E-COMMERCE ANALYTICS TECHNIQUES FOR IMPROVING SALES

BY

OKORIE, PROGRESS OBINNA

(17CG023198)

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY OTA, OGUN STATE.

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE BACHELOR OF SCIENCE (HONOURS) DEGREE IN COMPUTER SCIENCE

JULY 2021

CERTIFICATION

I hereby certify that this project was carried out by OKORIE, Progress Obinna in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ogun State, Nigeria, under my supervision.

1. Name: Dr. Ibukun Afolabi

   (**Supervisor)**

   Signature _____ Date _____

2. Name: Dr. Olufunke O. Oladipupo

   (**Head of Department)**

   Signature _____ Date _____

DEDICATION

I dedicate this project to God almighty, who has been there from the beginning and has seen me through till the end of this project. He is indeed excellent and worthy of being praised.

ACKNOWLEDGEMENT

Firstly, I would like to acknowledge God, who gave me strength, wisdom, and the ability to complete this project successfully.

My family for their encouragement and prayers. I hope to keep exceeding their expectations and make them proud.

My supervisor, Dr Ibukun Afolabi, I could not have asked for anyone better. She was always supportive, joyful, encouraging and an excellent mentor to me. I am forever grateful to you, Ma.

To Afrimash Limited, thank you for trusting me and embarking on this project journey with me. May God bless you all.

My friends and colleagues. I thank you all that I have gone through these past four years for the encouragement, joy, and uplifting you gave me throughout our journey together. The same way you have blessed me, may my good Lord bless you in a thousand folds.

TABLE OF CONTENTS

LIST OF EQUATIONS

LIST OF FIGURES

ABSTRACT

The internet over the years has turned into a marketplace due to the increasing number of e-commerce websites. The e-commerce market is a large market, but despite the size of the e-commerce, several companies have failed to survive worldwide. Some of these failures, which this project aims to solve, include low customer retention, lack of precise marketing strategy, and low website traffic and conversion.

In this project, various e-commerce analytics techniques were used to try to tackle these problems. Techniques used included descriptive analytics, predictive analytics using MLR (Multiple Linear Regression) and ANN (Artificial Neural Network), classification analytics using classification tree and ANN (Artificial Neural Network) and clustering using K-means algorithm.

The project proposes the implementation of the prediction, classification and clustering models developed to be applied to external customers to predict the total amount likely to be spent per order, classify the order status of each order and segment proposed customers according to profiles for target marketing, respectively.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND INFORMATION

E-commerce is a common word used to describe the purchase and sale of goods and services over the internet.

The increasing number of E-Commerce websites has turned the internet into an essential marketplace, with more and more people and investors scrambling to invest in it. It has even gone as far as causing a shift in the behaviour of customers who have come to prefer online shopping to traditional shopping, which has caused the number of transactions and the amount spent to keep growing steadily yearly (Guler & Tufan, 2013).

The E-commerce market is growing fast and in significant numbers. In Nigeria: in 2017, 28.81 million users spent $1.720 billion in the e-commerce market; in 2020, the number was even higher. It was 64.69 million and $4.942 billion. As it is reported in Statista in 2024, the numbers are expected to go even higher, up to 111.62 million and $8.917 billion (*Statista Market Forecast*, 2020).

Despite the e-commerce market size, as shown above, their success rate is not that encouraging, as several e-commerce companies have failed worldwide. An example of such a failure story is Zulily, Pets.com, and a host of others. Pet.com was launch in 1998 and was selling pet accessories. Pet.com later went bankrupt and closed completely a few years later. The bankruptcy resulted from a lack of customer retention and bad pricing and discounting (Guler & Tufan, 2013). Another is Zulily which was established in 2009. It was selling moms and children clothing and products. Zulily was shut down and later became active again. Its shutdown resulted from bad marketing strategies and a lack of customer retention (*Failory*, 2015).

Therefore, this project aims to address the challenges faced by E-commerce companies and how data analytics (E-commerce analytics) could solve them.

E-commerce businesses have one goal in common, which is the need to get more customers & increase sales. This is so because of the ever-growing need to show progress to their investors and increase their revenue and profit.

Social engagement has been a significant source of traffic, customer, and advertisement for e-commerce companies. E-commerce companies strive to improve their social media engagement because it generates traffic, customers and, in turn, leads to an increase in sales and revenue.

Businesses that do not identify and follow trends tend to fall behind; that's why companies look for trends to increase their website popularity. They discover trends that could be used in marketing campaigns to draw viewers' attention; they also monitor their website visitors' behaviour to identify trends to optimize their websites for better customer interaction.

Businesses need to identify their bestselling or most demanded products at different periods to prepare effectively for the period. The preparation might be by stocking enough of the product so that they will not run out of it, can also be to have marketing campaigns ready to promote the product during that period in which it is highly demanded.

E-commerce analytics is the process of gathering data from all parts that have an impact on an online store and using this information to analyze patterns and changes in customer behaviour so as to make data-driven decisions that will increase online sales. (*Ecommerce Analytics 101*, 2020).

The examples above highlight a few of the challenges e-commerce companies face that can be solved using e-commerce analytics. There are still other problems that e-commerce analytics can help an e-commerce company solve. Some organizations have found and explored this hidden gem called e-commerce analytics for their benefit and are soaring high, but others are yet to apply it for their growth; a famous example is Amazon.

E-commerce analytics provides unique tools to solve problems and distinguish an e-commerce company from a host of others. It is helpful in planning strategies for marketing, pricing, and retention of customers.

1.2 STATEMENT OF PROBLEM

The e-commerce market is a large and steady growing market. It is unfortunately very competitive; the competition is at a local level and a global level. Every e-commerce company expects to make a profit and grow its business; however, this is not so for some of them who have failed to tackle essential problems such as:

- Low customer retention: The lack of the ability to keep customers and turn them into loyal customers. This is a problem because, without customer loyalty, businesses would struggle, and it costs five times as much to gain a new client as it does to keep a current one.
- Low website traffic and conversion: Poor website traffic is when an e-commerce company fails to attract visitors to their website. Low website traffic conversion is when e-commerce companies fail to get customers to buy a product.
- Poor/lack of precise marketing strategy: A lack of precise marketing strategy is a marketing strategy that does not have a target audience. When an organization lacks a precise marketing strategy, it leads to a low return on investment.

The companies that failed to tackle these issues either shut down, went bankrupt, or were acquired by another company.

In this project, we will show how to solve these problems using e-commerce analytics.

1.3 AIM AND OBJECTIVE OF STUDY

This project aims to use e-commerce analytics techniques to improve the sales of an e-commerce business.

The objectives of the study are:

1. To retrieve relevant data from which analytics will be carried out on the above case study.
2. To analyze the data towards providing actionable recommendations that can be implemented to get more customers and increase sales.

3. To present the result of the analysis in the form of traceable recommendations to return on investments.

1.4 METHODOLOGY OF THE STUDY

The methodology for the project can be summarised as follows:

- Data Retrieval and pre-processing: Data were retrieved from Afrimash DEngage platform. The retrieved data was pre-processed based on the kind of data in question. Typical data pre-processing steps were followed for the structured data, such as correlation analysis and attribute reduction techniques. These pre-processing steps were selected based on the data mining technique in question.
- Data Analysis: The data analysis was carried out using appropriate prediction, classification, and clustering techniques. The models used were experimentally determined. The selected algorithms include Multiple Linear Regression (MLR), classification tree, Artificial Neural Networks (ANN) and K-means. This implementation of the analysis will be done using the R studio platform.
- Reporting and Recommendations: Different analytics reporting tools will be used to present the results, such as tableau and rapidminer.

1.5 SIGNIFICANCE OF STUDY

The study will provide actionable recommendations for agricultural-based e-commerce businesses to increase profit and increase return on investment. The awareness of using data to make decisions that will drive sales in e-commerce websites will be created, leading to more proactive and confident decisions.

The study will help understand the company's type of customers, behaviours, and demographics using e-commerce analytics to improve company strategies. When an e-commerce business can understand its customers, it can determine company strategies or realign them to meet its goals. It can also help in marketing campaigns to maximize conversion and bring about a higher return on investment.

It will also help the organizations measure the effectiveness of their marketing campaigns and have a better marketing strategy. Companies will save money using precise marketing strategies and realize a higher return on marketing campaigns.

It will also foster an increase in the amount of quality traffic towards the e-commerce website resulting in increased sales and higher profits.

The study will shine a light on the advantages of e-commerce analytics' and how it can contribute to the organization's growth and survival.

1.6 SCOPE OF STUDY

The case study used for the project is Afrimash Ltd (https://www.afrimash.com/). Afrimash is a privately held e-commerce company based in Oyo State, Nigeria.

It is one of the leading e-commerce marketplaces for agricultural items in Nigeria. Afrimash also offers other services like training and veterinary consultation.

1.7 ARRANGEMENT OF THESIS

Chapter 1 of the project contains an overview of the project, stating the problems, Aim & Objective, the methodology, the significance of the study, and its scope.

Chapter 2 Literature review

Chapter 3 Methodology

Chapter 4 Analysis result and implementation

Chapter 5 Summary, Conclusion, and Recommendation

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter focuses on a study of previous e-commerce analytics systems that have been created. The review identifies different approaches used in getting more customers, increasing sales, and discovering trends that can increase the popularity of the e-commerce website amongst its visitors. The review also highlights the challenges faced, the strengths and the effectiveness of each method.

2.2 REVIEW OF EXISTING SYSTEM

This section covers a surface study on the existing systems related to the methods used for the study.

2.2.1 Existing Systems in E-Commerce to Drive Sales

This section covers the surface study on the existing system related to getting more customers and increasing sales.

2.2.1.1 Loyalty of Young Female Arabic Customers Towards Recommendation Agents

This paper aims to use an e-commerce recommender agent to investigate the major factors that affect female online shoppers' loyalty creation.

This study's data came from a survey of female students at Imam Abdulrahman Bin Faisal University who used Amazon.com.

Structural Equation Modelling (SEM) and partial least squares were used to analyze the relationships between constructs. SEM integrates the advantages of factor, path, and multiple regression analysis to develop a vigorous approach for evaluating construct

relationships. Both above promoted the use of SmartPLS to analyze the findings of the survey.

This study's key contribution was introducing, creating, and analyzing a model to improve e-commerce customer loyalty through recommender systems.

Data distribution and selection were limitations of this study; only a single e-commerce platform was evaluated, and questionnaires were distributed to students at a single Saudi Arabian university.

Future research is needed to assess the effects of several recommendation techniques on user behaviour and look at the website's user interface the recommender system is hosted on(Ali et al., 2020).

2.2.1.2 Mining Customer Knowledge for Direct Selling and Marketing

This research recommends the usage of internet marketing in Taiwan's direct selling industry and cosmetics market.

The data used for this research was obtained from 638 questionnaires that were collected.

This research made use of association rules and cluster analysis as techniques for data mining.

The clustering algorithm used was the K-means clustering algorithm, and the Apriori algorithm was the methodology used for the association rule.

K-means was used to cluster data from databases into two categories: "customer has online shopping experience" and "customer has no online shopping experience". The consumer data/information was also subjected to the K-means algorithm, which was classified into two categories: "the customer has purchased the direct selling cosmetic" and "the customer has not purchased the direct selling cosmetic" (Liao et al., 2011).

2.2.1.3 Segmenting Customers in Online Stores Based on Factors That Affect the Customer's Intention to Purchase

This paper suggests a method for online stores to deliver personalised marketing to their customers by segmenting them based on psychographic data.

This study's data was gathered from a Korean customer survey who had previously purchased from an online store.

The method used in this paper were SOM (Self-Organized Map), k-means and k-nearest neighbour. SOM and K-means were integrated for the segmentation of customers. Five clusters were gained from the segmentation. Customers that were left out of the dataset for clustering were then classified using the K-nearest neighbour's process. The K-nearest neighbours method had an accuracy of 89.74% when k=1.

The paper's contribution was that it demonstrated that k-nearest neighbours' predictive accuracy was sufficient in this case to use our approach in promotional marketing.

Additional data mining methods should be studied further with a broader data collection (Hong & Kim, 2012).

2.2.1.4 Internet Customer Segmentation Using Web Log Data

This research aims to analyze web transaction log data to expose customer behaviour in the Internet channel and segment them.

The data used for this study was collected from an online pet shop. The number of observations in the dataset is 14,312.

This paper made use of cluster analysis for segmentation. The K-means algorithm with Euclidean distances was the clustering algorithm used. The full dataset had 4 clusters, and the member dataset had 6 clusters.

This paper contributes by conducting research alongside online retailers who provide online communities and analyzing the relationship between engagement in online communities and consumption by making use of transactional metrics.

This paper's limitations include the deletion of some variables due to the lack of raw data, and the collection of web-log data was done within a short period (10 days) (Rho et al., 2004).

2.2.1.5 Sales Analysis of E-Commerce Websites using Data Mining Techniques

This paper aims to use data mining methodologies to infer and obtain valuable data from e-commerce websites.

The dataset used for this study was mined from Flipkart.com. The dataset included attributes like product name, product price, quantity, type, online.

Methods used in this study for analysis are web mining and decision tree algorithms. ID3 (Iterative Dichotomiser) was the decision tree algorithm used.

Descriptive analysis was carried out on the data to show correlation and other descriptive statistics. ID3 was used to predict the online rating of a product. The model developed using the ID3 algorithm had an accuracy of 86.4780% (Bejju, 2016).

2.2.1.6 Customer Segmentation in A Large Database of An Online Customized Fashion Business

This project aims to look at data mining methodologies for consumer segmentation that can be used in the fashion industry to address marketing and manufacturing issues.

Data is from bivolino.com. It consists of 10,775 customer orders with attributes grouped like

- Product characteristics (Fabric type, Colour, Structure, Collar type)
- Demographic and biometric (Gender, Collar size, Age groups, BMI (body mass index))
- Geographic (Country/Nationality)
- Psychographic (Lifestyle)
- Behavioural (Price sensitivity)

1-K-Medoids algorithm (Clustering) and CN2-SD algorithm (Subgroup discovery) were used for the segmentation.

1-K-Medoid algorithm (Clustering Analysis) was done in two steps: Clustering analysis based on product characteristics and Clustering analysis based on customer characteristics.

Clustering analysis based on product characteristics was performed on a dataset containing 10,775 shirt orders with 29 attributes. The model produced six distinctive clusters. Overall, the clusters were into two: Fashion garments of the same type: Work shirts (Clusters 1, 2 and 4), which is 65% of the total order and Fashion shirts (Clusters 3, 5 and 6) which are 35% of the entire order. Analyzing further showed specific business dress codes and age condition customer choices.

The clustering analysis based on customer characteristics was supplemented with base variables pertaining to the dimensions of customer characterization: demographic, biometric, geographic, psychographic, and behavioural. The model also produced 6 clusters which can be grouped into three: Work shirt, party shirt and fashion shirt. We can see other details from the grouping, like men with an age range (25-34) favour slim-fitting shirts while men with an age range (45-54) prefer comfort fit shirts.

CN2-SD Algorithm (Subgroup discovery) was performed on 7,066 shirt orders which were characterized by 19 variables. It omitted orders from five countries and females. The result was evaluated according to measures like the size of subgroups and the subgroup distribution deviation in relation to the whole dataset. The model produced contained 54 rules, each of which described a population subgroup whose distribution differed from that of the entire population. The rules were classified into three categories: uninteresting, interesting for marketing, and exciting.

This paper's significant contribution is CN2-SD subgroup discovery used for the characterization of subgroups of observation with rare distributions was first used in this paper.

The work's limitations include the limitation of the method adopted, the nature of variables used (Categorical), and the algorithms' computational effort.

Additional research is required to establish heuristic approaches for determining the optimal number of clusters and implement and test other subgroup discovery algorithms (Quelhas et al., 2015).

2.2.1.7 The Determinants of Conversion Rates in SME E-Commerce Websites

This paper aims to define and analyze the factors that improve conversion rates or a combination of factors that will enhance conversion rates.

The paper's data comprises of 1184 daily observations from six small-business e-commerce websites in Italy that sell clothing, luggage, shoes, and related accessories. The data retrieved include:

- Conversion Rate (%)
- Free shipping (0 no, 1- yes)
- Free returns (0 -no, 1 - yes)
- Discounts
- Season (0 - regular, 1- sales)
- Speed of load (0 – 100)
- Luxury websites (0 - non-brand products, 1 - branded products)
- Week (0 - Saturday & Sunday, 1- Monday to Friday)

To achieve the paper's aim, exploratory regression analysis is used to determine which determinants are essential. Qualitative Comparative Analysis (QCA) highlights more details on the circumstances where conversion rates increase.

An OLS regression was used to test the impact independent variables have on conversion rates. Because the data contains many observations across time that are held inside the six retailers, standard errors were grouped to correct the correlations among the data from the same source. Two models were developed, and they both showed that all variables appear to affect conversion rates except free returns and day of the week product is purchased. Some variables had a difference in p-value, but the differences were $p < 0.05$. Model one had an R-squared of 0.647 and a MdAPE (Median Absolute Percentage Error) of 7.34, while

model two had an R-squared of 0.646 and a MdAPE (Median Absolute Percentage Error) of 7.36.

The individual independent variables impact on conversion rate are tested in regression analysis. A QCA is used to assess the impact of any combination of variables on the conversion rate. The first step in QCA is to identify all the different attributes that affect the conversion rate. Step two: Using a fuzzy-set calibration, model the degree to which other cases belong to a set, which varies from 0 to 1, including intermediate membership levels. The third step is to decide which attribute configurations will serve as appropriate conversion rate conditions. The consistency test must meet a minimum requirement of 0.75 to be deemed adequate. Step four: From the appropriate configurations, remove any redundant attributes. The consistency of the five configurations is significantly higher than the minimum threshold. The five configurations can be grouped into two groups. The effect of discounts associated with promotional seasons, and in most cases, free shipping is the focus of Group A. Luxury e-commerce is identified as Group B, which is described by a combination of luxury products and load speed.

A significant contribution made by the paper was the fuzzy analysis used provided a more detailed perspective.

The paper's limitations include: The dataset (Individual transactions that cannot be tracked information may be lost if aggregated data is used) and the paper was focused on funnel process, so some factors that affect the amount of traffic to the website were not considered.

Further research is needed to determine the result of other website quality factors on conversion rate and consider other factors that affect the purchasing decision (Fatta et al., 2018).

2.2.1.8 A Purchase-Based Market Segmentation Methodology

The paper aims to develop a purchase-based market segmentation methodology to solve the problem of market segmentation by traditional marketers.

The data required for the methodology proposed are transactional data containing customer id, transaction time, product item, quantity, and expenses.

A purchase-based segmentation (PBS) algorithm is built based on the similarity measure to perform market segmentation. The clustering algorithm used was the Genetic Algorithm (GA) to guarantee that customer in the same cluster has a very close purchase pattern. After segmentation, the RFM model is used to analyze the relative profitability of each customer cluster. The RFM profitability review uncovered additional marketing prospects and aided marketers in revising their marketing plans.

This paper's contribution was based on a series of experiments on GA's efficiency, which revealed that GA adoption resulted in a significantly better clustering quality within a reasonable computation time.

More work could be done to improve the proposed approach by applying fuzzy set theory to the unpredictable relationship between customers and clusters (Tsai & Chiu, 2004).

2.2.1.9 Comparing A Traditional Approach for Financial Brand Communication Analysis with A Big Data Analytics Technique

This paper compares the conventional approach to big data analytics techniques to improve the understanding of the opportunities provided for brand communication research in the financial sector by social data.

Data used for the research was gathered from the Twitter accounts of Spanish banks. In total, 7,598 tweets containing the hashtags of the banks under investigation were downloaded and filtered, leaving 7,398 tweets in the dataset.

The study contrasts two research methodologies: the conventional Periodic Evaluation of the Image (PEI) method and the sentiment analysis (Support Vector Machine (SVM)) technique used in big data analytics.

Sentiment Analysis is a data collection method that uses machine learning mechanisms to retrieve and analyze tweets, allowing for a faster and more open data interpretation than conventional interviews. The study is more time-consuming, and qualitative biases based on researcher actions exist.

The sentiment analysis algorithm used in this study was the Support Vector Machine (SVM). An average accuracy of 71.1% was achieved after training with data samples.

The two methodologies produced very similar results; however, the Big Data analytics approach allows for faster data analysis than the conventional method.

This study's analytical approach can be applied to future data analysis studies in the financial sector (Saura et al., 2019).

2.2.1.10 Machine Learning Approach to Auto-Tagging Online Content for Content Marketing Efficiency: A Comparative Analysis Between Methods and Content-Type

This paper aims to address the issue of unstructured data that is dispersed across platforms and in various formats, causing performance and user experience issues.

The dataset includes data on the article's content, title, publication date, and keywords. The dataset comprises of 21,709 web pages, which were scraped from Al Jazeera's website.

In this paper, they compared three machine learning (ML) techniques for multilabel classification, namely Neural Network (NN), K-Nearest Neighbour (KNN), Random Forest.

K-Nearest Neighbour (KNN) produced an average f-1 score of 0.577, and Random Forest had an average f-1 score of 0.458, but Neural Network outperforms them both, yielding an average F1 score of 0.627. The organization's YouTube content provided reasonable cross-

platform applicability. The built model will automatically tag 99.6% and 96.1% of unlabelled website content and unlabelled YouTube content, respectively.

The comparative assessment of Machine Learning models for multilabel content classification and cross-channel validation for a different category of content contributed to this paper's marketing literature.

Limitations of this paper include: The dataset was small, which limited the capacities of the model.

Future work may include identifying other types of online material, such as images and videos, to broaden the approach's cross-channel applicability (Salminen et al., 2019).

2.2.1.11 Extending Market Basket Analysis with Graph Mining Techniques: A Real Case

This research aims to solve the problem of finding a set of products that are sold together.

The data in this paper came from two Chilean retail chains. The Retail A chain provided about 238 million transactions, 160 thousand clients, and over 11 thousand SKUs. Also gathered approximately 128 million transactions, nearly 2 million customers, and 31 thousand separate SKUs from Retail chain B.

They contrasted K-means, SOM, and Apriori, which are conventional market basket analysis methods, to a novel market basket analysis method based on graph mining techniques built in this paper.

The novel approach contained Temporally Transactional Weighted Product Network, which was used to show the retailers or organization's information. Threshold and filter methodology was set up to reduce noisy data. To generate frequent item sets, overlapping community detection algorithms were used. This method was benchmarked using COPRA and SLPA algorithms.

Using the K-means algorithm, two representative clusters were obtained, with one of them concentrated with 93% of the products (around 14,000). The Apriori algorithm came up

with a few rules that had low support and confidence. On the other hand, the established novel approach had 30 clusters and an average of 7 products.

The result of the novel approach was more meaningful than that of the traditional techniques. The traditional technique produced meaningless results because of the huge amount of data and the data's spar nature.

Future research may focus on generating a fuzzy consumer profile using these product clusters/communities (Videla-cavieres & Ríos, 2014).

2.2.1.12 Web Usage Mining to Improve the Design of An E-Commerce Website: Orolivesur.Com

This study aims to improve the design of e-commerce websites using web usage mining.

The data for this study came from the Google Analytics of OrOliveSur.com, an e-commerce website. Eight thousand eight hundred thirty-two records in the dataset represent customers who used the website for more than one second.

Clustering with K-Means, association rule learning with the Apriori algorithm, and subgroup discovery with the NMEEF-SD algorithm were the methodologies used in this paper.

K-means algorithm obtained 5 clusters, and Apriori algorithm obtained six rules and NMEEF-SD algorithm 11 subgroups. The confidence of the six rules obtained from the Apriori algorithm was high, from 0.9 to 1.0 (Carmona et al., 2012).

2.2.1.13 The Effect of Relationship and Transactional Characteristics on Customer Retention in Emerging Online Markets

This study aims to see how emerging online markets retention is affected by trust and a variety of other transactional characteristics and relationships.

The datasets for this paper were obtained from an Indian online retailer. A consumer survey was used to assess confidence (and a few other variables) and the purchase transaction data from before and after the survey of the customer.

This research used a latent attrition model as its methodology. Customers' attrition is inferred in a probabilistic fashion using the latent attrition model, and the inferences are then used to evaluate the hypothesis proposed.

This paper's data modelling approach helps to probabilistically infer customer retention based on actual customer transaction activity rather than relying on customers' reported loyalty intent. According to the findings, firms can effectively monitor the relationship and transactional characteristics using data records, which has a systemic effect on retention. All of which are vital contributions of this paper.

This paper's drawback is that it only uses data gotten from a single retailer and doesn't address the impact of retailer rivalry.

Future research into consumer behaviour across various retailers will help to understand better the role of confidence in customer retention (Jaiswal et al., 2018).

2.2.1.14 Acquiring Customers Through Online Marketplaces? The Effect of Marketplace Sales on Sales in A Retailer's Own Channels

Using time-series category sales data, this paper aims to determine which of the two opposing forces (online marketplace and retailer's webshop/website) prevails.

This study's data came from a webshop run by a multinational retailer of refurbished electronics and used media that operates in a vast global marketplace.

The data collection contains aggregated daily sales data from each category on the retailer's website and on the marketplace and visits to the retailer's website in the company's largest country of operation.

The relationships between the online marketplace and the retailer's webshop were modelled using vector autoregressive (VAR) models. It demonstrates that revenues on a retailer's website increase due to marketplace sales. The rise is most significant in groupings with a wide range of products and low prices. It also demonstrated that attracting consumers via the marketplace could be less costly than other approaches.

By examining the effect that selling on an online marketplace has on a retailer's webshop, this paper contributes to the debate about if it is the right decision to sell on an online marketplace. It also closes a cell of the channel cross-elasticity matrix that has yet to be investigated: the effect of marketplace revenues on the website of the company.

The focal company's unique industry background could threaten the generalizability of the results, and the available data were restricted in multiple ways, preventing more thorough testing of the planned mediation process.

Future research into the impact of marketplace sales is needed (Maier & Wieringa, 2020).

2.2.1.15 Automating the Extraction of Static Content and Dynamic Behaviour from E-Commerce Websites

This paper aims to propose a technique for the extraction of information from an e-commerce website's structure, content, and typical users' actions and returning a reliable model representing the website content, page relationships, and archetypical users given an e-commerce website.

The data used for this paper was gotten from data pulled down from e-commerce websites with a web crawler's help. On the niche dedicated e-commerce website, the web crawler was used to build a web graph that included: 2,687 pages crawled (web graph nodes); 361,344 useable links were discovered (web graph edges). When the crawler was used on the General-Purpose E-commerce Website, the web graph gotten included: The total number of pages crawled (graph nodes) was 621,303, and the total number of valid edges identified was 11,044,225. There was a total of 50,000 unique users extracted.

Clustering was done using the k-means algorithm to find the website's archetypical users. This yielded five clusters, with 26 customers being excluded as outliers.

Future work may be done to use other pattern discovery techniques to find the website's archetypical user (Dias & Ferreira, 2017).

2.2.1.16 Mining Shopping Behaviour in The Taiwan Luxury Products Market

This paper aims to use data mining techniques to extract information from luxury product buyers in Taiwan.

The information in this paper was gotten from questionnaires sent out to members of the public who have purchased high-end goods. In this study, 1946 valid questionnaires were used. The questionnaire had six sections.

In this study, association rule was carried out using an apriori algorithm, and clustering analysis was done using the K-means algorithm.

K-means was used to segment data into "basic consumer information" and "consumer behaviour and reason". In this study, the system splits all objects or data with a high degree of likeness into various clusters; however, since there are significant variations between these subject clusters, the aim is to differentiate between different consumer groups' purchase product actions. K-means algorithm produced three clusters. Cluster 1 had a sample size of 772 and was named "pragmatism of consumption groups". Cluster 2 had a sample size of 637 and was called "maturity frequency of consumption groups". Cluster 3 had a sample size of 537 and was named "fashion of consumption groups".

The apriori algorithm was primarily used for evaluating the relationships among items or features that synchronously happen in databases. The Apriori algorithm performs cluster analysis on the individual k-means cluster groups (Wen et al., 2012).

2.2.1.17 Analysing Online Consumer Behaviour in Mobile and PC Devices: A Novel Web Usage Mining Approach

The aim of this paper is to look at and make a comparison between online customer behaviour on an e-commerce website on mobile and PC devices.

The data for this paper came from a well-known Israeli e-retailer with a diverse product offering. The web server log files used in this paper were gathered over 30 days.

This paper used the web usage methodology as its methodology. Data pre-processing, usage mining, and pattern analysis are the three phases of the technique.

Data processing's primary objective is to prepare session-level data and extract session characteristics from the raw data. It may also include episode recognition, which entails classifying subsets or sub-sequences of page views that are semantically or functionally linked into a concept hierarchy.

To find variations between users of mobile and PC computers browsing behaviour, usage mining employs both statistical analysis and sequential association rule mining. Differences in interaction indicators between the two platforms are investigated using T-test tests. Identifying frequent sequences in the collection of navigation route patterns representing a session using sequential association rule mining. Finally, we analyze the relationship between pattern and conversion in pattern analysis, which describes the underlying browsing motivation. The mean values of all commitment indicators vary significantly (p 0.001), according to T-test analyses.

The cross-platform scope of this study makes it a significant contribution to e-commerce research, demonstrating behavioural variations across different technological channels.

In this paper, the data specificity and techniques used for web usage mining were its limitations.

Future research could expand the method suggested and used in this study to other forms of data and procedures and better understand how information systems are used in non-commercial situations (Raphaeli et al., 2017).

2.2.1.18 Clustering Retail Products Based on Customer Behaviour

This paper aims to identify items based on the quantitative and qualitative characteristics using a solely data-driven method.

The data used in this paper is a real sample from the Czech Republic's supermarket chain drugstore industry. From the entire year of 2015, a sample of receipts with at least four items.

Individual purchasing data from one of the Czech Republic's drugstore chains forms the basis of the dataset. The dataset consisted of 10,608 baskets, each of which included ten best-selling items from the drugstore's ten most common categories, totalling 100 products.

A new approach for retail clustering items based on customer behaviour was used in this paper. The genetic algorithm was used to cluster the data.

Various models with various clusters were developed. The effects of the findings were compared to k-means, Ward's hierarchical clustering, and self-organized maps, which are all simple clustering algorithms. The approach used in this paper found slightly different subcategories than K-means or Ward's hierarchical clustering. As a result, it was suggested that the results of both methods be investigated for practical purposes (Cerny et al., 2017).

2.2.2 Existing Systems in E-Commerce to Increase Website Popularity

This section covers the surface study on the existing system related to methods used to discover trends that can help to increase the popularity of the e-commerce website among its visitors.

2.2.2.1 The Determinants of Conversion Rates in SME E-Commerce Websites

This paper aims to define and analyze the factors that improve conversion rates or a combination of factors that will enhance conversion rates.

The paper's data is made up of 1184 daily observations from six small-business e-commerce websites in Italy that sell clothing, luggage, shoes, and related accessories. The data retrieved include:

- Conversion Rate (%)
- Free shipping (0 - no, 1- yes)
- Free returns (0 - no, 1 - yes)
- Discounts
- Season (0 - regular, 1- sales)
- Speed of load (0 - 100)

- Luxury websites (0 - non-brand products, 1 - branded products)
- Week (0 - Saturday & Sunday, 1- Monday to Friday)

To achieve the paper's aim, exploratory regression analysis is used to determine which determinants are important. Qualitative Comparative Analysis (QCA) highlights more details on the conditions where conversion rates improve.

An OLS regression was used to test the impact independent variables have on conversion rates. Standard errors were clustered to correct the correlations among the data from the same source as the data involves multiple observations over time that is stored within the six retailers. Two models were developed, and they both showed that all variables appear to affect conversion rates except free returns and day of the week product is purchased. Some variables had a difference in p-value, but the differences were $p < 0.05$. Model one had an R-squared of 0.647 and a MdAPE (Median Absolute Percentage Error) of 7.34, while model two had an R-squared of 0.646 and a MdAPE (Median Absolute Percentage Error) of 7.36.

The individual independent variables impact on conversion rate are tested in regression analysis. A QCA is used to assess the effect of any combination of variables on the conversion rate. The first step in QCA is to identify all the different attributes that affect the conversion rate. Step two: Using a fuzzy-set calibration, model the degree to which other cases belong to a set, which varies from 0 to 1, including intermediate membership levels. The third step is to decide which attribute configurations will serve as appropriate conversion rate conditions. The consistency test must meet a minimum requirement of 0.75 to be deemed adequate. Step four: From the appropriate configurations, remove any redundant attributes. The consistency of the five configurations is significantly higher than the minimum threshold. The five configurations can be grouped into two groups. The effect of discounts associated with promotional seasons and, in most cases, free shipping is the focus of Group A. Luxury e-commerce is identified as Group B, which is described by a combination of luxury products and load speed.

A significant contribution made by the paper was the fuzzy analysis used provided a more detailed perspective.

The paper's limitations include: The dataset (Individual transactions that cannot be tracked information may be lost if aggregated data is used.) and the paper was focused on funnel process, so some factors that affect the amount of traffic to the website were not considered.

More research is required to determine the effect of other website quality factors on conversion rate and consider other factors that affect the purchasing decision (Fatta et al., 2018).

2.2.2.2 Classification of Design Parameters for E-Commerce Websites: A Novel Fuzzy Kano Approach

This paper aims to classify and analyze design parameters based on consumer expectations to evaluate the usability of e-commerce websites more thoroughly.

The data used in the paper came from a 147-person survey on an e-commerce website usability.

The Kano model and the fuzzy set were combined in this analysis.

The membership degrees of design parameters to the classes are determined using a Kano model combined with fuzzy sets. Since the Kano model has a flaw in that, it considers most responses when determining a parameter class, and so fuzzy sets are merged into it.

The study's limitation was that it only looked at the characteristics of Turkish youths.

Future research should concentrate on the effect of these criteria on e-commerce websites' usability, with people from diverse backgrounds participating (Ilbahar & Cebi, 2017).

2.2.2.3 Loyalty of Young Female Arabic Customers Towards Recommendation Agents: A New Model for B2C E-Commerce

This paper aims to use an e-commerce recommender agent to investigate the major factors that affect female online shoppers' loyalty creation.

This study's data came from a survey of female students at Imam Abdulrahman Bin Faisal University who used Amazon.com.

Structural Equation Modelling (SEM) and partial least squares were used to analyze the relationships between constructs. SEM integrates the advantages of factor, path, and multiple regression analysis to develop a robust approach for evaluating construct relationships. Both above promoted the use of SmartPLS to analyze the survey findings.

This study's key contribution was introducing, creating, and analysing a model to improve e-commerce customer loyalty through recommender systems.

Data distribution and selection were limitations of this study; only a single e-commerce platform was evaluated, and questionnaires were distributed to students at a single Saudi Arabian university.

Future research is required to compare the effects of various recommendation methods on user behaviour and look at the user interface design of the website the recommender system is hosted on (Ali et al., 2020).

2.2.2.4 Acquiring Customers Through Online Marketplaces? The Effect of Marketplace Sales on Sales in A Retailer's Own Channels

Using time-series category sales data, this paper aims to determine which of the two opposing forces (online marketplace and retailer's webshop/website) prevails.

This study's data came from a webshop run by a multinational retailer of refurbished electronics and used media that operates in a vast global marketplace.

The data collection contains aggregated daily sales data from each category on the retailer's website and on the marketplace and visits to the retailer's website in the company's major country of operation.

The relationships between the online marketplace and the retailer's webshop were modelled using vector autoregressive (VAR) models. It demonstrates that revenues on a retailer's website increase due to marketplace sales. The rise is most significant in categories with a wide range of products and low prices. It also demonstrated that attracting consumers via the marketplace could be less costly than other approaches.

By examining how selling on an online marketplace affects a retailer's webshop, this paper contributes to the debate about whether selling on an online marketplace is the right decision. It also closes a cell of the channel cross-elasticity matrix that has yet to be investigated: the effect of marketplace revenues on a company's website.

The focal company's unique industry background could threaten the generalizability of the results, and the available data were restricted in multiple ways, preventing more thorough testing of the planned mediation process.

Future research into the impact of marketplace sales is needed (Maier & Wieringa, 2020).

2.2.2.5 An Empirical Analysis of Factors That Influence Retail Website Visit Types

This research aims to analyse websites to establish a typology of website visit behaviours and identify factors that are correlated with each type of visit. Three retail websites are under investigation.

This paper's data came from a speciality fashion retailer in North America that sells across two channels: in-store and online and has three distinct brands. The data used is based on the website visits of a group of 2,186 customers who were new to the retailer at the time of the data collection.

The Latent Cluster Model (LCM) is used to develop typologies for each brand's website visit. "Touching base," "search/deliberation," "goal-directed," and online shopping "cart-only" visits were the four forms of visits that were consistent across all brands. One of the brands has a new form of visit known as "considered visits." Maximum likelihood was used to estimate LCM. The Average Weighted Evidence (AWE) value, which is based on the model's log-likelihood, determines the number of visit type clusters for each brand.

This research adds to established knowledge by proposing a refined website visit typology and undertaking an empirical review of the variables associated with these types of visits.

As for all other research, this one has certain limitations. The data used only covers searching at websites operated by the same retailer, and the analysis was purposefully restricted to new consumers of a brand.

The development of topology for both experienced and new customers should be the focus of future studies. Future research should look at how the proposed typology can be used to personalize the website content and whether this results in a positive result (Pallant et al., 2017).

2.2.2.6 What Factors Determine E-Satisfaction and Consumer Spending in E-Commerce Retailing?

This paper aims to examine customer satisfaction in the e-commerce industry. To pinpoint the factors that influence customer e-satisfaction and the correlation between customer satisfaction and consumer spending in e-commerce retailing.

The American customer satisfaction index (ASCI) in terms of the top 115 e-retailers in the United States, their market value in the United States spent by customers, and their consumer price index (CPI) in the United States are among the data used in this study.

Using regression analysis and panel data analysis, the study analyzes the influence of customer satisfaction on consumer spending in American-based e-commerce companies.

According to the data, customer satisfaction impacts consumer spending in American e-commerce companies. Furthermore, there is a link between customer satisfaction and consumer expenditure, with better e-satisfaction leading to higher e-commerce spending. The findings also show that there is a clear link between e-service efficiency, e-satisfaction, and e-loyalty and online spending by users.

Before conducting the regression analysis, the paper used the unit toot test to ensure that the time series data was stationary. For the entire sample, the R-Square and modified R Square coefficients of determination are 0.783 and 0.773, respectively.

The pooled model and random-effect model based on panel data analysis both passed the T-test, with period regression coefficients of 0.378 and 0.384 for the ACSI and CPI. This means that if a specific e-customer retailer's satisfaction changes by one unit, consumer spending in particular e-retailers located in the United States would rise by 37.8% and 38.4%, respectively, as calculated by the correlation study.

Future research could concentrate on customer complaints and how they are handled and the return policy of American-based e-commerce retailers, which is directly related to online customer satisfaction (Nisar & Prabhakar, 2017).

2.3 REVIEW OF EXISTING METHOD

This section explains the various methods used according to the literature review above and their findings.

2.3.1 Existing Methods in E-Commerce to Drive Sales

This section explains the various methods used according to the literature review above and their findings to get more customers and increase sales.

2.3.1.1 Structural Equation Modelling

This method was used by Ali et al. (2020) to analyze the relationships between constructs. SEM assessed the relationship of independent variables and dependent variables, which is recognized for its use in quantitative research by the research community. SEM combines the advantages of factor, path, multiple regression analysis, and a robust methodology to assess the relationships among constructs.

2.3.1.2 K-Means Algorithm

Carmona et al. (2012), Dias & Ferreira (2017), Hong & Kim (2012), Liao et al. (2011), Rho et al. (2004), Videla-cavieres & Ríos (2014) and Wen et al. (2012) all employed the use of K-means algorithm to develop models.

The K-means algorithm is a widely used and effective clustering algorithm. It was observed by Videla-cavieres & Ríos (2014) that k-means does not produce a meaningful result when the data is enormous and is spar in nature and, as such, discarded the use of the method.

To improve customer segmentation, Hong & Kim (2012) integrated SOM (Self Organizing Map) model and the K-means model to produce a more sophisticated and detailed customer segment. The k-means algorithm, according to Hong & Kim (2012), is not realistic for

massive datasets due to its space complexity, which varies depending on the size of the data set.

Wen et al. (2012) used the K-means algorithm to classify data into the specified grouping variables, then performed cluster analysis on each cluster group using the apriori algorithm.

2.3.1.3 K-Nearest Algorithm

K-Nearest algorithm was used by Hong & Kim (2012) and Salminen et al. (2019) to develop models.

K-Nearest Neighbour allocates points to data, compares them using a distance metric, and classifies them based on the nearest points' labels. Salminen et al. (2019) noted that K-Nearest Neighbour support multilabel classification, and before training, it's preferable to use a dimension reduction technique on the data. Salminen et al. (2019) computed the algorithm runtime of KNN, Random Forest and Neural Network with the same data and found out that KNN is the fastest.

Hong & Kim (2012) got a K-Nearest algorithm accuracy of 89.74%. The k-nearest neighbours' system, according to Hong & Kim (2012), helps to stop endless clustering by locating the nearest neighbours. By computing the pairwise distance between the segmented dataset and the new dataset, K-nearest neighbours infer segments of a new dataset.

2.3.1.4 ID3(Iterative Dichotomiser)

Bejju (2016) used ID3(Iterative Dichotomister) to predict the online rating with an accuracy of 86.4780%.

For the visualization of probabilistic business models, decision trees are used. ID3 is a simple decision tree algorithm that uses a top-down greedy approach to check each tree node attribute to construct a specified dataset decision tree.

2.3.1.5 1-K-Medoids

1-K-Medoids was employed by Quelhas et al. (2015) for the clustering analysis of product characteristics and customer characteristics.

The K-Medoids algorithm finds k clusters in n objects by: (1) randomly selecting a representative object (the medoid) for each cluster; (2) associating each remaining object with the medoid to which it is most similar; (3) updating the medoids by selecting the most representative object in each of the k clusters; and (4) repeating steps two and three until convergence or a stopping criterion is reached.

K-Medoids were employed instead of K-Means since K-Means can only be used for numerical results. The K-Medoids technique relies on object dissimilarities, which may be calculated using any dissimilarity function. As a result, it may be used to mixed data, such as the data in this project. Furthermore, it uses representative items as reference points, whereas the K-Means impacts may be imperceptible. The method takes the number of clusters, k, as an input parameter.

Despite being less susceptible to outliers than K-Means due to the use of the median rather than the mean, Quelhas et al. (2015) noted that the K-Medoids clustering algorithm still involves the a priori description of the number of clusters.

2.3.1.6 CN2-SD Algorithm

For customer segmentation by subgroup discovery, Quelhas et al. (2015) used the CN2-SD algorithm. The rules provided by the model define a subgroup of the population whose composition varies from that of the whole population.

To establish the unusualness of a rule, the difference between the proportion of instances in the subgroup that belonged to the class in the consequent of that rule and the proportion of the same class examples in all orders should be determined.

For subgroup discovery, the CN2-SD algorithm is an extension of the CN2. The CN2-SD algorithm was introduced by Quelhas et al. (2015) in two different variants. The first can only work with binary target variables, whereas the second works with categorical variables.

The use of the Weighted Relative Accuracy heuristic (WRAcc) accuracy measure to determine the consistency of the rules and the use of the weighted coverage algorithm are the most important improvements in the adaptation of CN2 for subgroup discovery dubbed CN2-SD.

The two most significant changes in the adaptation of CN2 for subgroup discovery referred to as CN2-SD are the Weighted Relative Accuracy heuristic (WRAcc) accuracy measure to evaluate the quality of the rules and the use of the weighted coverage algorithm.

2.3.1.7 OLS Regression

This was the method used by Fatta et al. (2018) to test the impact of independent variables on conversion rate with a MdAPE of 7.34% and 7.36% for the two models developed. They noted that OLS regression only test for the effects of individual independent variables and not a combination of independent variables, so they employed the use of Qualitative Comparative Analysis (QCA) as an addition to OLS regression analysis.

2.3.1.8 Support Vector Machine (SVM)

The sentiment analysis algorithm used by Saura et al. (2019) to demonstrate the possibilities provided by social data for brand communication analysis in the financial sector was the SVM.

The SVM is a supervised sentiment analysis tool that uses key factor classification and categorization. For Big Data analytics, SVM algorithms use machine learning and are trained to improve their results' integrity.

Before the tweets are classified into different sentiments (negative, neutral, and positive), the algorithm was trained to increase their accuracy. After Saura et al. (2019) trained the support vector machine algorithm, an accuracy of 71.1% was achieved. This accuracy level was above the conventional threshold of 60%.

The downside of sentiment analysis, according to Saura et al. (2019), is that it ignores the context in which the tweet is sent. Also, sentiment analysis is less applicable to data containing sarcasm, irony, or metaphor.

2.3.1.9 Neural Network

Neural Network was used by Salminen et al. (2019) for multilabel classification. To approximate a function from the input to the output, a neural network performs several matrix multiplications. With high-dimensional data, neural networks function better.

The Keras library is used to develop neural network architecture for Neural Networks (NN). Since Keras does not by default support scikit-learn levels of cross-validation, a custom class was created to cross-validate and test the neural network. Compared to the other two models, the NN model outperforms them all used by Salminen et al. (2019) with an Average F1 Score: 0.627. It was noted that the Neural network has no performance bottleneck.

It was noted by Salminen et al. (2019) determining the number of epochs used by a neural network model is an essential feature in optimizing it. In the Salminen et al. (2019) case, the best results are obtained after four epochs, after which the F1 Score starts to decline. An epoch is a single iteration of the neural network over the entire training set, during which the weights connecting each neuron are changed.

2.3.1.10 Random Forest

Random Forest was used by Salminen et al. (2019) for multilabel classification. The random forest divides the data into several decision trees or statistical data structures based on criteria that separate the label best and average them to provide a balanced forecast.

It was pointed out that before using the random forest to train the data, it is preferable to use a dimension reduction technique. In Salminen et al. (2019) study, the random forest had the least performance between the three machine learning classifiers used in the study. It yielded an average F1 score of 0.458.

2.3.1.11 Apriori Algorithm

Carmona et al. (2012), Liao et al. (2011) and Wen et al. (2012) made use of the Apriori algorithm as their association rule algorithm of choice.

Association rule learning is a descriptive data mining methodology that uses unsupervised learning to uncover interesting relationships between variables in large datasets.

The Apriori algorithm is a data mining technique comprised of association rules that are used to derive information from a customer database.

As a result, applying the association rules algorithm entails analyzing random data for synchronous relationships and using these relationships as a decision-making reference.

The Apriori algorithm generates an association rule of form X? Y, where X and Y are item sets. The aggregation of all items in the investigation's domain consists of a series of transactions. Apriori refers to a data set's training instances, items are binary attributes, and item sets are feature combinations. Its confidence and support determine the quality of an association rule established by Apriori.

Wen et al. (2012) used cluster analysis to analyze each of the K-means algorithm cluster groups.

2.3.1.12 NMEEF-SD

Carmona et al. (2012) used NMEEF-SD to extract knowledge from a dataset containing information about a user's history associated with an e-commerce website.

NMEEF-SD algorithm was used to find subgroups. The main goal of subgroup exploration is to uncover fascinating relationships in data pertaining to a particular property.

The NMEEF-SD algorithm is a fuzzy evolutionary approach that extracts informative fuzzy and/or crisp rules for the SD task based on the type of variables present in the problem. The NMEEF-SD algorithm is based on the NSGA-II method and can be used to discover subgroups using a wide range of quality measures. NMEEF-SD algorithm makes use of

quality measures of significance (SIGN), unusualness (UNUS), sensitivity (SENS) and fuzzy confidence (FCNF).

2.3.1.13 Latent Attrition Model

Jaiswal et al., 2018 used the latent attrition model to evaluate customer retention in a probabilistic way dependent on transactional patterns of individuals.

Split sample analysis on the latent attrition model was mentioned, which involves estimating the latent attrition model differently in different levels of variables.

With a 1.8% error rate, the latent attrition model was used to predict the cumulative number of customers' transactions.

Estimates of model parameters are used to evaluate hypotheses. There are posterior means and posterior intervals in parameter estimates.

2.3.1.14 Self-Organized Map (SOM)

Hong & Kim (2012) and Videla-cavieres & Ríos (2014) made use of the Self-organized map (SOM) for segmentation and set discovery.

The self-organizing map (SOM) is a neural network-based artificial intelligence system that learns from data through unsupervised learning.

To improve customer segmentation, Hong & Kim (2012) combined SOM (Self-Organizing Map) and K-means models to create more sophisticated and detailed customer segments.

2.3.1.15 Vector Autoregressive (VAR) Model

This method was used by Maier & Wieringa (2020) to model relationships between an online marketplace and retailer webshop/website. VAR models are suitable for modelling the complex dynamic interplay between endogenously related variables, which is why they were chosen. They indicated that VAR models work well with aggregated data which was used for the study carried out. The complex dynamic nature of the interactions between the online marketplace and the retailer's webshop/website necessitated VAR models' versatility,

which allowed for the detection of dynamic marketing phenomena such as complementary or cannibalizing effects between two channels.

2.3.1.16 Web Usage Mining

Raphaeli et al. (2017) used a novel web usage mining approach to analyze online consumer behaviour.

This innovative approach used sequential association rule mining and statistical analysis to detect variations between users of mobile and PC computers browsing behaviour.

The set of navigation path patterns representing a session was extracted using a sequential association rule mining algorithm to find frequent sequences. Sequential patterns are the sequences of items that regularly occur in an adequately significant proportion of sequence transactions, which capture the trail of frequently visited web pages in the order users visited them. The 'support' and 'confidence' tests are widely used to test the usefulness of a sequential pattern.

Differences in interaction indicators between the two platforms are investigated using T-test tests.

2.3.1.17 Genetic Algorithm

Cerny et al. (2017) employed the use of the genetic algorithm for the clustering of retail products.

Cerny et al. (2017) noted that several parameters are to be appropriately chosen to use a genetic algorithm. The first parameter is the population size. More potential clustering can be investigated with a greater population of people, resulting in a better solution. The algorithm selects new populations iteratively after the initial population is generated at random. A parameter refers to the number of iterations called the number of generations.

An individual represents each candidate solution. Individuals' chromosomes store information about candidate solutions. A number of individuals with the lowest cost

function values (referred to as the elite population) pass on to the next generation unchanged. Cerny et al. (2017) noted that the elite population's ratio is from 0 to 0.2.

To reveal the properties of the method used, several simulations are performed.

2.3.1.18 Market Basket Analysis

Videla-cavieres & Ríos (2014) developed a novel market basket analysis method based on graph mining techniques to generate frequent product itemsets from transactional data produced by a retail chain.

The methodology is focused on product networks and overlapping community detection as a detection algorithm for frequent itemset, according to Videla-cavieres & Ríos (2014).

The result was the creation of a network of products focused solely on transactions, with each product connected to the others because they appear in the same ticket from the same buyer opportunity. The network is referred to as a co–purchased product network. After that, a temporary series of filters was added to verify the consistency and stability of the communities discovered. Temporally Transactional Weighted Product Network is the name given to the created network.

The detection of overlapping populations is an extension of conventional community detection. Videla-cavieres & Ríos (2014) point out that subgraphs are not always disjoint, which means that a node may belong to more than one subgraph.

2.3.2 Existing Method in E-Commerce to Increase Website Popularity

This section explains the methodology used according to the literature review above and their findings to discover trends to increase the popularity of the e-commerce website amongst its visitors.

2.3.2.1 OLS Regression

This was the method used by Fatta et al. (2018) to test the impact of independent variables on conversion rate with a Median Absolute Percentage Error of 7.34% and 7.36% for the two models developed. They noted that OLS regression only tests for the effects of

individual independent variables and not a combination of independent variables, so they employed Qualitative Comparative Analysis (QCA) to complement the OLS regression analysis.

2.3.2.2 KANO Model

This method was used by Ilbahar & Cebi (2017) to classify the feature/design parameters of an e-commerce website that affects customer purchases. Since the Kano model considers most responses when determining the class of a parameter, so a fuzzy test was added to the Kano model to address this shortcoming. The membership degrees of design parameters to the classes are calculated using a Kano model and fuzzy logic combination.

2.3.2.3 Structural Equation Modelling

This method was used by Ali et al. (2020) to analyze the relationships between constructs. SEM assessed the relationship of independent variables and dependent variables, which is recognized for its use in quantitative research by the research community. To evaluate the relationships between constructs, SEM combines the advantages of factor analysis, path analysis, multiple regression analysis, and a robust methodology.

2.3.2.4 Vector AutoRegressive (VAR) Model

This method was used by Maier & Wieringa (2020) to model relationships between an online marketplace and retailer webshop/website. It was noted that VAR models are suited to model the dynamic interplay between endogenously related variables which was the reason for it being chosen. They noted that it works well with aggregated data which was used for the study carried out. The complex dynamic nature of the interactions between the online marketplace and the retailer's webshop/website necessitated VAR models' versatility, which allowed for the detection of dynamic marketing phenomena such as complementary or cannibalizing effects between two channels.

2.3.2.5 Latent Cluster Model (LCM)

Pallant et al. (2017) used this approach to create typologies for the types of visits to a brand's website. They explained why they chose the Latent Cluster Model over other clustering methods. To begin, LCM provides a statistical basis for determining how many clusters to incorporate in the final solution. Second, LCM makes no assumptions about linearity or normality. Finally, unlike other clustering methods, LCM allows for the direct inclusion of covariates in the model, unlike other clustering algorithms that require a post-hoc analysis of cluster profiles.

2.3.2.6 Regression Analysis

This method was used by Nisar & Prabhakar (2017) to analyse the effect of e-satisfaction in e-commerce retailers and the relationship between variables. It was noted that the standard regression model procedure was followed. Panel data model-based regression analysis, which is a type of statistical model, was then introduced. They indicated that panel data sets were made up of three separate models: pooled model, fixed effects model and random effect model, but that only the pooled and random effect models were used in the analysis.

2.4 REVIEW OF RELATED FINDINGS

The diverse nature of the above-reviewed methods in which notable models and attributes were used does not change the fact that there can be improvements or development of models. Finding the right attributes to use in e-commerce analytics to drive sales or increase website popularity can never be over-emphasised as new attributes are explored daily to get better results.

The methods reviewed had different use because of the purpose of the research carried out. This limits the ability to handpick the best method to carry out our study but tables options for the use of attributes/data.

While most of the literature reviews focus on the methods used, the attributes vary. There is an excellent opportunity to study and develop different models using a wide range of e-commerce attributes.

With the enormous amount of data stored by e-commerce companies, there is a high demand for studies that use them and turn them into actionable recommendations.

CHAPTER THREE

METHODOLOGY

3.1 INTRODUCTION

This chapter focuses on the description of the proposed research methodology using a research framework. The research framework shows the step-by-step processes of the research design. The various processes in this research framework will be elaborated upon.

3.2 RESEARCH DESIGN

This section covers the step-by-step processes involved in the research design. The research framework below is used to express diagrammatical the processes involved in the research. The framework used is referred to as CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. Procedures in the research framework would also be elaborated on in this section.

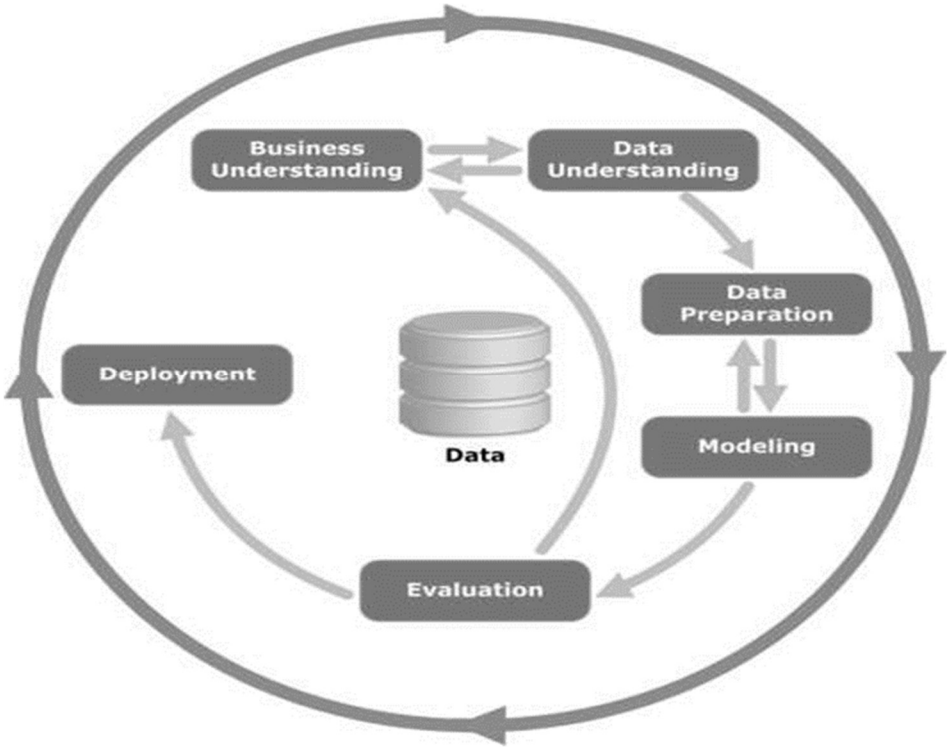


*Figure 3.1: Research Framework (*Big Data Analytics - Data Life Cycle*, 2021)*

3.2.1 Business Understanding

Step one is to understand the project goals and criteria from a business standpoint and translate that information to a data mining problem description. An initial plan is created to accomplish the goals (*Big Data Analytics - Data Life Cycle*, 2021). In this stage, you set your objectives, develop your project plan, and lay down the criteria for success.

3.2.2 Data Understanding

The data understanding process begins with data acquisition and continues with the task of becoming more acquainted with the data, identifying data quality issues, uncovering initial insights into the data, or discovering relevant subsets to create a hypothesis for hidden information (*Big Data Analytics - Data Life Cycle*, 2021).

Data acquisition refers to the activities involved in acquiring data generated by a source outside the organisation into the organisation for use. A data acquisition strategy must be defined before the acquisition of data. The data acquisition strategy is in stages:

1. Stage one – A specific purpose must be defined. Stating the issue, we are attempting to address or clarify.
2. Stage two – A clear cut understanding of the data. Which includes a thorough description of the data sample, an initial examination of the data and the evaluation of the quality of the data.
3. Stage three – involves the selection, cleaning, formatting, and merging of the data samples received.

There are two possible scenarios about the data source from which the data will be acquired, and they are Data sources with control over and Data sources without control over.

1. Data sources with control over - These data sources, you have control and the ability to monitor the data sources somehow. The control is over the data source and not the event in which the data is linked. This scenario is common in high-complexity or long-term projects in which a data model has been developed, and mechanisms

are in place to ensure its accuracy and integrity. This scenario is usually common in corporations and institutions.

2. Data sources without control over - It is the most common scenario in short-term or intermittent projects and projects, analyses, or long-term studies that use or need data other than their own from sources. Whatever the project, the strategy to acquire the data must be more accurate since any move outside its control will raise its research costs or even make the project untenable. Example Obtaining the number of covid-19 cases in a nation when the research team does not influence the development and distribution of the data(Silveira, 2020).

For this project, we will be dealing with the scenario of data sources we have control over. The data would be retrieved from Afrimash's DEngage platform. Data retrieved from Afrimash's DEngage platform contained transactional data, customer information data, product data, etc.

3.2.3 Data Preparation

Also known as Data pre-processing includes all activities that are performed to produce the final dataset, which will be used in the modelling tools (*Big Data Analytics - Data Life Cycle*, 2021). It is also the phase in which the data is transformed or encoded so that the algorithm can now easily interpret the data's features (Pranjal Pandey, 2019). Data preparation activities are liable to be repeated severally in no order. Tasks done are dependent on the dataset that is being worked on. Data quality assessment, data aggregation, data sampling, dimensionality reduction, and data encoding are all tasks in data preparation.

3.2.3.1 Data Quality Assessment

It is a method of evaluating data quality by looking for problems with the data, which may include missing, inconsistent, duplicate, unreliable, incomplete, or irrational data values. We would expound on a few methods to deal with them.

1. Missing Values: This is a variable that does not have a stored data value. It is prevalent to have missing values in your dataset. These values are often expressed as NaN or None. It can be resolved by eliminating rows with the missing data if the

dataset is big enough, estimating missing values, fill in missing values with whatsoever value that appears immediately after them in the same column, make use of the scikit-learn library's imputer class to fill in missing values with the data's (mean, median, mode) and to fill all null variables with 0 if working with numerical values.

2. Inconsistent Values: This is when the data entered contradicts other data in the dataset. E.g., Age field containing a phone number.
3. Duplicate Values: This is when a data object in the dataset is a duplicate of another.

3.2.3.2 Data Aggregation

It is the process of presenting data in a summarised form. It provides a higher-level view of the data as the aggregates are more stable than the individual data objects.

3.2.3.3 Data Normalization

Normalisation is the process of changing each original variable with a standardised version of the variable that has a unit variance. The effect of normalisation is that it gives all variables equal importance in terms of variability. Types of data normalisation techniques include:

1. Min-Max normalisation
2. Z-score normalisation
3. Normalisation by decimal scaling

3.2.3.4 Data Integration

It is the joining of data from various sources into a single store. The integration of various data sources may help reduce or eliminate redundancies in the data and improve the quality of the data mining and its speed.

3.2.3.5 Data Sampling

It is a statistical analytic approach that chooses, manipulates, and analyzes a representative group of data points in order to identify patterns and trends in a larger data set. Types of data sampling methods:

1. Simple random sampling: It involves choosing an entity at random from the whole population.
2. Stratified sampling: Samples are randomly chosen from subsets of the population formed based on common factors.
3. Cluster sampling: Samples are chosen randomly from clusters of the dataset based on a defining factor.
4. Systematic sampling: Sample is formed from the selection of intervals from the dataset. E.g., Selecting the 20th row of a 2000 row spreadsheet to create a sample size of 100 rows.
5. Multistage sampling: This approach, which is more complex than cluster sampling, divides the broader population into various clusters. Stage two clusters are then separated based on a secondary factor, and those clusters are sampled. This staging could proceed as further subsets are detected, grouped, and evaluated (Biscobing, 2018).

3.2.3.6 Dimensionality Reduction

It is the process of converting data from a high-dimensional space to a low-dimensional space in such a way that the low-dimensional representation retains any significant features of the original data, preferably such as its inherent dimension. For data mining methods to work efficiently, the dimension of a dataset, or the number of variables, must be lowered. Using domain expertise to eliminate or combine categories, data summaries to discover information overlap across variables, data conversion tools, and automated reduction procedures are some approaches to dimension reduction. Some data mining methodologies can also be used to reduce duplicate variables, such as regression models, classification, and regression trees.

3.2.3.7 Data Encoding

Is the transformation of categorical variables into numerical or binary variables to be accepted by the machine learning algorithm and still retain its meaning. Many modelling methods require categorical variables to be encoded, including linear regression, support vector machine, and neural networks.

3.2.3.8 Training/Validation/Test Split

It is used to split the dataset into two or three parts. This is done because all algorithms need to train, validated, and tested before being deployed.

- Training data: The part which is used to train a machine-learning algorithm to build a model.
- Validation data: The part of the dataset used to validate the various model fits(hyperparameter).
- Testing data: The part of the data used to test the model. It is used to measure how the machine learning algorithm model will perform with real-world data.
- Split ratio: This is the ratio in which data is split. It depends on the type of model being built and the dataset itself (Pranjal Pandey, 2019).

3.2.4 Modelling

Various modelling techniques are chosen and used in this stage, and their parameters are fine-tuned to achieve the best results. Usually, many techniques exist for the same data mining problem type. Some approaches have unique data format specifications.

3.2.4.1 Descriptive Analytics

It is the analysis of historical events/data to understand and discover trends in a business. There are different types of descriptive analytics, and each type tries to answer a particular question: Exploratory analytics answers the question of “WHY?” descriptive analytics answers the question of “WHAT?” and casual analytics answers the question of “CAUSE”. There are several tools used for descriptive analytics, such as:

1. Bar charts: This is used to compare groups using a single statistic. The length/height of the bar signifies the value of the statistic, and different bars represent the other groups being compared.
2. Line charts: Mainly used for time series data, where you would like to track changes over some time, long or short.
3. Scatter plot: Used to show the relationship between numerical variables. It can be used to identify outliers or a group of outliers in a dataset. It also indicates clusters in the dataset.
4. Boxplot: They are used for comparing subgroups of numerical data through their quartiles. Lines extending outside the box plot indicates variability outside the upper and lower quartiles.
5. Histogram: It is mainly used for identifying the shape of your distribution. It shows the frequency of the values (on the x-axis) using connected bars.
6. Heatmaps: A heatmap is a graphical representation of numerical data that uses colour to express values. Heatmaps are particularly effective in data mining for two purposes: showing correlation tables and showing missing values in data.
7. Geographical map: Used to represent your data geographically. Requires location data to be used (Shmueli et al., 2017).

3.2.4.2 Prediction Modelling

Predictive modelling is a technique used to predict the future by analysing historical and current data to produce a model to predict the future. In predictive modelling, data is gathered, a statistical model is developed, predictions are generated, and the model is verified (or altered) when new data becomes available. The two most widely used predictive modelling techniques include regression model (Simple Linear Regression and Multiple Linear Regression (MLR)) and Neural Networks. For this project, we would be applying Multiple Linear Regression and Neural Networks.

1. Multiple Linear Regression: This model is used to predict a numerical outcome variable Y (Target or dependent variable) from a set of predictors X_1, X_2, …, X_n (Independent or input variables). The hypothesis is that the function produced by the

model will produce the best possible prediction of the numerical outcome from the predictors (Shmueli et al., 2017):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Equation 3.1: MLR model function

Where:

$\beta_0, \ldots, \beta_n$ – Coefficients

ϵ – Noise or unexplained part

Y – Target variable

$X_1, \ldots, X_n$ – Independent variables

2. Neural Networks: Neural networks, often known as artificial neural networks, are classification and prediction models. The neural network is based on a model of biological activity in the brain, in which neurons communicate and learn through experience. Neural networks learn in the same manner as humans do. Neural networks' learning and memory qualities are like those of human learning and memory, and they also can generalise from particulars.

   Academics have examined a wide range of neural network designs, but so far, the multilayer feedforward networks have shown to be the most successful neural network applications in data mining. A feedforward network is a network that is completely linked and has a one-way flow with no cycles.

   These are networks with an input layer made up of nodes (also known as neurons) that merely accept predictor values and subsequent layers of nodes that receive input from the preceding levels. Nodes in each layer's outputs are inputs to nodes in the following layer. The final layer is known as the output layer. Hidden layers are those that exist between the input and output layers.

3.2.4.3 Classification Modelling

It is a type of prediction modelling that predicts/classifies categorical outcomes. A classification model tries to derive some conclusion from observable data. A classification model will attempt to predict the value of one or more outputs given one or more inputs. Examples of classification modelling techniques include classification tree, random forest, neural networks, decision tree and logistic regression etc. For this project, we will be making discussing classification trees and neural networks.

1. Classification Tree: This is the structural mapping of decisions that leads to the determination of an object's class. The branches of a classification tree represent attributes, while the leaves represent decisions. In practice, the decision-making process begins at the trunk and proceeds through the branches until it reaches a leaf.

   Classification trees are based on two main concepts. The first is the concept of recursive partitioning of the predictor variable space. The second is the concept of trimming based on validation data.

   A tree has two types of nodes. Nodes with successors are referred to as decision nodes, while nodes with no successors are known as terminal nodes and represent the partitioning of data by predictors (Shmueli et al., 2017).

2. Neural Networks: Neural networks, often known as artificial neural networks, can be used for classification and prediction modelling. The neural network is founded on the biological activity in the brain, in which neurons communicate and learn through experience. Neural networks learn in the same manner as humans do. Neural networks' learning and memory qualities are like those of human learning and memory, and they also have the ability to generalise from particulars.

   Academics have examined a wide range of neural network designs, but so far, the multilayer feedforward networks have shown to be the most successful neural network applications in data mining. A feedforward network is a network that is completely linked and has a one-way flow with no cycles.

These are networks with an input layer made up of nodes (also known as neurons) that merely accept predictor values and subsequent layers of nodes that receive input from the preceding levels. Nodes in each layer's outputs are inputs to nodes in the following layer. The final layer is known as the output layer. Hidden layers are those that exist between the input and output layers (Shmueli et al., 2017).

3.2.4.4 Clustering

Clustering is an unsupervised learning algorithm where the aim is to segment the data into a group of clusters to generate insight. There are different clustering approaches, but in this project, we will discuss two, which are Hierarchical clustering and K-means clustering (Non-hierarchical clustering).

1. Hierarchical clustering: Here, records are systematically grouped to form clusters based on distances between records and distances between clusters.
   Hierarchical clustering can also generate a graphical representation of the clustering process referred to as dendrogram.
   In hierarchical clustering, the measure of distance between clusters is implemented as:
   - Single linkage clustering: The minimum distance is used.
   - Complete linkage clustering: The greatest distance is utilized as the distance measure.
   - Average linkage clustering: The distance between the two clusters is the average distance.
   - Centroid linkage clustering: This is based on centroid distance, with clusters represented by their mean values for each variable, resulting in a vector of means.
2. Non-hierarchical clustering(K-means): In non-hierarchical clustering, the number of desired clusters, k, are specified and assigns each example to a cluster so that dispersion between each cluster is minimised.
   A measure of the distance between examples in the same cluster is the sum of the distances of the examples from their cluster centroid.

The choice of the number of clusters depends on previous knowledge, practical limitations or trying different clusters and comparing their results (Shmueli et al., 2017).

3.2.5 Evaluation

Model evaluation is the process of predicting the model's generalisation accuracy on future data. Model evaluation metrics are used to compare models in terms of model selection and prediction associated with a particular model, and a dataset is expected to be accurate.

In this project, the confusion matrix was used to evaluate the classification results

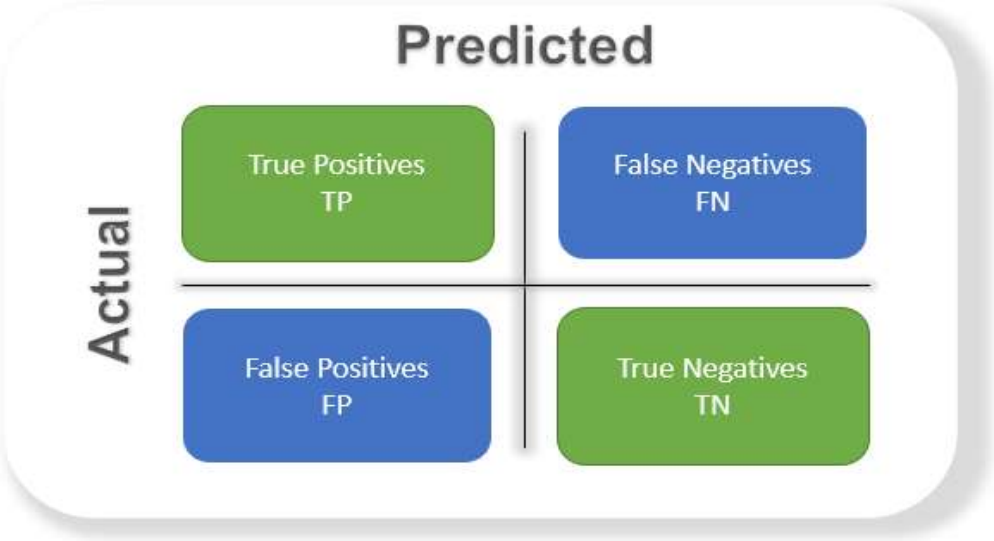


Figure 3.2: Confusion Matrix

- True positives (TP): Model predicted positive and was correct.
- False positives (FP): Model predicted positive while it is negative.
- True negatives (TN): The model predicted negative and was correct.
- False negatives (FN): Model predicted negative while it is positive.

Accuracy is the percentage of the total prediction which was correctly predicted.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Equation 3.2: Accuracy

Precision is the percentage of the rightly predicted positives to the number of positive predicted values.

$$\frac{TP}{TP + FP}$$

Equation 3.3: Precision

A recall is the percentage of the rightly predicted positives to the total number of positive values.

$$\frac{TP}{TP + FN}$$

Equation 3.4: Recall

Specificity/False Positive Rate is the percentage of the correctly predicted negatives to the total number of negative values.

$$\frac{TN}{TN + FP}$$

Equation 3.5: Specificity/False Positive Rate

F1 score is the harmonic mean of precision and recall. The higher the F1 score, the better.

$$F1 = 2 . \frac{Precision . Recall}{Precision + Recall} \Rightarrow \frac{2 . TP}{2 . TP + FN + FP}$$

Equation 3.6: F1 score

PR curve is the curve between precision and recall for threshold values. The higher the PR AUC value, the better. The PR AUC is the area under the graph (Nighania, 2018).

The ROC curve is a graphical representation of the output of a binary classifier system as its discriminating threshold is varied. The curve is constructed by plotting the true positive rate versus the false positive rate at different threshold settings.

Lift is a calculation of a predictive model's efficacy measured as the combination of outcomes obtained with and without the predictive model. Cumulative gains and lift charts are useful visual tools for evaluating model efficiency.

In this project, RMSE (Root Mean Square Error) is used to evaluate the prediction results. It is calculated by taking the square root of the absolute value of the correlation coefficient between real and expected values.

Gini coefficient is used for classification models. Gini = 2*AUC - 1, where AUC represents the region under the ROC curve. A Gini ratio greater than 60% indicates a decent one (*11 Important Model Evaluation Techniques Everyone Should Know*, 2016).

MAE (Mean Absolute Error) is the magnitude of the average absolute error.

The mean error is identical to MAE, with the exception that it retains the sign of the errors; therefore, negative errors cancel out positive errors of the same size. As a consequence, it reveals whether the forecasts are on average over-or under-predicting the outcome variable.

MPE (Mean Percentage Error) provides a percentage score based on how far predictions differ from actual values (on average), considering the direction of the error.

MAPE (Mean Absolute Percentage Error) metric provides a percentage score based on how far predictions differ (on average) from actual values.

3.2.6 Deployment

In some instances, the development of the model is not the end of the process. The knowledge obtained must be structured and delivered in a manner that is beneficial to the user.

The deployment phase may be as straightforward as producing a report or as complicated as applying a repeatable data scoring or data mining method, depending on the requirements(*Big Data Analytics - Data Life Cycle*, 2021)

Findings and recommendations can be reported to users in the form of a PowerPoint presentation, which can include data visualisations (using visualisation tools like Tableau etc.) for better explanation. Model built can also be integrated into websites or applications to be used.

3.3 DATA DESCRIPTION

For this project, the data used was extracted from the Afrimash DEngage platform. The datasets extracted from the Afrimash DEngage platform was from a period of six months, beginning from the 1st of January 2021 to the 1st of June 2021. The datasets include:

- Dataset of all contacts (called - master_contact)
- Dataset of all devices used to visit the site (called - master_device)
- Dataset of order event details
- Dataset of orders
- Dataset of order details

The following datasets above were combined into one dataset, which was the dataset used for the analysis.

3.4 CONCLUSION

The whole point of developing the CRISP-DM model was to help plan and organise a data analytics project. This is because it is very important to the success of the project. This chapter does just that by laying out the plan to be used to carry out our research analysis.

CHAPTER FOUR

IMPLEMENTATION AND RESULTS

4.1 INTRODUCTION

This chapter gives an overview of the data pre-processing, algorithms, implementation tools employed to execute the project's objectives. The implementation of the prediction, classification, clustering models/algorithms and their results would be explained and displayed.

4.2 DATA PREPROCESSING

Here we would describe all the processes taken to get the dataset that was used to build and train the machine learning models/algorithms for this project.

4.2.1 Data Description

The dataset used for this project was extracted from the Afrimash DEngage platform. They include:

1. Master_contact
   (contact_key, contact_status, email, name, surname, gender, segment, source, subscription_date, created_at, created_by, updated_at, updated_by, contact_type, email_subscriber, vendor,affiliate, shipping_city, shipping_state, email_permission, email_status, email_status_reason, is_email_active)
2. Master_device
   (device_id, token, contact_key, device_type, device_brand, os, os_version, browser, last_activity_date, created_at, created_by, updated_by, wp_subscription)
3. Orders
   (contact_key, order_id, order_date, order_status, order_source, item_count, total_amount, discounted_price, payment_method, shipping, coupon_code)

4. Order Event Details

   (order_id, discounted_price, unit_price, quantity, product_id, product_variant_id, key, event_date)
5. Order Event

   (camp_id, coupon_code, event_date, key, session_id, total_amount, shipping, payment_method, item_count, event_type, order_id, discounted_price, send_id)
6. Order details

   (product_variant_id, order_product_status, order_id, discounted_price, unit_price, quantity, product_id, contact_key)

These datasets were joined together to form a single dataset. Before the data pre-processing steps, the dataset contained 38 columns with 15709 rows. The attributes in the dataset include Order Id, Billing Country, Billing State, Browser, Created At, Date Registered, Device Brand, Device Id, Device Type, Email Permission, Email Status, Email Status Reason, Event Date, Event Type, Last Activity Date, Order Date, Order Source, Order Status, Os, Os Version, Payment Method, Product Id, Product Variant Id, Segment, Shipping Country, Shipping State, Subscription Date, Updated At, Wp Subscription, Created By, Discounted Price, Item Count, Quantity, Total Amount, Unit Price, Updated By, Shipping, Contact Key.

4.2.2 Data Integration

All the six datasets extracted from the Afrimash DEngage platform were joined together to form a single dataset that was used for analysis. The joining of the datasets was done using RapidMiner Studios.

First, the Order details, Order Event, Order event details and orders datasets were joined together to form order tables, as shown in the figure below.

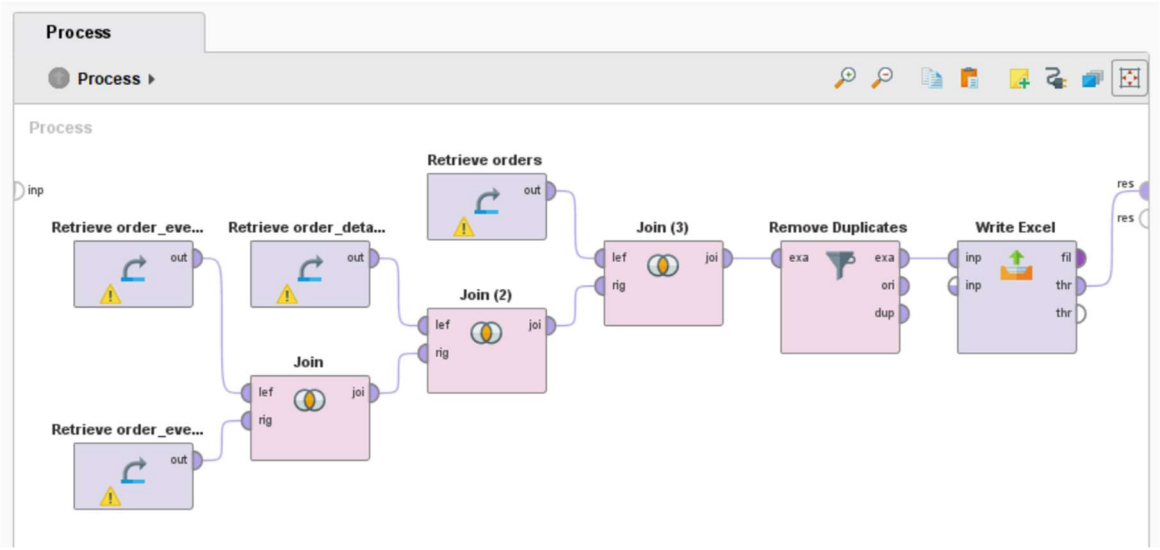


Figure 4.1: Order Table Join Process in RapidMiner Studio

The master_contact and master_device datasets were joined to form the customer table, as shown in figure 4.2.

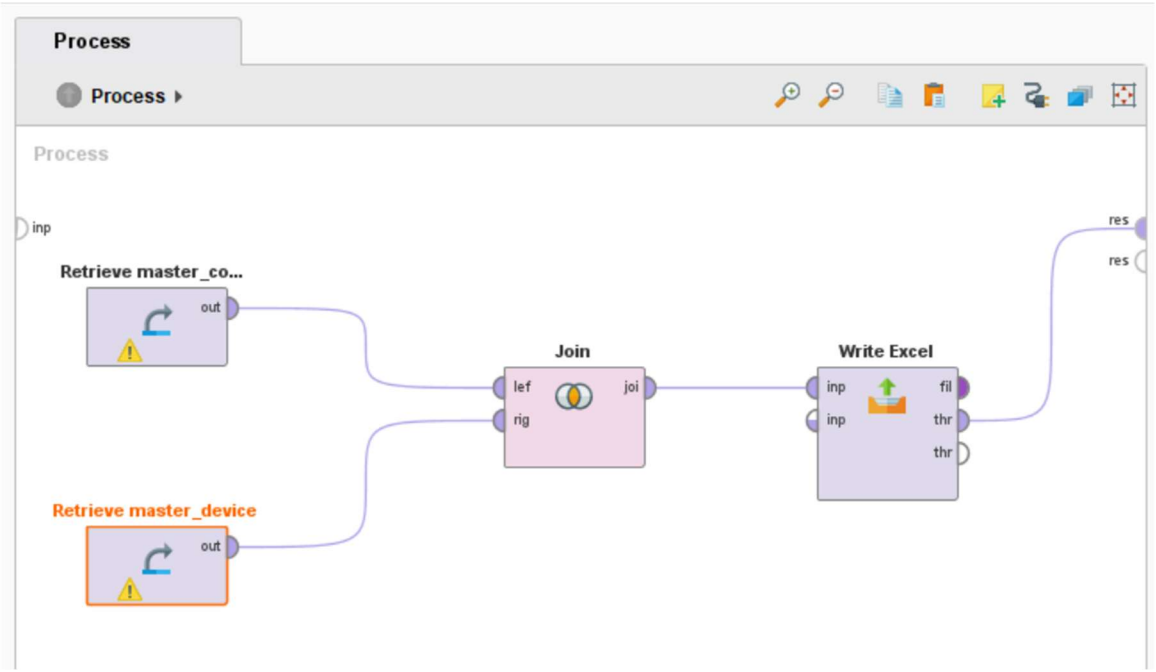


Figure 4.2: Customer Table Join Process in RapidMiner Studio

Finally, the customer table and order table were joined to give the final dataset, which was used for analysis as shown in figure 4.3 below.

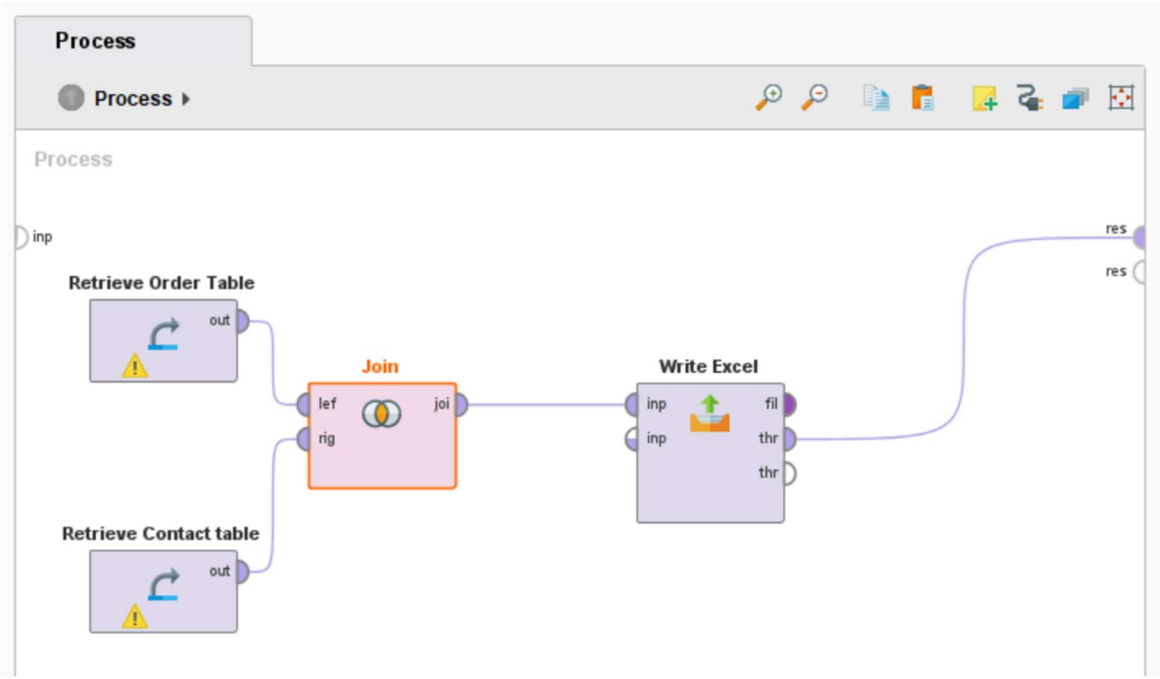


Figure 4.3: Final Table Join Process in RapidMiner Studio

4.2.3 Attribute Selection and Filtering

The dataset formed from joining was divided into two – those that have tried ordering from Afrimash and those who have not. This was done using the "Filter Example" operator in RapidMiner studios, as shown in the figure below.

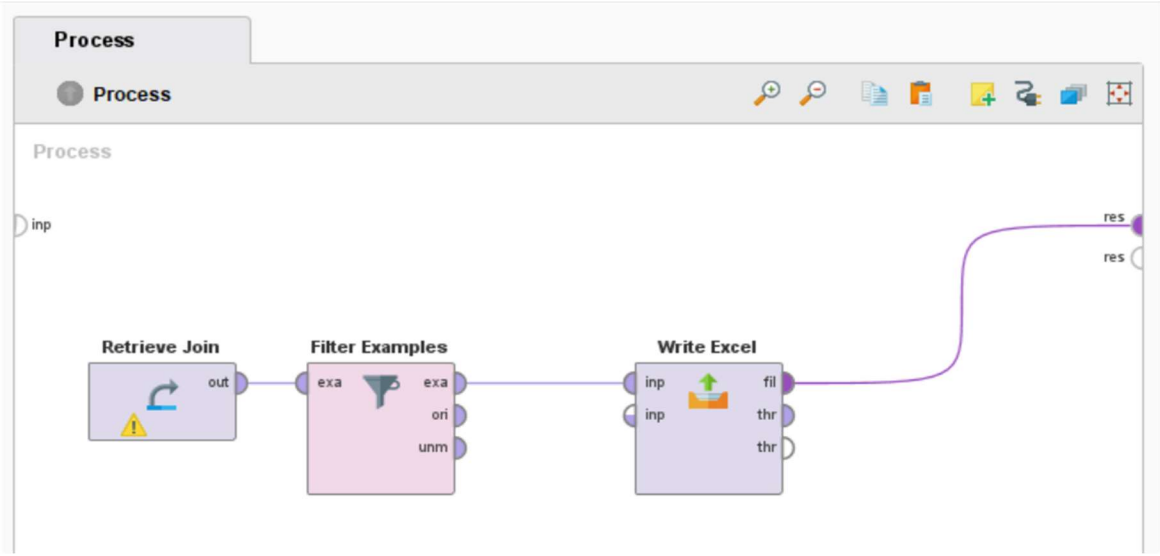


Figure 4.4: Filtering using RapidMiner Studios

After filtering, some columns in the datasets that were not needed were removed. This was done using the “Select Attribute” operator in RapidMiner studios, as shown in figure 4.5.

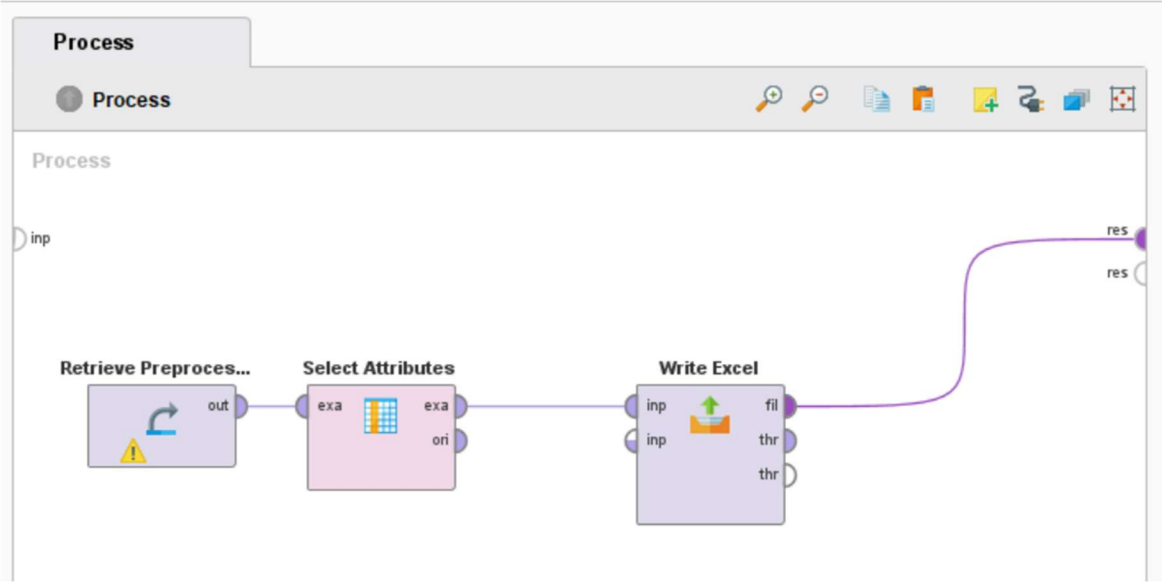


Figure 4.5: Selection of attributes using RapidMiner studios

4.2.4 Data Cleaning

Unfortunately for us, the data extracted wasn’t clean, and after data integration, it didn’t get better. Several data cleaning processes were used during this project, and we would look at them. The data cleaning done in this project were done in RapidMiner studio.

Duplicated data were removed from the datasets, as shown in figure 4.6 below. The duplicated data were primarily found in the order table.

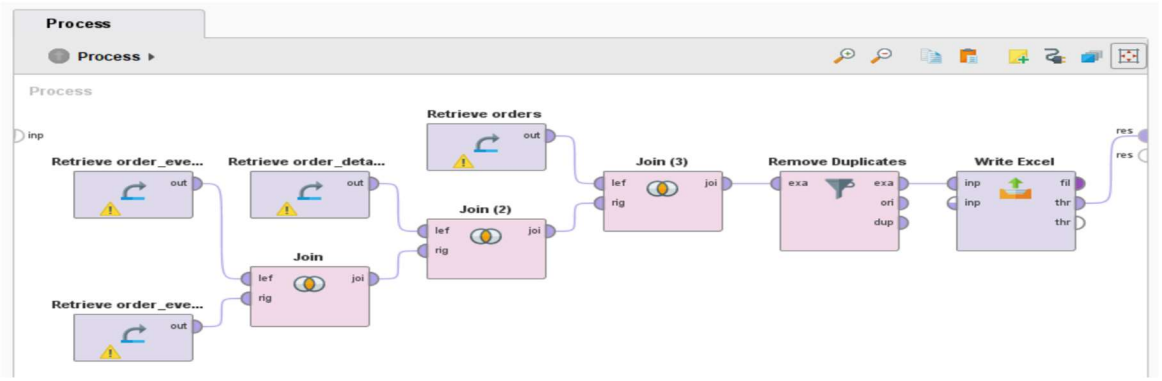


Figure 4.6: Removal of Duplicates from the Order Table in RapidMiner Studio

Some empty columns in the dataset were filled, as shown in the figure below, while others were deleted.

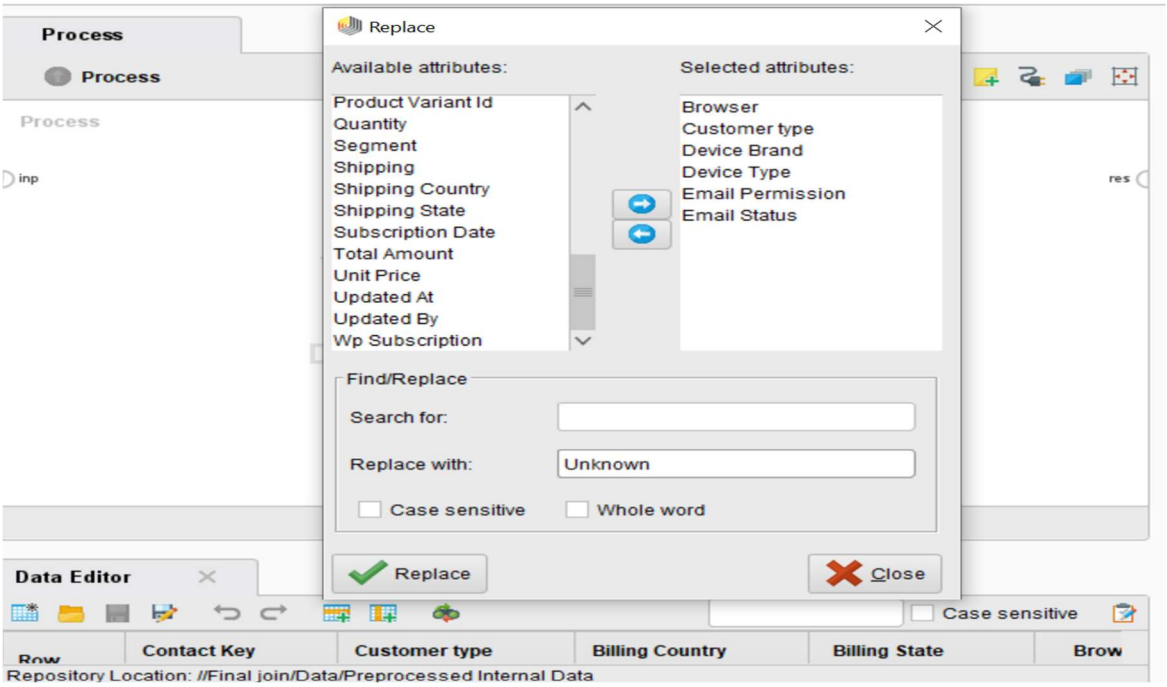


Figure 4.7: Filling of empty columns in RapidMiner Studio

Columns that contained abbreviations were replaced with the full meaning so that the data would be uniform, as shown in the figure below.

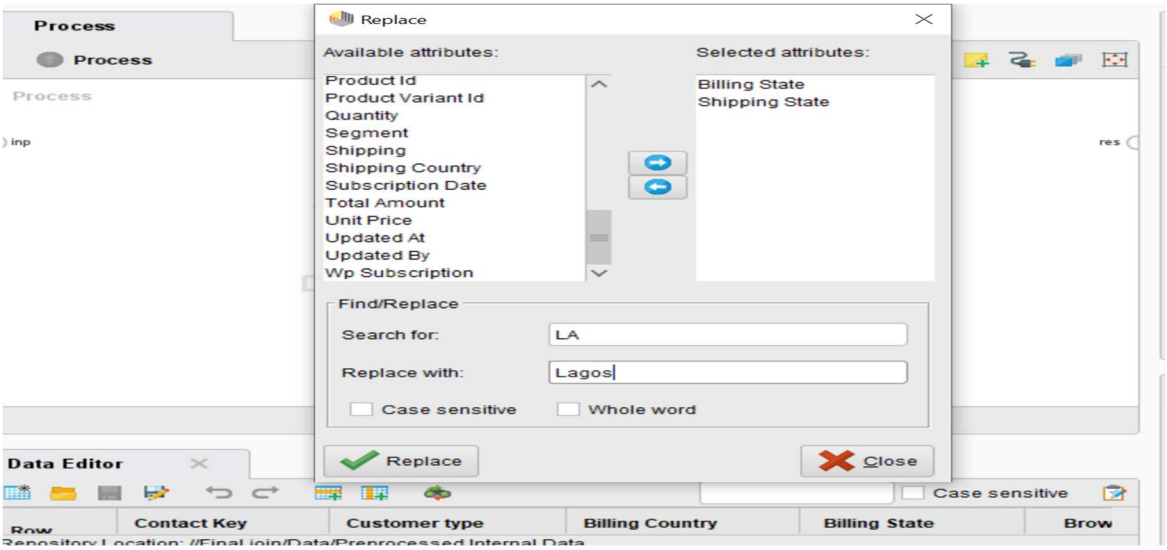


Figure 4.8: Replacing abbreviated words in dataset using RapidMiner Studio

4.2.5 General Descriptive Analytics

After most of the pre-processing steps were done to get the basic statistics of the dataset, we used the summary () method in R programming, which returns the statistics of the data as shown in figure 4.9 and figures 4.10 below.

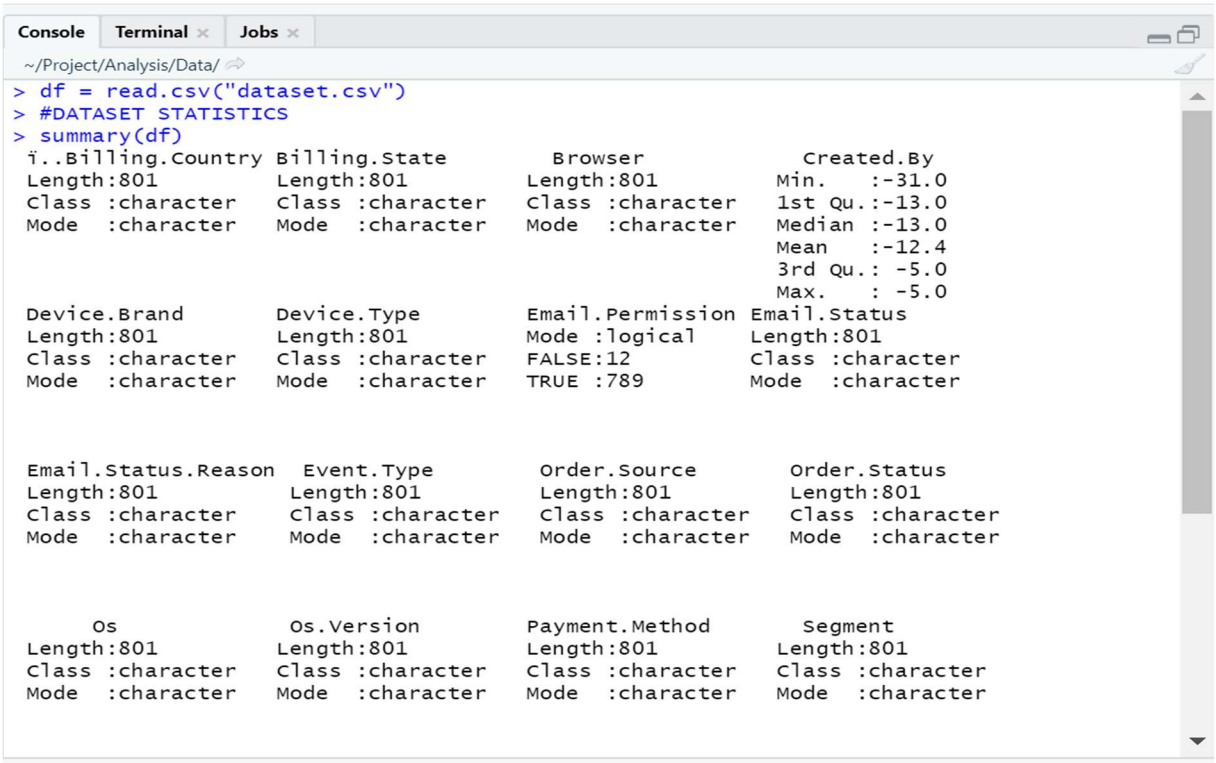


Figure 4.9: Implementing the summary () function in R

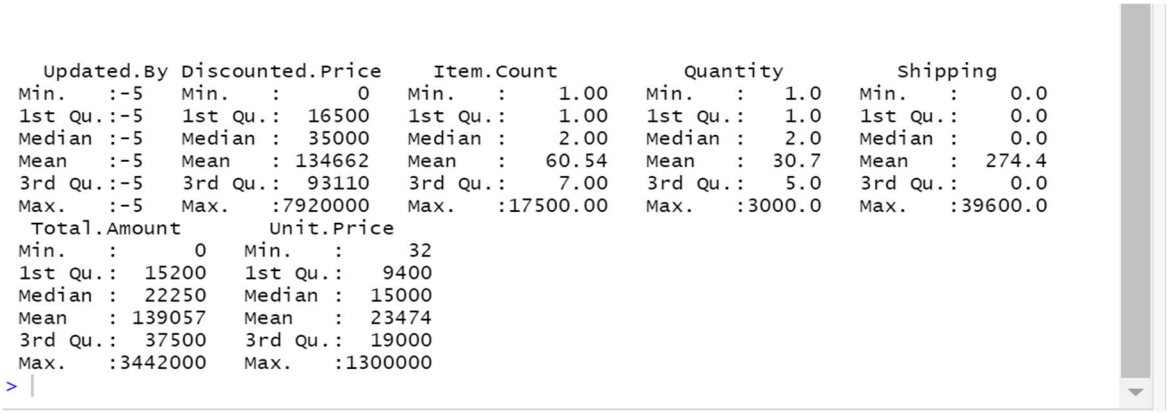


Figure 4.10: Implementing the summary () function (2) in R

4.2.6 Data Binning and Encoding Categorical Variable

Data binning was done after the data cleaning process to minimize small observation errors. Data binning was done using the group function in Tableau desktop, as shown in the figure below.

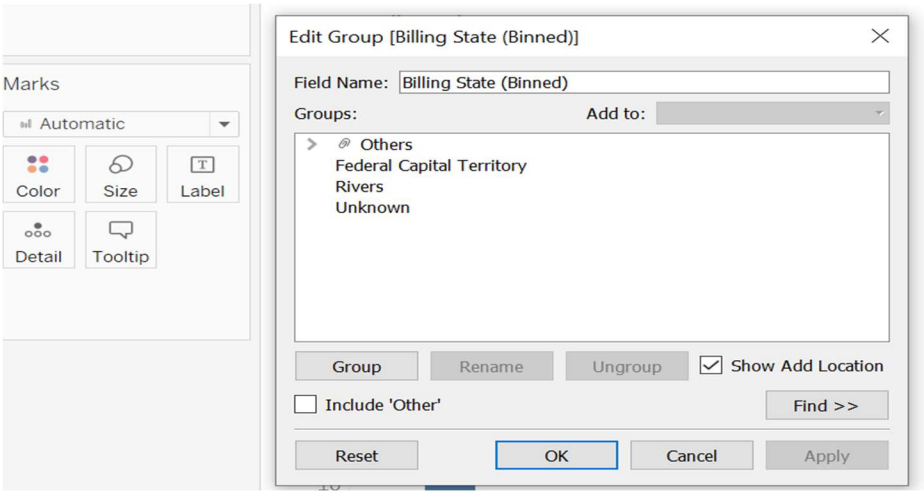


Figure 4.11: Binning of attributes in Tableau

After the data were binned, the categorical variables were converted to numerical variables. The conversion was done using the group function in Tableau desktop, as shown in the figure below.

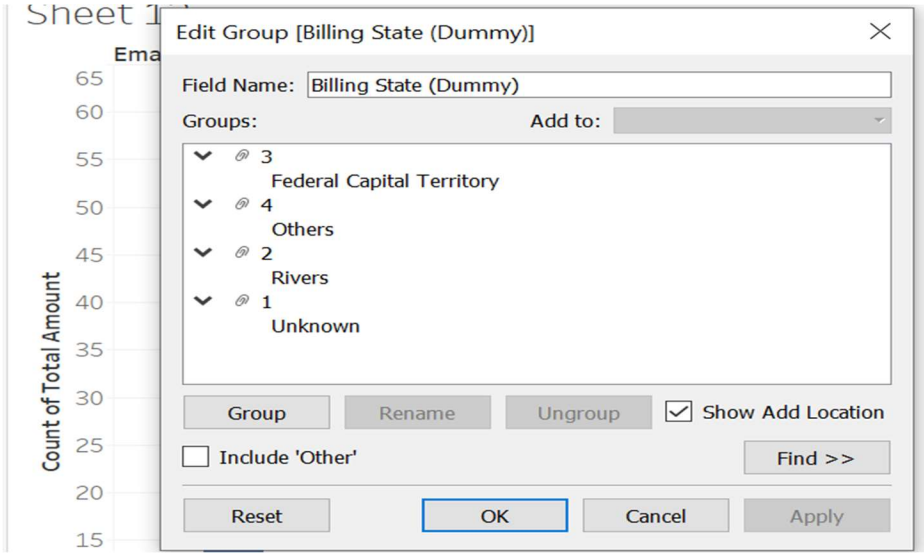


Figure 4.12: Encoding of categorical variables in Tableau

4.3 DESCRIPTIVE ANALYTICS

A little descriptive analytics was carried out on the dataset. The descriptive analytics was carried out using tableau to visualize the data. To start the visualization, we looked at the number of customers per region/state that has tried to place an order. It was visualized using a bubble plot and a map, as shown in figure 4.13 and figure 4.14, respectively, below.

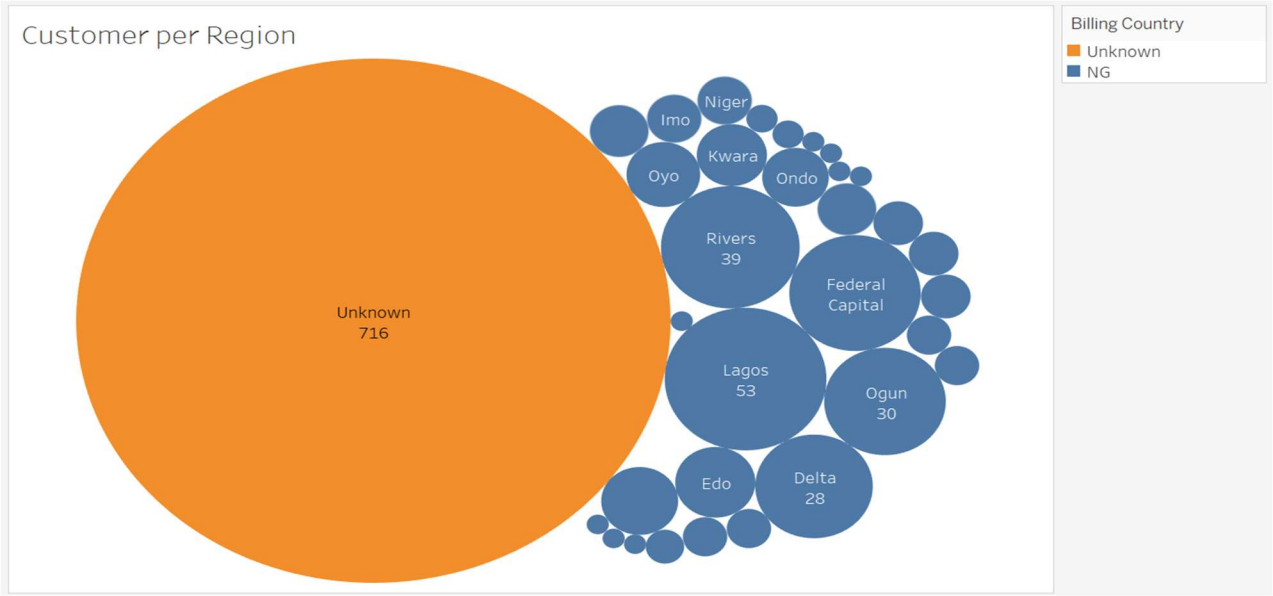


Figure 4.13: Customer per region bubble plot

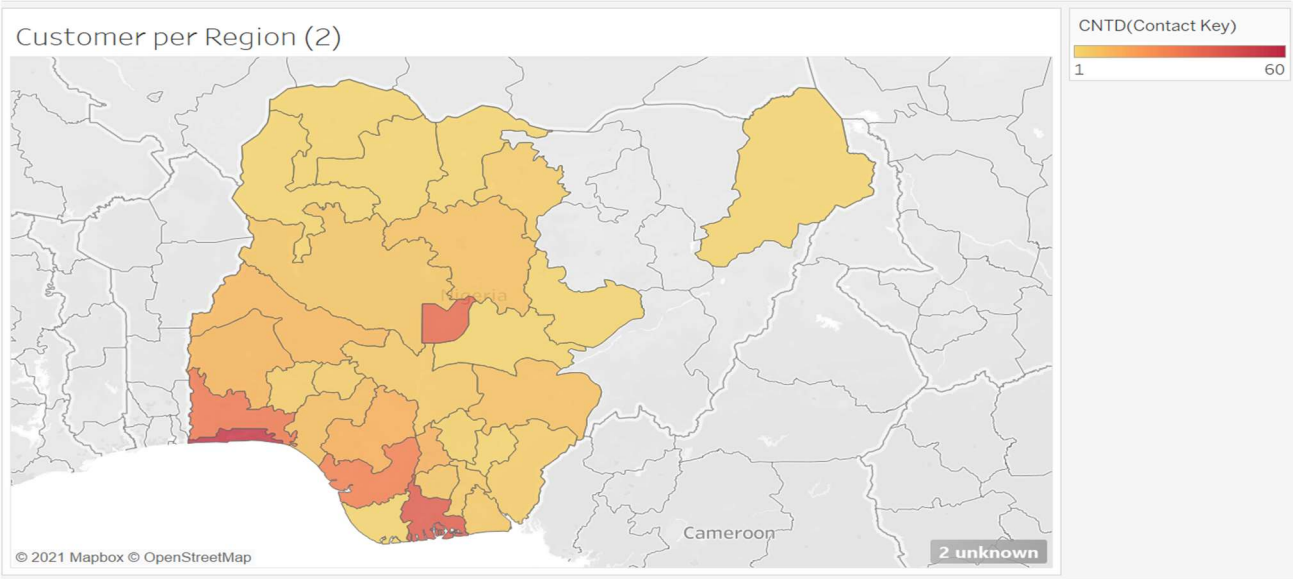


Figure 4.14: Customer per region map

Analysis of the amount spent per region was done. It was visualized using a bubble plot and a map, as shown in figure 4.15 and figure 4.16, respectively, below.

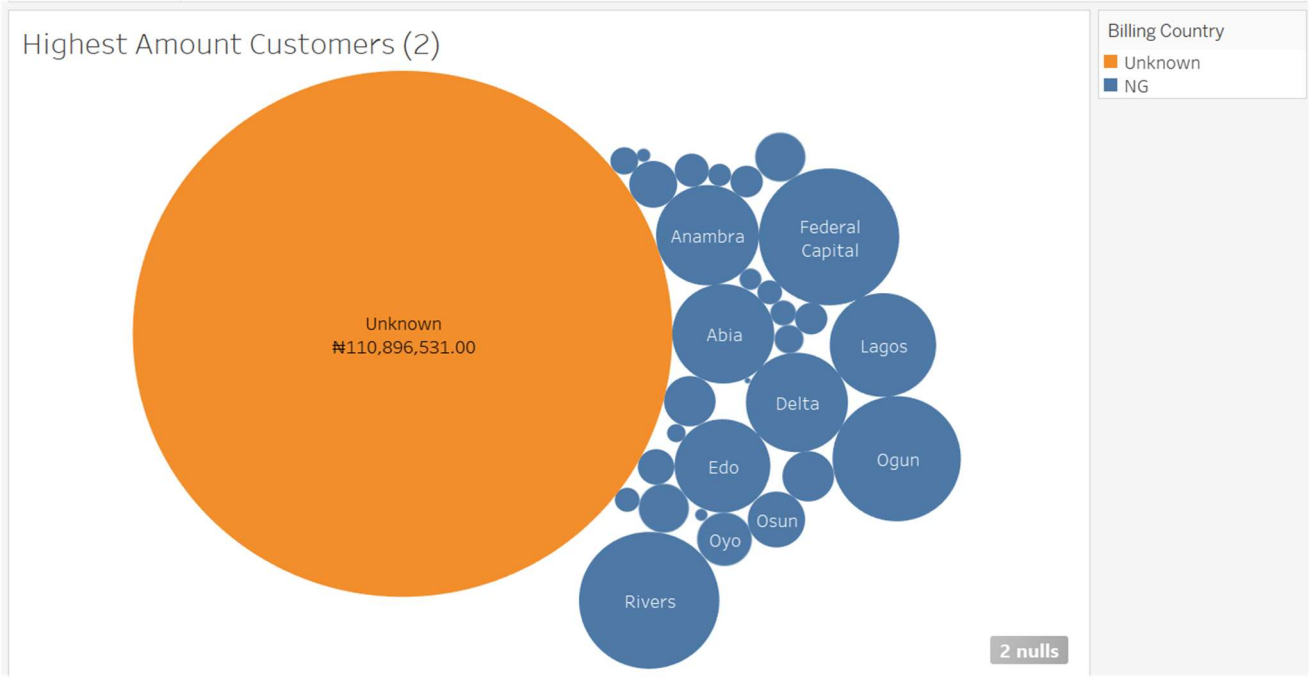


Figure 4.15: Amount spent per region bubble plot

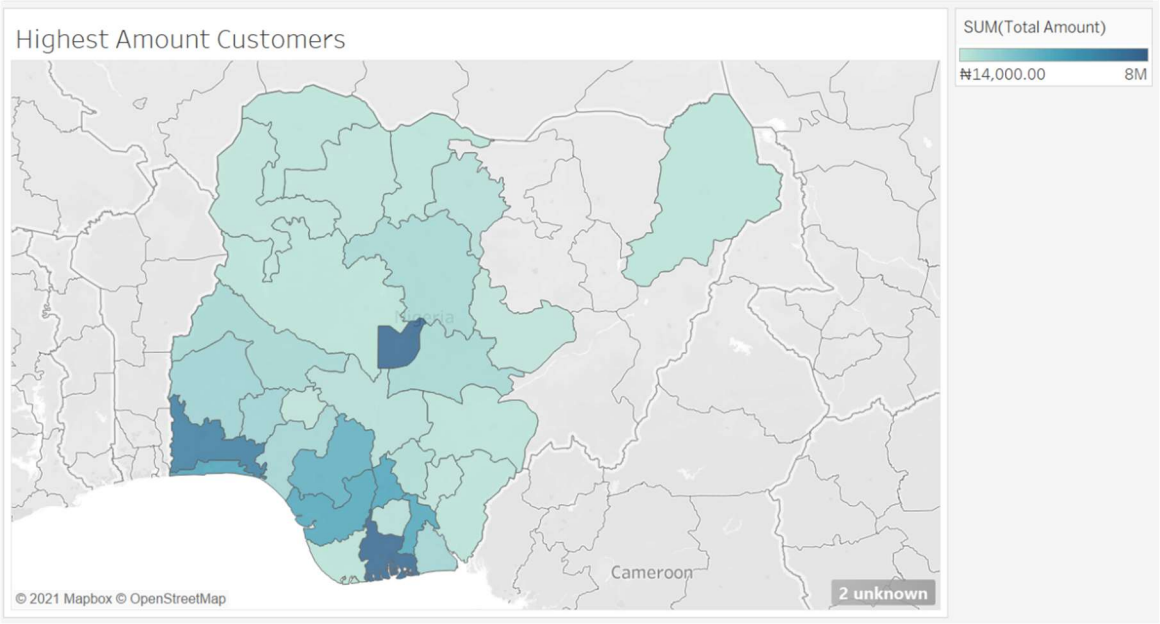


Figure 4.16: Amount spent per region map

Analysis of the total amount of sales and average sales amount throughout the study. It was visualized using a combination of bar and line charts, as shown in figure 4.17 below.

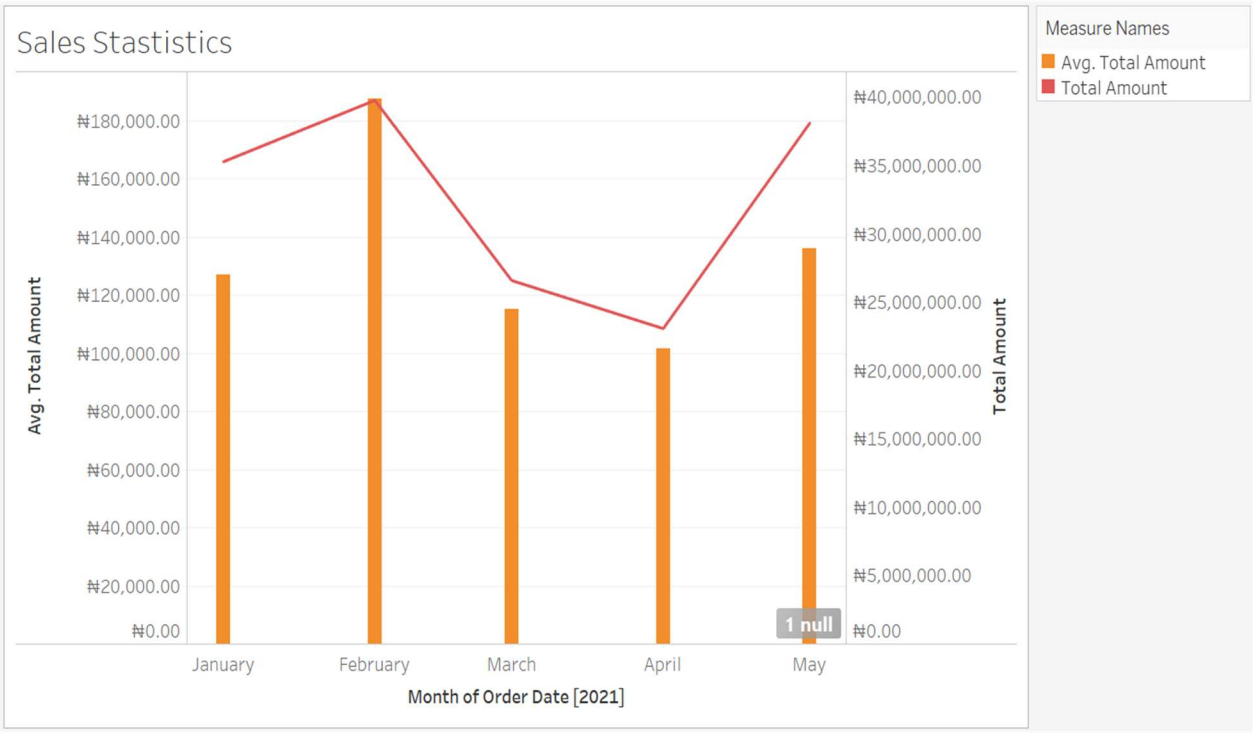


Figure 4.17: Total amount of sales and average sales amount over months

4.4 PREDICTION

In this project, we used MLR (Multiple Linear Regression) and ANN (Artificial Neural Network) to predict the possible amount to be spent by a customer for each order.

4.4.1 Visualization

Visualizations were carried out to determine if the predictors were suitable for prediction.

The relationship between the target variable/outcome to the categorical predictors were studied using boxplots and bar charts, as shown in figure 4.18 below. After plotting the boxplots, the categorical predictors which could be used was inferred and like in the case of the figure below; the categorical predictor was used.

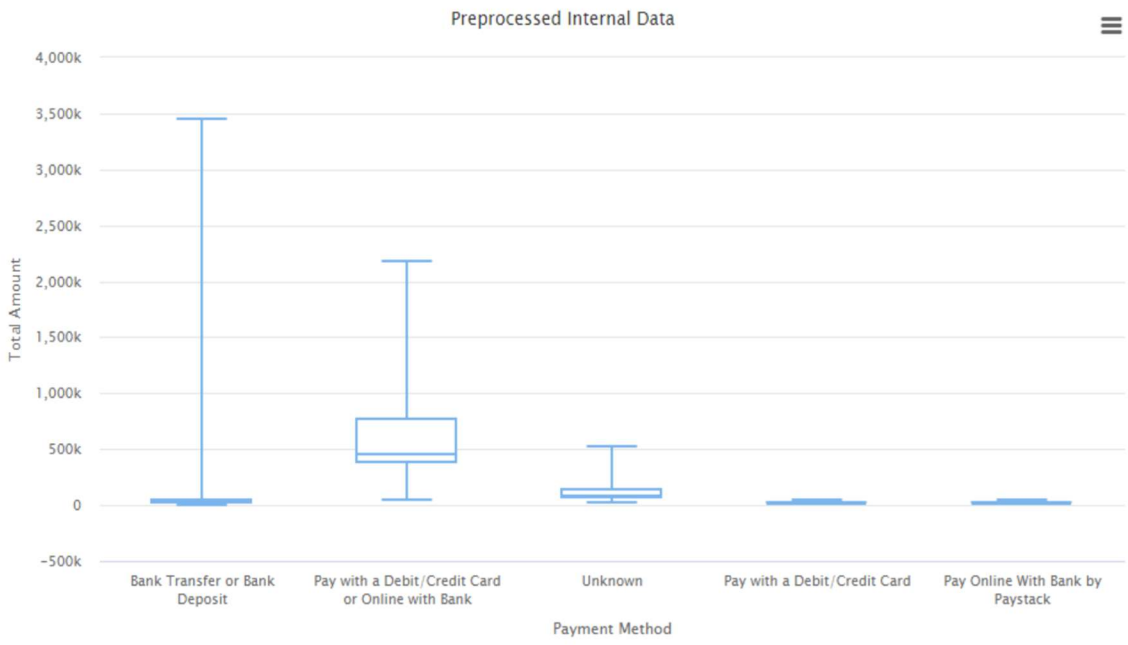


Figure 4.18: Payment Method against Total Amount

We then visualized the total amount using a histogram to see the distribution of the total amount, as shown in the figure below. From the histogram, it can be seen that most customers order amount was between 0 to 300k.

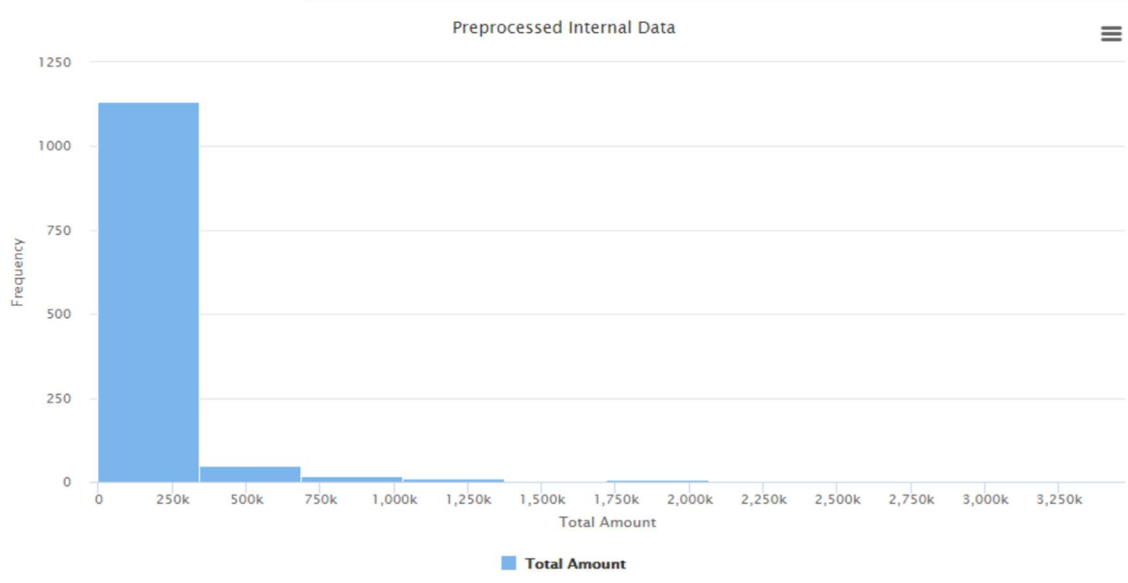


Figure 4.19: Distribution of total amount

4.4.2 Prediction using MLR (Multiple Linear Regression)

We used MLR (Multiple Linear Regression) in this project to predict the total amount to be spent in each order by a customer.

4.4.2.1 Modelling and Evaluation for MLR

The dataset was split into testing and training datasets. It was split 60:40 with 60% of the dataset for the model's training while 40% for testing of the model, as shown below.

```
> df <- read.csv("Dummy Data With State.csv")
> # Splitting the dataset into the Training set and Test set
> library (caTools)
Warning message:
package 'caTools' was built under R version 4.0.5
> set.seed(101)
> sample = sample.split(df$Total.Amount, SplitRatio = .60)
> train = subset(df, sample == TRUE)
> test  = subset(df, sample == FALSE)
```

Figure 4.20: Training and Testing split in R

After which, the model was ready to be trained. We trained the model using lm() to run a linear regression of the total amount on all the predictors in the training set.

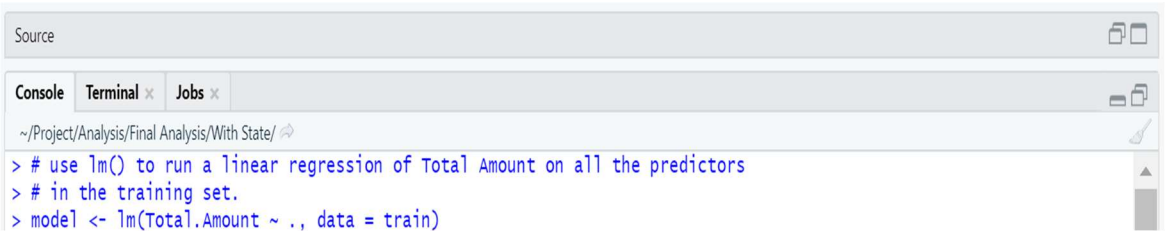
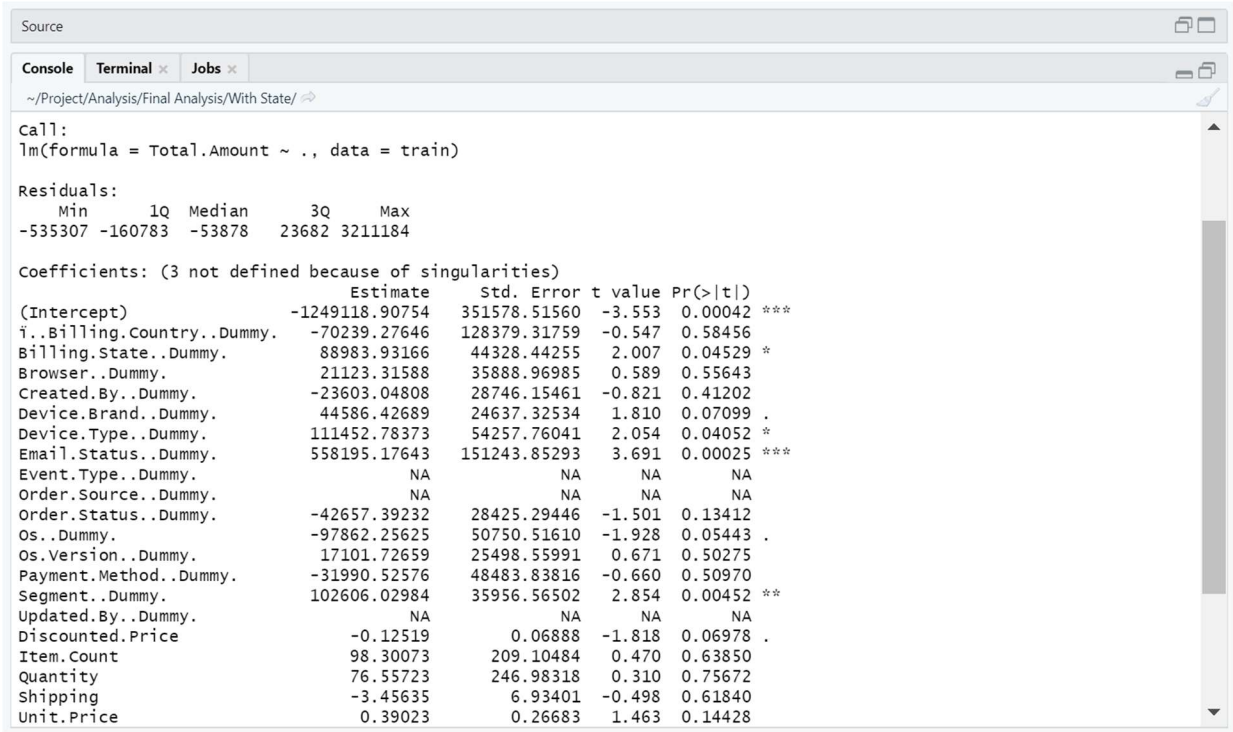


Figure 4.21: Training the model in R

Then we made use of summary() to find the statistics for the model shown in figure 2.23 below. We used options() to make sure the numbers are not displayed in scientific notations.

```
> # use options() to ensure numbers are not displayed in scientific notation.
> options(scipen = 999)
> summary(model)
```

Figure 4.22: Displaying statistics of model




```
Call:
lm(formula = Total.Amount ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-535307 -160783  -53878   23682 3211184

Coefficients: (3 not defined because of singularities)
                              Estimate     Std. Error t value Pr(>|t|)
(Intercept)                -1249118.90754   351578.51560  -3.553  0.00042 ***
ï..Billing.Country..Dummy.   -70239.27646   128379.31759  -0.547  0.58456
Billing.State..Dummy.         88983.93166    44328.44255   2.007  0.04529 *
Browser..Dummy.               21123.31588    35888.96985   0.589  0.55643
Created.By..Dummy.           -23603.04808    28746.15461  -0.821  0.41202
Device.Brand..Dummy.          44586.42689    24637.32534   1.810  0.07099 .
Device.Type..Dummy.          111452.78373    54257.76041   2.054  0.04052 *
Email.Status..Dummy.         558195.17643   151243.85293   3.691  0.00025 ***
Event.Type..Dummy.                    NA            NA      NA       NA
Order.Source..Dummy.                  NA            NA      NA       NA
Order.Status..Dummy.         -42657.39232    28425.29446  -1.501  0.13412
Os..Dummy.                   -97862.25625    50750.51610  -1.928  0.05443 .
Os.Version..Dummy.            17101.72659    25498.55991   0.671  0.50275
Payment.Method..Dummy.       -31990.52576    48483.83816  -0.660  0.50970
Segment..Dummy.              102606.02984    35956.56502   2.854  0.00452 **
Updated.By..Dummy.                    NA            NA      NA       NA
Discounted.Price                -0.12519       0.06888  -1.818  0.06978 .
Item.Count                      98.30073     209.10484   0.470  0.63850
Quantity                        76.55723     246.98318   0.310  0.75672
Shipping                        -3.45635       6.93401  -0.498  0.61840
Unit.Price                       0.39023       0.26683   1.463  0.14428
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 363200 on 462 degrees of freedom
Multiple R-squared:  0.1251,    Adjusted R-squared:  0.09291
F-statistic: 3.886 on 17 and 462 DF,  p-value: 0.0000003697
```

Figure 4.23: Statistics of the model (1)

Figure 4.24: Statistics of the model (2)

The equation of the regression model developed is shown below

$$Y = -1249118.90754 - 70239.27646x_1 + 88983.93166x_2 + 21123.31588x_3 - 23603.04808x_4 + 44586.4289x_5 + 111452.78373x_6 + 558195.17643x_7 - 42657.39232x_8 - 97862.25625x_9 + 17101.72659x_{10} - 31990.52576x_{11} + 102606.02984x_{12} - 0.12519x_{13} + 98.30073x_{14} + 76.55723x_{15} - 3.45635x_{16} + 0.39023x_{17}$$

Equation 4.1: Regression equation

We then evaluated the model using the test dataset. We used to predict() from the forecast library to make predictions on the test data and accuracy() to compute common accuracy

measures. It returned a Root Mean Square Error (RMSE) of 370,820 and an MAE (Mean Absolute Error) of 200,650.

```
> library(forecast)
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
Warning message:
package 'forecast' was built under R version 4.0.5
> # use predict() to make predictions on a new set.
> pred <- predict(model, test)
Warning message:
In predict.lm(model, test) :
  prediction from a rank-deficient fit may be misleading
> options(scipen=999, digits = 0)
```

Figure 4.25: Evaluating the model using the test dataset in R

```
> # use accuracy() to compute common accuracy measures.
> accuracy(pred, test$Total.Amount)
           ME   RMSE    MAE  MPE MAPE
Test set -5549 370820 200650 -822  929
>
```

Figure 4.26: Model accuracy scores

4.4.3 Prediction using ANN (Artificial Neural Networks)

In this project, we made use of ANN (Artificial Neural Network) to predict the total amount to be spent in each order by a customer.

4.4.3.1 Modelling and Evaluation for ANN

The attributes/predictors selected for the model were chosen from the MLR model with Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05. They include billing state dummy, device type dummy, email status dummy and segment dummy.

We first imported the libraries needed for this analysis which included: tidyverse, neuralnet, GGally.

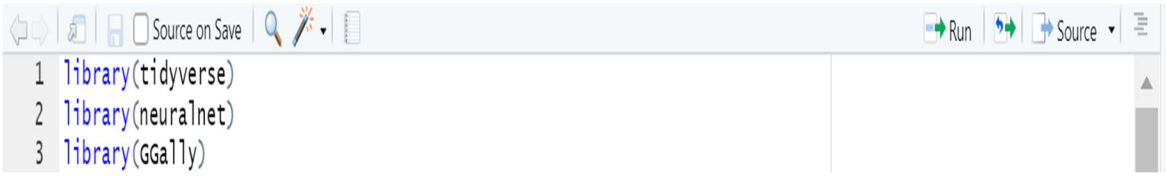


Figure 4.27: Imported libraries

After importing the dataset, we then normalized/scaled the data because ANN performs better with normalized data.

```
> # Scale the Data
> scale01 <- function(x){
+   (x - min(x)) / (max(x) - min(x))
+ }
> datum <- datum %>%
+   mutate_all(scale01)
```

Figure 4.28: Scaling/Normalization of the dataset in R

The dataset was split into testing and training datasets. It was split 60:40 with 60% of the dataset for the training of the model while 40% for testing of the model.

```
> set.seed(12345)
> datum_Train <- sample_frac(tbl = datum, replace = FALSE, size = 0.60)
> datum_Test <- anti_join(datum, datum_Train)
Joining, by = c("ï..Billing.State..Dummy.", "Device.Type..Dummy.", "Email.Status..Dummy.", "Segment..Dummy.", "Total.Amo
unt")
```

Figure 4.29: Training and testing split in R

To train the model, we made use of the neuralnet() from the neuralnet library. We set the hidden to five as it gives the best accuracy for the model.

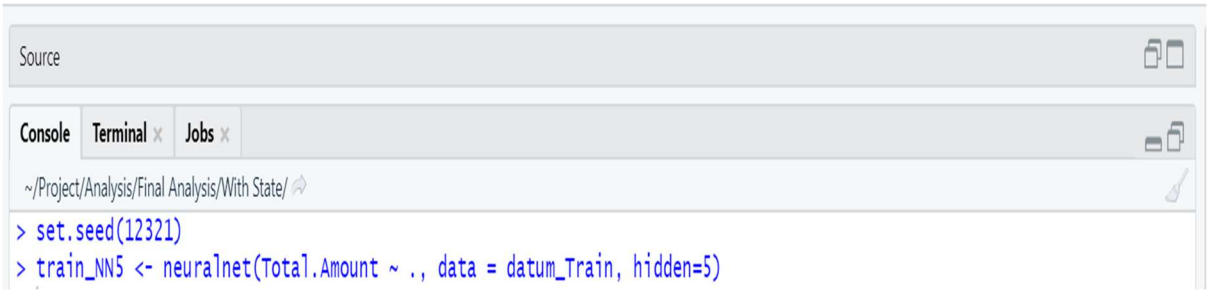


Figure 4.30: Training the model in R

We then evaluated the model using the test dataset. We made use of compute() to make predictions on the test data and manually calculated the RMSE and MAE. It returned a Root Mean Square Error (RMSE) of 0.1269182 and an MSE (Mean Squared Error) of 0.01610823.

```
> # Predict on test data
> pr5 <- compute(train_NN5, datum_Test)
>
```

Figure 4.31: Evaluating the model using the test dataset in R

```
> # Compute mean squared error
> pr.nn5 <- pr5$net.result * (max(datum$Total.Amount) - min(datum$Total.Amount)) + min(datum$Total.Amount)
> test.r5 <- (datum_Test$Total.Amount) * (max(datum$Total.Amount) - min(datum$Total.Amount)) + min(datum$Total.Amount)
> MSE.nn5 <- sum((test.r5 - pr.nn5)^2) / nrow(datum_Test)
> #mean squared error
> MSE.nn5
[1] 0.01610823
> #Root mean squared error
> RMSE5 <- sqrt(MSE.nn5)
> RMSE5
[1] 0.1269182
```

Figure 4.32: Model accuracy score

To view the diagram the model produced, we made use of the plot().

```
> #To view the diagram of the ANN
> plot(train_NN5)
>
```

Figure 4.33: To view the model diagram

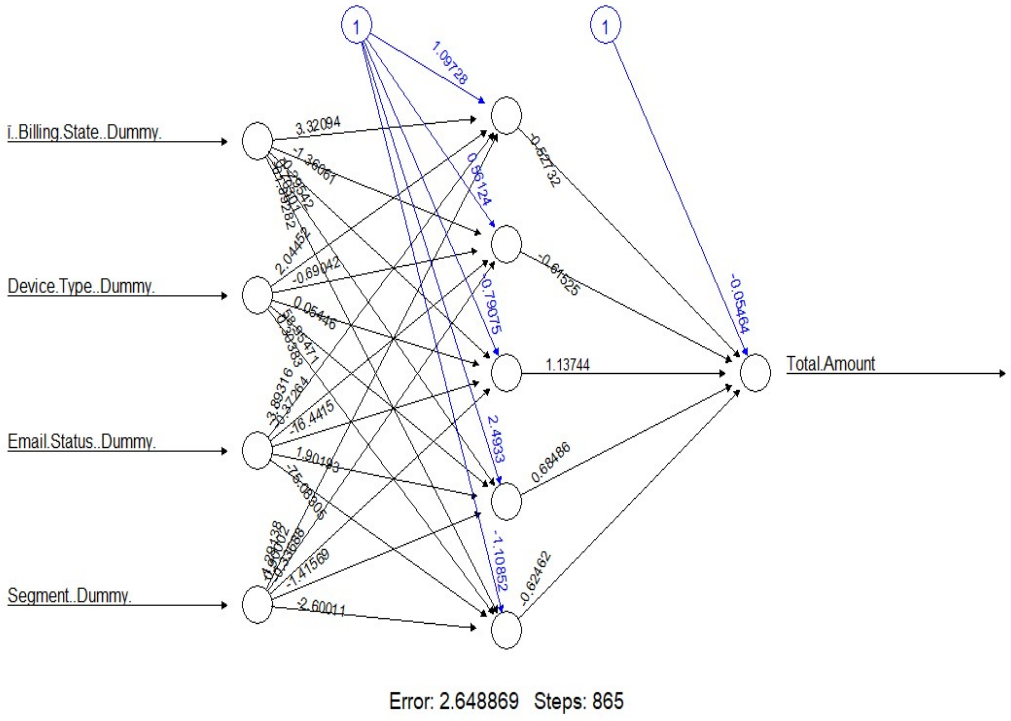


Figure 4.34: Model diagram with hidden = 5

4.5 CLASSIFICATION

In this project, we made use of a classification tree and ANN (Artificial Neural Network) to classify the order status for each order a customer wants to make.

4.5.1 Visualization

Visualizations were carried out to determine if the independent variables were suitable for classification.

The relationship between the target variable/outcome to the categorical predictors were studied using bar charts. The bar chart below in figure 4.35 shows the browsers which customers used. The unknown, chrome and chrome mobile are the most widely used compared to other browsers, so other browsers were binned to have a higher level of significance. This was done for other categorical predictors after visualization using a bar chart.

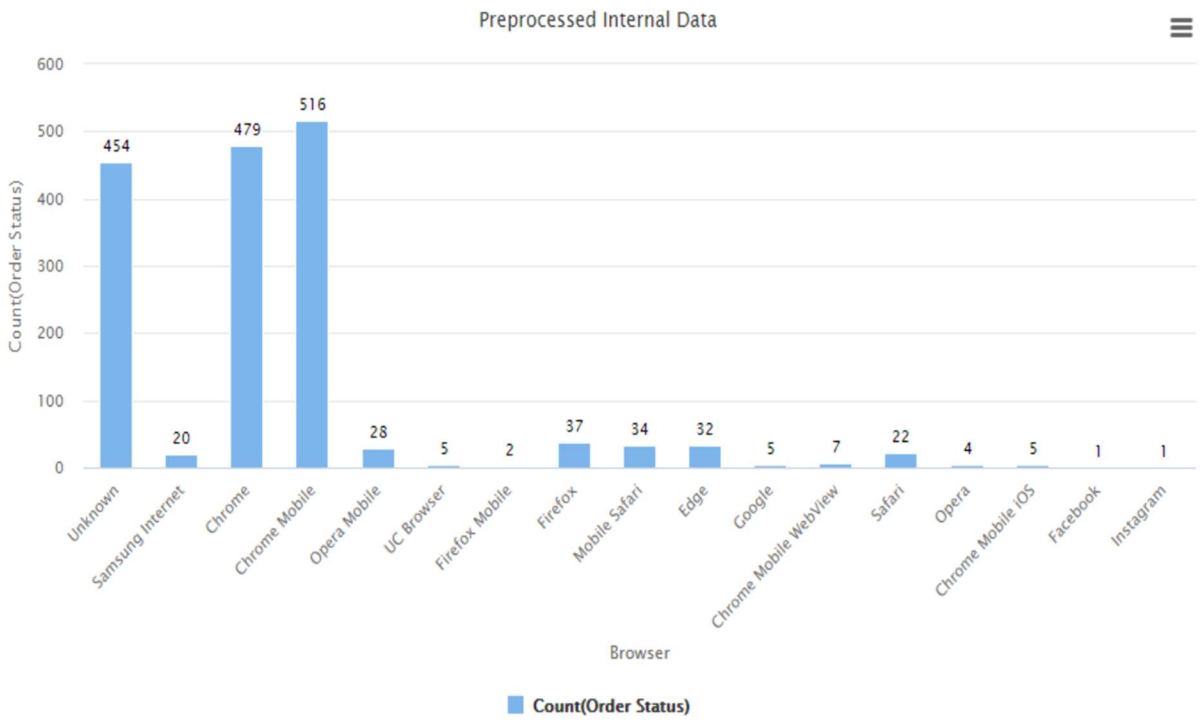


Figure 4.35: Order status against browser

4.5.2 Classification using Classification Tree

We used a classification tree in this project to classify the order status of orders placed by customers.

4.5.2.1 Modelling and Evaluation of Classification Tree

The libraries needed for this analysis included: rpart, rpart.plot, caret.

Figure 4.36: Imported libraries for classification tree

The dataset was split into testing and training datasets. It was split 60:40 with 60% of the dataset for the model's training while 40% for testing of the model.

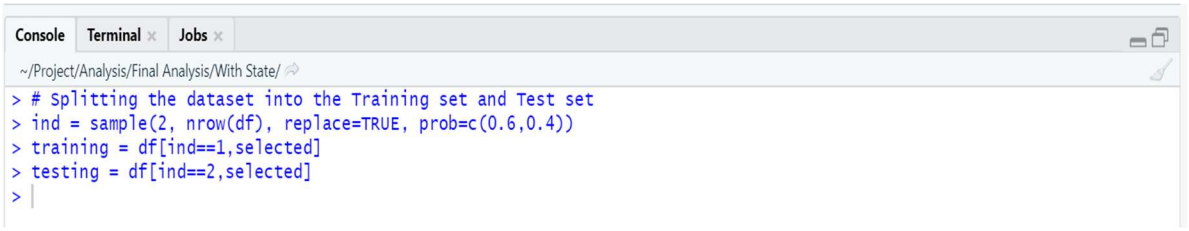


Figure 4.37: Training and testing split in R

To train the model, we used the rpart(); we then set the method to "class", as shown in the figure below.

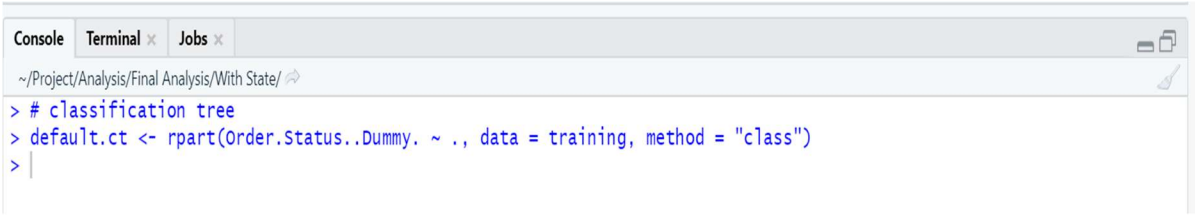


Figure 4.38: Training the model in R

We then visualized the default classification tree using prp().

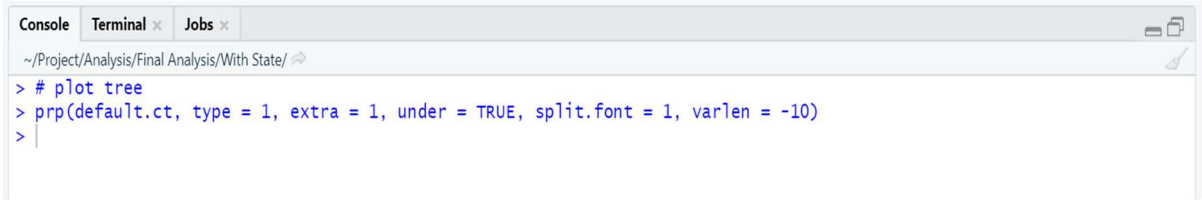


Figure 4.39: Plot default classification tree in R

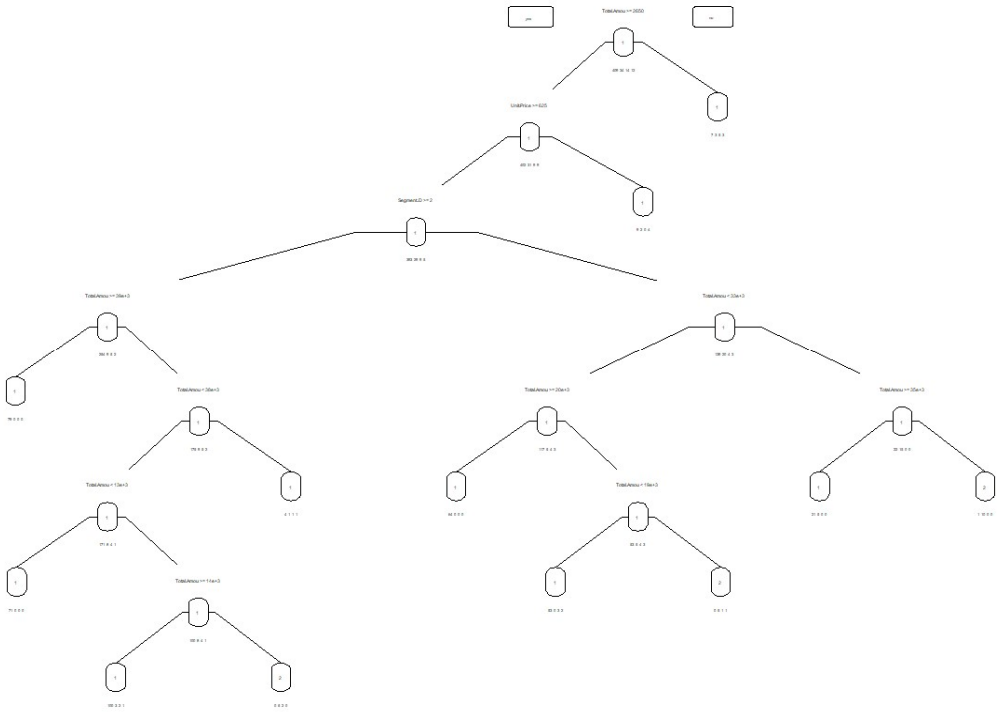


Figure 4.40: Default classification tree

We then evaluated the model using the test dataset. We made use of predict() to make predictions on the test data.

```
Console  Terminal  Jobs
~/Project/Analysis/Final Analysis/With State/
> #CODE FOR TESTING ACCURACY
> dpred.test <- predict(default.ct,testing,type = "class")
>
```

Figure 4.41: Predicting on test data

A confusion matrix was then used to determine accuracy. It gave an accuracy of 88.92%.

```
Console  Terminal  Jobs
~/Project/Analysis/Final Analysis/With State/
> #default tree: testing
> confusionMatrix(factor(dpred.test, levels = 1:4), factor(testing$Order.Status..Dummy., levels = 1:4))
```

Figure 4.42: Evaluating the model using a confusion matrix in R

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 281  19   8   7
         2   1   8   0   1
         3   0   0   0   0
         4   0   0   0   0

Overall Statistics

               Accuracy : 0.8892
                 95% CI : (0.85, 0.9212)
    No Information Rate : 0.8677
    P-Value [Acc > NIR] : 0.1428

                  Kappa : 0.292

 Mcnemar's Test P-Value : NA
```

Figure 4.43: Confusion matrix and statistics result

```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.9965  0.29630  0.00000  0.00000
Specificity            0.2093  0.99329  1.00000  1.00000
Pos Pred Value         0.8921  0.80000      NaN      NaN
Neg Pred Value         0.9000  0.93968  0.97538  0.97538
Prevalence             0.8677  0.08308  0.02462  0.02462
Detection Rate         0.8646  0.02462  0.00000  0.00000
Detection Prevalence   0.9692  0.03077  0.00000  0.00000
Balanced Accuracy      0.6029  0.64479  0.50000  0.50000
```

Figure 4.44: Statistics by class

We trained another model use the rpart() to produce a deeper classification tree. The number of leaves in the deeper classification tree was gotten using length()

```
Source
Console  Terminal  Jobs
~/Project/Analysis/Final Analysis/With State/
> deeper.ct <- rpart(Order.Status..Dummy. ~ ., data = training, method = "class", cp = 0, minsplit = 1)
> # count number of leaves
> length(deeper.ct$frame$var[deeper.ct$frame$var == "<leaf>"])
[1] 59
```

Figure 4.45: Training of model for deeper classification tree in R

We then visualized the deeper classification tree using prp().

```
Console  Terminal  Jobs
~/Project/Analysis/Final Analysis/With State/
> # plot tree
> prp(deeper.ct, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10,
+     box.col=ifelse(deeper.ct$frame$var == "<leaf>", 'gray', 'white'))
Warning message:
labs do not fit even at cex 0.15, there may be some overplotting
>
```

Figure 4.46: Plot deeper classification tree in R

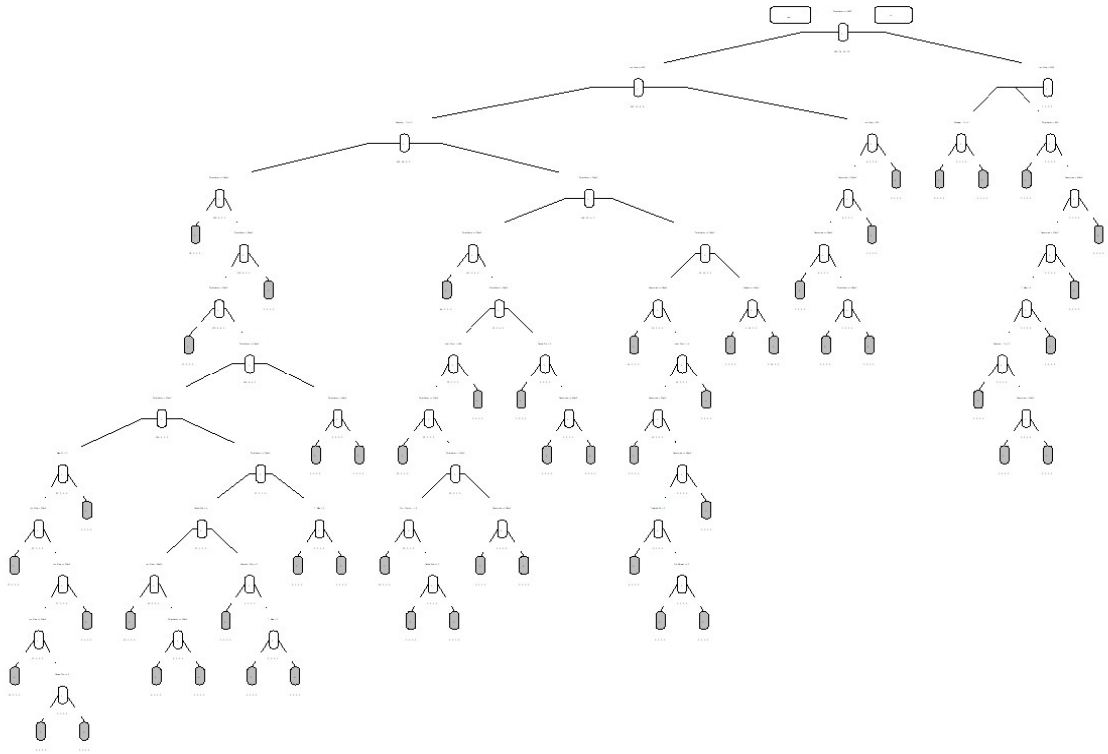


Figure 4.47: Deeper classification tree

We then evaluated the model using the test dataset. We made use of predict() to make predictions on the test data.

```
Console  Terminal  Jobs
~/Project/Analysis/Final Analysis/With State/
> #CODE FOR TESTING ACCURACY
> deeperpred.test <- predict(deeper.ct,testing,type = "class")
```

Figure 4.48: Predicting on test data in R

A confusion matrix was then used to determine accuracy. It gave an accuracy of 85.23%.

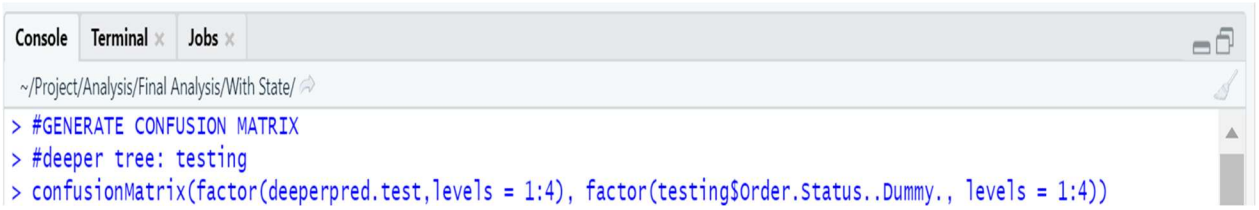


Figure 4.49: Evaluating model using a confusion matrix in R

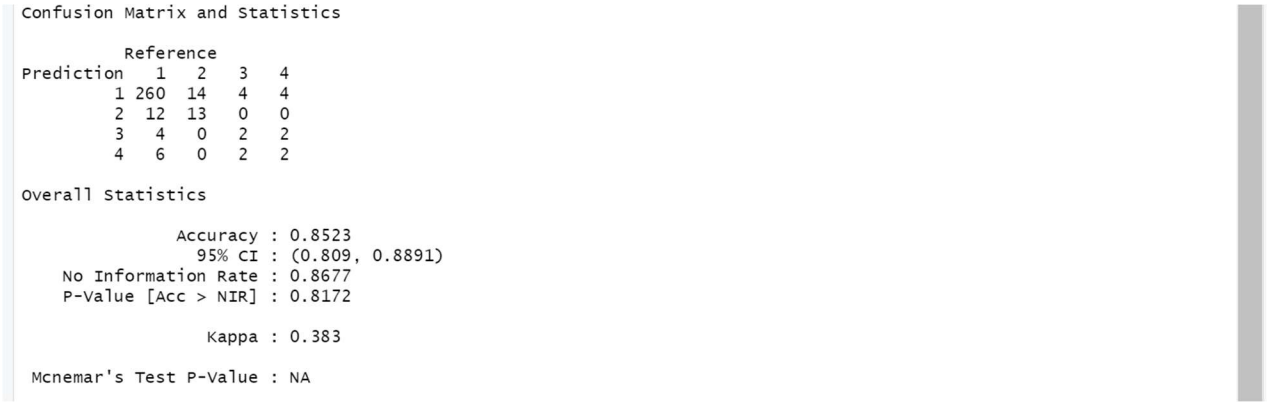


Figure 4.50: Confusion matrix and statistics result

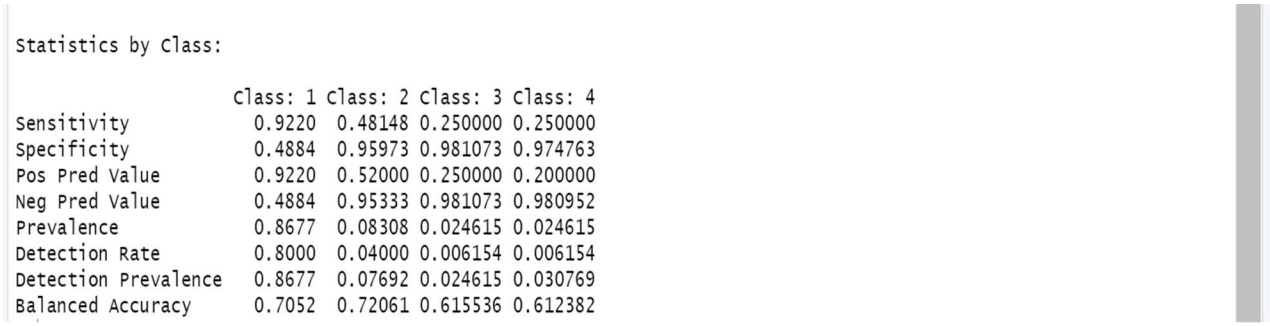


Figure 4.51: Statistics by class

4.5.3 Classification using ANN (Artificial Neural Networks)

In this project, we used ANN (Artificial Neural Network) to predict the order status of orders placed by customers.

4.5.3.1 Modelling and Evaluation for ANN

We first imported the libraries needed for this analysis which included: tidyverse, neuralnet, GGally.

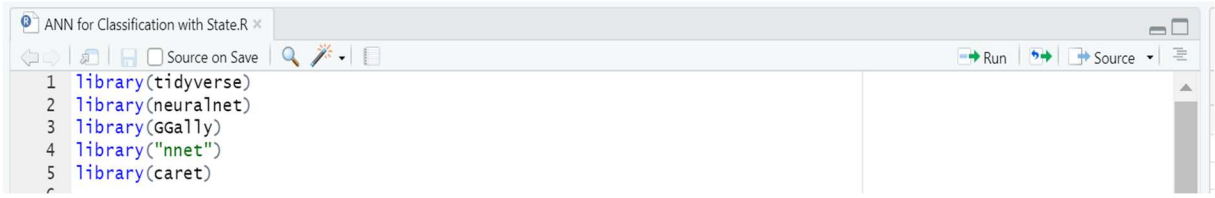


Figure 4.52: Imported libraries

The dataset was split into testing and training datasets. It was split 60:40 with 60% of the dataset for the model's training while 40% for testing the model.

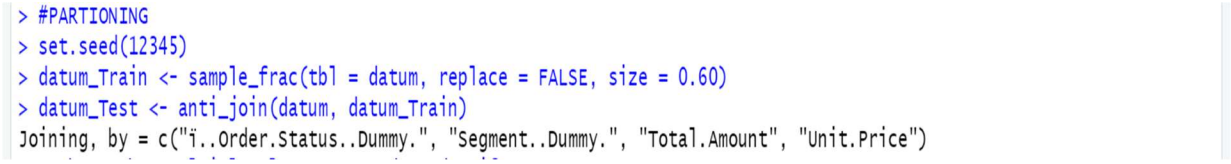


Figure 4.53: Test/Train split in R

After which, the target variable, order status, which had multiple classes, was dummified as shown below.

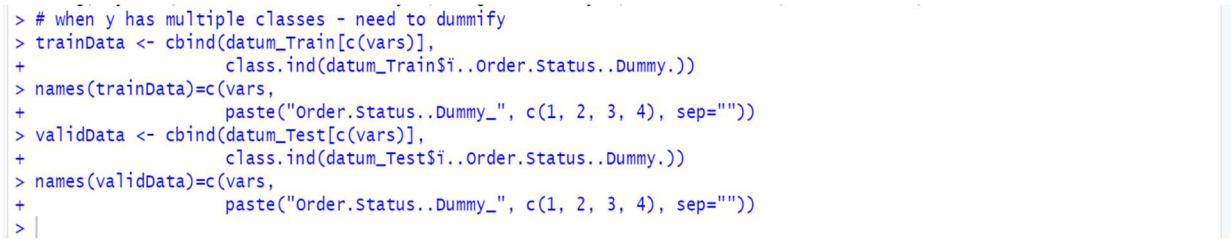


Figure 4.54: Dummification of the target variable in R

To train the model, we made use of the neuralnet() from the neuralnet library. We set the hidden to two as it gives the best accuracy for the model.

```
Console  Terminal  Jobs
~/Project/Analysis/Final Analysis/With State/
> #run nn with 2 hidden nodes
> #use hidden= with a vector of integers specifying number of hidden nodes in each layer
> nn <- neuralnet(Order.Status..Dummy_1 + Order.Status..Dummy_2 +
+                   Order.Status..Dummy_3 + Order.Status..Dummy_4 ~
+                   Segment..Dummy.+ Total.Amount+ Unit.Price,
+                   data = trainData, hidden = 2)
>
```

Figure 4.55: Model training in R

To view the diagram the model produced, we made use of the plot().

```
> #view diagram
> plot(nn)
>
```

Figure 4.56: To view the model diagram in R

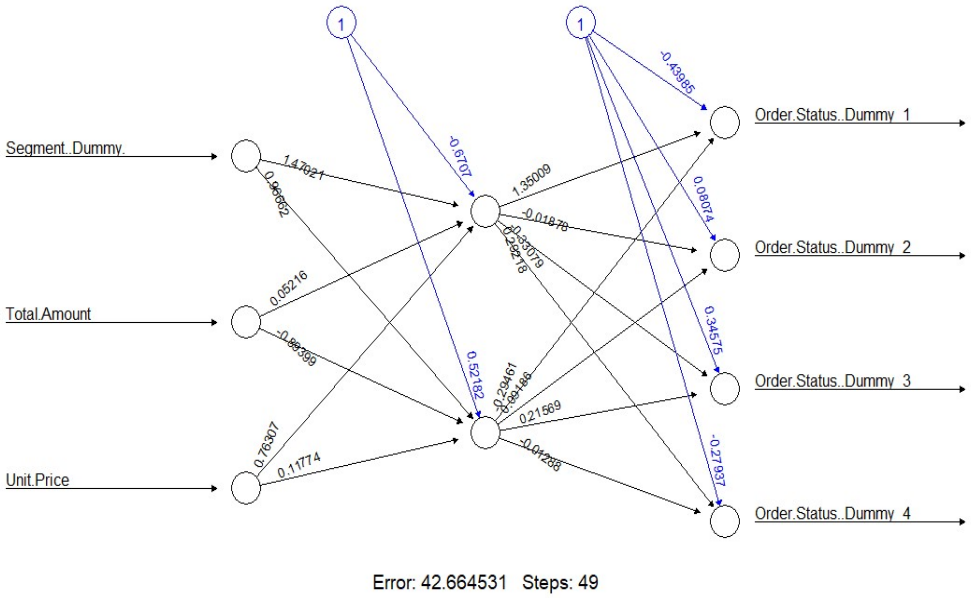


Figure 4.57: Model diagram with hidden = 2

We then evaluated the model using the test dataset.

```
> #testing data
> validation.prediction=compute(nn, validData)
> validation.class=apply(validation.prediction$net.result,1,which.max)
```

Figure 4.58: Predicting on test data in R

A confusion matrix was then used to determine accuracy. It gave an accuracy of 85.9%.

```
> confusionMatrix(factor(validation.class), factor(datum_Test$i..Order.Status..Dummy., levels = 1:4))
```

Figure 4.59: Evaluating model using a confusion matrix in R

```
Console   Terminal   Jobs
~/Project/Analysis/Final Analysis/With State/

          Reference
Prediction   1   2   3   4
         1 268  24   9  11
         2   0   0   0   0
         3   0   0   0   0
         4   0   0   0   0

Overall Statistics

               Accuracy : 0.859
                 95% CI : (0.8153, 0.8956)
    No Information Rate : 0.859
    P-Value [Acc > NIR] : 0.5401

                  Kappa : 0

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity             1.000  0.00000  0.00000  0.00000
Specificity             0.000  1.00000  1.00000  1.00000
Pos Pred Value          0.859      NaN      NaN      NaN
Neg Pred Value            NaN  0.92308  0.97115  0.96474
Prevalence              0.859  0.07692  0.02885  0.03526
Detection Rate          0.859  0.00000  0.00000  0.00000
Detection Prevalence    1.000  0.00000  0.00000  0.00000
Balanced Accuracy       0.500  0.50000  0.50000  0.50000
```

Figure 4.60: Confusion matrix and Statistics

4.6 CLUSTERING

4.6.1 Clustering using K-Means

In this project, we made use of the K-means algorithm to cluster customers. The model was developed using the K-means algorithm in RapidMiner studios.

4.6.1.1 Modelling and Evaluation for K-means

After the dataset was imported, it was normalized before it was passed into the clustering algorithm. The clustering algorithm used was K-means, and the number of clusters used was four as it gave the best accuracy.

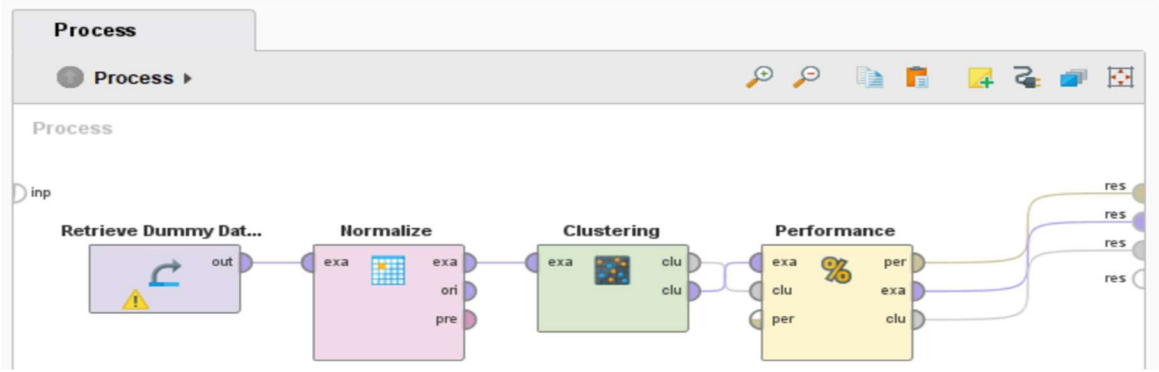


Figure 4.61: K-means clustering in RapidMiner studios

The cluster performance was obtained using the performance operator in RapidMiner. It calculated the average within centroid distance of each cluster and the Davies Bouldin value. Performance vector of the clusters:

Average within centroid distance: -11.710

Average within centroid distance_cluster_0: -10.009

Average within centroid distance_cluster_1: -9.032

Average within centroid distance_cluster_2: -12.876

Average within centroid distance_cluster_3: -16.617

Davies Bouldin: -1.488

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -11.710
Avg. within centroid distance_cluster_0: -10.009
Avg. within centroid distance_cluster_1: -9.032
Avg. within centroid distance_cluster_2: -12.876
Avg. within centroid distance_cluster_3: -16.617
Davies Bouldin: -1.488
```

Figure 4.62: Performance vector of the cluster

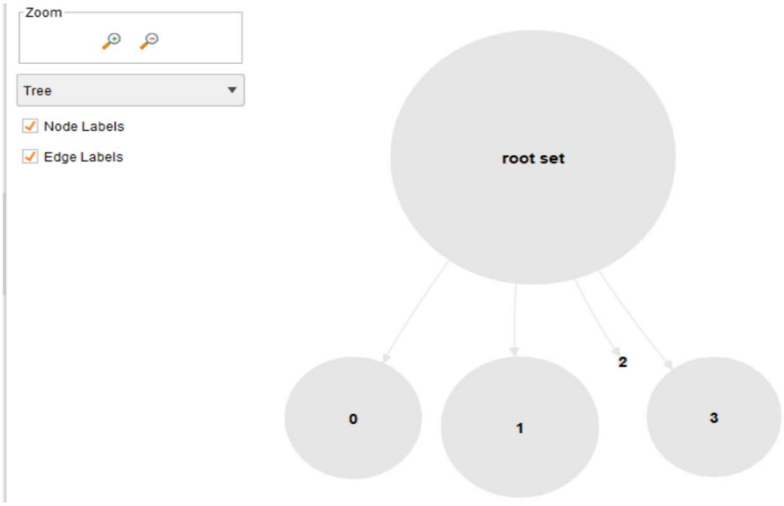


Figure 4.63: Graph of the clusters

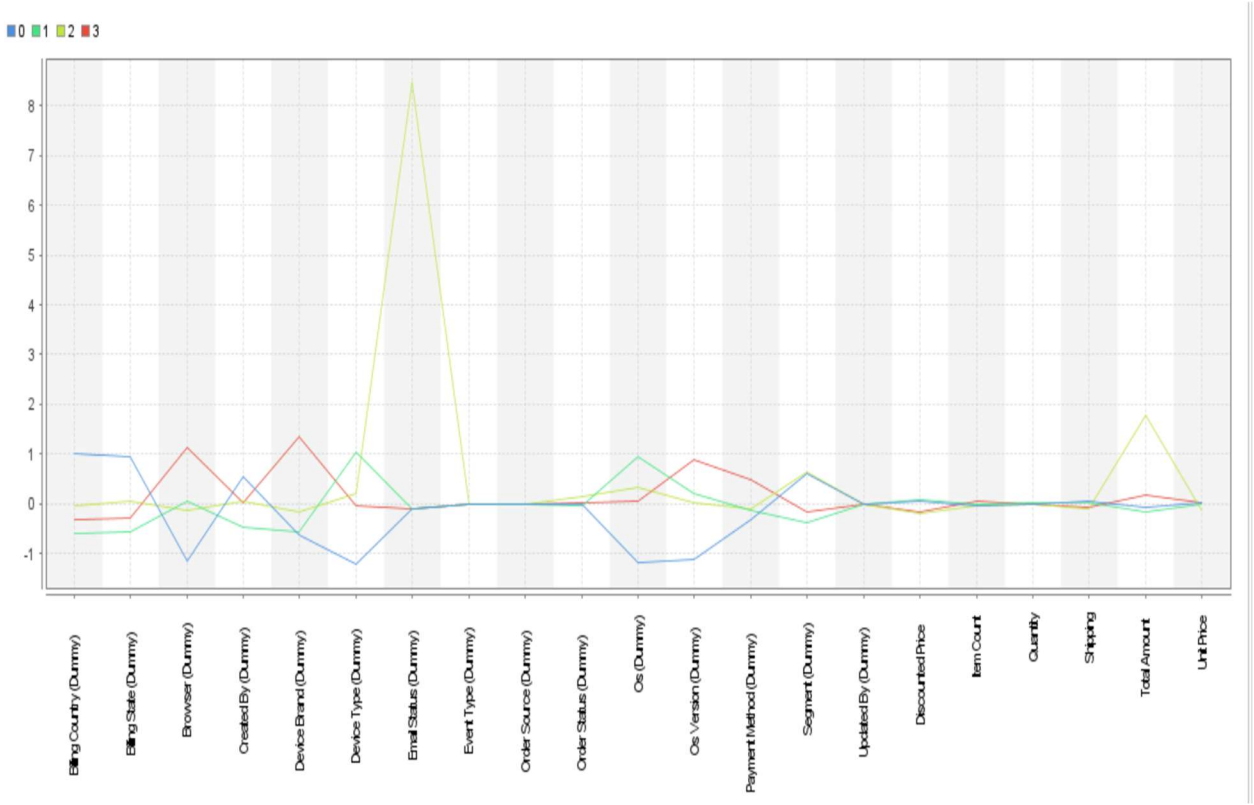


Figure 4.64: Plot of the clustering attributes

It can be seen from the plot of the clustering attributes in figure 4.64 above that there are few separations between the clusters. Each cluster has its own distinguishing attributes (profiles), which are:

- Cluster 0: Billing country (Dummy), Billing state (Dummy), Browser (Dummy), Order by (Dummy), Device type (Dummy), OS (Dummy), OS version (Dummy).
- Cluster 1: Billing country (Dummy), Billing state (Dummy), Created by (Dummy), Browser (Dummy), Order by (Dummy), Device type (Dummy), OS (Dummy), OS version (Dummy), Segment (Dummy).
- Cluster 2: Email status (Dummy), Total amount.
- Cluster 3: Billing country (Dummy), Billing state (Dummy), Browser (Dummy), Device brand (Dummy), Payment method (Dummy), OS version (Dummy).

4.7 RECOMMENDATION AND ANALYSIS

After the whole analysis is done with the recommendations and inferences derived from the analysis carried out.

4.7.1 Descriptive Analytics

According to the bubble plot and map, the state with the most customers and the highest amount spent on orders is from unidentified states known as "Unknown". Lagos and rivers state are 2nd and 3rd states respectively with most customers but far behind that of the unknown. While F.C.T and rivers state is 2nd and 3rd states respectively with the highest amount of spent on orders but far behind that of the unknown.

According to the analysis of the total amount of sales and average sales amount throughout the study, which was visualized using a combination of bar and line charts as shown in figure 4.15, we can see that the total amount of sales increases or decreases the average sales amount moves in the same direction.

4.7.2 Prediction and Classification Modelling

Based on the models developed for both prediction and classification, the recommendation is that the models developed should be applied to external customers data. The models developed for prediction should be applied to the external customer data to predict the total amount likely to be spent by the customer. This will allow the organization to be to forecast possible revenue from customers. While the model developed for classification should be

applied to the external customer data to classify the order status of each order. This would allow the company to know which customers are likely to complete an order.

4.7.3 Clustering

Based on the clustering results and profiles gotten, external customer data/external market can be cloned. Proposed customers can be segmented based on profiles acquired which can be used for target marketing. It can be used to target customers with the same description as a particular cluster.

CHAPTER FIVE

SUMMARY, RECOMMENDATION, AND CONCLUSION

5.1 SUMMARY

The study will provide actionable recommendations for agricultural-based e-commerce businesses to increase profit and increase return on investment.

Descriptive analytics was used in this study to get an understanding of the customers, their regions, total spending and the average sales and total sales per month

In the study, we used prediction algorithms such as MLR (Multiple Linear Regression) and ANN (Artificial Neural Network) to predict the total amount spent on each order placed by each customer.

Classification algorithms such as Classification tree and ANN (Artificial Neural Network) were used to classify the order status of each order by customers. To predict whether the order will be completed or cancelled.

K-means clustering algorithm was used to cluster/group the customers according to their attributes/characteristics to understand them better.

5.2 RECOMMENDATION

While the project tries to address e-commerce companies' problems concerning sales and customer retention, there is still room for improvement. Due to the limitation of the projects and challenges encountered during the implementation. The following recommendations and further works are key:

i. Models developed should be applied to external data/customers that are yet to order.
ii. More and better descriptive attributes of customers and orders should be collected for better prediction accuracy.
iii. More data spanning over the years should be provided.
iv. Data should be collected appropriately so most of it won’t be lost to data cleaning.

5.3 CONCLUSION

The study will help the company understand its customer's types, behaviours, and demographics to improve company strategies to get more sales. Using descriptive analytics, we can understand where the customers, their demographics etc. With clustering, we can tell the characteristics of different customers and how it affects the total amount spent on ordering and completing an order.

It will also help companies develop effective marketing strategies that target specified customers than a wide range of customers in the process of saving funds. From the models trained, they can predict customers who are likely to order from them.

REFERENCES

11 Important Model Evaluation Techniques Everyone Should Know. (2016). https://www.datasciencecentral.com/profiles/blogs/7-important-model-evaluation-error-metrics-everyone-should-know

Ali, R. A., Ibrahim, O., & Nilashi, M. (2020). Loyalty of young female Arabic customers towards recommendation agents : A new model for B2C E-commerce. *Technology in Society*, *61*(August 2019), 101253. https://doi.org/10.1016/j.techsoc.2020.101253

Bejju, A. (2016). Sales Analysis of E-Commerce Websites using Data Mining Techniques. *International Journal of Computer Applications*, *133*(5), 36–40.

Big Data Analytics - Data Life Cycle. (2021). https://www.tutorialspoint.com/big_data_analytics/big_data_analytics_lifecycle.htm

Biscobing, J. (2018). *What is data sampling? - Definition*. https://searchbusinessanalytics.techtarget.com/definition/data-sampling

Carmona, C. J., Ramírez-gallego, S., Torres, F., Bernal, E., Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website : OrOliveSur.com. *Expert Systems With Applications*, *39*(12), 11243–11249. https://doi.org/10.1016/j.eswa.2012.03.046

Cerny, M., Sokol, O., & Holy, V. (2017). Clustering retail products based on customer behaviour. *Applied Soft Computing*, *60*, 752–762. https://doi.org/10.1016/j.asoc.2017.02.004

Dias, J. P., & Ferreira, H. S. (2017). Automating the Extraction of Static Content and Dynamic Behaviour from e-Commerce Websites. *The 8th International Conference on Ambient Systems, Networks and Technologies(ANT 2017)*. https://doi.org/10.1016/j.procs.2017.05.355

Ecommerce analytics 101. (2020). https://supermetrics.com/blog/ecommerce-analytics

Failory. (2015). https://www.failory.com/cemetery/zulily

Fatta, D. Di, Patton, D., & Viglia, G. (2018). The determinants of conversion rates in SME e-commerce websites. *Journal of Retailing and Consumer Services*, *41*(October 2017), 161–168. https://doi.org/10.1016/j.jretconser.2017.12.008

Guler, B., & Tufan, K. (2013). Unsuccessful e-commerce stories Dotcom boom. *AICT 2013 - 7th International Conference on Application of Information and Communication Technologies, Conference Proceedings*, *February 2014*. https://doi.org/10.1109/ICAICT.2013.6722640

Hong, T., & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems With Applications*, *39*(2), 2127–2131. https://doi.org/10.1016/j.eswa.2011.07.114

Ilbahar, E., & Cebi, S. (2017). Classification of design parameters for E-commerce websites: A novel fuzzy Kano approach. *Telematics and Informatics*, *34*(September), 1814–1825. https://doi.org/10.1016/j.tele.2017.09.004

Jaiswal, A. K., Niraj, R., Hee, C., & Agarwal, M. K. (2018). The effect of relationship and transactional characteristics on customer retention in emerging online markets. *Journal of Business Research*, *92*(July), 25–35. https://doi.org/10.1016/j.jbusres.2018.07.007

Liao, S., Chen, Y., & Hsieh, H. (2011). Mining customer knowledge for direct selling and marketing. *Expert Systems With Applications*, *38*(5), 6059–6069. https://doi.org/10.1016/j.eswa.2010.11.007

Maier, E., & Wieringa, J. (2020). Acquiring customers through online marketplaces ? The effect of marketplace sales on sales in a retailer's own channels. *International Journal of Research in Marketing*, *xxxx*. https://doi.org/10.1016/j.ijresmar.2020.09.007

Nighania, K. (2018). *Various ways to evaluate a machine learning model's performance*. https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Nisar, T. M., & Prabhakar, G. (2017). What factors determine e-satisfaction and consumer spending in e-commerce retailing? *Journal of Retailing and Consumer Services*,

39(August), 135–144. https://doi.org/10.1016/j.jretconser.2017.07.010

Pallant, J. I., Danaher, P. J., Sands, S. J., & Danaher, T. S. (2017). An empirical analysis of factors that influence retail website visit types. *Journal of Retailing and Consumer Services*, *39*(April), 62–70. https://doi.org/10.1016/j.jretconser.2017.07.003

Pranjal Pandey. (2019). *Data Preprocessing : Concepts*. https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

Quelhas, P., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer Integrated Manufacturing*, *36*, 93–100. https://doi.org/10.1016/j.rcim.2014.12.014

Raphaeli, O., Goldstein, A., & Fink, L. (2017). Analyzing online consumer behavior in mobile and PC devices : A novel web usage mining approach. *Electronic Commerce Research and Applications*, *26*, 1–12. https://doi.org/10.1016/j.elerap.2017.09.003

Rho, J. J., Korea, S., Moon, B., Korea, S., Kim, Y., Korea, S., Yang, D., & Korea, S. (2004). Internet Customer Segmentation Using Web Log Data. *Journal of Business & Economics Research*, *2*(11), 59–74.

Salminen, J., Yoganathan, V., Corporan, J., Jansen, B. J., & Jung, S. (2019). Machine learning approach to auto-tagging online content for content marketing efficiency : A comparative analysis between methods and content type. *Journal of Business Research*, *101*(April), 203–217. https://doi.org/10.1016/j.jbusres.2019.04.018

Saura, J. R., Herráez, B. R., & Reyes-menendez, A. N. A. (2019). *Comparing a Traditional Approach for Financial Brand Communication Analysis With a Big Data Analytics Technique*. *7*. https://doi.org/10.1109/ACCESS.2019.2905301

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*.

Silveira, T. (2020). *Data acquisition, web scraping, and the KDD process: a practical study*

with COVID-19 data in Brazil. https://towardsdatascience.com/data-acquisition-web-scraping-and-the-kdd-process-a-practical-study-with-covid-19-data-in-brazil-f7f397e814b7

Statista Market Forecast. (2020). https://www.statista.com/outlook/243/160/ecommerce/nigeria

Tsai, C., & Chiu, C. (2004). A purchase-based market segmentation methodology. *Expert Systems With Applications*, *27*, 265–276. https://doi.org/10.1016/j.eswa.2004.02.005

Videla-cavieres, I. F., & Ríos, S. A. (2014). Extending market basket analysis with graph mining techniques : A real case. *Expert Systems With Applications*, *41*(4), 1928–1936. https://doi.org/10.1016/j.eswa.2013.08.088

Wen, C., Liao, S., Chang, W., & Hsu, P. (2012). Mining shopping behavior in the Taiwan luxury products market. *Expert Systems With Applications*, *39*(12), 11257–11268. https://doi.org/10.1016/j.eswa.2012.03.072

APPENDIX

PREDICTION USING MLR (MULTIPLE LINEAR REGRESSION) CODE

```
df <- read.csv("Dummy Data With State.csv")

# dividing the dataset into training and testing sets

library (caTools)

set.seed(101)

sample = sample.split(df$Total.Amount, SplitRatio = .60)

train = subset(df, sample == TRUE)

test = subset(df, sample == FALSE)

# model development

model <- lm (Total.Amount ~., data = train)

# To prevent numbers from being shown in scientific notation, use options ().

options (scipen = 999)

summary(model)

#CODE FOR PREDICTION AND MEASURING ACCURACY

library(forecast)

# To create predictions on a new set, use predict ().

pred <- predict (model, test)

options (scipen=999, digits = 0)

# To compute common accuracy measurements, use accuracy().

accuracy(pred, test$Total.Amount)
```

PREDICTION USING ANN (ARTIFICIAL NEURAL NETWORK) CODE

```
library(tidyverse)

library(neuralnet)

library(GGally)

# Importing the dataset

datum= read.csv('Dummy Data With State for ANN Prediction.csv')

#Scatterplot matrix

ggpairs(datum, title = "Scatterplot Matrix of the Features of the Data Set")

# Scale the Data

scale01 <- function(x){

  (x - min(x)) / (max(x) - min(x))

}

datum <- datum %>%

  mutate_all(scale01)

set.seed(12345)

datum_Train <- sample_frac(tbl = datum, replace = FALSE, size = 0.60)

datum_Test <- anti_join(datum, datum_Train)

#HIDDEN = 2

set.seed(12321)

train_NN2 <- neuralnet(Total.Amount ~ ., data = datum_Train, hidden=2)

# Predict on test data
```

```
pr <- compute(train_NN2, datum_Test)

# Compute mean squared error

pr.nn <- pr$net.result * (max(datum$Total.Amount) - min(datum$Total.Amount)) +
min(datum$Total.Amount)

test.r <- (datum_Test$Total.Amount) * (max(datum$Total.Amount) -
min(datum$Total.Amount)) + min(datum$Total.Amount)

MSE.nn <- sum((test.r - pr.nn)^2) / nrow(datum_Test)

# mean squared error

MSE.nn

# Calculate the root mean squared error

RMSE <- sqrt(MSE.nn)

RMSE

#To view the diagram of the ANN

plot(train_NN2)

# Plot regression line

plot(datum_Test$Total.Amount, pr.nn, col = "red",

    main = 'Real vs Predicted for HIDDEN = 2')

abline(0, 1, lwd = 2)

#HIDDEN = 5

set.seed(12321)

train_NN5 <- neuralnet(Total.Amount ~ ., data = datum_Train, hidden=5)
```

```
# Predict on test data

pr5 <- compute(train_NN5, datum_Test)

# Compute mean squared error

pr.nn5 <- pr5$net.result * (max(datum$Total.Amount) - min(datum$Total.Amount)) +
min(datum$Total.Amount)

test.r5 <- (datum_Test$Total.Amount) * (max(datum$Total.Amount) -
min(datum$Total.Amount)) + min(datum$Total.Amount)

MSE.nn5 <- sum((test.r5 - pr.nn5)^2) / nrow(datum_Test)

#mean squared error

MSE.nn5

# Calculate the root mean squared error

RMSE5 <- sqrt(MSE.nn5)

RMSE5

#To view the diagram of the ANN

plot(train_NN5)

# Plot regression line

plot(datum_Test$Total.Amount, pr.nn5, col = "blue",

    main = 'Real vs Predicted for HIDDEN = 5')

abline(0, 1, lwd = 2)
```

CLASSIFICATION USING CLASSIFICATION TREE CODE

```
library(rpart)
```

```
library(rpart.plot)

library(caret)

df <- read.csv("Dummy Data With State.csv")

selected = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21)

# dividing the dataset into training and testing sets

ind = sample(2, nrow(df), replace=TRUE, prob=c(0.6,0.4))

training = df[ind==1,selected]

testing = df[ind==2,selected]

dim(training)

dim(testing)

# default classification tree

default.ct <- rpart(Order.Status..Dummy. ~ ., data = training, method = "class")

# tree plot

prp(default.ct, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10)

#CODE FOR TESTING ACCURACY

dpred.train <- predict(default.ct,training,type = "class")

dpred.test <- predict(default.ct,testing,type = "class")

str(testing$Order.Status..Dummy.)

str(dpred.test)

#GENERATE CONFUSION MATRIX

#default tree: training
```

```
confusionMatrix(dpred.train, as.factor(training$Order.Status..Dummy.))

#default tree: testing

confusionMatrix(factor(dpred.test, levels = 1:4), factor(testing$Order.Status..Dummy., levels = 1:4))

#CODE FOR CREATING A DEEPER CLASSIFICATION TREE

deeper.ct <- rpart(Order.Status..Dummy. ~ ., data = training, method = "class", cp = 0, minsplit = 1)

# count number of leaves

length(deeper.ct$frame$var[deeper.ct$frame$var == "<leaf>"])

# plot tree

prp(deeper.ct, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10,

   box.col=ifelse(deeper.ct$frame$var == "<leaf>", 'gray', 'white'))

#CODE FOR TESTING ACCURACY

# classify records in the testing data.

#set argument type = "class" in predict() to generate predicted class membership.

deeperpred.train <- predict(deeper.ct,training,type = "class")

deeperpred.test <- predict(deeper.ct,testing,type = "class")

#GENERATE CONFUSION MATRIX

#deeper tree: training

confusionMatrix(deeperpred.train, as.factor(training$Order.Status..Dummy.))

#deeper tree: testing
```

```
confusionMatrix(factor(deeperpred.test,levels= 1:4), factor(testing$Order.Status..Dummy.,
levels = 1:4))
```

CLASSIFICATION USING ANN (ARTIFICIAL NEURAL NETWORKS) CODE

```
library(tidyverse)

library(neuralnet)

library(GGally)

library("nnet")

library(caret)

# Importing the dataset

datum= read.csv('Dummy Data With State for ANN Classification.csv')

vars=c("Segment..Dummy.", "Total.Amount", "Unit.Price")

#PARTIONING

set.seed(12345)

datum_Train <- sample_frac(tbl = datum, replace = FALSE, size = 0.60)

datum_Test <- anti_join(datum, datum_Train)

# dummify y

trainData <- cbind(datum_Train[c(vars)],

                   class.ind(datum_Train$ï..Order.Status..Dummy.))

names(trainData)=c(vars,

                   paste("Order.Status..Dummy_", c(1, 2, 3, 4), sep=""))

validData <- cbind(datum_Test[c(vars)],
```

```
                class.ind(datum_Test$ï..Order.Status..Dummy.))
names(validData)=c(vars,
                paste("Order.Status..Dummy_", c(1, 2, 3, 4), sep=""))
# hidden = 2
nn <- neuralnet(Order.Status..Dummy_1 + Order.Status..Dummy_2 +
                Order.Status..Dummy_3 + Order.Status..Dummy_4 ~
                Segment..Dummy.+ Total.Amount+ Unit.Price,
                data = trainData, hidden = 2)
plot(nn)
training.prediction=compute(nn, trainData)
training.class=apply(training.prediction$net.result,1,which.max)
confusionMatrix(factor(training.class),factor(datum_Train$ï..Order.Status..Dummy.,
levels = 1:4))
#testing data
validation.prediction=compute(nn, validData)
validation.class=apply(validation.prediction$net.result,1,which.max)
confusionMatrix(factor(validation.class),factor(datum_Test$ï..Order.Status..Dummy.,
levels = 1:4))
```