SEGMENTATION VIA PRINCIPAL COMPONENT ANALYSIS FOR PERCEPTRON CLASSIFICATION

KHAMIS MWERO MANENO

MASTER OF SCIENCE

(Information Technology)

JOMO KENYATTA UNIVERSITY OF

AGRICULTURE AND TECHNOLOGY

2022

Segmentation via Principal Component Analysis for Perceptron Classification

Khamis Mwero Maneno

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Technology of the Jomo Kenyatta University of Agriculture and Technology

2022

DECLARATION

This thesis is my original work and has not been presented for a degree in any other university

Signature……………………………………………….Date………………………

Khamis Mwero Maneno

This thesis has been submitted for examination with our approval as University Supervisors;

Signature……………………………………………. Date…………………………….

Dr. Richard Rimiru, PhD

JKUAT, Kenya

Signature……………………………………………. Date…………………………….

Dr. Calvins Otieno, PhD

M.U, Kenya

DEDICATION.

I dedicate this thesis to God Almighty my creator, my strong pillar, my source of inspiration, wisdom, knowledge and understanding. He has been the source of my strength throughout this program.

ACKNOWLEDGEMENT

I am thankful and fortunate enough to get constant encouragement, support and prudent guidance from my supervisors Dr. Richard Rimiru and Dr. Calvins Otieno of Information Technology Department which abetted me in successfully completing my thesis. Their willingness to give their time so generously has been very much appreciated. I do thank them for their assistance in keeping my progress on schedule. I would also like to sincerely thank my family, for their consistence support and encouragement along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

ANN	Artificial neural networks
BI	Business Intelligence
BIC	Bayesian Information Criterion.
CDR	Call Detail Records
CPA	Customer portfolio analysis.
CRM	Customer Relationship Management
DT	Decision Trees
ETL	Extraction, Transformation and Loading
FCM	Fuzzy C-Means
KDD	Knowledge Discovery in Databases
KVQ	Kohonen vector quantization
LR	Logistic Regression
OLAP	On-Line Analytical Processing
PCA	Principal Component Analysis
RFM	Regency, frequency and monetary
SOM	Self-Organizing Map
SPM	Sequential pattern mining

SVC	Support Vector Clustering
SVM	Support Vector Machines
UCAM	Unique Clustering through Affinity Measure

ABSTRACT.

In today's competitive environment, companies must identify their most profitable customer groups and the groups that have the biggest potential to become as such. By identifying these critical groups, they can target their actions, such as launching tailored products and target one-to-one marketing to meet customer expectations. With the profound advancements in clustering algorithms, segmentation has emerged as the method of choice for isolating the various groups of interest. However, the quality of segments of the groups of interest is affected by the type of input data to the clustering algorithms resulting from high dimensionality. In this study, Principal Component Analysis was usedto solve the high dimensionality of data problem in the the subscriber data. Principal component analysis was used to reduce the nine variables in the subscriber data to five. The factored data was then used to cluster the various customers into segments. The elbow criterion was used to determine the optimum number of clusters. The data was then clustered via several methods; K-means, Fuzzy C-means , Partioning About MediodsPAM, and Hierarchal Clustering. Results showed that k-means was not just the simplest method but also performed best with dimensionally reduced data. By using real case data, the study was able to verify that dimensional reduction can be applied before clustering algorithms. The dimension reduction of telecom data can thus be solved via Principal Component Analysis. The study was extended to include the classification of new subscribers basing on the dimensionally reduced data. For that purpose, a perceptron neural network was developed. Using the k-means clusters as targets, a perceptron capable of classifying was created and validated. The perceptron was able to classify new subscribers with acceptable accuracy. The perceptron model was validated and found to be accurate with an R^2 of 0.9999, Root Mean Square Error of 0.01813, Sum of Square Error of 1.0947 in all the data. Dimension reduction via Principal Component Analysis can, therefore, be used to achieve the segmentation of existing customers. The use of a perceptron is also important for automating the process of customer classification. Companies can therefore easily identify profitable customers from both old and new customers.

CHAPTER ONE

INTRODUCTION

1.1 Background

The telephony market provides a vibrant example of how increased competition and the exponential growth in the volume of data over the internet, has made it difficulty in predicting customer segments and pattern using traditional mathematical models. A 2016 report from the communication Authority of Kenya (CAK) provided insight into the Kenyan mobile sector by providing statics related the mobile money, penetration sector in the country for all the different telecommunication service providers. According to the report, Kenya's mobile penetration rose to 38.3 million subscriptions from 37.7 million the previous quarter, an increase of 3.5 million. Subsequently, mobile penetration grew by 1.5 percentage points during the period under review to stand at 89.2% up from 87.7% recorded the previous quarter

In terms of subscriber market share gains, Safaricom rose from 64.7% to 65.6% reflecting a change of 0.9% which directly translated to an increase from 24.4 to 25.1 million subscribers. Airtel, on the other hand, lost 1.7% market share to stand at 17.5%. According to Ngugi (2017), this was because of fresh SIM card regulations imposed by the regulator. This led to the disconnection of over 500,000 of Airtel's subscribers. Its total, subscriber base stood at 6.7 million, down from 7.2 million subscribers. Orange recorded a 0.1% increase in market share to 12.5% with 4.8 million subscribers from 4.6 million subs. Equity Bank's Equitel gained 0.7% market share to 4.4% with 1.6 Million subscribers (Ngugi & Komo, 2017). In that period, a new operator, Sema Mobile, acquired an MVNO license and managed \net 158 subscribers. In general, the growth in the telecommunication sector has been observed across most of the telecommunication companies. According to Premkumar (2017), the primary drive for growth in the telecommunications and the information technology sectors is associated with the speed of new technology implementation, which extends the market potential by introducing new services, and developing new capabilities to key players, as well as reducing their costs (Premkumar & Rajan, 2017). Additional factors affecting the competition and growth in the sector, are the world-wide de-regulation and privatization and the government efforts to change the monopoly position of the national communication carriers. The use of cell phones in Kenya continues to grow and this is seen not only through the upturn in the number of subscribers and providers, but also in the kinds of services that are provided. It is an ever-expanding industry; there is always room for innovation and growth and of course, a plethora of services to cater to the needs of every single customer (Ndambuki, Bowen & Karau, 2017).

Because of the large user base, the telecommunications sector is characterized with tremendously large volumes of data. These large data sets associated with mobile subscribers in the telecommunications have attracted a lot of research with scholars aiming at understanding and making use of the different data sets. Over many year's segmentation has become a fundamental concept both in marketing theory and in its practices.

Understanding customer's behaviors has key marketing strategies as it facilitates the development of targeted marketing programs (Mudogo, 2019) Basing on their traits, customers can thus be grouped into different segments. Customers may be thus be segmented on either demographic, geographic or behavioral traits. Consequently, each member of the segment has similar needs, wants and attributes.

In a much broader sense, different scholars have defined customer segmentation; Osman (2019) defines it as the division of customers into similar groups based on various features like their spending or buying traits. Hoegele, Schmidt & Torgler (2016) defines customer segmentation as process of dividing a large set of heterogeneous customers into groups that have similar characteristics, behavior or needs. Also, Kwach, (Nisbet et al., 2017) also in a similar way defines it as the development of meaningful customer groups in which various members have similar behaviors and traits. Like previous researchers, this research, customer segmentation will encompass the subdivision of customers into groups with similar characteristics

The main goal of dividing customers into groups is to enable companies to understand their market much better so as to increase their profitability through reducing operational costs in cases of the geographic. According to (Osman, 2019) segmentation provides a multidimensional visualization of the market data structure. This is key to a company as it enables them to identify customers who of strategic importance and consequently more profitable. Features that can be used for segmentation include demographic factors such as age, job status, marital status, gender. Geographic attributes include country and region. Geographic segments are very important as they enable a company to decide where to establish new branches. Other traits that can also be used are the psychographic attributes like life-style and Usage attributes that inform the companies on the frequency with which a given product is bought and in what quantity (Feizabadi & Shrivastava, 2018; Omamo et al., 2018). Segmentation is therefore very key in the development of customer profiles (Maruotti et al., 2019)

According to Greff et al., (2016), Segmentation aims at separating similar and dissimilar objects with the external market. The process requires that customer data be collected and analyzed which enables companies to identify the reliability and loyalty there by enabling them to increase their revenues. Segmentation plays a fundamental role in providing an enabling frame work for customized marketing. It’s also important for business decision support where by credit can be extended to a specific group of customers because of a particular trait like their buying characteristics. According to (Arora & Malik, 2015), this approach is key to unraveling the forces of demand and supply by eliminating some latent dependencies and associations that exist between/amongst customers and products.

According to Yen-Chung Liu et al., (2017), the process of segmentation largely depends on data that spans domains like their purchasing sequence data, product value data, customer profile data and their transaction data. The type of data to be used for segmentation process is highly dependent on the end user requirements for the data segments. For example, Yen-Chung Liu et al., (2017) proposed the use of customer purchasing sequence data as input data for clustering data on the basis of purchasing power. However, the limitation of this approach was highlighted in the works of (Osman, 2019)who alluded that the approach was unable to reflect the dynamic nature of customer data. Following the works of Chuang and Wong (2013), Yen-Liang Che et al., (2016) argued that profile data is also most likely no longer very useful given the complexity of the real-world data. To counter the predicaments of the works of Yen-Liang che et. al, Lee, M. K., Verma, R., & Roth, A. (2017) investigated the use of transaction data and was able to verify that it was able to reveal customer preferences and their needs and recommended that research should focus on the use of transaction data. The current diversity in psychophysical attributes, demographic attributes have limited the effectiveness of several segmentation methods and models (Aluizio F.R. Araujo, Victor O. Antonino, 2020). In order to overcome this limitation, current models for marketing have therefore focused on the use of transaction data. Segmentation has been proven as very essential in identifying customer behavior. Proper identification and profiling customers ensure that companies can discover how to reach their customers more effectively. Understanding customer behavior better companies can provide customized services and products (Maruotti et al., 2019) Therefore, segmentation has proven to be a valuable source of information for marketing, sales and business planning. Traditionally, companies conduct segmentation based on different market research and customer surveys. With the recent development of techniques for solving big data problems, the approaches for customers have evolved significantly rendering a number of traditional market analysis techniques inefficient and even obsolete in a number of cases (Simeon at el., 2015).

Also, there has been the emergency of large data repositories which has created the need to extract insightful information from these databases via a technique often referred to as data mining (Feizabadi & Shrivastava, 2018) The exponential growth in the volume of data over the internet becomes difficult to predict customer segment and pattern using traditional mathematical models. For companies to stay competitive, it’s important that they able to make decisions early in time. The huge volume of data exists, but companies starve for knowledge. Companies have therefore turned to these large databases to look for answers that can help them become more competitive by becoming more knowledgeable about their customers a process known as business intelligence.

According to Chen et al. (2019), Business Intelligence (BI) refers to the process of taking a large volume of data, analyzing it and presenting a set of high-level reports that contains important information that can be used to make decisions and support business actions. The process of Business Intelligence (BI) largely involves data extraction software, Transformation and Loading (EATL), data warehousing, database and query reporting and on-line analytical processing (OLAP). The insights provided by BI are fundamental to answering questions that can enable a company to understand long term progressions and also analyze the significance of the different improvement strategies (Kanyuga, 2019).

A term closely related to Business intelligence is data mining, which according to Zheng (2015) refers to the process of deriving insightful knowledge from large volumes of raw data. The acquired can then be made to use by applying to various applications that span several domains including market analysis. Nisbet et al. (2017) describes data mining as a single step in the process of knowledge discovery from a give data set. The data mining step primality involves extraction of useful information that wasn’t previously known and making it comprehensible. Data mining therefore directly improves the profitability of a given company through identification of the various business needs.

The main aim of Data mining in aims at identifying new, valid, important, explicable relationships and patterns in existing data sets. The process of identifying these patterns is what is now known under different names like knowledge extraction, information archelogy, and information discovery. The term data mining is commonly used among statisticians and database scholars. The term is closely related to Knowledge discovery in Databases (KDD) and essentially refers to extraction of useful information from a given database where by in this case data mining is simply a sub process. Other steps involved the KDD, include a data preparation process that involves data cleaning, a selection of appropriate variables, data interpretation and visualization. Data mining is only applicable when the goal of KDD is to unearth useful information. Data mining by itself is simply an extension of traditional data analysis approaches and statistical methodologies and draws from analytical, Machine learning, AI and data visualization approaches (Shmueli & Lichtendahl, 2017).

Basing on the underlying goal of data mining, the process of can be divided into two. Firstly, data mining can follow a supervised/predictive approach. In this method, knowledge is extracted from a given data set based on recognizing patterns from an external source which form a target data set. The supervised learning methodology therefore requires prior existence of known classes that serve as targets. The other data mining approach may follow an unsupervised learning approach (Femina & Sudheep, 2015) In the unsupervised learning approach, predefined classes aren’t necessary. Clustering techniques which group a set of inter-related objects are used. Some of the unsupervised learning techniques commonly used include k-Means algorithm, k-Nearest Neighbor algorithm, Self-Organizing Map (SOM) and Principal Component Analysis (PCA).

Data mining techniques in their different variants either supervised or unsupervised have already been used by several scholars like scholar (Dursun and Caber, 2016; Safari et al., 2016; Wang et al., 2016) for purposes of data segmentation. For example, Dursun and Caber (2016) was able to use k-means for profiling profitable hotel customers. Safari et al., (2016) developed a method that used Fuzzy c-means and fuzzy AHP for customer segmentation. In a similar study, Wang et al. (2016) was able to develop bi-clustering-based market segmentation which was also novel approach. More recently Wang and Zhang, (2020), used Self organizing maps for clustering high dimensionally Telecom data The works of these scholars are evident on the wide use of segmentation approaches in the field of data mining. It’s clear that apart from using existing methods to segment data, research is also ongoing in the development of new algorithms for data segmentation and also on the dimension reduction of the data.

The success of data mining methods is highly dependent on two main factors; a business problem that can be solved using data mining and an appropriate data sets for the implementation of the methodologies Bose and Chen, 2015; Gucdemir and Selim 2015 have affirmed the adherence of telecommunication sector to the perquisites. According to Deepali et al. (2017), was able to show that the telecommunications industry is confronted with many problems that can be broadly grouped into three broad categories: Firstly, there exits Business level problems that pose trivia's related to better understanding and predicting the behavior of different customer behavior, identification of customer needs, customer-oriented supply of new services, and improvement of business processes. On this level client-oriented data is used mined for solutions to a given trivia. The second level of problem exists at Product or service level like web mining. On this level service-oriented data is mined. Examples of data mining at this level can be found in cases where a customer's internet data usage and traffic mined. Basing on the type of information a customer's frequently access on the internet, tailored content can be provided. Providing the proper web links for target customer can be rewarding as user's internet data utilization increases thereby increasing profitability for the telecommunication. The third level of predicaments exits at Network and information infrastructure analysis level, at this level data mining can be used to predict fault detection, support network management and improve resource planning. The three problem categories are already a subject of research by different scholars (Ahmad et al., 2019; Peker et al., 2017; Suryadi & Kim, 2019).

The urgent need to solve the problems at the various levels using innovative, powerful methods fueled by the availability of large quantities of high-quality data accounts for the dominance and success of different data mining techniques. Therefore, with focus on the understanding customer behaviors, which has been classified as a level problem in the previous paragraph. This research will contribute a new model that uses k-means clustering algorithm to segment customer in mobile service providers industries.

1.2 Problem Statment

Customer segmentation is a key area in business intelligence that aggregate customer into groups of similar traits and behaviour such as demographic, geographical and behavioural. Its goal is to know the customer better and to apply the knowledge to increase profit, reduce operational cost and enhance customer services.

Previously, different customer clustering approaches have been proposed for segmentation, In a study by Chuang and Wong (2016), the SOM and SVM methodologies, much success haven't been found for solving customer segmentation due to the type of input data to the clustering algorithms and the high dimensionality of the input data. The findings were confirmed in the works of Yen-Liang Che et al., (2016) and later by Lee, M. K., Verma, R., & Roth, A. (2017).

Lee et al (2018). Proposed a method that makes use of transactional data and showed that it can reveal customer preferences and needs and after some period of time, it even reveals trends making it a more powerful input for clustering. As affirmed by the latest research, k-means algorithm improves performance in segmentation compared to other clustering algorithm, this was substantiated by (vibin vijay at el., 2017). K-means algorithm was acknowledged to be efficient in segmentation, as evidenced work done by (Ramirez-Ortiz et al., (2015). For this study, a factored transactional data in combination with an appropriate clustering algorithm, like k-means and a classifier such as perceptron will be used to develop the segmentation system.

However, it is not easy as expected since k-means algorithm requires the optimal clusters be known prior to the clustering. This field of determining the optimum clusters for clustering methodologies is extremely contentious and a number of methods have been proposed to overcome this underlying limitation. The number of clusters is extremely important as it can lead to misleading results aren't selected appropriately. Two overcome this challenge, researchers have widely adopted using clues package in R to automate and evaluate the process of optimal cluster selection. This research will therefore also adopt the clues package in R for optimum cluster selection.

1.3 Objectives of the study

1.3.1 General objective

To examine the segmentation of mobile telecom subscribers in Kenya via Principal Component Analysis for Perceptron Classification.

1.3.2 Specific objectives

i. To establish the best number of input variables through dimension reduction using Principal Component Analysis.
ii. To determine optimal number of clusters (the value of parameter k) that would represent the categories that exist in the data using the elbow criterion.
iii. To segment the data into the optimum clusters (k) using the k-means clusters and validate clustering results using silhouette index.
iv. To develop a perceptron model for classifying new customers into the established
v. Segments.

1.4 Research questions

i. What variables can best imitate the given data if the data is to be dimensionally resized using principle component analysis?
ii. What is the optimum number of clusters (value of parameter k) that should be generated from the data set?
iii. How well does the K-means perform in terms of accuracy when compared other highly used clustering methods like fuzzy c-means, PAM and Hierarchal clustering?

iv. How valid and accurate is the perceptron in assigning new customers to the different segments?

1.5 Significance of the study

The results of this research will be very crucial to all stakeholders especially the players in telecommunication industry where the competition is very intense. As affirmed by the latest research, k-means algorithm improves performance in segmentation compared to other clustering algorithm, this was substantiated by (vibin vijay at el., 2017). K-means algorithm was acknowledged to be efficient in segmentation, as evidenced work done by (Ramirez-Ortiz et al., (2015). Peker et al., (2017) concluded that K-means is much superior compared to Self Organizing Maps (SOM) when it comes to segmentation of data set that is distinct or well separated from each other. K-means algorithm was accredited to be fast, robust, easier to understand and relatively efficient this was verified by Wang andZhang (2020)

Based on its efficiency, robustness and accuracy in segmentation business operators will be able to segment their customers based on their behavior hence concentrating on the most reliable market and untapped market. By segmenting the customer properly, companies will be to develop customized marketing approaches that suite each market segment. The results are also key for decision support as they are key to offsetting forces of demand and supply through credit risk management. This is achieved by eliminating interdependencies that exist between and among both products and customers. This essentially refers to the act of identifying loyal customers who have high expenditures and grouping them. Credits can therefore be extended to them ensuring that they have purchasing power and consequently offsetting demand that would otherwise not exist without credit. This study will enable business operators to know the customer behavior better and to apply that knowledge to increase profit margins, reduce operational cost, and enhance customer service. Understanding customer behavior better will enable business operators identifying loyal customer, enabling them to focus of customer retention initiatives that key to a company's profitability. This research will also add to the body of literature concerning customer segmentation using k-means algorithm.

1.6 Scope of the study

The main focus of this research will be on telecommunication industries and will emphasize on the subscriber call expenditure across different times of day: - total day calls charge, evening call charges, night call charges, international calls and the Total cumulative charges as the information will be taken as inputs in k-means for founding customer segments.

1.7 Contribution of the Study

This study is crucial to all the players in telecommunication industry where the competition for subscribers is fierce and firms needs to stay profitable. Understanding and grouping customers in segments becomes not just crucial for market campaigns but also provides a competitive advantage through targeted service delivery. To address the customer segmentation challenge in the current Kenyan telecommunication industry, this research has made contributions that transverse several domains. Firstly this research has confirmed the existence of four customer categories which has four long been used as a basis for service packages like starter, standard, basic and premium packages. Secondly, this research has for the first time segmented the different clients of the telecommunication industry using novel K-means clusters that have been validated. This research has also affirmed the strength of the K-means over other methods like FCM, PAM in terms of efficiency, robustness and accuracy for solving segmentation problems. By segmenting the customer properly, this research has shown the ability of a business to develop and customize products suitable for each of its customer segments. As a third contribution, this research has developed a novel perceptron model that can be used to directly classify new clients into the different segments. This research has therefore not just provided insight into the customer segments but has also contributed a method to assign customers to the different segments.

1.8 Limitation of the study

The main limitations of this study is inherited from the k-means algorithm that requires the optimal clusters be known prior to the clustering. This field of determining the optimum clusters for clustering methodologies is extremely contentious and a number of methods have been proposed to overcome this underlying limitation. The number of clusters is extremely important as it can lead to misleading results if not selected appropriately. To overcome this challenge, researchers have widely adopted using clues package in R to automate and evaluate the process of optimal cluster selection. This research therefore adopted the clues package R for optimum cluster selection.

The effects of the recently passed General Data Protection Regulation (GDPR) hampered the efforts of the researcher to obtain real time data as custodians of personal data from telephone companies are barred to share Call Details records without legal consent as it may amount to data breach. The data collected for the study is a real data set donated to the machine learning repository for research and further advancement in the field of machine learning and artificial intelligence. The data is however historical and may have been collected at different times, therefore quality and accuracy of research could have been compromised. The time lag was assumed and therefore the findings are not an accurate reflection of the present.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This section gives brief overview of data mining, discussing various models available, followed by their functionalities, then discussing the relationship between the field of data mining and it's closely related counterpart Business intelligence with regard to their application to segmentation. Previous research on those who succeeded and failed in segmentation, then critique of the aforementioned then finally this will result to the gap of the research.

2.2 Business intelligence and data mining

Business organizations are adopting data-driven strategies to generate more profits out of their business. Growing startups are investing a lot of funds in data economy to maximize profits of the business group by developing intelligent tools backed by machine learning and artificial intelligence. The kind of business intelligence (BI) tool depends on factors like business goals, size, model, technology, etc. In this paper, the architecture of BI tool and decision process was discussed with a focus on customer segmentation, based on user behavior.

Business intelligence (BI) is the process of taking large volume of data, analyzing that data and presenting a high-level set of reports that concentrates the important of that data into the basis of business actions, supporting managers to make daily decisions (Femina & Sudheep, 2015). BI includes much software for Extraction, Transformation and Loading (ETL), data warehousing, database query and reporting, multidimensional or on-line analytical processing (OLAP).

The decision making cannot be based on archaic information only; it can be based on current information too, for example, executives cannot afford to make decisions based on financial statements which compare last months results to a budget created up to a year ago. Do they need information that helps them quickly answer the basic questions, for instance, what was the sale last year? What continues to sell in this year? What costs can be cut without causing long-term ham?

Business intelligence system provides the ability to answer the critical questions by turning the massive amount of data from operational systems into a format that is more readable and easier to understand (Greff et al., 2016) BI software's allows the organization and even departments inside it to analyze current and long-term trends and gives continuous feedback on the decision's effectiveness.

The BI and Data Mining applications in any industry depend on two main factors: the availability of business problems that could be successfully approached and solved with the help of BI and Data Mining technologies, and the availability of data for the implementation of such technologies. In most of the analyzed literature sources, these two main prerequisites are entirely fulfilled for the Telecommunications sector. The Telecommunications industry is confronted with many business problems that need urgent handling by using innovative, powerful methods and tools and is in possession of large quantities of high-quality data that is a key success factor for BI and Data Mining applications.

The data multidimensionality is discussed by (Osman, 2019) and is considered as one of the most important factors for the wide variety of BI and Data Mining applications. The availability of tremendously large volumes of Telecommunications data is a very important reason for the heighted interests in the use of Data Mining and Business Intelligence. Some of its applications are summarized in Table1.

Table 1: Fields in which Data mining and Business intelligence has been used.

Areas	Business Problems Addressed
1. Marketing, Sales, and CRM	• Generating customer profiles from call detail records. • Profiles for marketing purposes. • Measuring customer value and retaining profitable • customers• • Maximizing the profit obtained from each customer. • Acquiring new customers. • Segmentation Analysis. • Segmentation management:
2. Fraud Detection	• Identification of potentially fraudulent users and their atypical usage patterns (subscription fraud) • Detecting attempts to gain fraudulent entry to customer accounts (super-imposed fraud) • Discovering unusual patterns that may need special attention such as busy hour, frustrated call attempts, switch and route congestion patterns, etc.
3.Network Management	• Network fault identification; Alarm correlation (for relating multiple alarms to a single fault) • Network fault prediction • Identifying and comparing data traffic • System workload management • Resource usage management • User group behavior

2.2 Customer segmentation and data mining techniques

Suryadi (2019), defines customer segmentation as the use of stored data to divide customers into similar groups based on various traits. Also, Hoegele, Schmidt & Torgler (2016) defines customer segmentation as process of dividing a large set of heterogeneous customers into groups that have similar traits, behavior or needs. Sarvari et al. (2016) defines segmentation as the process of developing meaningful customer groups that are similarly based on individual explanation characteristics and behavior. In other words, customer segmentation is also described as the process of dividing customers into homogeneous groups based on shared or common attributes.

The goal of segmentation is to know the customer better and to apply that knowledge to increase profitability, reduce operational cost, and enhance customer service. Segmentation can provide a multidimensional view of the customer for better treatment strategy (Feizabadi & Shrivastava, 2018). By using customer's segmentation, a company there for be able to directly identify its strategic customers and maximize its profitability. Scholars have shown that customers can be segmented basing on three main characteristics that include, geographic location like country and region, demographic traits like age, gender, job and marital status and Psychographic characteristics like life-style (Feizabadi & Shrivastava, 2018; Nisbet et al., 2017)Customer segmentation is used to build the customers' profiles which makes up the core of a customer-centric information system (Bose and Chen, 2015).

The main target of segmentation is to separate the objects that are homogeneous and heterogeneous with the external market (Peker et al., 2017). Segmentation requires the collection, organization and analysis of customer data. Segmentation is vital to many organizations' strategic marketing plan since goods and services can no longer be produced and retailed without bearing in mind variations in the customers' needs and preferences (Lee et al., 2015)

With proper segmentations of a customer's data, it is possible to identify the reliability and loyalty of customers in order to increase the revenue of the organization. Customer segmentation avails the knowledge that enables business to customize market programs. According to the results of Arora & Malik, 2015, customer segmentation can be divided into groups which can repay or default there credit obligation. This is important as it enables companies to identify which customers can receive credit and which one's can't.

According to Yen-Chung Liu et al., (2017 segmentation approaches can be characterized basing on customer data into four main categories; the first form of categorization makes use of customer purchasing sequence data. The second customer categorization uses customer value data. This is assessed basing on the amounts of products purchased by a given customers. The third classification uses profile data of a given customer, the data in this category includes customer age, gender. The fourth class makes used of transaction data. Greff et al., (2016) suggested customer value data approach for segmenting customer. Che et al. (2016) concluded that profile data is no longer plays a significant role in the current, increasingly complicated world. Lee, M. K., Verma, R., & Roth, A. (2017) affirmed that transaction data can reveal customer preferences, needs and behavior. The diversity of customer needs and buying behavior, influenced by lifestyle, income levels or age, makes past segmentation approaches less effective consequently, current models for marketing segmentation are often based on customer behavior inferred from transaction data. Segmentation has been proven as very essential in identifying customer behavior. The identification and profiling the most important segments according to the behavior, companies can discover how to reach their customers more effectively. Understanding customer behavior better companies can provide customized services and products (Nisbet et al., 2017). Therefore, segmentation has proven to be a valuable source of information for marketing, sales and business planning. Traditionally, companies conduct segmentation based on different market research and customer surveys. According to Simeon Ozuomba at el., 2015, the paradigm shift in the field of segmentation is strongly attributed to the recent developments in the field of big data and machine learning. The large repositories and data ware houses have by themselves fueled the widespread of data mining as companies seek to gain insight into the behavior of customers so as to gain a competitive advantage (Lee et al., 2015). The exponential growth in the volume of data over the internet, becomes difficult to predict customer segment and pattern using traditional mathematical models. Timely identification of newly emerging trends from the huge volume of data plays a major role in the business process and decision making. Huge volume of data exists, but companies starve for knowledge. Thus, Business Intelligence and Data mining tools are needed to overcome this knowledge scarcity and analyzing the data then presenting set of reports that will assist business operators in making daily decisions. Business intelligence (BI) is the process of taking large volume of data, analyzing that data and presenting a high-level set of reports that concentrates the important of that data into the basis of business actions, supporting managers to make daily decisions (Arora & Malik, 2015) BI includes much software for Extraction, Transformation and Loading (ETL), data warehousing, database query and reporting, on-line analytical processing (OLAP).Business intelligence system provides the ability to answer the critical questions by turning the massive amount of data from operational systems into a format that is more readable and easier to understand (Greff et al., 2016). BI software's allows the organization and even departments inside it to analyze current and long-term trends and gives continuous feedback on the decision's effectiveness.

Data mining techniques help to overcome this knowledge scarcity (Deepali Kamthania al el., 2017). Zheng (2015) defines data mining as the process of mining, gaining, knowledge from large volume of raw data. This knowledge gained can be used for applications ranging from market analysis among others. Data mining is the process of extracting valid, useful, previously unknown, and ultimately comprehensible knowledge from large database. Data mining is considered as a step in the whole process of knowledge discovery. Data mining techniques can be used efficiently in any business application that involves data, such as increasing the business unit and overall profitability, understanding customer desires and needs, identifying profitable customers and acquiring new ones (Peker et al., 2017).

The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing.

The term “data mining” is primarily used by statisticians, database researchers, and the business communities. The term KDD (Knowledge Discovery in Databases) refers to the overall process of discovering useful knowledge from data, where data mining is a step in this process. The steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results ensure that valuable information is derived from the data. Data mining is an extension of traditional data analysis and statistical approaches as it incorporates analytical techniques drawn from various disciplines like AI, machine learning, OLAP, data visualization among others (Shmueli & Lichtendahl, 2017).

It is into two main classes according to their goals; they include - Supervised/Predictive learning and Unsupervised learning. Supervised and unsupervised learning are two quite different techniques of learning as the names suggest, supervised learning involves learning with some supervision from external sources whereas unsupervised learning does not. Supervised learning involves classification techniques which have a set of predefined classes and want to know which class a new object belongs to (Peker et al., 2017). These includes decision tree, Bayesian network classifier and so on.

Unsupervised learning involves clustering techniques which try to group a set of objects and find whether there is some relationship between the objects. This from of learning in this research is classified under unsupervised learning. Unsupervised is dominated by clustering algorithms like K-means Self-organizing map. These methodologies without prior knowledge extract information inform of clusters and iterate until when the clusters are correct by a selected accuracy criterion. Each cluster is formed such that the member have similar characteristics whilst different clusters represents other traits. Previous study shows different authors who proposed supervised and others supervised approaches for segmentation. Verster et al., (2017) applied k-means using information customer value data. Perez et al., (2018) investigated the balancing efforts and benefits of k-means. Larose (2014) used SVC for segmentation. (Kohonen 2013) analyzed Self Organizing Maps for customer segmentation. Bose & Chen (2010) applied both K-means and Kohonen vector quantization (KVQ) to group customers based on the attributes. Mehta et al., (2017) applied both K-means and SOM have for segmentation.AL Zahrani et al., (2015) investigated hierarchical agglomerative cluster analysis to reviewed two clusters in which one cluster has low fruit consumption and second one includes high sweet consumption. Peker et al. (2017) described the use of Kernel-means for segmentation.

2.3 Concept of Data Mining

The most perilous aspect of customers' behavior analysis is the discovery of hidden and useful customers' behavioral patterns. This process entails historic and transactional data. Increased business activities, competition, and affordability assist organizations to generate huge transaction data repositories. Data mining enables industries to discover useful knowledge for providing customer with oriented services. It exploits well-established statistical and machine learning techniques to construct models that predict customers 'behavior patterns. Further still the process of data mining can be described as the construction of the set of rules and mathematical equations that can be used to identify patterns, understand them and use them to predict behavior as can be observed in figure 2. In its basic form, it can be classified into two broad classes; Supervised and unsupervised learning. As the names suggest, supervised learning makes use of external assistance inform of a target data set to extract data from a new data set whereas unsupervised learning directly seeks information from a data set without assistance from an external data set. Supervised learning involves classification techniques which have a set of predefined classes and want to know which class a new object belongs to (Braha, 2013). Unsupervised learning involves clustering techniques which try to group a set of objects and find whether there is some relationship between the objects.

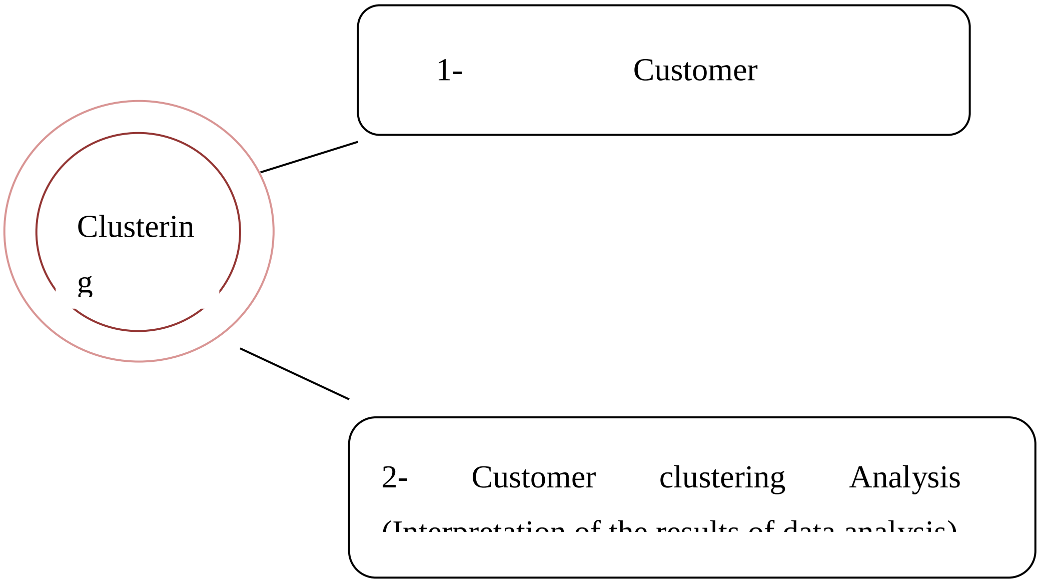

1- Customer
Clusterin
g
2- Customer clustering Analysis
(Interpretation of the results of data analysis)

Figure 2.1: Clustering Phase

Today, the knowledge discovery process constitutes automated technologies that integrates the process with commercial data warehouses, and presents the outputs in the desired format for easy comprehensibility by users. Hence, the key dimensions of segmentation are supported by different Data Mining models such as association, forecasting, regression, sequence discovery and visualization. Data Mining is therefore a multidisciplinary approach that unifies several methods (statistics, mathematics and machine learning) to translate the desired knowledge into rules and models, useful for problem solving and decision making. The process of knowledge discovery commences with the acquisition of customers' information to form a data warehouse. The data warehouse in turn provides a foundation for the customer information. The process of knowledge discovery can be automated using data mining in which case the data mining only serves as a step within the overall major process. The entire knowledge discovery process id summarized below:

- Define and understand the application realm and anticipated business objectives.
- Create a target dataset – selection of a subset of variables and data samples to be mined.
- Clean and data pre-processing in order to reduce the occurrence of noise or outliers and selecting information needed for model formation.
- Reduce and project the data for meaningful discovery.
- Deploy data mining tools and systems for classification, clustering and so on.
- Selection of the data mining tool/ research technique.
- Developing clusters in the data for the purpose of unearthing new knowledge.

Several DM techniques and tools have been proposed for segmentation, these include visualization, neural networks, genetic algorithms, fuzzy logic, rules induction, decision trees, and clustering. K-means was used by Nisbet et al. (2017) K-means algorithm for clustering Taiwanese healthcare institutions based on customer value assessment. Farshid Abdi and Shaghayegh Abolmakarem (2018) applied clustering and classification techniques in Customer Behavior Mining Framework (CBMF). Wighton et al., (2015) in a similar study used random walker algorithm for segmentation. Alshammari, (2019) used 3D depth information to RGB color images to improve segmentation of pigmented and non-pigmented skin lesion. Komati et al., (2016) suggested for an improved version for the JSEG color image segmentation algorithm, combining the classical JSEG algorithm with local fractal operator

Different algorithms have been used by a number of scholars for purposes of segmentation like

Nisbet et al. (2017) used Hierarchical clustering technique for modeling the reliability of power systems Also Aluizio (2020) used Self-organizing maps (SOM) and k-means methods then compared the results to get the appropriate strategies. Nisbet et al.

(2017)proposed watershed-based algorithm for automatic segmentation in dermoscopy images. Lee et al. (2015) presents an automatic skin cancer segmentation system using neural network.

2.3.1 Classification techniques for customer segmentation

Classification analysis is the process of finding a model (or function) that describes and distinguishes data classes or concepts, by being able to use the model to predict the class of objects whose classes are unknown. By using classification, it is possible to organize data in each class. Classification uses known class labels to bring together the objects in an orderly pattern. Classification is one of the most commonly used supervised modeling techniques. In classification, a user needs to divide data into segments and then make distinct non-overlapping groups. For dividing data into groups, a user needs to have certain information about the data to be divided into segments.

The main objective of classification is to identify the characteristics that enable an object to belong to a given group. The characteristics of each group can be mastered and used to map new objects to the different classes. According to Al-Mashraie et al., (2020), the different patterns mined clustering can be directly used with predictive models like a perceptron. The process is involving two stages; the primary data is first spilt into validation, testing and a training data set. The model can then be fully developed so that it can handle new unknown data sets. Classification tasks have been carried out for various purposes in CRM domain. Ocumpaugh et al., (2014) adopted decision tree to classify the customers and develop strategy based on customer life time value. Shmueli & Lichtendahl, (2017) identified the slope of the customer lifecycle based on Bayesian network classifier. The author illustrated Bayesian network classifiers as useful tool in the toolbox of CRM analysts in application of identifying the slope of the customer lifecycle of long-life customers.

De Keyser et al. (2015) adopted decision tree to explore the potential relationship between important influential factors and customer loyalty. The findings from segmentations can be used to construct relation trees that can help explore relations like the relationship between a customer and purchase amounts, demographic and behavioral characteristics, with special attention to the characteristics of high-and low-spending customers.

2.3.2 Clustering techniques for customer segmentation

The working of Clustering approaches follows a similar objective of maximizing the similarity between objects in the same class (intra-class similarity) while minimizing similarity between objects in the same class (inter-class similarity). The difference between clustering and classification mainly lies the nature of data that each uses to discover classes (Nisbet et al., 2017). It’s a common practice to modify results from clustering by variable since several variables may not be relevant and may not cause significant result. After clusters have been created, they can be used to train a classification algorithm to train new data. The most common algorithms often used for clustering are K-means and Kohonen feature maps. A cluster is an assemblage of data with similar characteristics. Clustering can be used to group customers with similar behavior and to make business decisions in industry to as unsupervised learning (Greff et al., 2016) Unsupervised learning is a process of classification with an unknown target, that is, the class of each case is unknown. The aim is to segment the cases into disjoint classes that are homogenous with respect to the inputs.

Clustering studies have no dependent variables. Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a

De Keyser et al. (2015) applied segmentation of customers of Trade Promotion Organization of Iran using a proposed distance function which measures dissimilarities among export baskets of different countries based on association rules concepts. Later, to suggest the best strategy for promoting each segment, each cluster is analyzed using

RFM model. Variables used for segmentation criteria are “the value of the group commodities”, “the type of group-commodities” and “the correlation between export group-commodities”.

Greff et al. (2016) elaborates the use of clustering to segment customer profiles of a retail store. The study concluded that the K-Means clustering allows retailers to increase customer understanding and make knowledge-driven decisions to provide personalized and efficient customer service. Huang et al., (Lee et al., 2015) applied K-means method, Fuzzy C-means clustering method and bagged clustering algorithm to analyze customer value for a hunting store in Taiwan and finally concluded that bagged clustering algorithm outperforms the other two methods.

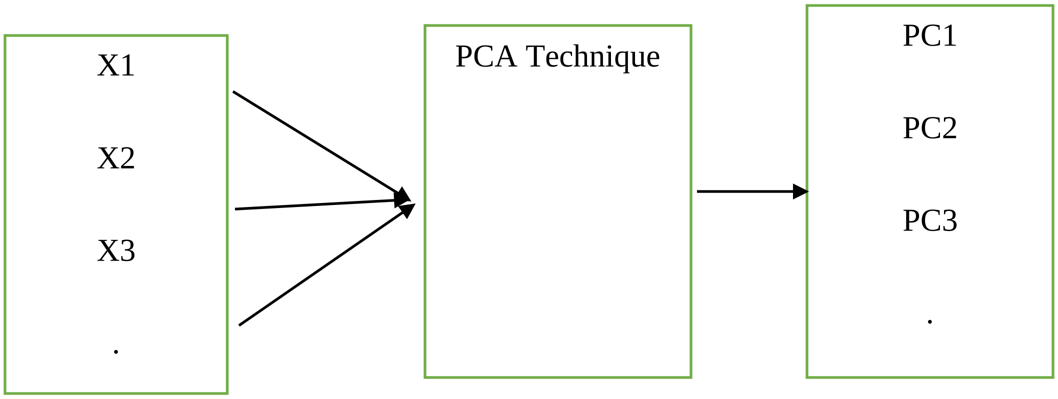
2.3.3 Principle component analysis

PCA technique is based on the assumption that variables are linearly related. PCA Analysis is like having a different viewpoint for the same data set. By changing the coordinate system to the centroid of the data and rotating the axes, the view point can be varied. For a set of n variables bounded in the domain f(X) = X {1…n} PCA calculates a set of n linear combinations of the variables PC_n where n = {1…n} such that

i. Variation in the new data set formulated from PCA is same as that in primary data set
ii. The first PC contains the most variance possible, i.e. as much variance as can be captured in a single axis.
iii. The second PC is orthogonal to the first one (their correlation is 0), and contains as much of the remaining variance as possible.
iv. The third PC is orthogonal to all previous PC's and also contains the most variance possible.

The above process is accomplished by calculating a matrix of coefficients where columns are referred to as eigenvectors of the variance-covariance or of the correlation matrix of the data set. The process of data transformation from high to low dimension is illustrated as shown in figure 2

Input (High dimension data) PCA transformation (output) reduced dimension



Where, $n \leq m$

Figure 2: PCA Technique for transformation of data to low dimension

2.4 K-means Model

This is the simplest unsupervised learning clustering algorithm as it mainly involves grouping data into predefined groups. The k in name stands for the number of clusters selected. An observation is assigned to a specific cluster for which its distance to the cluster mean is the smallest. The main function of the algorithm involves finding the k-means. It begins with defining the initial set of means then subsequent classification is based on their distances to the centers, the clusters ‘mean is computed again and then reclassification is done based on the new set of means. The process is repeated until the there’s minimum variation in the cluster means for successive iterations (Mehmanpazir, and Asadi 2017). This is done until all cluster centers are established. The different data values are then allocated to a permanent cluster. For a set of observation X_n where n is a real multi-dimensional real vector, K means partitions that into clusters

$K \leq n$ via an algorithm presented below.

K-Means algorithm

Simplified simulation flow of k-means algorithm

```
Begin
Inputs:
X = ($x_1$, $x_2$......, $x_n$)
Determine:
Clusters –k
Initial Centroids -$C_1$, $C_2$..., $C_k$
Assign each input to the cluster with the closest centroid
Determine:
Update Centroids -$C_1$, $C_2$......, $C_k$
Repeat:
Until Centroids don’t change significantly (specified threshold value)
Output:
Final Stable Centroids -$C_1$, $C_2$......, $C_k$
End
```

2.4.1 Cluster validation

How good a cluster is can be directly assessed by a cluster validation procedure in order to eliminate the possibility of clustering random data. Cluster validation is also important for comparing different clustering methods. Several scholars (Nisbet et al., 2017; X. Wang & Huang, 2015) have shown that there are two methods of cluster validation. Firstly, cluster may be validated using internal validation cluster approach which measures the goodness of clustering without external information. The second method involves external cluster validation in which the results are compared to results of known clusters in which each group member is well labeled. In an external method, true are clusters are compared to the clusters from the cluster algorithm. The error can therefore be directly estimated. This method is mainly used with supervised clustering approaches (Maruotti et al., 2019)

Cluster validation can still be categorized into three basing on the type of dataset involved. Three indices can be directly inferred, External, internal and relative indices. External validation can be used in cases in which input-output data sets exist. On the other hand, internal indices like silhouette indices can be used in cases where a single dataset can be used (only input data set). Relative indices are useful incases were clustering methods need to compared and the results may make use of both internal and external indices.

The two most common methods for evaluating the goodness of clustering are the silhouette index and the Davies Bouldin index. The silhouette index measures the extents of clusters and measures the distance between them. A plot of silhouette indices therefore reveals how close each point is to its neighbor.

For this experiment the silhouette index will be used for validation due to graphical display and easy interpretation.

Silhouette index is a standard method for assessing and comparing different cluster algorithms whether supervised or unsupervised. The Silhouette value measures cohesion and separation in clusters. The values of the silhouette index range from -1 and +1 with -1 showing poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

2.5 Related Work

Existing work show different studies for both supervised and un supervised techniques applied in segmentation using various data approach, which involved application of novel clustering methods including Adaptive k-means (Suryadi & Kim, 2019), spectral clustering (Vercamer et al., 2015), self-organizing Maps (Verdu et al., 2016), subspace clustering (Femina & Sudheep, 2015).Yen-Chung Liu et al., (2017) investigated customer purchasing sequence data to cluster customers. Deligiannis et al., (2019) suggested customer value data approach for segmenting customer. Lee, M. K., Verma, R., & Roth, A. (2017) affirmed that transaction data can reveal customer preferences, needs and behavior.

Shereen Fouad et al., (2017) in their paper epithelium and stroma identification in histopathological images used unsupervised and semi-supervised super pixel-based segmentation. Arumawadu, Rathnayaka and Illangarathne (2015), proposed k-means clustering algorithm to evaluate customers' profitability in a telecom industry in Sri Lanka. Their results revealed the customers' profitability were mainly categorized into three levels. Sivasankar and Vijaya (2017) combined supervised and unsupervised techniques for segmentation, various unsupervised learning techniques were comparatively studied. Some of the algorithms include fuzzy C-means (FCM), Probabilistic fuzzy C-means (PFCM), k-means clustering (k-means) were used to segment customers' similarity patterns within a cluster for customer segmentation. They divided the clusters into training and testing using Holdout method. Their results showed that, combining k-means clustering with decision tree helps to improve the result for various segments. More recent researchers have also been applied for clustering in segmenting individual residential customers (McLoughlin et al., 2015). Aluizio (2020) analyzed Self Organizing Maps and acknowledged that, it is not easy to implement better segments using SOM as expected since, information might be lost if a high dimensional data is mapped onto an array that has a resolution too small to display the fine structures in the input data. On the other hand, according to Greff et al., (2016), to solve the issue of high dimensions for data segmentation, declared that SVC algorithm does not take care of dimensionality reduction, these models are weak and insufficient in segmentation. Md.Rakib Hassan et al., (2017), acknowledged the fact that a lot of work had been in the use of data techniques to improve marketing strategies through market segmentation. The author's also pointed out the fact that it the several studies, conclusions hadn't been drawn as to what algorithms yielded the best accuracy. T. Verster et al., (2017) in their paper a semi-supervised segmentation algorithm as applied to k-means using information value, avowed that computationally, clustering techniques are significantly more complex than k-means clustering. Perez et al., (2018) investigated the balancing efforts and benefits of k-means, stated that K-means is computing time reduction and solutions quality in terms of large and complex masses of data (Big data realms). Madhusmita Sahu et al., (2017) in compared the performance of k-means to other clustering algorithms and found that K-means gave the best results. Segmentation using K-means clustering, results showed that k-means is accurate in terms of performance (Rakesh Kumar., 2016). R. Spurgen Rathish et al., (2017) in their paper for the Comparison of Segmentation based on Threshold and K-means method underscored that k-means is more accuracy, less running time and high resolution. This was also accentuated by Yang et al., (2015)

Manzano et al., (2015) used evolving limitations in K-means algorithm in data mining and their removal.it was stated that K-means is popular because it is conceptually simple, computationally fast and memory efficient, but various types of limitations in k means algorithm that makes extraction difficult compared to other clustering and data mining methods.

Peker et al. (2017) used a mixture of strategies, proposed a customer classification technique which was used for facilitating cross-selling in a mobile telecom industry using different methods. Firstly, classification techniques like artificial neural networks (ANN), logistic regression (LR), and decision trees (DT) were applied one by one to predict the future purchase of products. This proved more complicated and time consuming, it further required advanced knowledge in data mining process. Secondly, each model produced the result and then applied genetic algorithm on the combined results of all models and applied K-means clustering of customers to establish whether customer would purchase a new product. The framework was tested in a mobile telecom company in Korea. The model produced excellent results for cross-selling. Based on simplicity, fast and efficient nature makes K-means outstands the unsurpassed data mining technique when it comes to customer segmentation.

Ramirez-Ortiz et al., (2015) used a simple and efficient implementation of Lloyd's k means clustering algorithm, which researchers call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. The study further found that as the number of clusters, k becomes greater, the performance of SOM algorithm becomes lower compared to K-mean algorithm. In their paper, researchers proposed a Lloyd's k means clustering algorithm for segmentation since it aided in improving the efficiency.

Amin et al. (2019) explored business opportunities from mobile services data of customers using an inter-cluster analysis approach. The authors utilize customer clustering technique to discover the customer behavior who assure to mobile services. The clustering was generally performed in services, revenue, usage and user categories attributes. In this study, K-means and Kohonen vector quantization (KVQ), were used to group customers based on the attributes. Then inter-cluster analysis was performed on the generated clusters and evaluated the scattering of customers among the dissimilar group of attributes. Customer transaction data was collected from mobile telecom operator that is in Hong Kong. Data was partitioned in to four clusters. As a result, this was found that K-means inclined to create clusters with a slighter variation in intra-cluster distribution and KVQ inclined toward creating clusters with slighter average intra-cluster distribution. Furthermore, the study found that the performance of K-means algorithm was more competent than that of KVQ and hierarchical clustering algorithms.

Mehta et al., (2017) found that both K-means and SOM have linear complexity, O (ndk) and O (nd) respectively. In which case n represents the data samples size, d size of dimensions and k the clusters size. Both methods have a good ability to scale data with comparable performance. Another method is SVC that has a much higher computational complexity, where n is the number of support vectors instead of the sample size. Vibin Vijay at el., (2017) in their paper Variance Based Moving K-Means algorithm, presented the performance of the proposed data clustering method on four datasets chosen from the field of image analysis, bioinformatics, remote sensing, and the stock market respectively. The details of experimentation along with quantitative as well as qualitative results including a comparison with seven state of the art algorithms namely K-Means (KM), Moving K-Means (MKM), Fuzzy C-Means (FCM), Adaptive Moving K- Means (AMKM), Enhanced Moving K-Means-1 (EMKM1), Enhanced Moving K-Means-2 (EMKM2) and K-Means clustering was presented ,K-Means and VMKM produced the best image segmentation with similar lowest MSE(mean squared error).

Liu (2017) presented an optimal K-Means Clustering algorithms, employed in many bioinformatics tasks, including categorization of protein sequences and analysis of gene-expression data. Segmentation was performed with several attributes, the high-dimensionality affects performance of the algorithms and quality of the clusters, so the fewer the dimensions are, the faster the execution and more compact distribution of data samples are, it was found that K-means algorithm handles the high-dimensionality of data during the iterative steps (Calvet et al., 2016)

Peker et al., (2017) described the use of Kernel-means, which was an extension of the standard means clustering algorithm that identified nonlinearly separable clusters. It proposed by the global kernel-means algorithm, a deterministic and incremental approach to kernel-based clustering. This also added one cluster at each stage, through a global search procedure consisting of several executions of kernel -means from suitable initializations. This algorithm did not depend on cluster initialization, identified nonlinearly separable clusters, and, due to its incremental nature and search procedure, located near-optimal solutions avoiding poor local minima. Furthermore, two modifications were developed to reduce the computational cost that did not significantly affect the solution quality. The proposed methods were extended to handle weighted data points, which enable their application to graph partitioning. This experiment with several data sets and the proposed approach compared favorably to K –means with random restarts. It concluded that K-means is much superior compared to SOM when it comes to segmentation of data set that is distinct or well separated from each other.

Crespo & Weber (2015) analyzed K-Means clustering algorithm and found that quality of the resultant cluster was based on the initial seeds where it was selected either sequentially or randomly. For real time large database, it was difficult to predict the number of cluster and initial seeds accurately. To overcome this drawback, new algorithms were proposed. They were Unique Clustering through Affinity Measure (UCAM) and it worked without fixing initial seeds, number of resultant clusters to be obtained and unique clustering was obtained with the help of affinity measures.

The real-world customer transaction data can contain outliers and missing values. These are often problematic for the clustering as they distort the results and, in this way, weaken the quality of the clusters. SOM visualizes the data set into a two-dimensional map which is a useful aid when detecting outliers. Outliers can be spotted from the map as they usually have the longest distance to all other clusters. When finding such a group the data miner can either remove outliers directly or further investigate if this group has some interesting characteristics. In the case of missing values, SOM can simply be tuned to leave out missing values from training, however it is not easy to implement better segments using SOM as expected since, information might be lost if a complex set of data is mapped onto an array that has a resolution too small to display the fine structures in the input data (Geng et al., 2015). For K-means, the algorithm is based on Isotropic, due to its isotropic nature (i.e. its tendency to form clusters that are spherical, it is common practice to standardize(i.e. transform to have zero mean and unit variance) all input variables for k-means analysis Verster et al., (2017), hence providing better segments. Consequently, with the non-spherical clusters, PD-clustering can be applied on the data since it’s a flexible method that can be used with non-spherical clusters data then k-means applied for segmentation Paul D. McNicholas et al., (2017)

K-means has been use by several scholars, for example, Namvar et al. (2017) used it to classify customer loyalty based on RFM values. Still using RFM values, Verma (2017)combined k-means and rough set theory to cluster customers. Yang et al. (2016) identified purchasing patterns based on sequential patterns. Wang et al. (2016) proposed a method for customer’s segmentation given by the nature of the products purchased by customers. This method was based on clustering techniques, which enable segmenting customers according to their lifestyles. The author segmented customers of a European retailing company according to their lifestyle and proposed promotional policies tailored to customers from each segment, aiming to reinforce loyal relationships and increase sales. The author used the VARCLUS algorithm, integrated in SAS software, to cluster the products. The methodology also involved the inference of the lifestyle corresponding to each cluster of products, by analyzing the type of products included in each cluster. In all cases, K-means proves to be fast, robust and easier to understand and relatively efficient: O (tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.

2.6 Critiques of existing literature

The field of clustering has seen the use of several clustering algorithms in combination with several input data sets. However several scholars have highlighted key issues that exist in the current approaches especially with the use of clustering as method for data segmentation. In the works of Yen-Chung Liu et al., (2017), he proposed the use of customer purchasing sequence data as inputs however several studies later revealed that approach was sufficient. In light of the challenge presented by several authors, in an investigative was carried out by Beheshtian-Ardakani et al. (2018) who tried to use customer value data and the method most was found to be successful. Later studies by Lee, M. K., Verma, R., & Roth, A. (2017) showed that the method proposed by Chuang Wong was not sufficient as it didn't account to for the dynamic nature of customer value data which changes at almost every purchase and therefore isn't reflective of the customer. Lee et al. then proposed a method that makes use of transactional data and showed that it can reveal customer preferences and needs and after some period of time, it even reveals trends making it a more powerful input for clustering. The works of these scholars where adopted by Vijay at el., (2017) in their paper Variance Based Moving K-Means algorithm where they compared the different clustering methods with up to four random data sets. The details of experimentation along with quantitative as well as qualitative results including a comparison with seven state of the art algorithms namely K-Means (KM), Moving K-Means (MKM), Fuzzy C-Means (FCM), Adaptive Moving K- Means (AMKM), Enhanced Moving K-Means-1 (EMKM1), Enhanced Moving K-Means-2 (EMKM2) and K-Means clustering was presented ,K-Means and VMKM produced the best customer segmentation with similar lowest MSE(mean squared error).

Dullaghan & Rozaki, (2017) explained that logistics regressions, classification, clustering and decision tree are very successful data mining techniques for determining the customer segmentation and in their proposal, they use survival analysis and hazard function instead citing several weaknesses of logistic regression, classification and clustering, they further insisted that K-means performs better in customer segmentation compared with Fuzzy and other predictive models.

Peker et al., (2017) described the use of K-means to segment data set that is distinct or well separated from each. It concluded that K-means is much superior compared to

SOM when it comes to segmentation of data set that is distinct or well separated from each other.

Aluizio (2020) analyzed Self Organizing Maps and acknowledged that, it is not easy to implement better segments using SOM as expected since, information might be lost if high dimensional data is mapped onto an array that has a resolution too small to display the fine structures in the input data. On the other hand, according to Dullaghan (2017) to solve the issue of high dimensions for data segmentation, declared that SVC algorithm does not take care of dimensionality reduction for segmentation. These models were found to be weak and insufficient in segmentation.

Shabana et al., (2016) investigated K-Means algorithm for segmentation where, Principal component analysis and linear transformation were used for dimensionality reduction and initial centroid was computed, then K-Means was applied. It was found that K-means produced better segments compared to other clustering algorithm. Verster et al., (2017) in their paper a semi-supervised segmentation algorithm as applied to k-means using information value, accredited that computationally, clustering techniques are significantly more complex than k-means clustering. Perez et al., (2018) investigated the balancing efforts and benefits of k-means, ascribed that K-means is computing time reduction and solutions quality in terms of large and complex masses of data (Big data realms). Madhusmita Sahu et al., (2017) in their paper Parametric Comparison of K-means and other Clustering algorithms on Performance of behavior segmentation. K-means was found to be the best performing algorithm in comparison to others algorithms. Using K-means in segmentation, results showed that k-means is accurate in terms of performance Rakesh Kumar et al., (2016). R. Spurgen. Rathish et al., (2017) in their paper for the Comparison of Segmentation based on Threshold and K-means method, acknowledged that k-means is more accuracy, less running time and high resolution.

From the literature review, both K-means and SOM are similar in the sense that they calculate distances and minimize some error rate to impose good cluster quality.

However, with SOM accomplishment of better segments is not easy as information might be lost if a multifaceted set of data is mapped on to an array of too small resolution to exhibit the High resolution in the input data. K-means is limited to data in numeric form and can only be useful in cases where mean or median of a dataset can be calculated. In the case of categorical data, numerous techniques are available to convert categorical variables to numeric: e.g. single standardized target rates, using weights of evidence, optimal scores using correspondence analysis (Verster et al., 2017). Additionally, K-means can only find spherical clusters, therefore in the instance of non-spherical clusters PD-clustering can be applied on the data since it's a flexible method that can be used with non-spherical clusters, then k-means applied for better segmentation (Paul D. McNicholas et al.,2017)

To improve the performance of K-means algorithm, an improvement was made to the distance measure so as to improve the cluster symmetry. The method was found to be both simple and efficient and was in later literature like Verster et al. (2017), Ramirez-Ortiz et al., 2015 and Perez et al. (2018) referred to as the filter algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. As the number of clusters, k becomes greater, the performance of SOM algorithm becomes lower compared to K-mean algorithm in segmentation (Ramirez-Ortiz et al., (2015). On dimension reduction, the result provides new insights to the observed effectiveness of K-means, K-means algorithm handles the high-dimensionality of data during the iterative steps. Grounding on the nature of K-means to work with feature extraction techniques such as multiple correspondence analysis and principal component analysis, K-means outstands to be the best data mining technique when it comes to Segmentation (Verster et al., 2017).

As has been shown by researchers, (Ramirez-Ortiz et al., 2015; Perez et al., 2018) that PCA provides, it's the best low-dimensional approximation of the data technique. Also according to several studies (Peker et al., 2017), the results confirm the effectiveness of K-means for segmentation. K-means clustering improves segmentation performance in comparison to other clustering algorithms. Therefore, basing on the data structure and required results, K-means presents the most appropriate methodology for segmentation. Hence, this study will utilize this strength which is a missing gap in the literature review

2.7 Research Gaps

Globally wide range of research has been done concerning segmentation of customer. This is clear from research in different States in USA, like Bravo et al., (2017), (Calvet et al., 2016), Ahmadi & Weimar (2015) in Pakistan. In a comparative study of the k-means algorithm by Ramirez-Ortiz et al., (2015) they found the performance of SOM algorithm becomes lower compared to K-mean algorithm as the number of cluster k increased.

Verster et al., (2017) in their paper A semi-supervised segmentation algorithm as applied to k-means using information value, avowed that computationally, clustering techniques are significantly more complex than k-means clustering which is simplest and most commonly used clustering algorithm. However, irrespective of the customer data type approach used, the issue of high-dimensionality is an impediment in segmentation (Shabana et al., 2016)

Yang et al. (2016), to, showed that SVC does not take care of dimensionality reduction analyzed Self Organizing Maps and acknowledged that, it is not easy to implement better segments using SOM as expected since, information might be lost if a high dimensional data is mapped onto an array that has a resolution too small to display the fine structures in the input data., these models are weak and insufficient in segmentation. Madhusmita Sahu et al., (2017) confirmed that k-means gave out better segments using customer transaction data. Due to the strength of k-means to produce better segments, the flexibility of K-means to work with feature extraction techniques in handling dimensionality for data segmentation and the few studies which used this approach as evidenced by the latest research in the literature review, it emanates out clear that this is the gap the study intends to utilize.

CHAPTER 3

METHODOLOGY

3.1 Research Design

The proposed framework as shown in Figure 3 1,is driven by unsupervised algorithms specifically, k-means. Our methodological workflow consists of several stages: customers' transaction dataset acquisition from a repository, creating a target dataset i.e. selection of a subset of variables and data samples to be mined ,detection and removal of outliers, dimension reduction and transformation using the principal Component analysis, customer clustering, cluster validity analysis using silhouette index, comparison of the results from various algorithm, interpretation of the results and design distinguishable marketing strategies based on different behavior of their mobile subscribers to improve their marketing result and revenue.

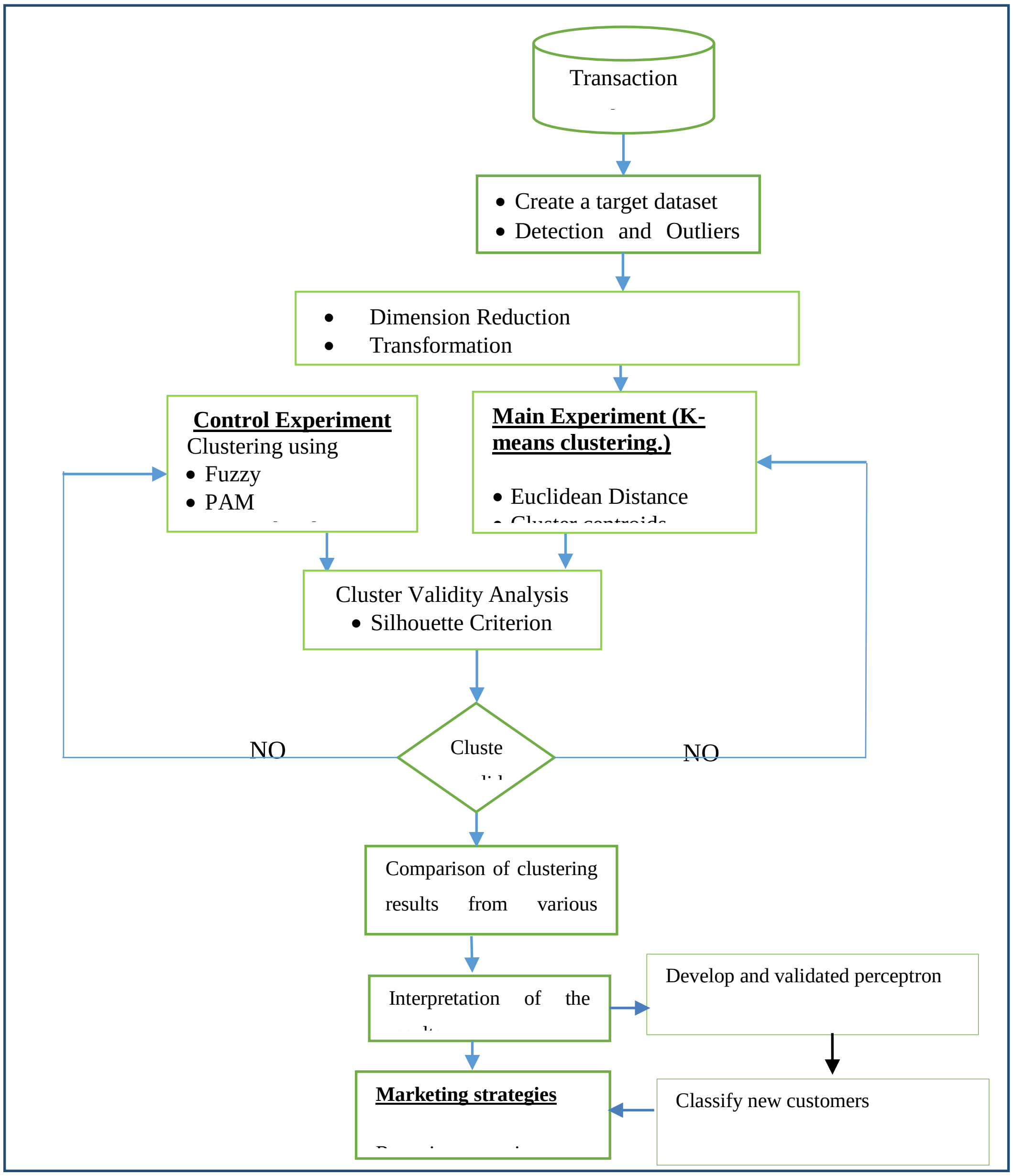


Figure 2.2: Flowchart for research framework

3.2 Sample Population

In this study, the entire group of persons with characteristic which was of interest to the study, intricate the data set which contained call record details which has 23 variables and 3333 subscribers.

3.3 Data source

Data relevant to the analysis was extracted from the machine learning repository. The data collected consisted of: -Demographic data: State, area code, phone number and transactional behavior data like day charges, evening charges, night charges, night calls, day calls, evening calls. Only variables that were relevant for segmentation were retained while the variables not relevant for clustering were disregarded.

3.4 Data Collection

Collection of data for this study was through the use of data mining techniques since the traditional methods of extracting knowledge and useful information from data are heavily dependent upon manual analysis and interpretation, which is often time-consuming, costly, and error-prone. The amount of data collected and stored in operational transaction systems and databases has been increasing at an unprecedented rate, traditional methods have become impractical for handling such amounts of data, and therefore, more sophisticated and affordable data mining tools need to emerge. Recent development of online payment systems have enabled transactions to be directly carried out on the internet. The transactions performed on the internet directly produce lots of data inform of logs which can be mined to yield insightful information. The use of data collected from the user via internet forms has become a trend for a number of online marketing companies. Several companies that sell products on the internet like amazon, Alibaba have directly adopted data mining so as tailor products to customers to searchers. Data mining techniques were more appropriate for this process of data collection in this research as compared to traditional since amount of data collected and stored in operational transaction systems and databases has been increasing at an unprecedented rate.

3.5 Perceptron Classifier

A perceptron model responsible for assigning new subscribers to the different clusters was created. A perceptron was used because of its robustness as its capable of learning from the current cluster allocations. An input data set consisting of total day charge, total evening charge, total night charge, and the total international charge were used against target clusters. The data set was partitioned into standard parts of 70% training data, 15% validation data, and 15% testing data. Testing data is critical for determining the generalization capability of the neural network. The different data sets for each group were selected randomly via random sampling. The random sampling helped to eliminate selection biases. One of the main challenges faced in creating perceptron is determining the number of neurons that yield the least error. A common method that has been adopted is the use of optimization method which involve plotting the number of neurons against the model accuracy in the testing data

CHAPTER FOUR

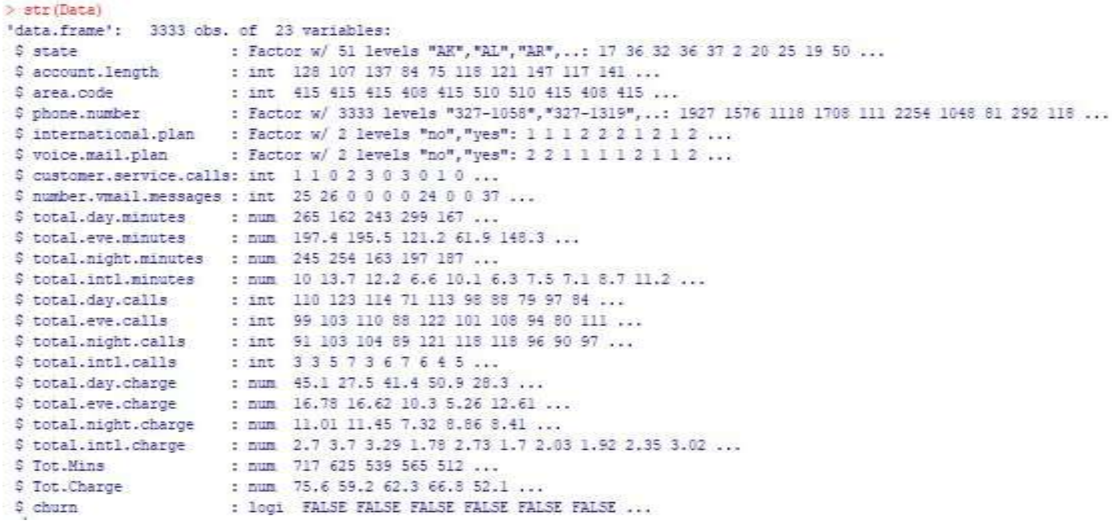
DATA ANALYSIS AND PRESENTATION OF RESULTS

4.1 Introduction

This chapter details the steps of how the methodology in chapter three was used to generate results. The results and finding of the experiments are also presented.

4.2 Data Description

Suitable dataset is crucial for any work as the dataset is a primary thing required for the experiment and results. The Data used for this experiment was obtained from the machine learning repository, Segmentation and Churn in telecom's dataset source (https://bigml.com/user/francisco/gallery/dataset/5163ad540c0b5e5b22000383). The data set has 23variables, 3333 instances and no missing data. Data pre-processing was done and data which wasn't useful for the purposes of segmentation were eliminated. A screen shot of the 23 variables characteristics is shown in (Figure 4). Attempts to segment mobile subscribers have previously been carried as has been discussed in the previous chapter (literature review). Two categories of variables have been used, those that depend on the number of call minutes like day call minutes, evening call minutes, international call minutes. The other forms of variables depend on the call charge. They directly linked to the revenues of a company. Clusters created from call cost data provide insight into the customer expenditures and can therefore be very important for dividing the customers into sub-groups like starter and premium customers. Five call cost data variables available in the current data set include, day, evening, night, international and total call cost charge. The five variables will therefore be adopted for adopted for this research



```
> str(Data)
'data.frame':   3333 obs. of  23 variables:
 $ state                 : Factor w/ 51 levels "AK","AL","AR",..: 17 36 32 36 37 2 20 25 19 50 ...
 $ account.length        : int  128 107 137 84 75 118 121 147 117 141 ...
 $ area.code             : int  415 415 415 408 415 510 510 415 408 415 ...
 $ phone.number          : Factor w/ 3333 levels "327-1058","327-1319",..: 1927 1576 1118 1708 111 2254 1048 81 292 118 ...
 $ international.plan    : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 1 2 ...
 $ voice.mail.plan       : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 1 2 ...
 $ customer.service.calls: int  1 1 0 2 3 0 3 0 1 0 ...
 $ number.vmail.messages : int  25 26 0 0 0 0 24 0 0 37 ...
 $ total.day.minutes     : num  265 162 243 299 167 ...
 $ total.eve.minutes     : num  197.4 195.5 121.2 61.9 148.3 ...
 $ total.night.minutes   : num  245 254 163 197 187 ...
 $ total.intl.minutes    : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
 $ total.day.calls       : int  110 123 114 71 113 98 88 79 97 84 ...
 $ total.eve.calls       : int  99 103 110 88 122 101 108 94 80 111 ...
 $ total.night.calls     : int  91 103 104 89 121 118 118 96 90 97 ...
 $ total.intl.calls      : int  3 3 5 7 3 6 7 6 4 5 ...
 $ total.day.charge      : num  45.1 27.5 41.4 50.9 28.3 ...
 $ total.eve.charge      : num  16.78 16.62 10.3 5.26 12.61 ...
 $ total.night.charge    : num  11.01 11.45 7.32 8.86 8.41 ...
 $ total.intl.charge     : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
 $ Tot.Mins              : num  717 625 539 565 512 ...
 $ Tot.Charge            : num  75.6 59.2 62.3 66.8 52.1 ...
 $ churn                 : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Figure 3.1: Variable attributes

In this chapter, factor analysis and feature extraction was conducted using Principal Component Analysis (PCA) so as to detect underlying data structures in order to present it in low dimensional space. The factors that led to most variability where identified using R. Normality check was done, so as to affirm whether the dataset approximates a normal distribution. To test how well the normal distribution fits the data, the Kernel density estimator is used to evaluate the probability of the various points in the data.. Optimal cluster selection is important in implementation of the k-mean algorithm and therefore a much emphasis was given to this stage. The elbow technique was used to estimate the appropriate cluster number. After finding the optimal value of k, the k-mean algorithm was applied to the transformed low dimensional data using customer transactional data across different times of day to generate groups around the calculated centroids. These centroids formed the basis of the new customer segments according to their transactional data and a heat map plotted to give a visual contrast of the dissimilarities among the generated clusters.

The Silhouette index was then used to validate and determine the goodness of clustering. Later the same data used to create new segments using different clustering algorithms (FUZZY, PAM, Hierarchical) in order to compare and determine the best algorithm.

4.2.1 Dimension reduction using Principal component analysis.

In this objective the data was examined to confirm if it was suitable for factor analysis before multidimensionality reduction, Kaiser-Meyer-Olin and Bartlett Test of Sphericity were used .For the dimensionality reduction, feature selection using Principal Component Analysis (PCA) was used as one of the factor analysis technique.

4.2.2 Kaiser-Meyer-Olin and Bartlett Test of Sphericity

Before conducting factor analysis, it is important to check if the data is suitable for a factor analysis, the most commonly used test in factor analysis is Kaiser-Meyer-Olkin (KMO) & Bartlett's Test of Sphericity It measures sampling adequacy by checking the case to variable ratio for the analysis being conducted. The KMO index ranges from 0 to 1, the accepted index is over 0.5. Bartlett's Test tests the hypothesis whether the correlation matrix is an identity matrix, which would indicate that the variables used are unrelated and therefore unsuitable for structure detection. Small values (less than 0.05) of the significance level indicate that a factor analysis may be useful with your data. The data used in this experiment had a score of 0.61 in the Kaiser Meyer Olkin Test meaning the case to variable sampling was adequate and the Bartlett's test significance level was below 0.00 meaning a factor analysis is suitable to unearth the data structure. The results of this test are as illustrated in the table 2:

Table 4.1: KMO and Bartlett test of sphericity

KMO and Bartlett's Test		
KMO Measure of Sampling Adequacy		.610
Bartlett's Test of Sphericity	Approx. Chi-Square	184015.071
	Df	36
	Sig.	.000

4.2.3 Feature selection using Principal Component Analysis (PCA)

Factor analysis is a techniques used to reduce a large number of variables into fewer numbers of factors. The most commonly used method of factoring is Principal Component Analysis (PCA). PCA extracts the variance sequentially and loads them to corresponding factors. According to Kaiser Criterion (Kaiser,1958) if the Eigen value is greater than 1 we consider that variable as a factor, if less than one the variable isn't considered a factor. The data with 9 variables was loaded into R to determine the principal components and the following results were extracted as shown in table 3.

Table 4.2: Feature selection using Principal Component Analysis

Component	Eigenvalues	% of Variance	Cumulative %
1	2.043	22.697	22.697
2	2.026	22.510	45.207
3	1.984	22.042	67.249
4	1.948	21.645	88.894
5	.999	11.105	100.000
6	7.255E-6	8.061E-5	100.000
7	7.847E-7	8.719E-6	100.000
8	2.237E-7	2.486E-6	100.000
9	4.775E-8	5.306E-7	100.000

As it can be seen in table 3, the Eigen values above one are only components 1 to 5, meaning these five components are the principal factors and cumulatively explain 100% of the variability in the data. The 9 original components have been successfully reduced to 5 components without losing information contained in the Data set. The data has therefore been transformed into a low dimensional space with five components

4.2.4 Distribution of data in the input variables

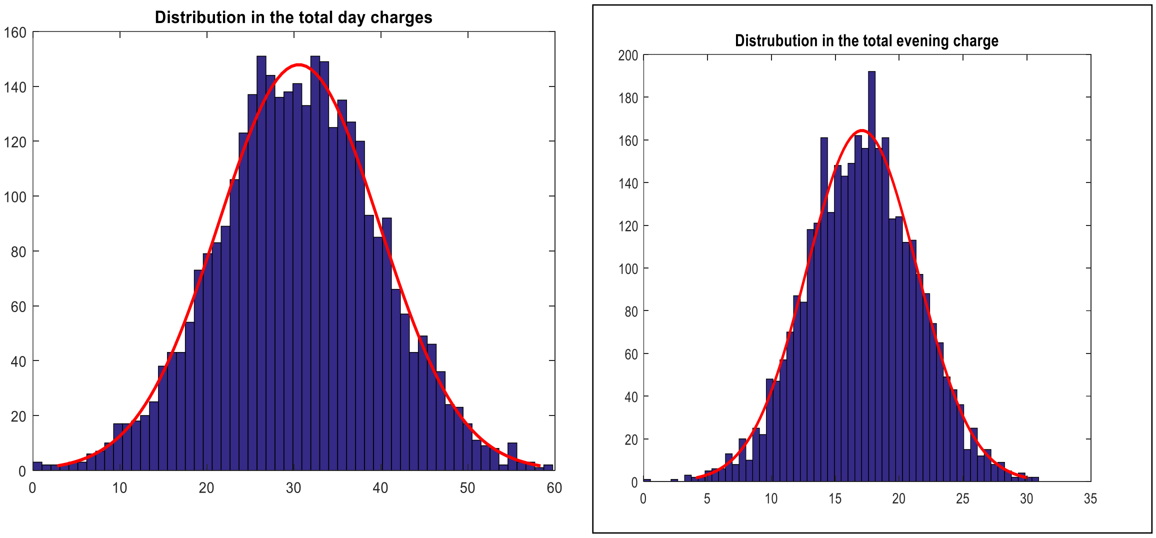


Figure 4.2: Normal distribution 1. Plot for total day charge and 2. Plot for total evening charges

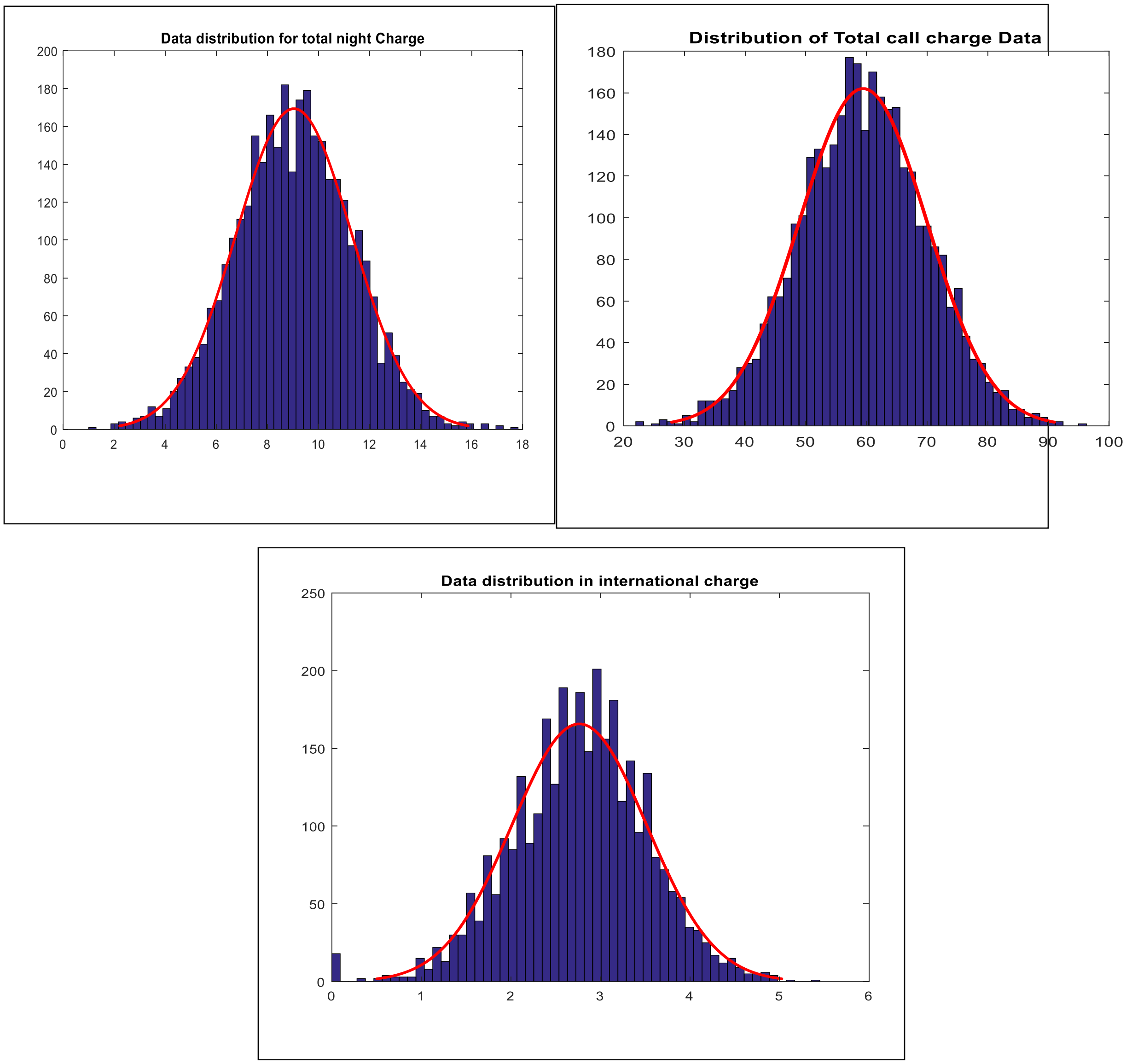
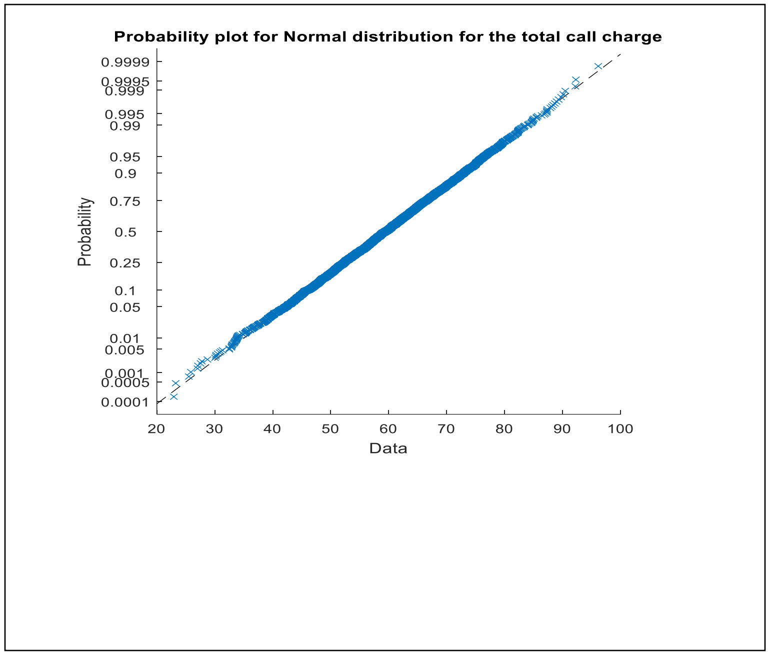


Figure 4.3: Normal distribution 1. Plot for total night charges, 2. Plot for total call charges and 3. Plot for total international charges

The plots 1-5 represent the distribution in the various input variables. As can be observed the data approximates a normal distribution. To test how well the normal distribution fits the data, the Kernel density estimator is used to evaluate the probability of the various points in the data. The degree to which each of the points approximates the normal distribution is then compared. The plots for the Kernel density are plotted in the plots (1-5) below. It’s a known statically fact that Natural phenomenon are expected to follow a normal distribution. The data points that out of the bounds become outliers of natural distribution. Whether to retain or keep outliers point depends on either subjective judgement or objective statistical schemes (Verma, R., & Roth, A. (2017). The subjective scheme involves a decision where the researcher studies the different outliers. The method is only effective for few number of outliers. The objective method involves using statistical tests like Dixon’s Q-tests, grabb t-test, Kdensity and the error margin test. All the tests depend on eliminating the values that deviate from the normal distribution. The Kdensity method was used in this research and figures 11-14 provide a visualization of the outliers. Eliminating the values that deviate from the normal distribution ensures the input variables follow a normal distribution thus ensuring that the data directly replicates natural phenomena. It’s also important the input variables follow the normal distribution because the validity of key statistical tests and classification techniques are dependent on the normality assumption.

4.2.5 Kernel density for normality check in the input variables



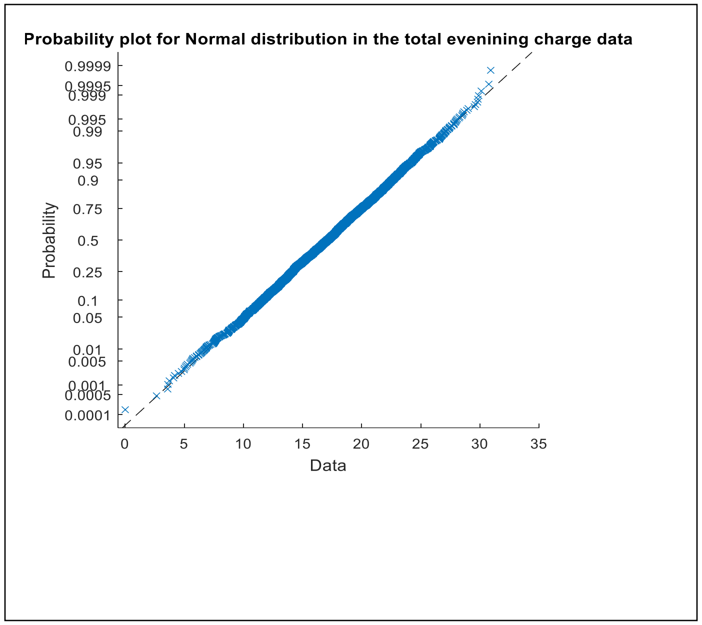


Figure 4.4: Kernel density for normality check, 1. Plot for total day charges and 2. Plot for total evening charges

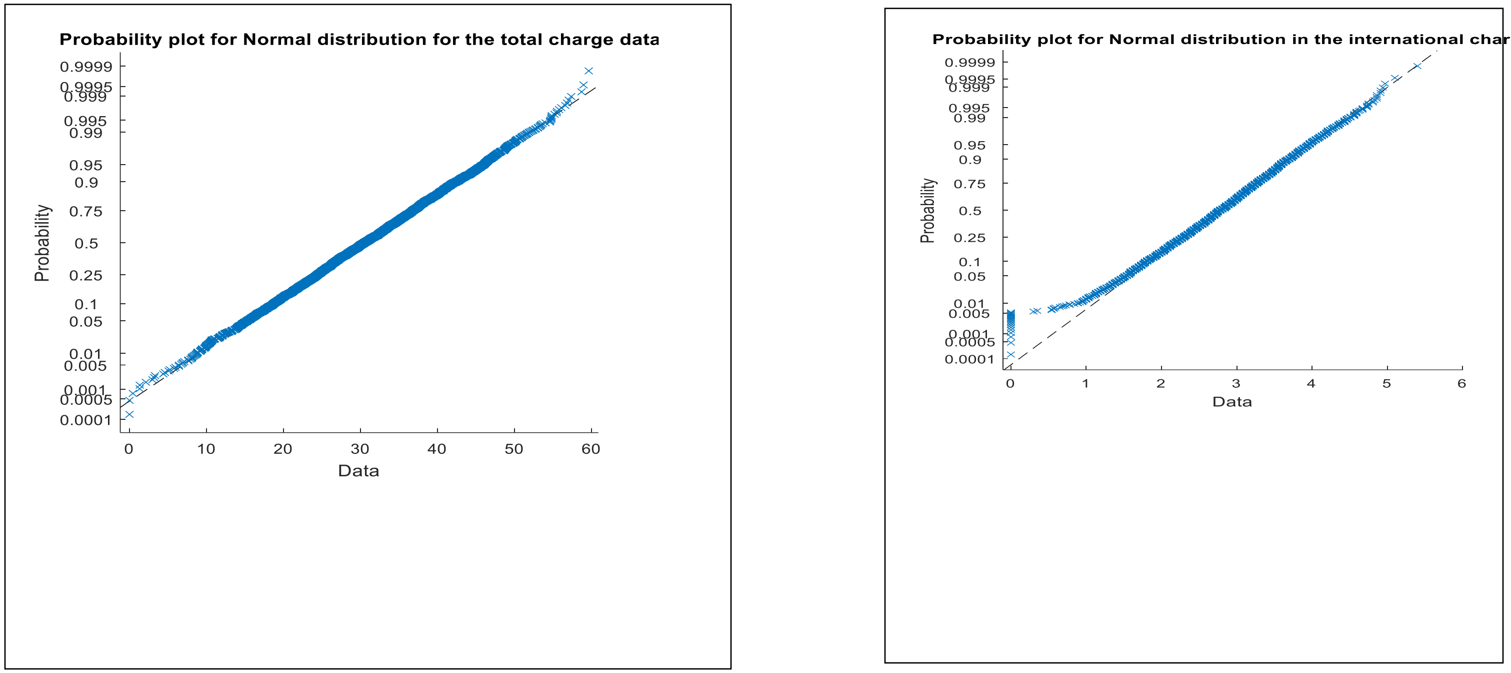


Figure 4.5: Kernel density for normality check, 1. Plot for total call charges and 2. Plot for total international charges

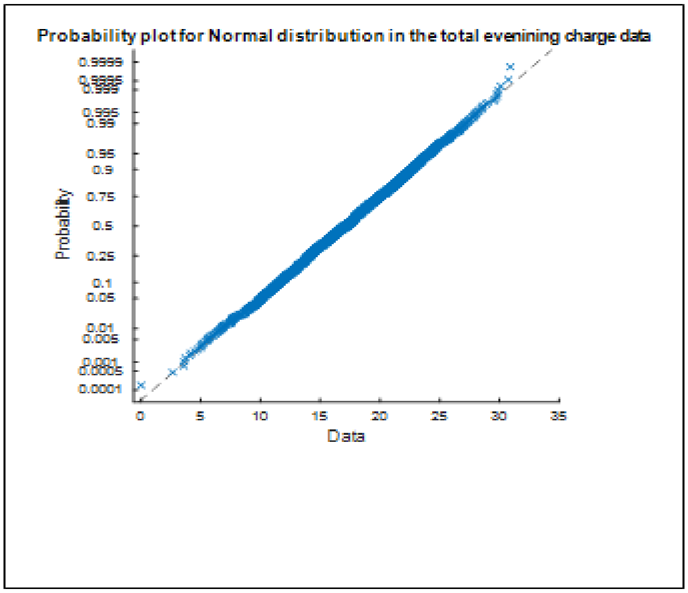


Figure 4.6: Kernel density for normality check, plot for total night charges

The kdensity plots 1-5 show how well the data fits the normal distribution. The data the deviates from the Normal distribution can be directly observed. To ensure that the Inherent variability within a given data set isn't eliminated, several researchers (Verma, R., & Roth, A. 2017; Dudoit et al., 2002; Tseng et al., 2005) recommended the use of an alpha of 0.05 so as to ensure data variability is maintained. An alpha factor of 0.05 will thus be used for this research. This will ensure that the input data sets used for building k-means clusters approximates a normal distribution at 95% confidence bounds. The alpha conditions ensures that the data satisfies the normality conditions within the bounds and also maintains variability within the same bound. The next section discusses the elimination of data points that deviate from the alpha 0.05 condition as outliers. Only the data that satisfies the normality condition has been studied as validity of the clustering techniques used within this research require the condition to be satisfied.

4.2.6 Identifying and Handling Outliers

An outlier is an extreme value or data that diverse drastically from a given average or mean of the dataset. A box and whisker plot is a graphical display summarizes data basing on five statistical quantities, the median of the data, the lower and upper quartiles (25% and 75%) and the minimum and maximum values. The Table summaries these statistics for the current dataset

Table 4.3: Variable summary statistics

	Tot.Charge	**total.day.charge**	**total.eve.charge**	**total.night.charge**	**total.intl.charge**
Min	22.93	0.00	0.00	1.40	0.00
1st Quartile	52.38	24.43	14.16	7.52	2.30
Median	59.47	30.50	17.12	9.05	2.78
3rd Quartile	66.48	36.79	20.00	10.59	3.20
Max	96.15	59.64	30.91	17.77	5.40

The box and whisker plot is an effective way to investigate the distribution of a set of data. For example, skewness can be identified from the box and whisker as the display does not make any assumptions about the distribution of the data.

4.3 Determine the value of parameter K (Number of clusters) using the Elbow criterion.

The elbow criterion determines shows how much of the variance in the data is explained as number of clusters is increased. The optimum number of cluster is established at the point where a further increase in the number of clusters doesn’t cause any significant improvement in the total variance than can be explained by the clusters. From the plot of total within variation and the number of clusters K, it can be observed that the first cluster. The significance of adding other clusters then begins very exponentially as an elbow plot. This was successfully achieved and the elbow identified at point k=4 (where the line graph cuts off) as illustrated in figure 16.

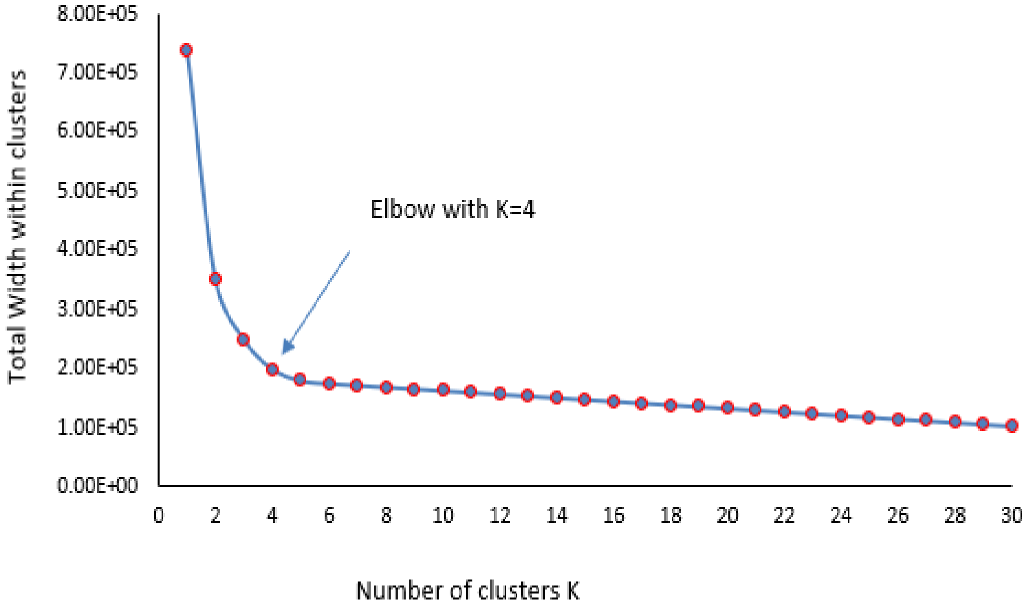


Figure 4.7: Elbow plot for the estimation of optimal clusters.

4.4 Using customer charges as inputs for segmentation.

In this objective, Customer charges were taken as input variables in k-means for segmentation. These include the subscriber call expenditure across different times of day.

4.4.1 Segmentation using k-means algorithm.

Using K-means algorithm customers are clustered according to the subscriber call expenditure across different times of day: - total day calls charge, evening call charges, night call charges, international calls and the Total cumulative charge as shown in table 5:

Cluster means:

Table 4.4: K-Means cluster centers

Cluster	Tot.Charges	total.day.charges	total.eve.charges	total.night.charge	total.intl.charge
4	**63.67109**	**34.21543**	**17.55374**	**9.132650**	**2.769274**
3	**54.10947**	**25.92775**	**16.43019**	**8.973573**	**2.777960**
2	**42.81664**	**16.29342**	**15.41868**	**8.400486**	**2.704062**
1	**74.75687**	**43.73540**	**18.75639**	**9.488785**	**2.776285**

The different charges for the 3333 subscribers in the experiment were plotted on a cluster plot as shown in the figure below.

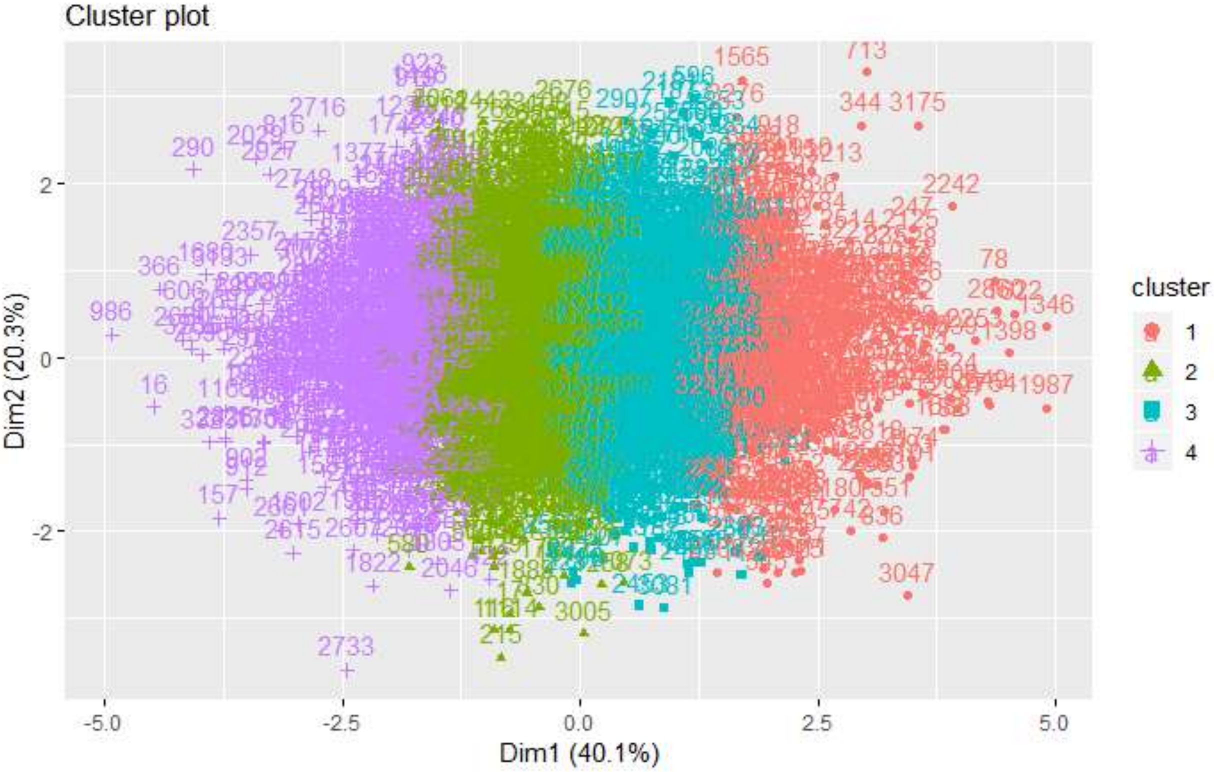


Figure 4.8: Cluster plot for the K-Mean algorithm

4.4.2 Using Customer charges as inputs in FCM Algorithm for segmentation

A control experiment was conducted on the same data set using a soft clustering algorithm (FCM algorithm) and the resultant cluster vector plotted on a scatter plot figure 8.As it is evident from the plot, the algorithm couldn't clearly separate points in cluster 2 indicated by color green and cluster 4 indicated by color purple on the scatter plot.

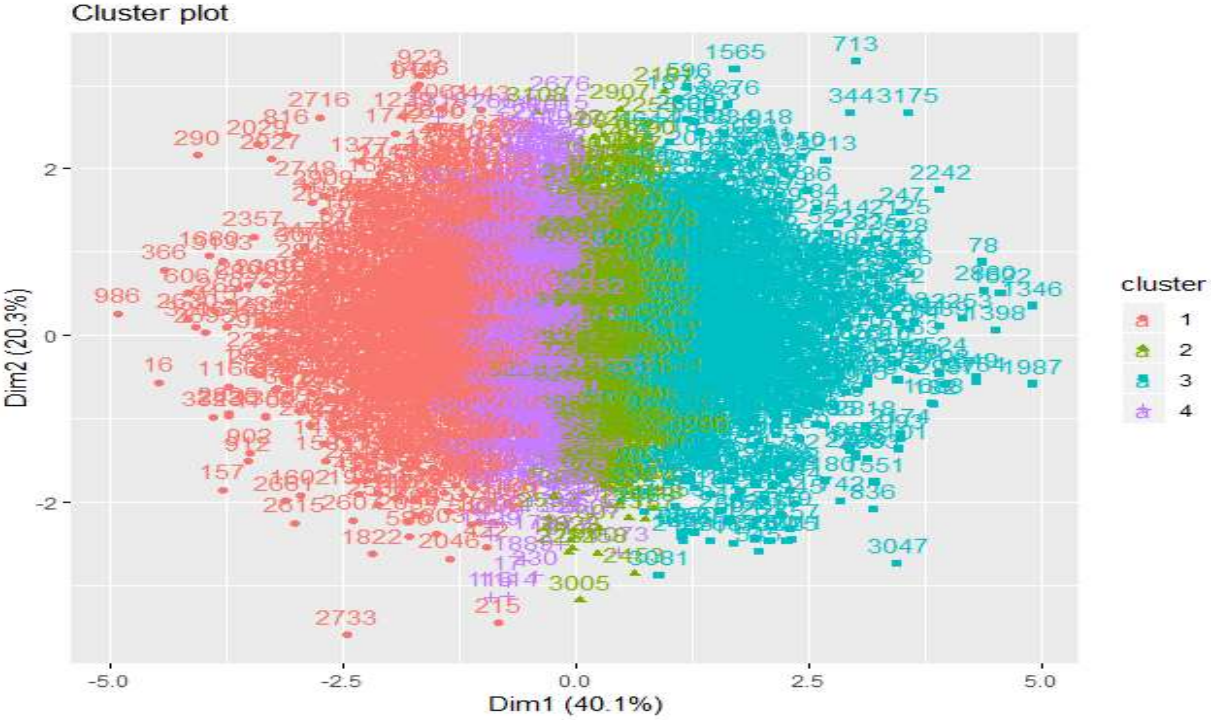


Figure 4.9: Cluster plot for FCM algorithm

4.4.3 Using Customer charges as inputs in PAM Algorithm for segmentation

PAM is an acronym for “partition around medoids”. The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters. It is similar to the K-Means algorithm but instead calculates medoids as the cluster center. The data was run through the algorithm and plotted on a scatter plot as shown below in figure 19.

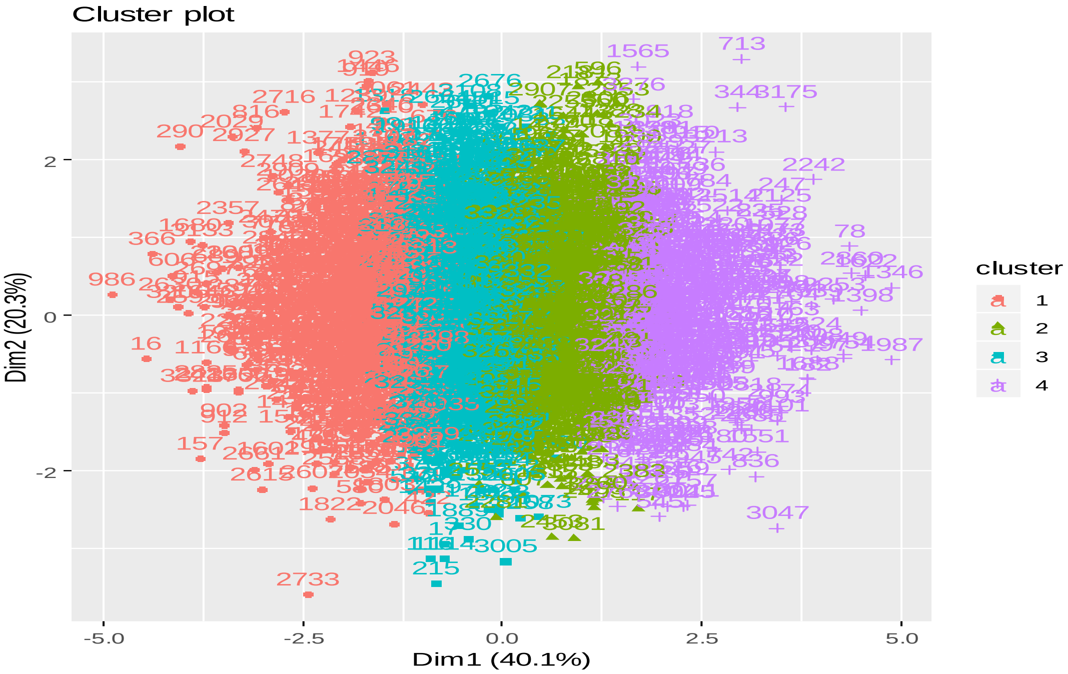


Figure 4.10: Cluster plot for PAM algorithm

4.4.4 Using Customer charges as inputs in Hierarchical Clustering Algorithm for segmentation

Hierarchical clustering is an algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters, where each cluster is distinct from another cluster, and the objects within each cluster are broadly similar to each other. The customer transactional charges were run through this algorithm and consequently used to come up with the dendrogram in figure 20 which shows the hierarchical relationship between the clusters:

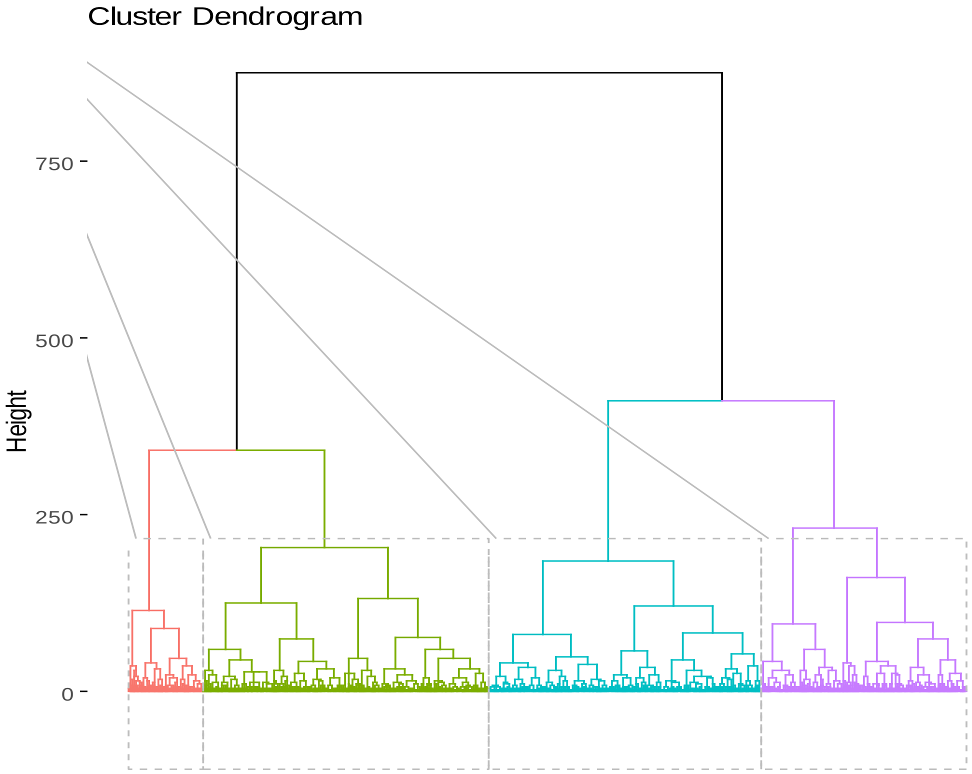


Figure 4.11: Cluster Dendrogram for Hierarchical algorithm

4.5 Development of peceptron

After the various clusters where formualted and forunalted in the data in the third objective, the need to model the clusters was imperative. The perceptron model was responsible for assignining the new customers to the different clusters. A perceptron was used because of it's robustances because its capable of learning from new data.The data set which consisting of the paramaters seleted like the total day charge, total evening charge, total night charge and total international charge where used as inputs for the perceptron while the clusters where used as targets. The data was partinioned into standard neural network partions of 70% training data, 15% training data and 15% testing data. The testing data is critical for determining the generalizatin capability of the neural network. The the different data sets for each group where selected randomly via a data random sampling command (Rand). The random sampling helps to eliminate slection biases. A percptron is classical two layer artificial neural network in it's design consisting of a hidden and output output layer. These models have been shown to have good generalization capabilities by different scholars as presented in literature. The main advantage of perceptrons over it's counter parts the multi layer network which in recent times constitute the deep learning neural networks is that it yields accurate models and at a lower computational cost. it's acceptable accuracy and lower computational cost has been responsbile for it's usage wide usage by a number of scholars (Mehmanpazir, and Asadi 2017). and now spans to include this research.

4.5.1 Determininination of neuron's in the Hidden layers

One of the main challenges faced in creating perceptron is determining the number of neurons that yield the least error. A common method that has been adopted is the use of optimization method which involves plotting the number of neurons aganist the model accuracy in the testing data. According to Maheshwari (2014), for a model that has 3-5 inputs, the optimum range of nuerons can be found between 2 and 15. As the number of nuerons is increased the accuracy of the model convergence to a point where a further increase in the number of neurons has no effect on the model accuracy. The curve for the model neuron vs. model accuracy in the testing data is presented in the figure 21

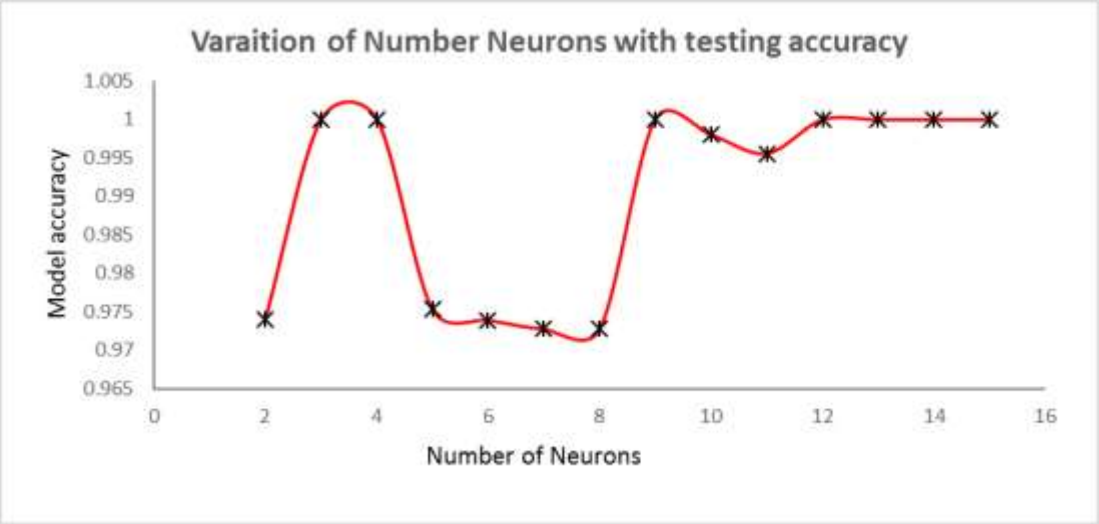


Figure 4.12: Variation of testing accuracy with number of neurons

Different models with different number of nuerons in the hidden layer were developed and there accuracy in terms of R-square measured. As expected, the model is consistent with the works of Maheshwari (2014) which predict saturation in the number of neurons beyound a given number of neurons. It's clear from the figure 20, that after 12 neurons a further increase in the number of neurons doesn't afftect the model accuracy signifying the onset of saturation in the number of neurons. Also from the figure, it can be observed that can be increasing the number of neurons leads to two maximum turning. The turning points directly concide with 4 and 9 nuerons. Even though both models yield highest accuarcies as measured by their R-square in the testing data, to determine the best model with good generalzation capbilities, another accuracy measurement like RMSE was used along side convergency characteristics of the different models.

The distribution of the errors in the different data sets is preseneted in the figure 22 for the model with 4 neurons. The errors can be seen to have a very small dispersion. As has been shown by Sheu et al., (2009) a low dispersion warrants that results of a given model are reproducable and also implies freedom of a model from systematic errors.

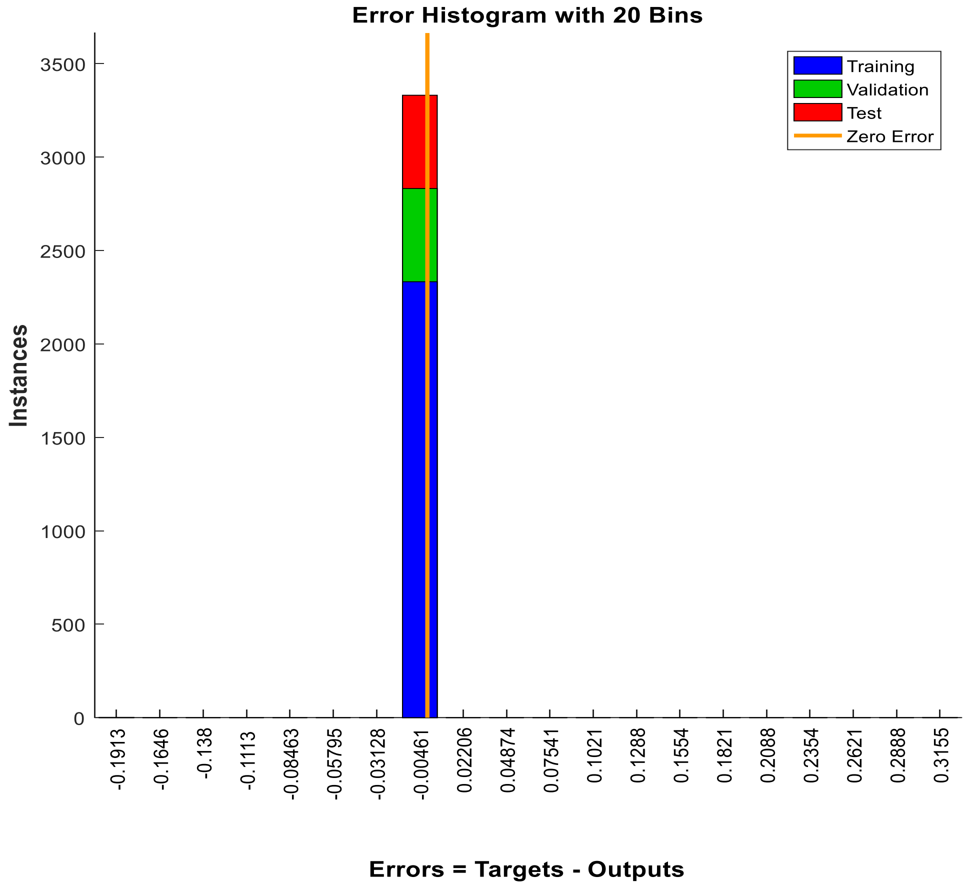


Figure 4.13: Error distribution in the data sets

However, for models that continously learns from data, it's important that the model not be overfitted so as it can yield consistent results even when new data is presented, the convergency characteristics of the model during training provide insight into this attribute. Models which have overfitting charatersitics can be observed from there convergency characteristics. Therefore even if the model with 4 neurons has an acceptable accuracy and good reproducability characteristic, it's has an overfitting characteristic. It can be observed from the figure 23 that the accuracy in the the different data sets improves at the same rate. A little past the 25th training epoch the error decay rate in the different data set begins to vary being fastest in the training data set. For the validation and testing data set, the error decay rate also increases but is faster in the later. However at the 34th epoch, the model error in the testing data begind to increase signifying the onset of overfitting in the model. The onset of overfitting of in model directly indicates limitations in the generalization capabilities as has been pointed out by o scholars Chuang and Wong (2013. The model with 9 was also evaluated on the same criterion and discussed in the sections that follow

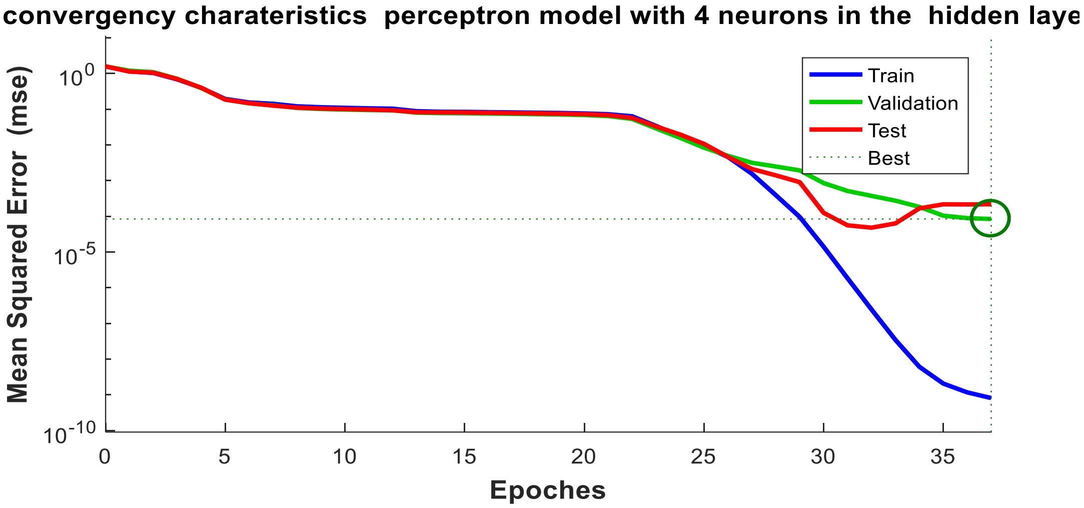


Figure 4.14: Convergence characteristic of 4-neuron perceptron

The second model that had a maximum peak accuracy was the model that had 9 neurons. The model also has errors with a very a low distribution as can be observed in the error histogram. The model also yields results which are reproducible. Even though the errors are equally perturbed about the mean, signifying existence of a degree of systematic errors, the error magnitude associated with the model is very small with a very small dispersion. Small error magnitudes and extremely low dispersion characteristics as observed in the figure 24 often have a negligible effect on the reproducibility and reliability of a given model. However, as has been highlighted by a number (Tseng et al., 2005) the extent of 'negligible' is subjective and can lead to improper model selection. A way round this subjectivity is presented in the works of Tiwari, 2007. According to

Tiwari, to eliminate subjectivity associated with reliability, the scholar recommends that the models in question be assessed on a basis of their MSE and convergence characteristics. Therefore to objectively select a model, the MSE and convergence characteristics where considered and discussed in the section that follow

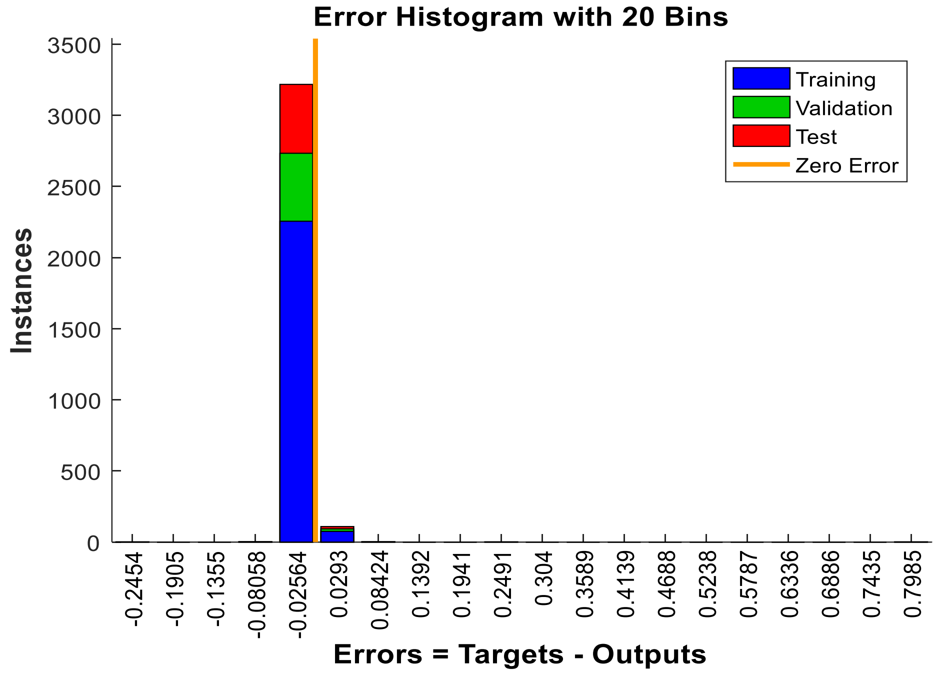


Figure 4.15: Error distribution in perceptron with 9 neurons

The MSE and Convergence are presented in the 25 below for the different data sets indicated by the color legend.

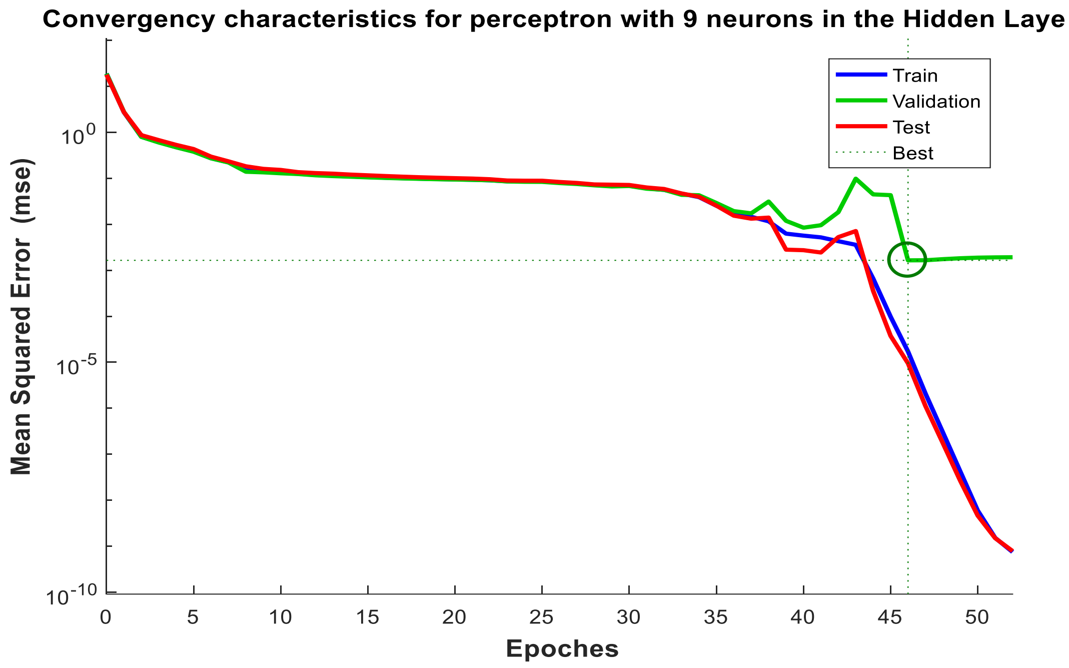


Figure 4.16: Convergence characteristic for 9-Neuron perceptron

The MSE in the all the data set increases as the number of training epochs increases. A break in the error reduction rate for all the data sets occurs slightly beyond the 35th epoch. The break in the error rate is characterized by an increment in both the validation and testing data set. In the case of the testing data set the rate begins to reduce after the 45th epoch. Beyond that point, the error decays at a faster rate than in the training data set and keeps reducing with the training data set. This signifies a much better generalization capability in the model. Therefore since both models are reliable and have a similar accuracy characteristics as discussed earlier, the generalization evaluation reveals that the model with 9 neurons has better generalization characteristics making it the better model of the two and will thus be adopted for this research

CHAPTER FIVE

DISCUSSION OF RESULTS

5.1 K-means clustering results.

The indices used to validate clusters measure the compactness, connectedness and separation of the clusters which is achieved via an internal index that yields the goodness of clustering. (Tseng et al., 2005), the results are evaluated using quantities and features inherent in the data set. The most commonly used indices for cluster validation are the Davies Bouldin index and the Silhouette index, in this experiment Silhouette index was used as it is the most commonly used index for unsupervised classification (Starczewski A., 2015). This validation technique measures goodness of the clusters by calculating the average distance between clusters, therefore the higher the Silhouette index the better the clustering (Rousseeuw, 1987).The Silhouette analysis for the K-mean algorithm are shown in the table 6.

Table 5.1: Silhouette results for K-Means Algorithm

Cluster	Cluster size	ave.sil.width
1	**1142**	**0.30**
2	**453**	**0.34**
3	**568**	**0.34**
4	**1170**	**0.31**

The information in table was plotted using a silhouette plot (graph 1) to get a visual representation of the clustering. The average silhouette index *(Si)* was 0.31 meaning the average distance between clusters was 0.31.

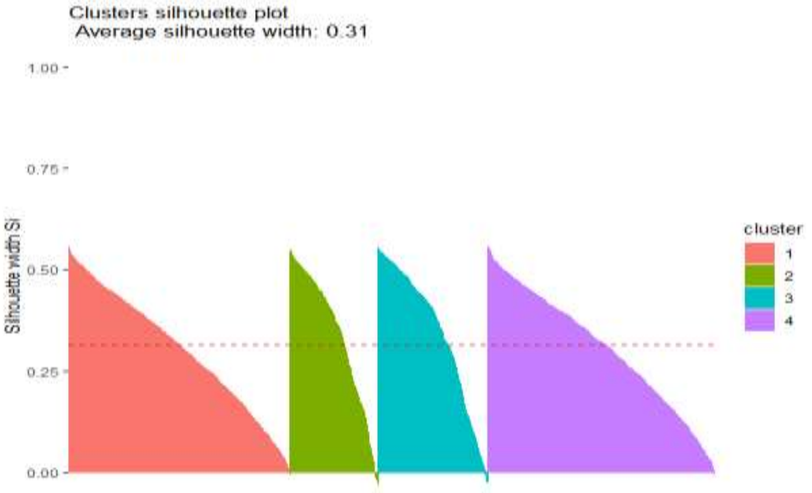


Figure 5.1: Silhouette plot for K-Means algorithm

The plot for the different silhouette index is important as it helps to identify points which have been wrongly clustered. The silhouette index of wrongly clustered points is negative and can therefore be easily visualized. An analysis using the R cluster package was done to identify the wrongly clustered points. This was used in coming up with a clustering error percentage for the algorithm using the following computation:-

Error% = (Total Number of Points with –ve Si/ Total number of Points Clustered) x 100

There were 34 points with –ve silhouette width out of a Total of 3333 clustered points, giving a clustering error rate of 1.02% for the K-Means algorithm.

5.2 Validation of the (FUZZY) FCM-algorithm clustering results.

The clustering of the FCM (FUZZY) algorithm were validated in order to make a comparison of the clustering abilities between the hard-clustering K-means algorithm and the soft clustering FCM algorithm. The average silhouette index (*Si)* for the 4 clusters was 0.22 as shown in the table 7, meaning the average distance between clusters of the FCM algorithm was 0.22.

Table 5.2: Silhouette results for Fuzzy algorithm

Cluster	Cluster size	ave.sil. width
1	**915**	**0.20**
2	**716**	**0.27**
3	**931**	**0.18**
4	**771**	**0.26**

A silhouette plot was developed using the values of table 7 in order to know how many points were wrongly clustered as it was done for the K-Mean algorithm. The R cluster package was used to identify the wrongly clustered points in order to calculate the Error rate.

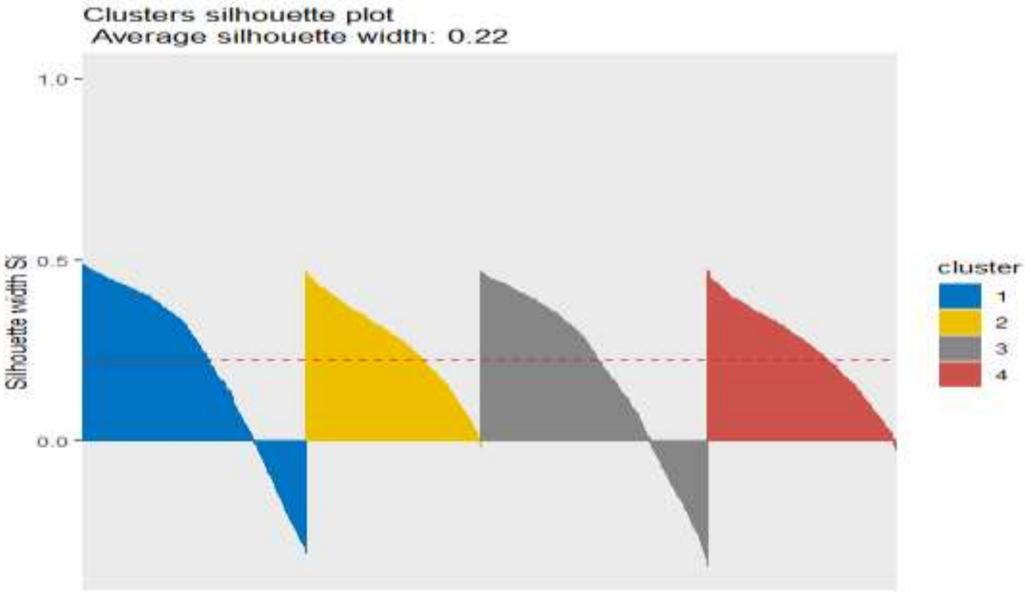


Figure 5.2: Silhouette plot for FCM algorithm

As it can be seen from the Silhouette plot, a number of points in cluster 1 and 3 had a negative Silhouette index. The clustering error rate for the FCM algorithm was calculated using the same formula as used in the K-Mean algorithm:

Error% = (Total Number of Points with –ve Si/ Total number of Points Clustered)

x 100

There was a total of 465 points with a –ve Silhouette width out of the 3333 points clustered giving an error rate of 13.951% for the FCM algorithm.

5.3 Validation of the PAM-algorithm clustering results.

The clustering of the PAM algorithm average silhouette index (*Si)* for the 4 clusters was 0.30 as shown in the table 8, meaning the average distance between clusters of the PAM algorithm was 0.30.

Table 5.3: Silhouette results for PAM algorithm

Cluster	**Cluster size**	**ave.sil. width**
1	**726**	**0.29**
2	**1033**	**0.29**
3	**1072**	**0.31**
4	**502**	**0.29**

A silhouette plot was developed using the values of table 8 in order to know how many points were wrongly clustered as it was done for the K-Mean algorithm.

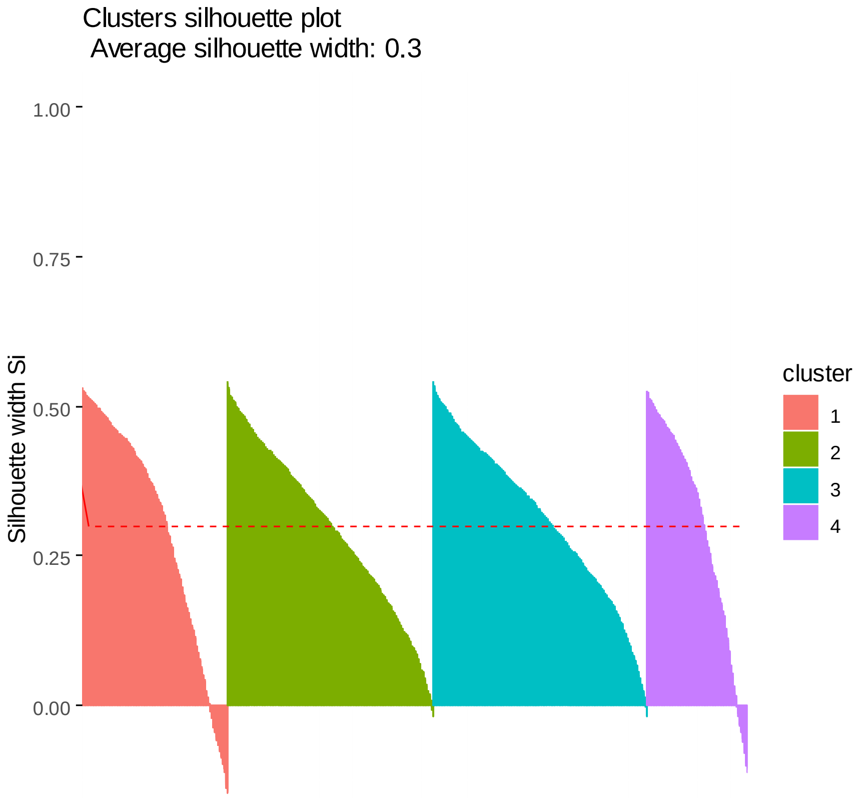


Figure 5.3 : Silhouette plot for PAM algorithm

As it can be seen from the Silhouette plot, a number of points in cluster 1 and 4 had a few points with negative Silhouette index. The clustering error rate for the PAM algorithm was calculated using the same formula as used in the K-Mean algorithm:

Error% = (Total Number of Points with –ve Si/ Total number of Points Clustered) x 100

There was a total of 154 points with a –ve Silhouette width out of the 3333 points clustered giving an error rate of 4.621% for the PAM algorithm.

5.4 Validation of the Hierarchical-algorithm clustering results.

The clustering of the hierarchical algorithm average silhouette index (*Si)* for the 4 clusters was 0.27 as shown in the table 9 , meaning the average distance between clusters of the algorithm was 0.27

Table 5.4: Silhouette results for Hierarchical algorithm

Cluster	Cluster size	ave.sil. width
1	**814**	**0.23**
2	**1142**	**0.24**
3	**1084**	**0.31**
4	**293**	**0.41**

A silhouette plot was developed using the values of table 9 in order to know how many points were wrongly clustered as it was done for the K-Mean algorithm.

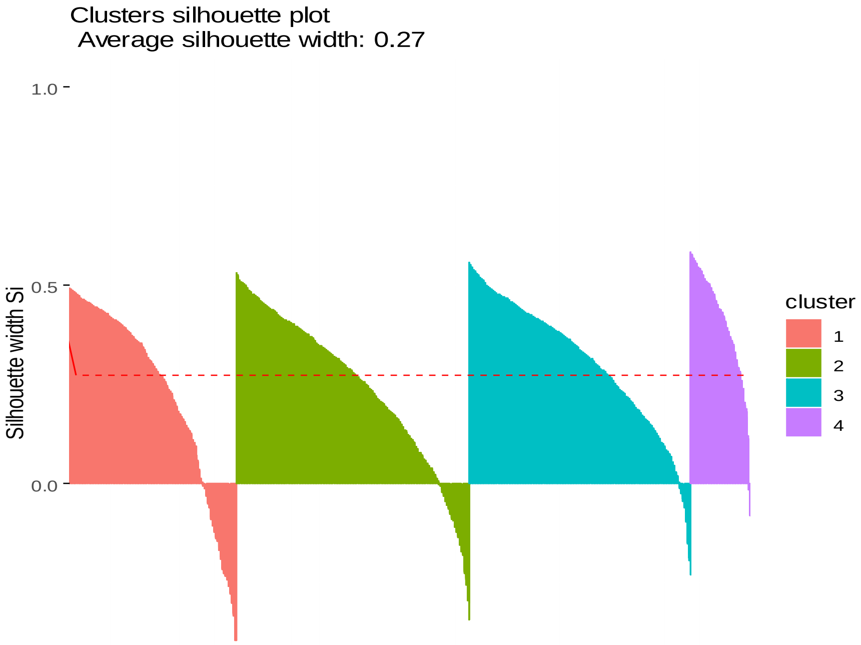


Figure 5.4: Silhouette plot for Hierarchical algorithm

As it can be seen from the Silhouette plot, a number of points in cluster 1, 2 and 3 had points with negative Silhouette index. The clustering error rate for the Hierarchical algorithm was calculated using the same formula as used in the K-Mean algorithm:

Error% = (Total Number of Points with –ve Si/ Total number of Points Clustered) x 100

There was a total of 366 points with a –ve Silhouette width out of the 3333 points clustered giving an error rate of 10.981% for the hierarchical algorithm.

5.5 Comparison of clustering results.

A comparison was done to compare the clustering quality of the four algorithms. As shown in table 10, K-Means algorithm outperforms all the other algorithms as evidenced by a higher silhouette index meaning the clusters were better partitioned and a very low Error rate (1%) meaning the algorithm performed exceptionally well in clustering the different points into their respective clusters.

The Algorithms were later ranked with the best being the algorithm with a high Silhouette index (highly spaced clusters) and low Error rate (low clustering error).

Table 5.5: Comparison of clustering results

Rank	Algorithm	Number of Clusters	Silhouette Index *(Si)*	Error Rate
1	K- Means	4	0.31	1.09%
2	PAM	4	0.30	4.62%
3	Hierarchical	4	0.27	10.98%
4	FCM	4	0.22	13.95%

5.6 Model performance Evaluation

After building the perceptron classifiers, its ability to allocate the data to the correct K-means cluster data set was evaluated. The K-means clusters generated by the perceptron where then validated by comparing them to the actual k-means clusters. The validity of the model was measured by obtaining the correlation between the two actual and predicated data set as measured by R-square, Root mean square error (RMSE) and Sum of square Error (SSE).

The testing data set always comprises of unseen data sets, were unseen data set consists of data not previously used for training or validating the model. The testing set for this model was made of 500 unseen samples and were used to simulate the model the classification performance of the model for each sample is presented in the Graph 5

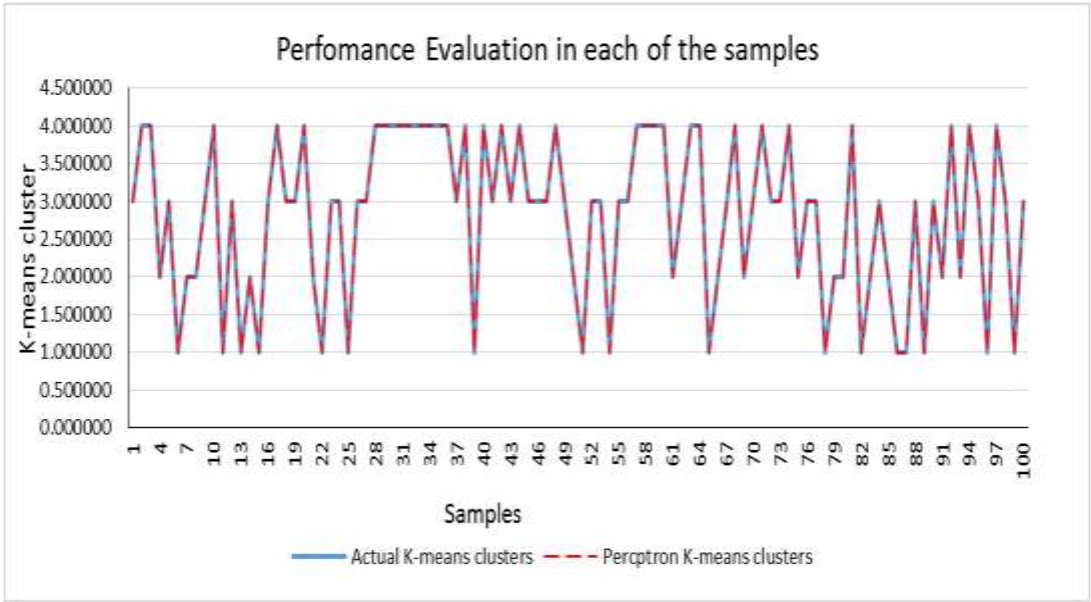


Figure 5.5: Performance evaluation in each sample for testing data set

The performance evaluation in individual data set shows that the perceptron accurately classifies input data sets into the right data set. The blue solid line represents the actual k-means generated from the previous objective. The red dotted line represents the k-means clusters from the perceptron. The direct overlay of the lines indicates a strong correlation between the results. This is important for quantifying each new customer into a given consumer group.

5.7 Model Validation

The accuracy of the model was measured by obtaining the fit between actual and predicated data sets. The graph 6 shows the correlation in the testing data set while the graph 7 shows the correlation for all data sets. The table 2 summarizes the accuracy measurements for the different model.

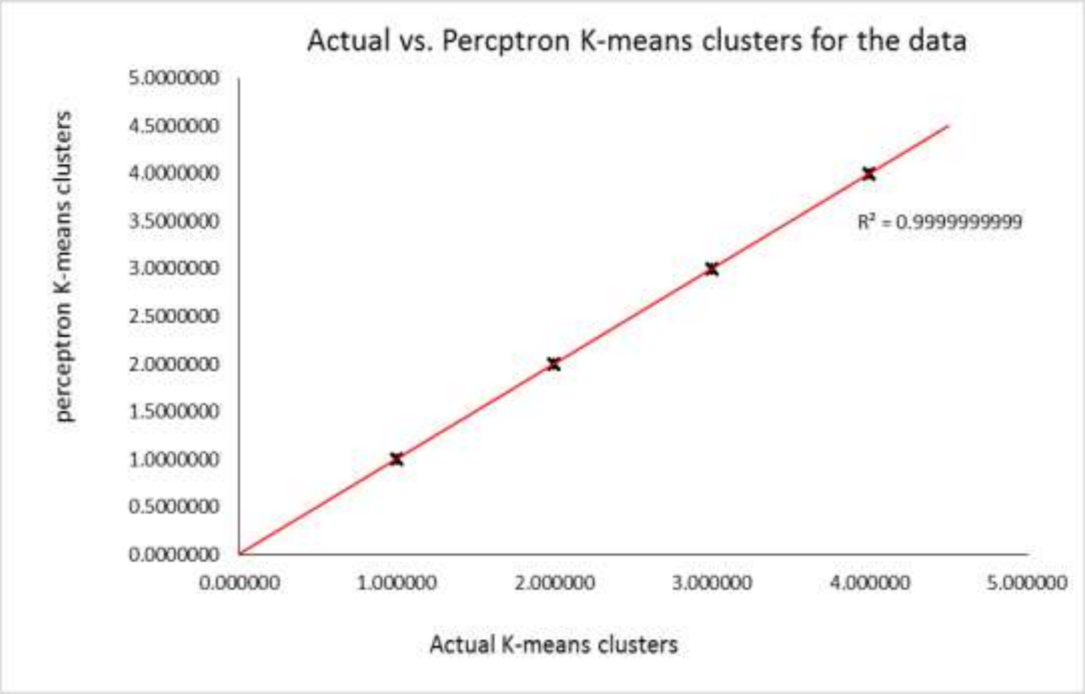


Figure 5.6: goodness of fit between perceptron classifier and actual clusters in the training data set

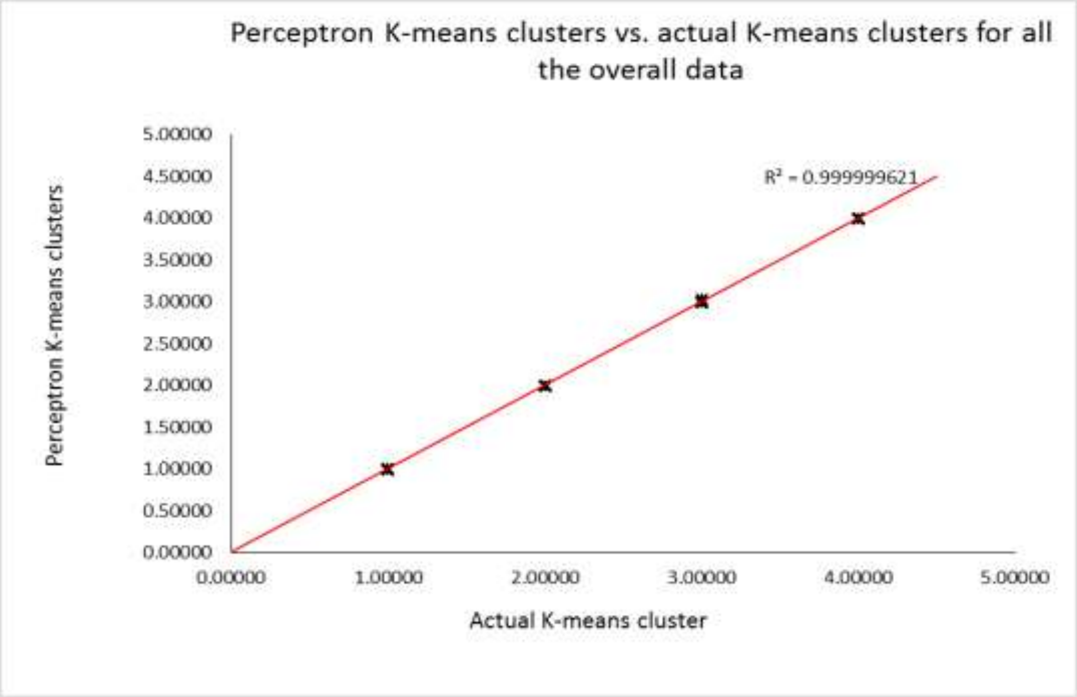


Figure 5.1: Goodness fit in all data

Table 5.6: summary of Goodness of fit

	R-square	**RMSE**	**SSE**
Testing data Set	0.99999999	0.01810	0.1638
Overall data set	0.99999621	0.01813	1.0947

It can be observed from the table that basing on the testing data which was composed of unseen samples, the model has very strong generalization characterization. The different accuracy measures affirm that the k-means clusters were successfully modeled by the perceptron in both the testing data and the overall data set that consists of training, validation and testing data. At 0.9999999 percent the error margin can be attributed to random errors which is consistent with histograms discussed on in the previous sections. The RMSE very close to zero also indicate a high accuracy in each of the individual data sets that where tested.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Summary of the Main Findings

The first objective was achieved through reduction of the nine transactional variables into five variables without losing variability in the data. The second objective required demanded that the optimum clusters in the data be determined. The Objective was achieved by using the elbow criterion where the total within sum of squares was plotted against varying number of clusters between 2 and 30, the elbow point was achieved at the point where k = 4 therefore providing the optimal number of clusters for the algorithm. A desktop review of the call packages available in the local Kenyan market confirmed the existence of 4 kinds of customers on the Kenyan Market. The four kinds of customer segments is consistent with the optimum segments established by this research. The current Kenyan market has 4 main packages for the different Customer packages, a starter client package, Standard client package, Super client package and Premium client's package.

The third objective required the determination of the best clustering methods. For this research, different top clustering methodologies were developed and ranked and there accuracy for optimal number of clusters determined from objective two basing on the Silhouette indices compared. Silhouette indices the accuracy of a clustering algorithm by comparing the cluster separation width and cluster misclassification error rate. For the four clusters, results showed that K-means yielded the least clustering error of 1.09%, compared to 4.62% in the PAM, 10.98% in the hierarchal and 13.95% for the FCM

To get a clear picture of the contrast in the difference existent between the different clusters a heat map figure 10 was plotted using the values of the call charges across different times of day. The colour values range from dark green (for low values) to red (for high values).

It is also clear to the decision makers that there is no major difference across the 4 segments in charges for calls made in the evening, night time and International calls. This information can be useful during product development by coming up with tariffs and offers that would encourage more people to make calls during this periods. A major difference is evident when analyzing total day charges as there are four different groups. The telephone company can utilize this information to know the core business operation hours therefore focus on maximizing profits and customer experience during these hours and being cognizant of the fact that subscribers are most sensitive to price changes during these hours.

As a fourth and final objective, a perceptron model was designed and validated to allocate new clients to the different K-means clusters determined from the means. Much as clustering by itself provides insight into the behavior of the different clients, it's important that new clients are also assigned to a given segment. The model was successful developed with a goodness of fit R-square 0.9999999, RMSE=0.01810, SSE=0.1638 in the testing Data and R-Square of 0.99999621, RMSE=0.01813, SSE=1.0947 in all the data.

6.2 Recommendation for Future Work

Due to the dynamic nature of data and consumer preferences which change rapidly, the use of historic data as used in this research may be outdated. To keep abreast with the quick changes, In future we recommend other researchers to use web and mobile based systems using the Online Analytical Processing capabilities of R inknitr and shiny apps packages provided by R studio to cluster and visualize the status in real-time thereby responding appropriately to the constantly changing consumer preferences.

6.3 Conclusion

The main goal of this research to segment customers using k-means algorithms and use a perceptron to segment new customers. To achieve the main goal, four objectives where executed. Firstly, the optimum number of variables to be used as inputs for data segmentation was established via dimension reduction using principal Component Analysis objective was achieved through reduction of the nine transactional variables into five variables without losing variability

The second objective involved determining the optimum number of using an elbow criterion. For this research the optimum number of clusters was found to four.

In a third objective, the current customer data was classified into the developed clusters using the k-means algorithm. The new k-Means algorithm was then compared to previously used methods like FCM and the method of K-means was found to yield better results.

In a final objective, a perceptron model was developed. The model was then used to classify new customers basing in the clusters developed in the chapter three. The performance of the perceptron to classify new customers was then evaluated. The perceptron was found to classify new customers accurately.

Through clustering, hidden relationships among variables that were not obvious to researchers were identified hence contributing knowledge in the field of unsupervised machine learning which would serve as a preliminary point for future research. K-means offered better results and can lead to an improvement in enhancing customer experience therefore reducing customer churn to competitors and consequently drive profitability in the firm. This research in cluster analysis has demonstrated how researchers can combine multidimensionality reduction and clustering algorithms to explore data and reveal the underlying structure of objects hence leveraging on the power of data analytics to drive growth and efficiency in business.

REFERENCES

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1).

Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. *Computers and Industrial Engineering*, *144*(April), 106476.

Alshammari, M. T. (2019). Editorial Preface From the Desk of Managing Editor… Associate Editors. *IJACSA - International Journal of Advanced Computer Science and Applications*, *10*(9), 204–208. Retrieved from http://www.ijacsa.thesai.orgwww.ijacsa.thesai.orgwww.ijacsa.thesai.org

Aluizio F.R. Araujo, Victor O. Antonino, K. L. P. G. (2020). Self-organizing subspace clustering for high-dimensional and multi-view data. *Neural Networks*, 109231.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer Churn Prediction in Telecommunication Industry under Uncertain Situation. *Journal of Business Research*, *94*, 290–301. Retrieved from http://repository.uwl.ac.uk/ id/eprint/4800/1/Customer churn prediction.pdf

Arora, D., & Malik, P. (2015). Analytics: Key to go from generating big data to deriving business value. *Proceedings - 2015 IEEE 1st International Conference on Big Data Computing Service and Applications, BigDataService 2015*.

Beheshtian-Ardakani, A., Fathian, M., & Gholamian, M. (2018). A novel model for product bundling and direct marketing in e-commerce based on market segmentation. *Decision Science Letters*, *7*(1), 39–54.

Calvet, L., Ferrer, A., Gomes, M. I., Juan, A. A., & Masip, D. (2016). Combining statistical learning with metaheuristics for the Multi-Depot Vehicle Routing Problem with market segmentation. *Computers and Industrial Engineering*, *94*, 93–104.

Chen, H., Zhang, L., Chu, X., & Yan, B. (2019). Smartphone customer segmentation based on the usage pattern. *Advanced Engineering Informatics*.

De Keyser, A., Lemon, K. N., Klaus, P., & Keiningham, T. L. (2015). A Framework for Understanding and Managing the Customer Experience. *Marketing Science Institute Working Paper Series 2015, Report No. 15-121, Forthcoming*.

Deligiannis, A., Argyriou, C., & Kourtesis, D. (2019). Predictive Personalization of Conversational Customer Communications with Data Protection by Design. *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume*, 305–308.

Dullaghan, C., & Rozaki, E. (2017). Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers. *International Journal of Data Mining & Knowledge Management Process*, *7*(1), 13–24.

Feizabadi, J., & Shrivastava, A. (2018). Does AI-enabled demand forecasting improve supply chain efficiency? *Innovation Strategies*.

Femina, B. T., & Sudheep, E. M. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*(1), 236–247.

Greff, K., Rasmus, A., Berglund, M., Hao, T. H., Schmidhuber, J., & Valpola, H. (2016). Tagger: Deep unsupervised perceptual grouping. *Advances in Neural Information Processing Systems*.

Kanyuga, L. (2019). Influence of Strategic Innovation on Performance of Telecommunication Firms : A case of Safaricom Company Lillian Kanyuga Influence of Strategic Innovation on Performance of Telecommunication Firms : A case of Safaricom Company. *Journal of Strategic Managment*, *3*(1), 21–39.

Lee, M. K., Verma, R., & Roth, A. (2015). Understanding customer value in technology-enabled services: A numerical taxonomy based on usage and utility. *Service Science*, *7*(3), 227–248.

Liu, Y.-C., & Chen, Y.-L. (2017). Customer Clustering Based on Customer Purchasing Sequence Data. *International Journal of Engineering Research and Applications*, *07*(01), 49–58.

Maruotti, A., Bulla, J., & Mark, T. (2019). Assessing the influence of marketing activities on customer behaviors: a dynamic clustering approach. *Metron*, *77*(1), 19–42.

Mudogo, e. K. (2019). *Technological innovation and performance of telecommunication companies in Kenya*. University of Nairobi.

Namvar, A., Ghazanfari, M., & Naderpour, M. (2017). A customer segmentation framework for targeted marketing in telecommunication. *Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2017*, *2018-Janua*, 1–6.

Ngugi, I. K., & Komo, L. W. (2017). Case study 9: M-Pesa: A renowned disruptive innovation from Kenya. In *Strategic Marketing Cases in Emerging Markets* (pp. 117–128). Springer International Publishing.

Nisbet, R., Miner, G., & Yale, K. (2017). Handbook of statistical analysis and data mining applications. In *Handbook of Statistical Analysis and Data Mining Applications*.

Omamo, A. O., Rodriguez, A. J., & Muliaro, J. W. (2018). A Systems Dynamics Model for Mobile Industry Governance in the Context of the Kenyan Vision 2030. *International Journal of System Dynamics Applications*.

Osman, A. S. (2019). Data mining techniques. *International Journal of Data Science Researcg*, *25*(2), 545.

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence and Planning*.

Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes*.

Shabana, K. M., Wilson, J., & Chaudhury, S. (2016). A Multi-view Non-parametric Clustering Approach to Mobile Subscriber Segmentation. *Proceedings - CBI 2016: 18th IEEE Conference on Business Informatics*, *1*, 173–181.

Suryadi, D., & Kim, H. M. (2019). A Data-Driven Methodology to Construct Customer Choice Sets Using Online Data and Customer Reviews. *Journal of Mechanical Design*, *141*(11), 1–12.

Verma, S. (2017). Big data and advance analytics: Architecture, techniques, applications, and challenges. *International Journal of Business Analytics*, *4*(4), 21–47.

Wang, S., Wang, D., Li, C., Li, Y., & Ding, G. (2016). Clustering by fast search and find of density peaks with data field. *Chinese Journal of Electronics*, *25*(3), 397–402.

Wang, S., & Zhang, X. (2020). Analysis of Self-Organizing Maps (SOM) Methods for Cell Clustering with High-Dimensional OAM Collected Data. *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, 229–233.

Wang, X., & Huang, J. Z. (2015). Editorial: Uncertainty in learning from big data. *Fuzzy Sets and Systems*.

Yang, J., Liu, C., Teng, M., Liao, M., & Xiong, H. (2016). Buyer targeting optimization: A unified customer segmentation perspective. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, 1262–1271.

Zheng, Y. (2015). Trajectory data mining: an overview. ACM Transactions on Intelligent Systems and Technology (TIST), 6(3), 29.

APPENDICES

Appendix I: Source code

RCODES USE

```
# selecting OPTIMUM CLUSTERS K

#Loading required Packages###

>install.packages("mclust")

> library(mclust)

library(ggplot)

library(dplyr)

set_wd<- ("C:\\Data()")

Data<-read.csv("C:\\DATA.csv")

df=select(Data,-c(1:16,21,23))

df=select(Data,-c(1:16,21,23))

##Scaling and Performing PCA

charge.pca<- prcomp(df, center = TRUE,scale. = TRUE)

> summary(charge.pca)

Importance of components:

                          PC1        PC2       PC3       PC4
PC5
```

```
Standard deviation      1.4152 1.0076 1.0023 0.9887 1.568e-
15

Proportion   of   Variance   0.4005   0.2031   0.2009   0.1955
0.000e+00

Cumulative   Proportion     0.4005   0.6036   0.8045   1.0000
1.000e+00

#PlottingElbow#

kmean_withinss<-function(k) {

cluster<-kmeans(df,k)

return(cluster$tot.withinss)

}

max_k<-30

wss<-sapply(1:max_k,kmean_withinss)

elbow<-data.frame(1:max_k,wss)

ggplot(elbow,aes(x=x2.max_k,y=wss))+geom_point()+geom_line(
)+scale_x_continuous(breaks=seq(1,20,by=5))

plot(1:max_k,wss,type="b",pch=19,    frame=TRUE,xlab="Number
of Clusters K",ylab="Tot.within cluster ss")


>kmm = kmeans(df,4,nstart = 2,iter.max = 30)
```

```
#ELBOW VALIDATION

d_clust<-      Mclust(as.matrix(df),      G=1:30,modelNames
=mclust.options("emModelNames"))

>d_clust$BIC

>  plot(d_clust)

Model-based clustering plots:

1: BIC

#FUZZY CLUSTERING

library(factoextra

> library(cluster)

>fanny(df, 4, metric = "euclidean", stand = FALSE)

Fuzzy Clustering object of class 'fanny' :



Available components:

 [1] "membership"  "coeff"       "memb.exp"    "clustering"
"k.crisp"

 [6]  "objective"    "convergence"  "diss"            "call"
"silinfo"

[11] "data"
```

```
>res.fanny<- fanny(df, 4)

>fviz_cluster(res.fanny, geom = "point", ellipse.type =
"norm", repel = FALSE,palette = "jco", ggtheme =
theme_minimal(),legend = "right")

fviz_silhouette(res.fanny, palette = "jco",ggtheme =
theme_minimal())

sil<- silhouette(res.fanny, dist(df))

>neg_sil_index<- which(sil[, "sil_width"] < 0)

>sil[neg_sil_index, , drop = FALSE]
```

#K-MEANS CLUSTERING

```
> res.km<-kmeans(df,4)

res.km$cluster

fviz_cluster(res.km, df, ellipse.type = "point")

>fviz_cluster(res.km, df, geom = "point", ellipse.type =
"norm", repel = FALSE,palette = "jco", ggtheme =
theme_minimal(),legend = "right")

K-means clustering with 4 clusters of sizes 452, 1139,
1169, 573


  require("cluster")
```

```
>sil<- silhouette(res.km$cluster, dist(df))

>fviz_silhouette(sil)


#WRONGLY CLUSTERED POINTS

>neg_sil_index<- which(sil[, "sil_width"] < 0)

>sil[neg_sil_index, , drop = FALSE]

      cluster neighbor     sil_width


#HEAT MAP

> cluster <- c(1:4)

> center_df1 <- data.frame(cluster, center)

> center_reshape1 <- gather(center_df1, features, values,
total.day.charge: Tot.Charge)

> head(center_reshape1)

>library(RColorBrewer)

>hm.palette<-colorRampPalette(rev(brewer.pal(10,
'RdYlGn')),space='Lab')

>ggplot(data = center_reshape1, aes(x = features, y =
cluster, fill = values))+scale_y_continuous(breaks = seq(1,
4, by = 1))+ geom_tile() +coord_equal()
```

```
+scale_fill_gradientn(colours = hm.palette(90))
+theme_classic()
```

#PAM clustering#

```
pam.res <- eclust(df, "pam", k = 4, graph = FALSE)
pam.res$cluster
fviz_cluster(pam.res, geom = "point", frame.type = "norm")
fviz_silhouette(pam.res)
     sil<- silhouette(pam.res$cluster, dist(df))
     >neg_sil_index<- which(sil[, "sil_width"] < 0)
     >sil[neg_sil_index, , drop = FALSE]
```

#Hierarchical clustering#

```
res.hc<- eclust(df, "hclust", k = 4,
                method = "complete", graph = FALSE)
head(res.hc$cluster, 3333)
fviz_dend(res.hc, rect = TRUE, show_labels = FALSE)
fviz_silhouette(res.hc)
     sil<- silhouette(res.hc$cluster, dist(df))
```

```
>neg_sil_index<- which(sil[, "sil_width"] < 0)

>sil[neg_sil_index, , drop = FALSE]
```

# Treating the outliers#

Capping

For missing values that lie outside the 1.5 * *IQR* limits, were capped by replacing those observations outside the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.

```
x <- df$Tot.Charge

>qnt<- quantile(x, probs=c(.25, .75), na.rm = T)

> caps <- quantile(x, probs=c(.05, .95), na.rm = T)

> H <- 1.5 * IQR(x, na.rm = T)

>x[x < (qnt[1] - H)] <- caps[1]

>x[x > (qnt[2] + H)] <- caps[2]

>NewsummaryStatistics after Capping(x)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  31.34   52.38   59.47   59.47   66.48   87.35
```