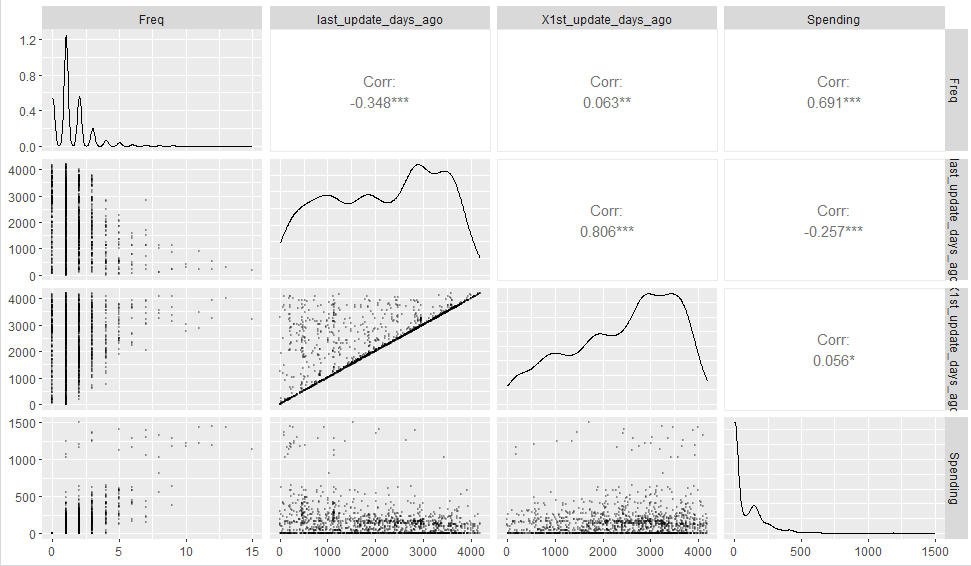


Fig 15: Pair plot.

Now in Fig 15, The pair plot will display scatterplots of each pairwise combination of numeric variables along with histograms of each variable's distribution along the diagonal. This visualization allows us to quickly identify patterns, correlations, and potential outliers in our data.

3. Dimension Reduction:

All the variables in the dataset are related to each other considered all are important variables for the analysis and in the modeling task. If any variables are irrelevant or redundant, then we can remove such columns from the data set. As everything is relevant to each other there is no dimension reduction.

4. Data partitioning methods:

As per the requirements data need to be partitioned into 3 partitions training data, validation data, and holdout data

*Training partition:*

The training partition, typically the largest partition, contains the data to build the various models we are examining. The same training partition is generally used to develop multiple models. We will apply a train data set for model fitting.

*Validation partition:*

The validation partition is used to assess the predictive performance of each model so that we can compare models and choose the best one. In some algorithms, the validation partition may be used in an automated fashion to tune parameters and improve the model.

*Holdout partition:*

The holdout partition is used to assess the performance of the chosen model with the new data. The more models we test, the more likely it is that one of them will be particularly effective in modeling the noise in the validation data. Applying the model to the holdout data, which has not been seen before, will provide an unbiased estimate of how well the model will perform with new data.

A screenshot of a computer code

Description automatically generated

Fig 16: Partition data

In above figure 16 data partition is done in three, training partition to 800 records which is 40% of the data, validation partition to 700 records which is 35% of the data, and holdout partition to 500 records which is 25% of the data. All three partitions have 24 columns, excluding sequence.

5. Classifier model selection:

*5.1 Logistic regression model:*

The main R functions used for Logistic regression are glm in the caret package. We also use package glmnet for variable selection.

Logistic regression extends the idea of linear regression to the situation where the outcome variable Y is categorical. We can think of a categorical variable as diving the records into classes.

For example, if Y denotes a recommendation on a purchase as a product, we have a categorical variable with two categories. We can think of each of the customers in the data set as belonging to one of two classes, purchase or not purchase product.

Logistic regression can be used for classifying a new record, where its class is unknown, into one of the classes, based on the values of its predictor variables (called classification).

Fitting logistic regression model to training data

A screenshot of a computer

Description automatically generated

Fig 17: Fitting logistic regression model to training data

In Figure 17 we are fitting the logistic regression model for the training partition.

*Estimated model:*

Fig 18 presents the output from running a logistic regression using the 24 predictors on the training data.

Ignoring the p-value for the coefficients, a model based on all 24 predictors has an estimated logistic equation.

Y= b0+b1x1+b2x2+……+bpxp

Logit (Purchase = yes) =

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

A screenshot of a computer

Description automatically generated

Fig 18: logistic regression model for purchase (Training data)

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

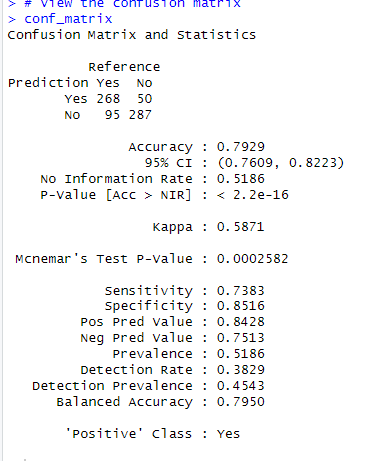


Fig 19: confusion matrix

Accuracy: The overall proportion of correctly predicted instances, is 79.29%. It indicates the model's ability to classify both "Yes" and "No" correctly.

Sensitivity (True Positive Rate): The proportion of actual positive cases (purchases) that were correctly identified by the model, which is 73.83%.

Specificity (True Negative Rate): The proportion of actual negative cases (non-purchases) that were correctly identified by the model, which is 85.16%.

*5.2 Improving model performance with forward stepwise regression:*

Now we can employ the step function to perform forward selection of important features/predictors. It works for both LM and glm models. On the other hand, there are various criteria to evaluate a model. Commonly used criteria include AIC, BIC, Adjusted R2, etc. The step function argument direction allows this control (default is both, which will select the better result from either backward or forward selection).

Forward stepwise selection is a technique used to select a subset of predictor variables that best predict the outcome variable. It starts with an empty model and sequentially adds predictors based on some criterion until no more improvements can be made.

To improve the model with forward stepwise regression, we use step\_null and step\_full to improve the model. step\_null represents a null model with only the intercept, while Step\_full includes all predictors except for Spending. These models can be used as starting points for further model building, such as stepwise selection, where predictors are added or removed based on certain criteria to improve model fit or interpretability.

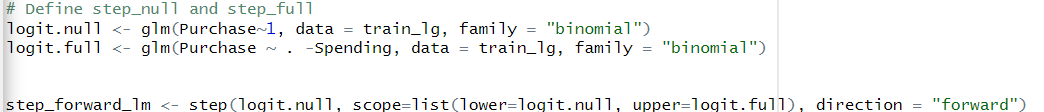
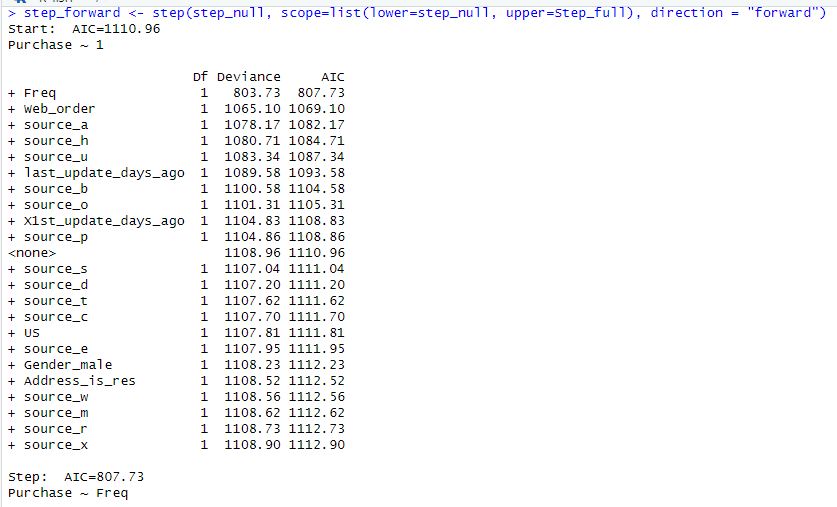
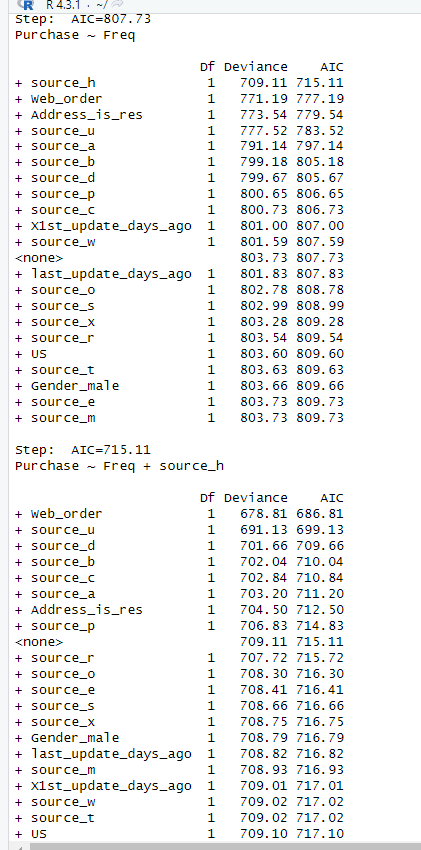
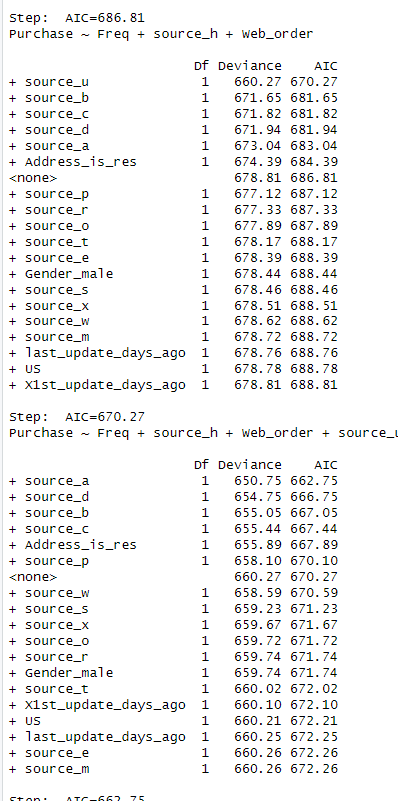
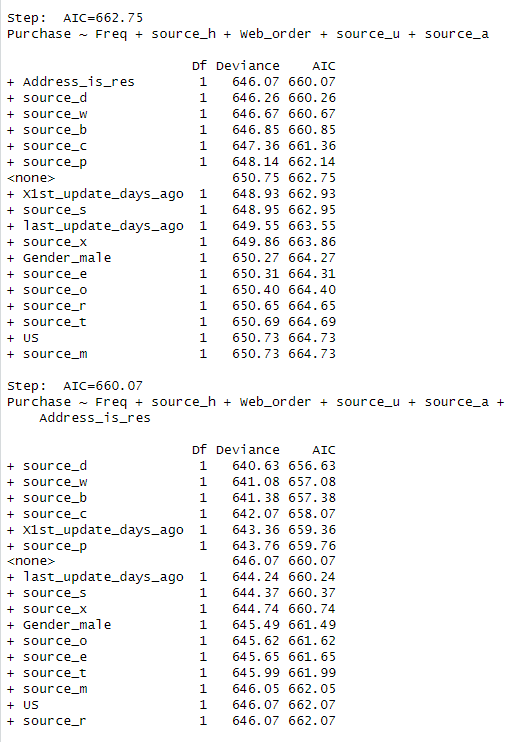


Fig 20: fitting a forward regression model

Model performing step-by-step procedure end their evaluation process it will start with an empty shell and add one after one



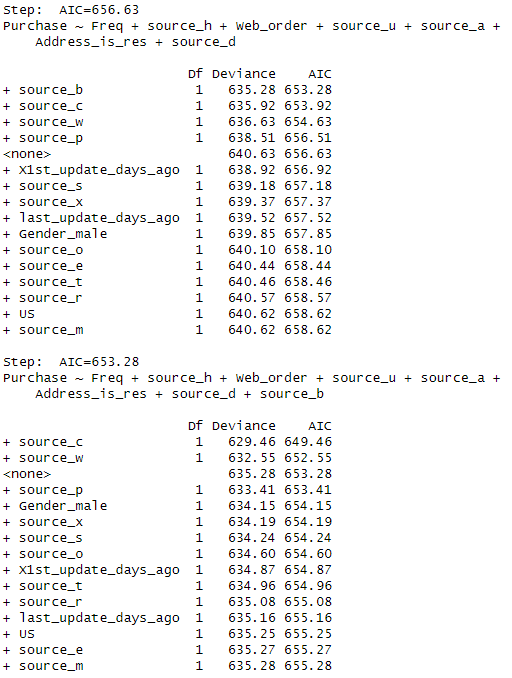
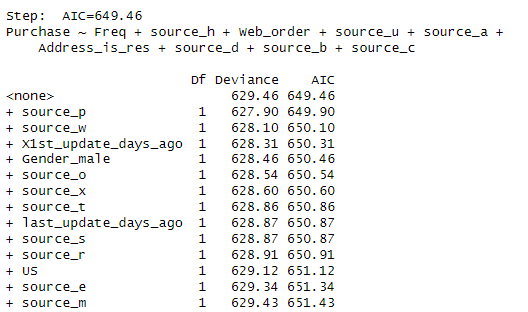
 

Fig 21: Step-by-step procedure for forward stepwise model

Summary for forward stepwise with estimated equation

Purchase = -3.1404 + 2.2955 \*Freq -3.8428\* source\_h + 1.0928\* Web\_order + 1.1544 \* source\_u + 0.7376\* source\_a - 0.7420\* Address\_is\_res -1.7553\* source\_d -1.2149 \* source\_b -0.9808\* source\_c

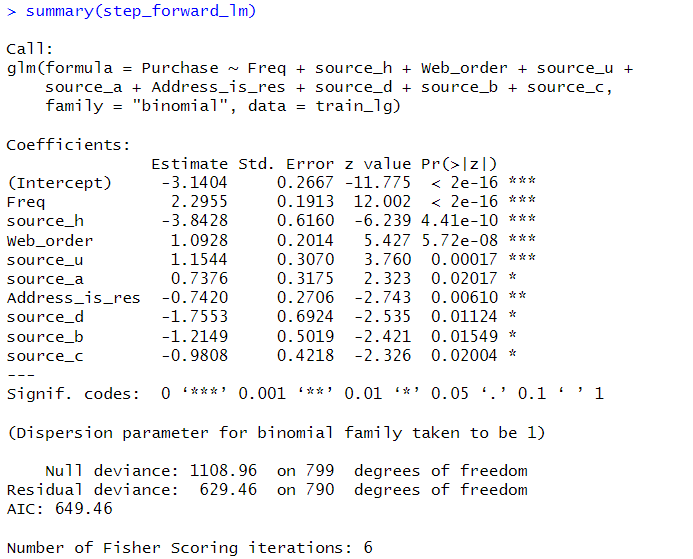


Fig 22: Summary of forward stepwise model

Fitting the model with the validation data.

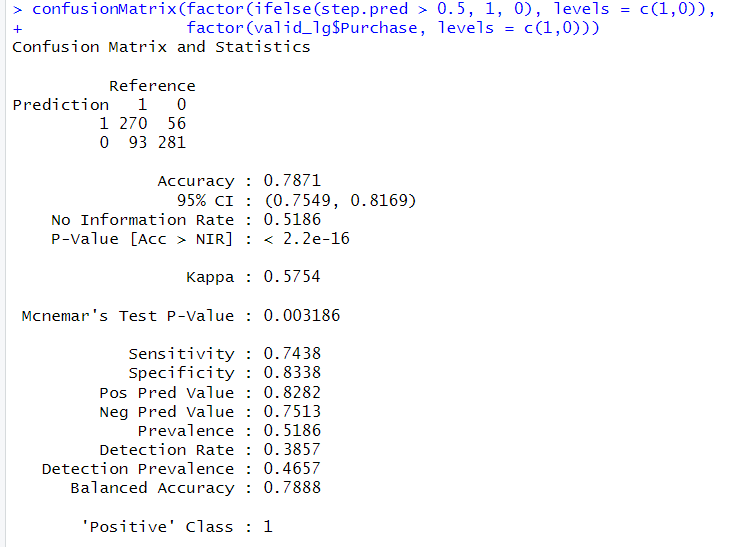


Fig 23: confusion matrix of forward stepwise model

* *Accuracy:* The model's accuracy is 0.7871, indicating that it correctly predicts the outcome approximately 78.71% of the time.
* *Sensitivity:* The proportion of actual positive cases that the model correctly identifies. In this case, it's 0.7438, indicating that the model identifies around 74.38% of actual positive cases (true positives).
* *Specificity:* The proportion of actual negative cases that the model correctly identifies. Here, it's 0.8338, indicating the model correctly identifies around 83.38% of actual negative cases (true negatives).

*5.3 Classification and regression trees:*

In the classification and regression tree model, we use the part and rpart. Plot package in R to build and visualize classification and regression tree. The packages randomForest, xgboost, and adabag and used to train the ensemble model.

If one had to choose a classification technique that performs well across a wide range of situations without requiring much effort from the analyst while being readily understandable by the consumer of the analysis. Trees, also called decision trees, can be used for both classification and prediction. We discuss the classification procedure first, and then in later sections, we show how the procedure can be extended to the prediction of a numerical outcome. A tree for classifying North Point customers who are spending on their purchase or not spending on their purchase, based on information such as their web\_orders, 15 sources, US residence, and gender.

*Tree structure:*

We have two types of nodes in a tree: decision (= splitting) nodes and leaf nodes that have successors are called decision nodes because if we were to use a tree to classify a new record for which we knew only the values of the predictor. we would "drop" the record down the tree so that at each decision node, the appropriate branch is taken until we get to a node that has no successors. Such nodes are called the leaf nodes (or terminal nodes or leaves of the tree) and represent the partitioning of the data by predictors.

It is useful to note that the types of trees grown by R's rpart () function, also known as CART or binary trees, have the property that the number of leaf nodes is exactly one more than the number of decision nodes.

*Fit classification tree for training model:*

A computer screen shot of a computer code

Description automatically generated

Fig 24: model fitting for training data

The fitting model for training data and Fig 24 below show the tree structure for the training data.

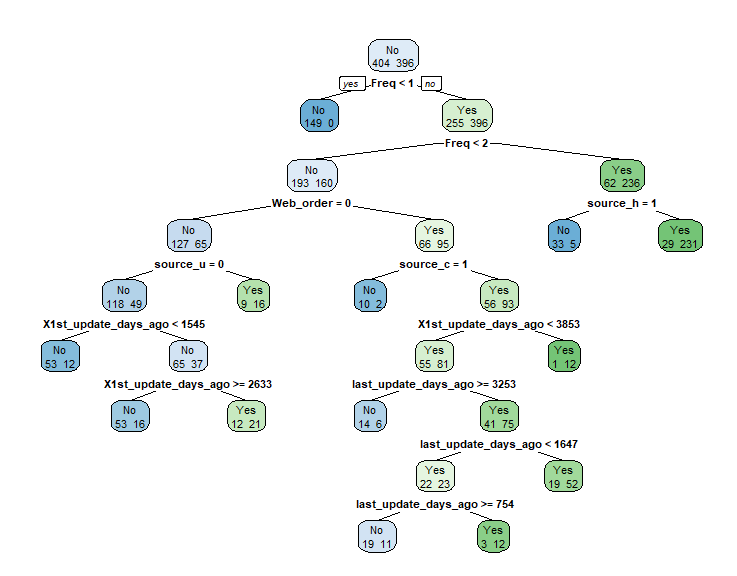


Fig 25: Tree structure for training data.

*Predicting model with validation partition:*

A white background with many numbers

Description automatically generated

Fig 26: predicting data with validation partition

Calculating accuracy for the model

A screenshot of a computer

Description automatically generated

Fig 27: calculating accuracy for the model.

*Accuracy:* The proportion of correctly predicted instances, is *77%*. It indicates the model's ability to classify both "Yes" and "No" correctly.

*Sensitivity (True Positive Rate):* The proportion of actual positive cases (purchases) that were correctly identified by the model, which is *75.48%*.

*Specificity (True Negative Rate):* The proportion of actual negative cases (non-purchases) that were correctly identified by the model, which is *78.64%*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Positive ‘Class |
| Logistic regression | 0.7929 | 0.7383 | 0.8516 | Yes |
| Forward selection logistic regression model | 0.7871 | 0.7438 | 0.8338 | Yes |
| classification tree | 0.77 | 0.7548 | 0.7864 | Yes |

Since the accuracy among the models is similar, it's prudent to focus on sensitivity, especially given the context of our classification task.

*Logistic Regression (0.7383):* This model correctly identifies about 73.83% of actual positive cases (purchases), indicating its ability to capture true positives.

*Forward stepwise regression (0.7438):* This model shows a sensitivity of 74.38%, which is better than logistic regression and indicates its capability to capture true positive instances.

*Classification Tree (0.7548):* This model shows a sensitivity of 75.48%, which is comparable to logistic regression and indicates its capability to capture true positive instances.

Among all the models forward logistic regression is the best model with a sensitivity of 74.38% we are considering this as the best classification model.

6. Model fitting:

*6.1 linear regression model:*

Linear regression is a statistical method used to model the relationship between one or more independent variables (predictors) and a dependent variable (response). It assumes that the relationship between the independent variables and the dependent variable is linear.

In linear regression, the goal is to find the best-fitting linear equation that describes the relationship between the independent variable X and the dependent variable Y. The linear equation can be represented as:

Y=a+bx

Where:

* Y is the dependent variable (response)
* X is the independent variables (predictors)
* a is the y-intercept (constant term)
* b is the coefficients (slope) of the independent variables

The goal of linear regression is to estimate the coefficients of a and b that minimize the sum of squared differences between the observed values of the dependent variable and the values predicted by the linear equation. Linear regression can be performed using various methods, including the ordinary least squares (OLS) method, which minimizes the sum of squared differences, or through other optimization techniques. In R, linear regression can be conducted using the *lm ()* function, which fits a linear model to the data.

The *summary ()* function provides detailed information about the fitted linear regression model, including coefficient estimates, standard errors, t-statistics, p-values, and R-squared values, among others.

As per the business requirements, we have used only purchase data which means those who have purchased the products have 1000 records and partitioned them with 40% train data, 35% validation data, and 25% holdout data. As per the requirement we just need Train data and validation data and we are applying it to this model.

Using already partitioned data from the 4th step and taking only purchase data from it.

A screen shot of a computer

Description automatically generated

A computer screen shot of a person

Description automatically generated

Fig 28. Taking only purchase data out of train data and validation data

Out of 1000 records, we got 396 purchase records from a train data set and 363 purchase records for validation data set.

Let us see the data distribution with the help of a histogram, it provides a visual representation of the distribution of data. By displaying the frequency or count of data points within specific intervals (bins), histograms allow analysts to quickly grasp the shape, central tendency, spread, and skewness of the data.



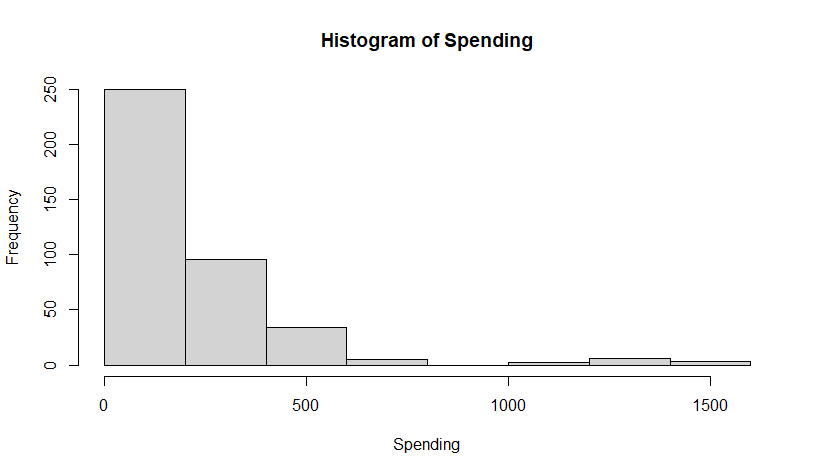


Fig 29: Histogram for spending frequency.

By the histogram, we can see that the majority of the data is part of the left-hand side whereas data like this is known as right screw data.

Applying linear regression model for Train data.

A screenshot of a computer program

Description automatically generated

Fig 30: Linear regression model for Train data

We can estimate the linear regression equation with the help of summary ()

A screenshot of a computer

Description automatically generated

Fig 31: Summary of linear regression

The estimated equation for the linear regression model can be constructed using the coefficients of summary output.

Therefore, the estimated equation is:

Spending = 48.17842 −26.14585\*US +43.84363\*source\_a + 20.75689\* source\_c -13.89153\* source\_b

-24.24353\* source\_d + 51.60606\* source\_e -44.83192\* source\_m -23.58561\* source\_o

+ 88.47928\* source\_h + 48.11398\* source\_r -2.25763\* source\_s +17.66805\* source\_t

+ 36.91298 ​\* source\_u+ 35.80236 \* source\_p + 21.06996\* source\_x + 99.36790\* source\_w

+ 79.41246\* Freq -0.03040\* last\_update\_days\_ago + 0.02347\* X1st\_update\_days\_ago

-17.20026\* Web\_order + 0.96519\* Gender\_male -79.35149\* Address\_is\_res

Predicting model with the validation data

A screenshot of a computer screen

Description automatically generated

Fig 32: predicting value with validation data

The forecast::accuracy() function is typically used to assess the accuracy of forecasting models, particularly time series forecasting models. It computes various accuracy measures such as Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), etc., to evaluate the performance of the forecasted values against the actual values

A close-up of numbers

Description automatically generated

Fig 33: Accuracy of forecasting models

The RMSE value of 167.2498 suggests that, on average, our model's predictions are off by approximately 167.25 units of spending when compared to the actual spending values in the validation set.

*6.2 Forward stepwise:*

Forward stepwise regression is a variable selection method used in linear regression modeling. It starts with an empty model and sequentially adds predictors that most improve the model fit until no further improvement is observed.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated

Fig 34: Stepwise forward regression

Begin with an empty model and consider adding predictors one by one. At each step, select the predictor that results in the greatest reduction in AIC. Repeat until no further reduction in AIC is observed or until a stopping criterion is met. The process stops when no further improvement in AIC is observed or when all predictors have been added. The output shows the AIC values at each step and the selected predictors. The final model includes predictors Freq, Address\_is\_res, source\_w, last\_update\_days\_ago, and US.

A screenshot of a computer program

Description automatically generated

Fig 35: Summary of linear regression forward stepwise model

The equation for the forward stepwise regression model is as follows:

Spending = 99.23266 + 85.050049\*Freq -74.528555\*Address\_is\_res +49.131531\*Source\_w

-0.017099\*Last\_update\_days\_ago -34.231941\*US

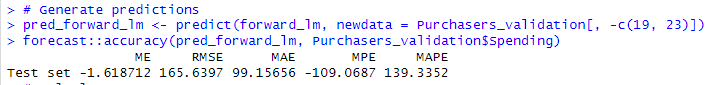


Fig 36: Generating prediction and forecast measure for forward stepwise

It measures the typical deviation of the predicted values from the actual values. A lower RMSE indicates better model performance. Here, the RMSE is 165.6397, indicating moderate prediction errors. our model's predictions are off by approximately 165.6397 units of spending when compared to the actual spending values in the validation set.

*6.3 Regression trees:*

In the regression tree model, we use the part and rpart. Plot package in R to build and visualize regression tree.

A regression tree is a type of decision tree used in statistical modeling and machine learning for predicting continuous numerical values. It's a predictive modeling technique that recursively partitions the data into subsets based on the values of input features, to minimize the variance of the target variable within each subset.

*Tree structure:*

We have two types of nodes in a tree: decision (= splitting) nodes and leaf nodes Nodes that have successors are called decision nodes because if we were to use a tree to classify a new record for which we knew only the values of the predictor. we would "drop" the record down the tree so that at each decision node, the appropriate branch is taken until we get to a node that has no successors. Such nodes are called the leaf nodes (or terminal nodes or leaves of the tree) and represent the partitioning of the data by predictors.

Regression trees are powerful because they can capture nonlinear relationships between the input features and the target variable and can handle interactions between features. However, they are prone to overfitting, especially when the tree depth is not properly controlled. Techniques like pruning, which involves removing parts of the tree that do not provide significant predictive power, can help mitigate overfitting and improve generalization performance.

*Fit decision tree for training model:*

A screenshot of a computer code

Description automatically generated

Fig 37: decision tree model fitting for training data

The fitting model for training data and Fig 56 below show the tree structure for the training data.

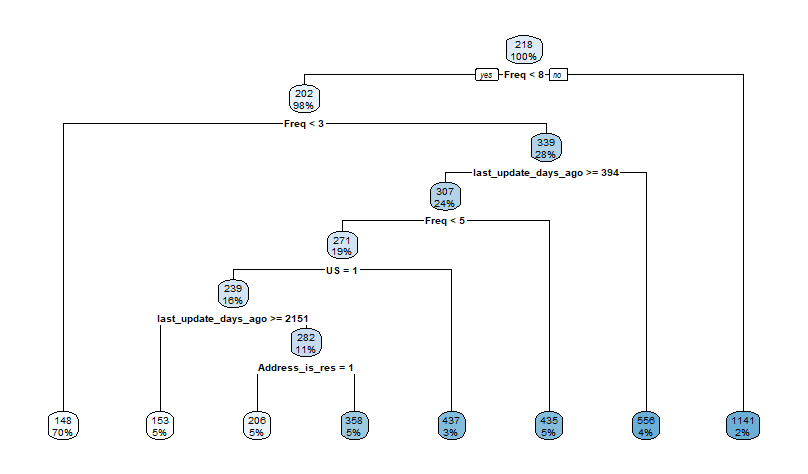


Fig 38: Tree structure for training data.

*Predicting model with validation partition:*

A white background with black text

Description automatically generated

Fig 39: Summary of predicted values

It appears that the predictions generated by the regression tree model (Regtree\_predict) are not varying much, as evidenced by the summary statistics showing a median and quartiles all at the same value of 147.5. On the other hand, the summary statistics for the 'Spending' variable in Purchasers\_validation demonstrate more variability, with different quartiles and a wider range. Evaluate the performance of the regression tree model on the validation dataset.

*Calculating accuracy for the model*

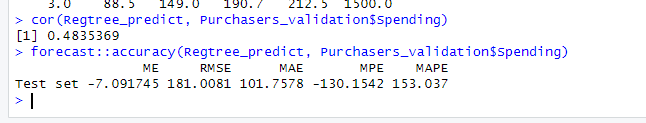


Fig 40: calculating accuracy for the model.

The correlation coefficient between the predictions (Regtree\_predict) and the actual 'Spending' values from the validation dataset is 0.4835369. This value indicates a moderate positive correlation between the predicted and actual spending amounts.

ME (Mean Error): The mean error is -7.091745, indicating that, on average, the predictions underestimate the actual spending by approximately $7.09. RMSE (Root Mean Squared Error): The root mean squared error is 181.0081, which represents the square root of the average squared differences between the predicted and actual spending amounts. This metric quantifies the overall accuracy of the predictions, with lower values indicating better performance.

*6.4 Improving model performance:*

The M5P algorithm represents a pruned model tree for predicting the 'Spending' variable based on various predictor variables from our training dataset. The M5P algorithm generates a pruned model tree that combines decision trees with linear regression models to predict the target variable.

A screenshot of a computer code

Description automatically generated

Fig 41: Improving model with M5P

The summary of the model explains the correlation coefficient, mean absolute error, root mean squared error, Relative absolute error, and root relative square error of the model.

A computer screen shot of a error

Description automatically generated

Fig 42: summary of the M5P model

A close-up of a text

Description automatically generated

Fig 43: correlation and forecast of the model

The correlation coefficient between the predictions (improve\_predict) and the actual 'Spending' values from the validation dataset is 0.5599385. This value indicates a moderate to strong positive correlation between the predicted and actual spending amounts.

ME (Mean Error): The mean error is -1.139432, indicating that, on average, the predictions are slightly underestimated by approximately $1.14.

RMSE (Root Mean Squared Error): The root mean squared error is 166.9304, which represents the square root of the average squared differences between the predicted and actual spending amounts. Lower RMSE values indicate better predictive accuracy.

Overall, the improved model shows better performance compared to the previous model in terms of correlation, mean absolute error, and other accuracy metrics. However, there may still be room for further refinement and improvement.

Accuracy comparison:

|  |  |
| --- | --- |
|  | RMSE |
| Linear Regression Model | 167.2498 |
| Forward Stepwise Regression Model | 165.6397 |
| Regression Tree Model | 181.0081 |
| Improved M5P Model | 166.9304 |

*Linear Regression Model:* This model has an associated value of 167.2498, which is presumably the error or loss metric we are using for comparison.

*Forward Stepwise Regression Model:* This model has a lower value of 165.6397 compared to the Linear Regression Model, suggesting it performs slightly better according to the chosen metric.

*Regression Tree Model:* This model has the highest value among the listed models at 181.0081, indicating it might not perform as well as the linear regression-based models according to the chosen evaluation metric.

*Improved M5P Model:* This model performs better than most with a value of 166.9304, suggesting it's competitive with the stepwise regression models in terms of performance.

We are selecting a forward stepwise linear regression model to provide easily interpretable coefficients that represent the relationship between each independent variable and the dependent variable. This can be advantageous for understanding the impact of predictors on the outcome. Decision trees, on the other hand, provide a hierarchical structure of decision rules, which might be less intuitive but can still offer insights into variable importance.

7. Report models performance:

The summary of the classification models is as below.

*Logistic regression:*

* *Accuracy:* The model's accuracy is 0.7929, which means it correctly predicts the outcome around 79.29% of the time.
* *Sensitivity:* The proportion of actual positive cases that the model correctly identifies. In this case, it's 0.7383, indicating that the model identifies around 73.83% of actual positive cases (true positives).
* *Specificity:* The proportion of actual negative cases that the model correctly identifies. Here, it's 0.8516, meaning the model correctly identifies around 85.16% of actual negative cases (true negatives).

Improving model performance with forward stepwise regression:

* *Accuracy:* The model's accuracy is 0.7871, indicating that it correctly predicts the outcome approximately 78.71% of the time.
* *Sensitivity:* The proportion of actual positive cases that the model correctly identifies. In this case, it's 0.7438, indicating that the model identifies around 74.38% of actual positive cases (true positives).
* *Specificity:* The proportion of actual negative cases that the model correctly identifies. Here, it's 0.8338, indicating the model correctly identifies around 83.38% of actual negative cases (true negatives).

Classification tree:

* *Accuracy:* The model's accuracy is 0.77, which means it correctly predicts the outcome around 77% of the time.
* Sensitivity: The proportion of actual positive cases that the model correctly identifies. In this case, it's 0.7548, indicating that the model identifies around 75.48% of actual positive cases (true positives).
* Specificity: The proportion of actual negative cases that the model correctly identifies. Here, it's 0.7864, meaning the model correctly identifies around 78.64% of actual negative cases (true negatives).

However, according to sensitivity wise classification model is performing well whereas the classification tree selects the customers based on ranking. Logistic regression is a widely used statistical method for analyzing datasets in which there are one or more independent variables that determine an outcome. It provides easily interpretable results. The coefficients represent the relationship between the independent variables and the log odds of the outcome. This makes it easy to understand how each predictor variable influences the probability of the outcome. The logistic regression model is the probability of a binary outcome (e.g., yes/no, 0/1) as a function of the predictor variables. It gives outputs between 0 and 1, which can be interpreted as probabilities. Of the customer who is buying who is not buying the product.

Here's a summary of the performance of the linear regression models and the regression tree model.

*Linear Regression Model*:

* Mean Absolute Error (MAE): 100.5464
* Root Mean Squared Error (RMSE): 167.2498
* Mean Percentage Error (MPE): -107.5604%
* Mean Absolute Percentage Error (MAPE): 138.8443%

*Forward Stepwise Regression Model:*

* Mean Absolute Error (MAE): 99.15656
* Root Mean Squared Error (RMSE): 165.6397
* Mean Percentage Error (MPE): -109.0687%
* Mean Absolute Percentage Error (MAPE): 139.3352%

Forward stepwise regression models show slightly lower MAE and RMSE compared to the linear regression model.

*Regression Tree Model:*

* Correlation (Correlation Coefficient): 0.4835369
* Mean Absolute Error (MAE): 101.7578
* Root Mean Squared Error (RMSE): 181.0081
* Mean Percentage Error (MPE): -130.1542%
* Mean Absolute Percentage Error (MAPE): 153.037%

*Improved M5P Model:*

* Correlation (Correlation Coefficient): 0.5599385
* Mean Absolute Error (MAE): 99.27658
* Root Mean Squared Error (RMSE): 166.9304
* Mean Percentage Error (MPE): -106.412%
* Mean Absolute Percentage Error (MAPE): 136.5006%

We are selecting a forward stepwise linear regression model to provide easily interpretable coefficients that represent the relationship between each independent variable and the dependent variable. This can be advantageous for understanding the impact of predictors on the outcome. Decision trees, on the other hand, provide a hierarchical structure of decision rules, which might be less intuitive but can still offer insights into variable importance.

By this we are considering forward stepwise linear regression as the best model.

8. Model evaluation:

Model evaluation is essential in machine learning as it helps to determine how well a model performs on specific tasks. By using evaluation metrics, developers can compare different models to choose the best one, identify areas where a model may need improvement, and ensure the model works well on new, unseen data. This process is crucial for tuning the model, ensuring it meets industry standards, and building trust with users by providing evidence of its effectiveness and fairness. Overall, model evaluation is a key step in developing reliable and efficient machine learning models.

*Improving model performance with forward stepwise regression:*

Among all other classification models, we can say that the forward stepwise regression model is the best model with a sensitivity of 74.38%, this model performs slightly better in identifying positive cases compared to the previous models. Linear regression models provide easily interpretable coefficients that represent the relationship between each independent variable and the dependent variable. Comparison between actual value and predicted value

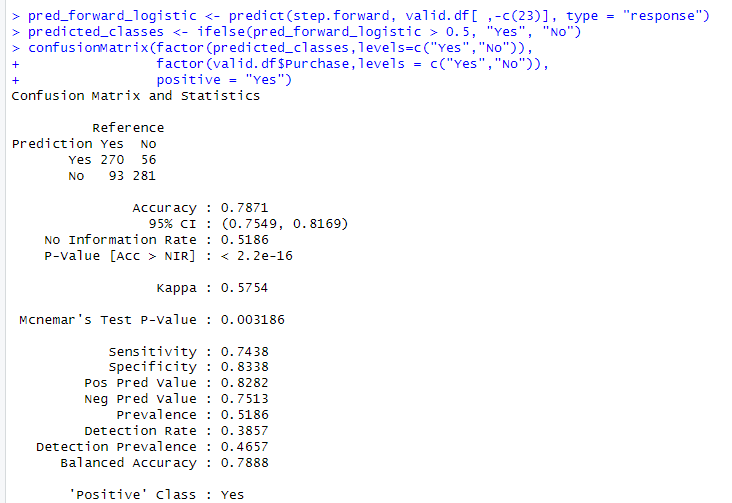


Fig 44: Evaluating model performance for validation data

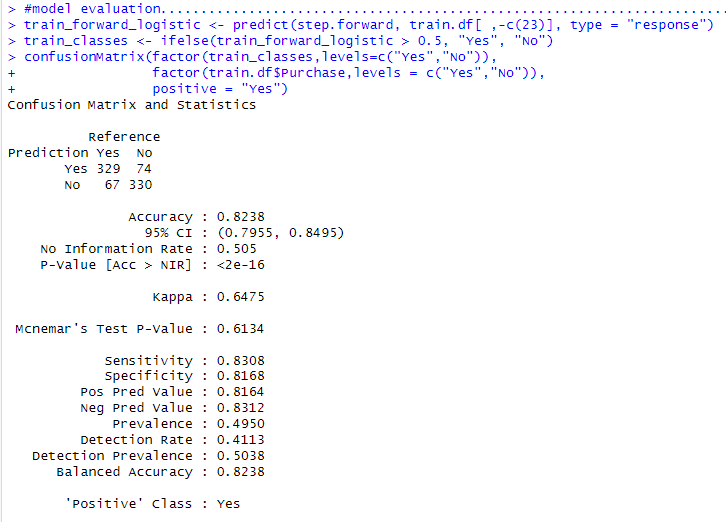


Fig 45: Evaluating model performance for train data

*Validation Data:*

* Accuracy: 0.7871
* Sensitivity (True Positive Rate): 0.7438
* Specificity (True Negative Rate): 0.8338
* 'Positive' Class: Yes

*Training Data:*

* Accuracy: 0.8238
* Sensitivity (True Positive Rate): 0.8308
* Specificity (True Negative Rate): 0.8168
* 'Positive' Class: Yes

Sensitivity, specificity, and positive predictive value are slightly higher on the training dataset compared to the validation dataset, indicating better performance in correctly identifying positive cases ('Yes' class) in the training dataset.

*Forward stepwise linear regression model:*

Forward Stepwise Regression Model is the best model which is approximately 165.6397 units of spending when compared to the actual spending values in the validation set. As our main goal is to get more customers with the spending values this model is the best model for this data set.

Let’s compare the forward stepwise linear regression model on both the training and validation datasets.

A close-up of a computer screen

Description automatically generated

Fig 46: comparing training data with validation data

*Training Data*:

* Mean Absolute Error (MAE): 106.576
* Root Mean Squared Error (RMSE): 168.9053
* Mean Percentage Error (MPE): -92.63857%
* Mean Absolute Percentage Error (MAPE): 121.6101%

*Validation Data:*

* Mean Absolute Error (MAE): 99.15656
* Root Mean Squared Error (RMSE): 165.6397
* Mean Percentage Error (MPE): -109.0687%
* Mean Absolute Percentage Error (MAPE): 139.3352%

The RMSE measures the average deviation of the predicted values from the observed values, indicating the model's predictive accuracy. Lower values of RMSE indicate better model performance. It seems that the model performs better on the training set compared to the validation set, which could suggest overfitting or issues with generalization to new data. Further analysis and possibly model refinement may be needed to address these issues. Due to the different data set partitions, there is a chance of overfitting there I very little unit difference between the train dataset and the valid dataset. If we use cross-validation, then we can overcome it.

*Adding columns to calculate expected spending:*

Adding a column to the data frame with the predicted probability to Predicted probabilities allows us to assess the performance of the selected model. we can compare these probabilities with the actual outcomes to evaluate how well our model predicts the target variable.

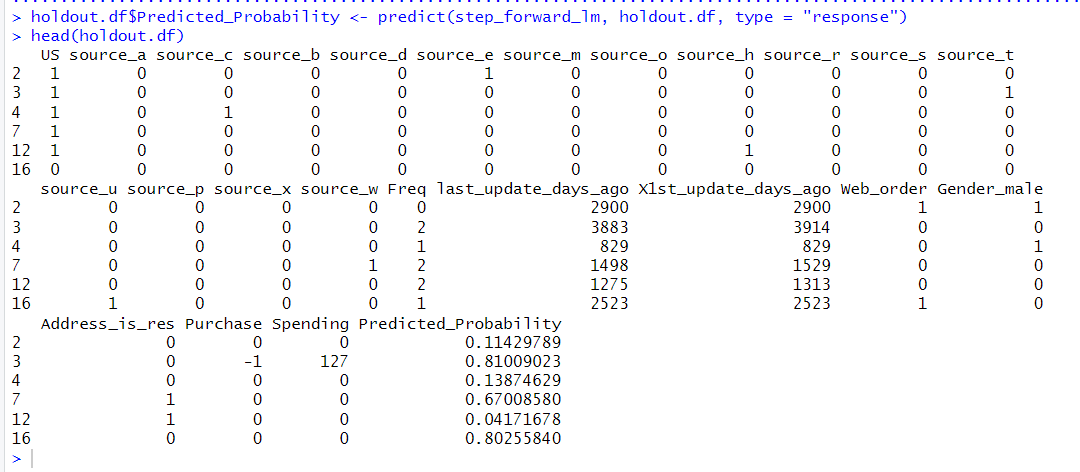


Fig 47: adding predicted probability column

Adding another column with the Predicted spending can be valuable in financial forecasting and budgeting. For businesses, understanding customers' predicted spending patterns helps in planning marketing campaigns, inventory management, and resource allocation.

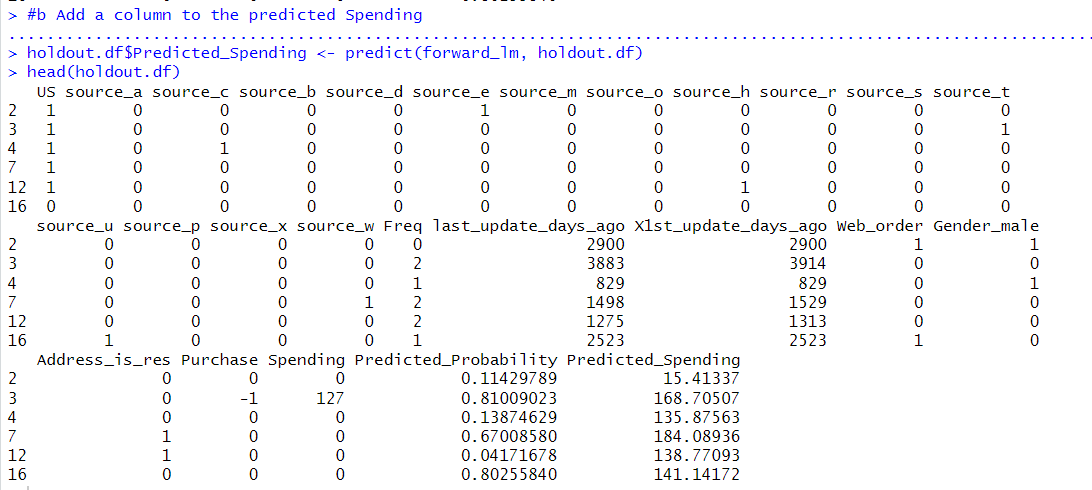


Fig 48: Adding another column with the predicted spending

A close up of a text

Description automatically generated

Fig 49: defining original purchase rate

A column for the adjusted probability of purchase is added because, if there is any imbalance in the data set, meaning that purchasers are significantly underrepresented compared to non-purchasers, the selected model may produce biased predictions. Adjusting the probabilities helps to focus on the effects of this imbalance data.

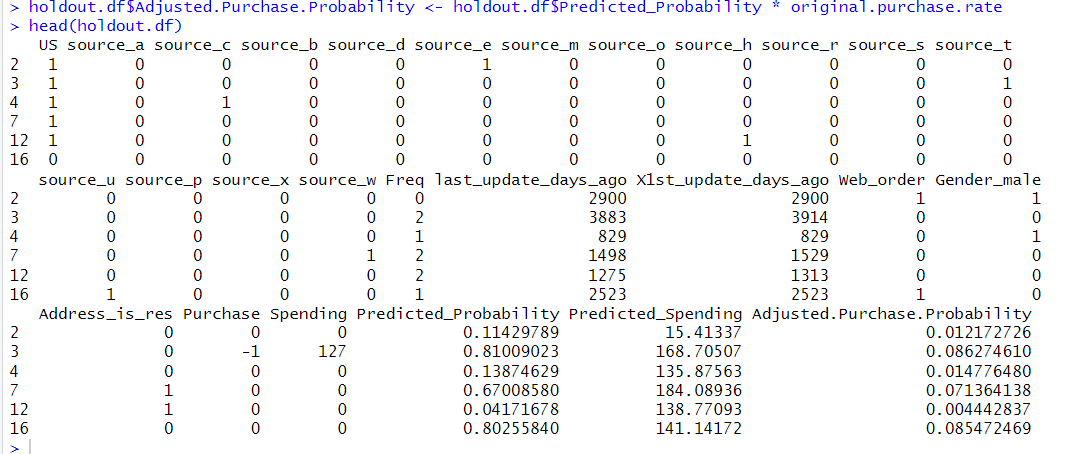


Fig 50: Adding a column for adjusted probability of purchase

Adding a column for Expected Spending provides valuable insights for business planning and resource allocation. By estimating the potential revenue generated from customers or prospects, businesses can make informed decisions about inventory management, marketing strategies, and overall budgeting.

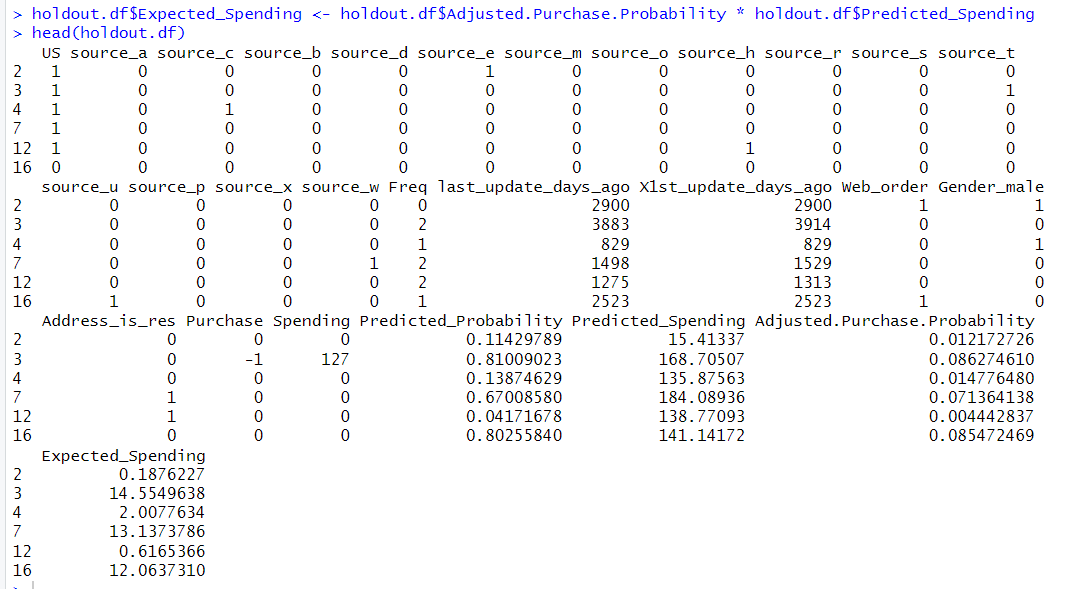


Fig 51: Adding a column for Expected Spending

*Cumulative gains chart:*

The Cumulative Gains Chart provides a clear visual representation of how well our predictive model performs compared to a baseline model or random guessing. It allows us to see the cumulative gain achieved by targeting cumulative expected spending as a function of records targeted.

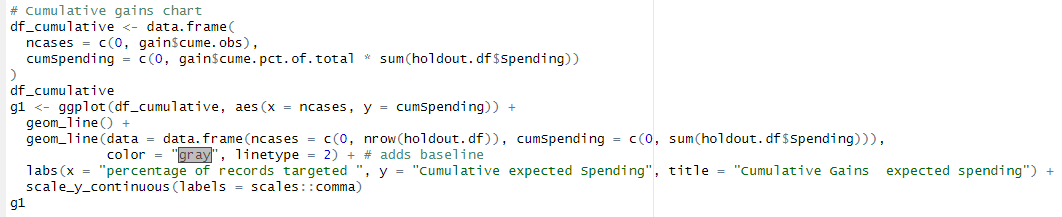


Fig 52: creating cumulative gains chart

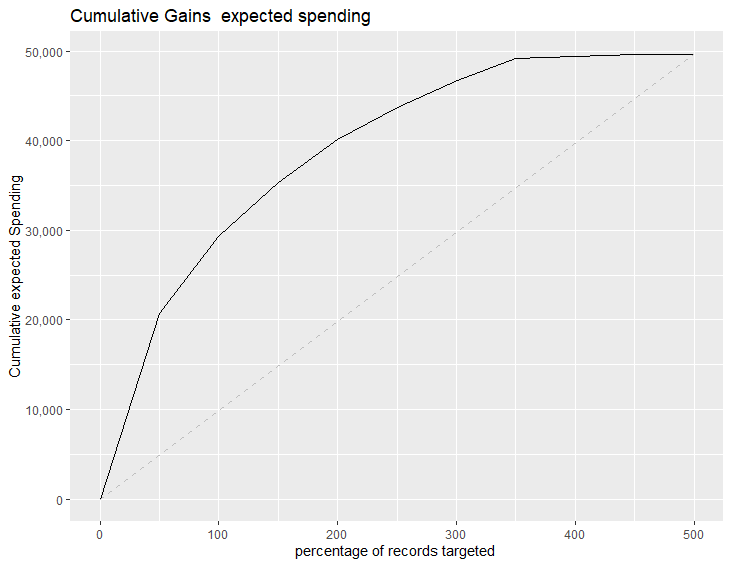


Fig 53: cumulative chat

By observing the cumulative gains chart, we can evaluate how well our predictive model performs in identifying positive cases like how many customers are likely to make a purchase. The chart shows the cumulative percentage of the spending as we traverse through the observations, sorted by some scoring metric or probability generated by our model. we can determine whether our campaign successfully reaches high-spending customers earlier in the process, thus allowing us to optimize resource allocation and mailing strategies.

*Decile-wise lift:*

Decile-wise lift chart is a way to visualize the performance of a model at different quantiles of the predicted probability. The x-axis of the chart is divided into ten deciles, which represent equal-sized portions of the data ranked by predicted probability. The y-axis shows the lift, which is the ratio of the conversion rate in a given decile to the average conversion rate across all deciles.

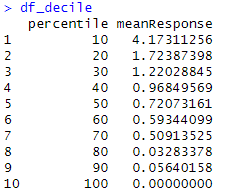


Fig 54: Mean response for Decile-wise lift

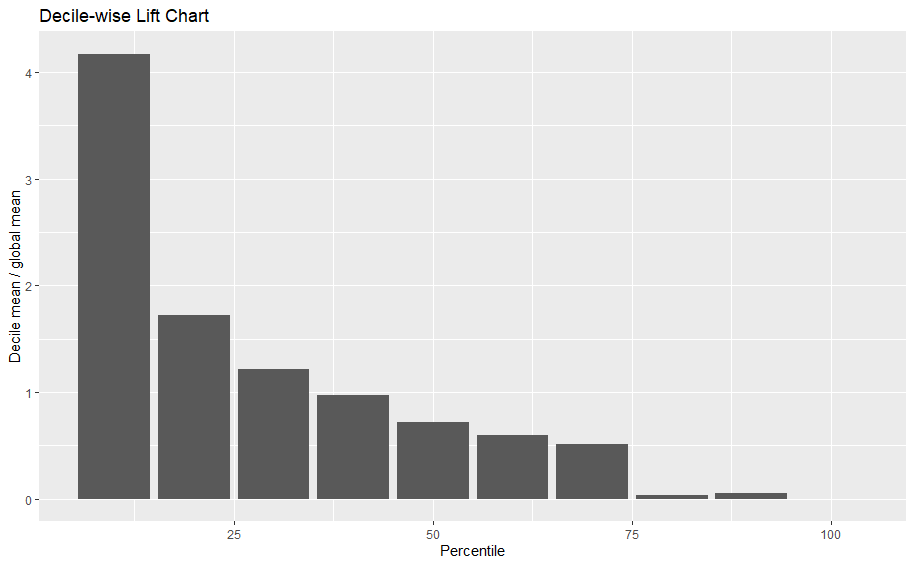


Fig 55: Decile-wise lift chat

*Profit analysis:*

The lift chart and decile chart, based on the application of the chosen model to the data, clearly show that the model outperforms the baseline in terms of lift. The decile chart reveals that targeting just the top 10% of customers predicted by the model as highest spenders can yield profits 4.17 times greater. Therefore, instead of distributing product lists to all 180,000 customers and potentially earning lower profits, North-Point Company can utilize this model to selectively target customers, thereby maximizing profits while reducing expenditures.

9. Conclusion:

By understanding the business goals, exploring the data by checking null values and zeros and we did not find any null values or zeros. We have explored the data and analysed binary variables and their distribution and analysed numerical variables with the help of a bar graph. Did bivariant analysis with target column and other columns, also correlation matrix for numerical variables and pair plot to understand the correlation variables visually. Partition of data is done into 3 partitions and the fit model for training data and test model with validation data with different models are used to evaluate the model performance. While various classification models are employed, logistic regression stands out for its interpretability and ability to provide clear insights into customer behavior. The coefficients in logistic regression represent the impact of independent variables on the probability of the outcome, making it a valuable tool for understanding customer purchasing patterns. Business owner can choose from the predicted values.

Now we have built model with linear regression model for purchase data and improved the model with a forward and backward stepwise. Compared to all these models’ linear regression forward stepwise is the best model. Also, we have created a decision tree model for the purchase data and to improve the model we have done the M5P model. Reported all model performances explained the best model evaluated the model and compared both train data and validation data for classification we have considered the C50 model as the best model and for regression tree model is the best model. We have added a predicted probability column, predicted spending column, adjusted probability of purchase column, and expected spending column and created a cumulative gain chat to provide a clear visual representation of how well our predictive model performs compared to a baseline model or random guessing.

By seeing the Decile-wise lift chart we can understand that the top 10% of customers have the highest mean value 4.17. whereas if we mail to the top 50% of customers, we can see a good rise in business profits rather than sending it to all 180,000 customers. We can see the more profit.