IJCSI International Journal of Computer Science Issues, Volume 15, Issue 5, September 2018
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.5281/zenodo.1467663

Customer Behaviour Segmentation among Mobile Service Providers in Kenya using K-Means Algorithm

Khamis Mwero Manero[1], Richard Rimiru[2] and Calvins Otieno[3]

[1] JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY
P.O. Box 62,000 – 00200 NAIROBI, KENYA

[2] JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY
P.O. Box 62,000 – 00200 NAIROBI, KENYA

[3] JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY
P.O. Box 62,000 – 00200 NAIROBI, KENYA

Abstract

In today's competitive environment, operators are investing in understanding their customers better, especially their most profitable customer groups and the groups that have the biggest potential to become such. By segmenting customers based on their behavior, operators can better target their actions, such as launching tailored products and target one-to-one marketing, to meet the customer expectations. The general objective of the study is to provide customer behavior segmentation in mobile telecommunication markets using K-means Algorithm. The specific objectives include to handle multidimensionality data using K-means algorithm with Principal component analysis, to determine the value of parameter K (number of clusters) using stability plot before clustering, to use financial variables (mean monthly charges) for each frequently used service as inputs in k-means for Segmentation, to evaluate Clustering results and determine the most profitable segment using completely randomized design (CRD).The experiment to achieve the objectives was being done on R software. Results show that Cluster 3 and 1 are the most profitable segments. The operators often need to design distinguishable marketing strategy based on different behavior of their mobile subscribers in order to improve their marketing result and revenue.

Keywords: *K-means, Customer segmentation, Algorithm*

1. Introduction

Over the years, the increase in competition amongst businesses and the availability of large historical data repositories have prompted the widespread applications of data mining techniques in uncovering valuable and strategic information buried in organizations' databases (Wei et al., 2013). Technological improvement has enabled the telecom companies to store record of customers. Analyzing the historical data help the companies to discover the behavioral patterns of existing

customers which could result in a significant impact on predicting the future customers' behavior. Customer portfolio analysis (CPA) is an effective tool to investigate the customer behavior. The aim of CPA is to segment customers into groups (Thakur and Workman2016). Customer segmentation is the use of past data to divide customers into similar groups based on various features (Hsu et al. 2012). Using the customer segmentation process, the company will be able to identify the customers who are strategically important and profitable. These customers can be categorized into two main classes including high future lifetime value customers or high-volume customers (Buttle and Maklan 2015). Competitors are ready to provide the same services and products with higher quality and lower prices. Customers simply will leave the company for lower costs or higher quality (Keramati et al. 2014). Losing customers also leads to opportunity costs because of decreases in sales (Verbekeet al. 2011). The previous studies have shown that retaining the current valuable customers of the organization is much cheaper than attracting new ones. Data mining is the process of extracting meaningful information from a dataset and presenting it in a human understandable format for decision support. The data mining techniques intersect areas such as statistics, artificial intelligence, machine learning and database systems. The applications of data mining include but not limited to bioinformatics, weather forecasting, fraud detection, financial analysis and customer segmentation.

Customer segmentation is the subdivision of a business customer base into groups called customer segments such that each customer segment consists of customers who share similar market characteristics. This segmentation is based on factors that can directly or indirectly influence market or business such as products preferences or expectations, locations, behavior and so on (Gupta, 2014).

2. Business intelligence and data mining

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 (Maheshwari, 2014). BI includes much software for Extraction, Transformation and Loading (ETL), data warehousing, database query and reporting, multidimensional/on-line analytical processing (OLAP) data analysis and data mining.

The decision making cannot be based on archaic information only; it can be with the aid of current information also, 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? How can cash be conserved? 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 (Braha, 2013). 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.

Telecommunications data is generated in the many different operating systems used and could be classified into three main types. First, customer contractual data – personal data about the customers, including name and address, service plan, contact information, credit score, family income and payment history. Second, call detail data –detailed call records, including the originating and terminating phone numbers, the date and time of the call and the duration of the call customer transactions, and billing data, from which knowledge could be extracted at customer level, about the customers calling behavior and third network data – data resulting from the operation of the extremely complex configurations of equipment, minimally including a timestamp, a string that uniquely identifies the hardware or software component generating the message and a code that explains why the message is being generated. The identified three types of Telecommunications data are described by (Witten et al., 2016). The data multidimensionality is discussed by (Floh et al., 2014) 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, described above, is a very important reason for the recent very high interests in BI and Data Mining in the Telecommunications industry.

3.0 Customer segmentation and data mining techniques

Customer segmentation is defined as the practice of classifying customer base into distinct groups. In other words, customer segmentation is also described as the process of dividing customers into homogeneous groups on the basis of 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 (Frenkel et al., 2013).

Segmentation requires the collection, organization and analysis of customer data. With proper segmentations of a customer's data, it is possible to identify the reliability/loyalty of customers so as to increase the revenue of the organization. Segmentation is the process of developing meaningful customer groups that are similarly based on individual explanation characteristics and behavior. Kwach, Flora & Rajagopal (2014) explained that there are two main segmentation approaches: the first type of segmentation is the process of segmenting the customers based on understanding the needs of the end user which is called needs-based segmentation. The second type of segmentation, characteristics-based segmentation, is the process of segmenting customers based on their characteristics, attitudes or behaviors.

Aggarwal & Reddy (2013) have suggested the use of segmentation models based on volume of sales, meaning that marketing efforts should focus on customers engaged in a considerable number of transactions. This approached, called ''heavy half theory'', highlighted that one half of the customers can account for up to 80% of total sales. During the 1970's, the validity of the multivariate approaches used to identify the variables that affect deal proneness was criticized, which motivated the development of enhanced theoretical models of consumer behavior. Lee, Verma & Roth (2015) developed a generalizable psychographic segmentation model that divided the market into groups based on social class, lifestyle and personality characteristics.

However, practical implementation difficulties of this complex segmentation model were widely noted segmentation is also viewed as a method to have more targeted communication with the customers; and the process of segmentation describes the characteristics of the customer's groups (called segments or clusters) within the data. The diversity of customer needs and buying behavior, influenced by lifestyle, income levels or age, makes past segmentation approaches less effective. Therefore, current models for marketing segmentation are often based on customer behavior inferred from transaction records or surveys. The resulting data is then explored with data mining techniques; such as cluster analysis. They surveyed the applications of data mining for segmentation purposes. In most studies, customer segmentation is mentioned as the ideal way to obtain customer profitability through careful customer targeting. Lee, Verma & Roth (2015) have also identified market segments for a new computer, system based on the use of cluster analysis with data from a customer survey. The segmentation was supported by the rate of importance given to the product attributes.

Wang & Fan (2014) in their book "Data Mining: A hand on approach for business professionals" summaries and explain the current field of data mining and talks about some popular tools on the market that could be of use to anyone who is considering data mining. He analyzed the data mining software called Knowledge Seeker, which uses the decision tree approach to data mining and concluded that data could be grouped in optimal ways and this can be very useful if you are looking at market segmentation studies. While summarizing the typical industries that make use of data mining as a tool, author stated that direct mail and mailing is another area where data mining is widely used, almost all types of retailers' use direct marketing, and their main concern is to have information about customer segmentation, which in data mining is a clustering problem. One of the big challenges with traditional survey-based market research is that it provides a lot of information about a few customers. However, to use the results of market research effectively often requires understanding the characteristics of all customers. That is, market research may find interesting segments of customers (Guha & Mishra, 2016). These then need to be projected onto the existing customer base using available data. Behavioral data can be particularly useful for this; such behavioral data is typically summarized from transaction and billing histories. One requirement of the market research is that customers need to be identified so the behavior of the market research participants is known. In today's competitive markets, this approach is not sufficient and efficient. On the contrary, organizations need to have a complete view of their customers in order to gain a competitive advantage. They also need to focus on their customers‟ needs, wants, attitudes, behaviors, preferences, and perceptions, and to analyze relevant data to identify the underlying segments. The identification of groups with unique characteristics will enable the organization to manage and target them more effectively with, among other things, customized product offerings and promotions (Witten et al., 2016). Marketers can use data mining methods to discover the attributes that can predict customers' responses to offers and promotion programs. Thus, customer acquisition is possible by matching to those attributes for converting non-customer to respond to new offers and promotions. However, in segmentation normally the group of people is of similar needs, characteristics or behavior.

Thus, for customer acquisition similar type of segment needs to be targeted with similar offers and promotion programs. Customer segmentation is a popular application of data mining with established customers. A segmentation project starts with the definition of the business objectives

and ends with the delivery of differentiated marketing strategies for the segments. There are many different segmentation types based on the specific criteria or attributes used for segmentation. Specifically, customers can be segmented according to their value. The type of segmentation used depends on the specific business objective. There are various segmentation types according to the segmentation criteria used. Particularly, customers can be segmented according to their value, socio-demographic and life-stage information, and their behavioral, need/attitudinal, and loyalty characteristics. The type of segmentation used depends on the specific business objective and your target. Different criteria and segmentation methods are appropriate for different situations and business objectives (Witten et al., 2016). In behavioral segmentation, customers are grouped by behavioral and usage characteristics. Although behavioral segments can be created with business rules, this approach has inherent disadvantages. It can efficiently handle only a few segmentation fields and its objectivity is questionable as it is based on the personal perceptions of a business expert. Data mining on the other hand can create data-driven behavioral segments. Clustering algorithms can analyze behavioral data, identify the natural groupings of customers, and suggest a solution founded on observed data patterns. Provided the data mining models are properly built, they can uncover groups with distinct profiles and characteristics and lead to rich segmentation schemes with business meaning and value.

In general, the application of a cluster model is required to reveal the segments, particularly if we need to combine a large number of segmentation attributes. As opposed to business rules, a cluster model is able to manage a large number of attributes and reveal data-driven segments which are not known in advance (Guha & Mishra, 2016). Data mining can also be used for the development of segmentation schemes based on the current or expected value of the customers. These segments are necessary in order to prioritize customer handling and marketing interventions according to the importance of each customer. Moreover, since a vital part of a segmentation project is insight into the derived clusters and an understanding of their meaning, we will also propose ways for profiling the clusters and for outlining their differentiating characteristics. One way to find behavioral segments is to use the clustering techniques described in. This method leads to clusters of similar customers but it may be hard to understand how these clusters relate to the business. More typically, a business would like to perform a segmentation that places every customer into some easily described segment (Guha & Mishra, 2016). Often, these segments are built with respect to a marketing goal such as subscription renewal or high spending levels. Decision tree techniques described in are ideal for this sort of segmentation.

4.0 Empirical Analysis

Hickman et al., (2013) investigated K-Means. In this paper principal component analysis and linear transformation is used for dimensionality reduction and initial centroid is computed, and then applied to K-Means clustering algorithm. Only K-means could allow the use of multiple algorithms like PCA and linear transformation something which is not present in other clustering methods like Self-organizing maps (SOM). Due to few studies which has used this approach, it is more appropriate to use it in Kenyan context and the flexibility of K-means is the gap this study intends to utilize.

Manzano et al., (2015) used evolving limitations in K-means algorithm in data mining and their removal". Manzano et al., (2015) states that K-means is popular because it is conceptually simple and is 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.

Ahn et al. (2011) using this mixture of strategies proposes a customer classification technique which may be 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) are applied one by one to predict the future purchase of products. This proves more complicated and time consuming, it further requires advanced knowledge in data mining process. Secondly, each model produces the result and then applied genetic algorithm on the combined results of all models and applying K-means clustering of customers to establish whether customer would purchase a new product. The framework is tested in a mobile telecom company in Korea. Therefore, the model produced excellent results for cross-selling. On the basis of simplicity and fast and efficient nature makes K-means outstanding data mining methods when it comes to customer segmentation.

Ramirez-Ortiz et al., (2015) used a Modified K-Means Algorithm for Circular Invariant Clustering. They found that several important pattern recognition applications are based on feature vector extraction and vector clustering. Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation. A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry. The Researchers present 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 this paper, Researchers propose a modified version of the K-means algorithm to cluster data since it will help in improving the efficiency of the data mining process.

Bose & Chen (2010) Explore 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 is generally performed in services, revenue, usage and user categories attributes. In this study data mining techniques, K-means and Kohonen vector quantization (KVQ), are used to group customers based on the attributes. Then inter-cluster analysis is 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 located in Hong Kong. There are 50,000 records in database for a period of one year and dataset range was 1st Sep 2004 to 1st Sep 2005. Data was partitioned in to four clusters. As a result, this is 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 and EM algorithms is better than that of hierarchical clustering algorithm. From the literature review only, limited studies have been done using K-means more specifically in Kenyan context, hence this will be a great opportunity to use the K-means algorithms. Therefore, this study concluded that K-means is a much better option than KVQ and hierarchical clustering methods.

Mehta et al., (2017) found that both K-means and SOM have linear complexity, O (ndk) and O nd) respectively, where n is the number of data samples, d is the number of dimensions and k the number of clusters. Hence, both methods scale to large data sets and have good performance. The third method, SVC has computational complexity of where n is the number of support vectors, not data samples. The main stepping-stone for SVC with large data sets is locating the Lagrange multiplier that is used for finding the boundary values of the clusters. However, there are proposed ways to overcome this issue such as the Sequential Minimal Optimization (SMO) algorithm. Consequently, all three algorithms are able to scale for large data sets and performance should not be an issue

Rodriguez & Laio (2014) presented an optimal K-Means Clustering algorithms, employed in many bioinformatics tasks, including categorization of protein sequences and analysis of gene-expression data. Although these algorithms are routinely applied, many of them suffered from the following limitations: (i) relying on predetermined parameters tuning, such as a-priori knowledge regarding the number of clusters; (ii) involving nondeterministic procedures that yield inconsistent outcomes. Thus, a framework that addressed these shortcomings was desirable. A data driven framework was provided that included two interrelated steps. The first one was SVD based dimension reduction and the second was an automated tuning of the algorithm's parameter(s). The dimension reduction step is efficiently adjusted for very large datasets. The optimal parameter setting was identified according to the internal evaluation criterion known as Bayesian Information Criterion (BIC). This framework can incorporate most clustering algorithms and improve their performance. When segmentation is performed with several attributes the data is often hard to visualize and understand. For the behavioral segmentation, several attributes were identified for clustering such as the average revenue per customer, visit count per different occasion and download volume per different occasion (Rodriguez & Laio, 2014). This high-dimensionality affects the 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, K-means algorithm handles the high-dimensionality of data during the iterative steps.

Peker et al. (2017) described the use of Kernel-means was an extension of the standard means clustering algorithm that identifies 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. In order 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. Unlike SOM, K-means does not incorporate any functionality for handling outliers - instead outliers affect the clustering as they are treated equally with all other samples. Inherently this will lead to poorer clustering results depending on a number of outliers in the data set. Neither does K-means take care of the missing values, so the only approach is often to discard such values manually before applying the clustering algorithm. On the other hand, SVC has a way to handle outliers and missing values automatically by using a soft margin constant (C), which leaves these data points outside the enclosing sphere. Consequently, if the algorithm should be trusted to handle outliers and missing values automatically, SOM and SVC have ways to handle this, whereas K-means does not.

Hosseini et al. (2010) adopted K-means algorithm to classify the customer loyalty based on RFM values. Cheng and Chen (2009) used K-means and rough set theory to segment customer value based on RFM values. Chen et al. (2009) identified purchasing patterns based on sequential patterns. Migueis et al. (2012) proposed a method for customer's segmentation given by the nature of the products purchased by customers. This method is 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.

5.0 Concept of Data Mining

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 (Braha, 2013).

Zheng (2015) defines data mining as the process of mining, gaining, knowledge from large volume of raw data. It is one of the great influences in information industry especially in telecommunication field as a whole in recent years due to the wide availability of huge amounts of data and needs for turning such data into useful information and knowledge, this knowledge gained can be used for applications ranging from market analysis among others.

Data mining is learning from data and consist of a set of rules and equations that can be used to identify useful data patterns, understand and predict behavior. It can be 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 (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. 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 particular step in this process. The steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge 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).

6.0 Classification techniques for customer segmentation

Classification analysis is also known as supervised classification. Classification analysis is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. By using classification, it is possible to organize data in a given class. The classification uses given class labels to organize the objects in the data collection in an orderly manner. Classification model 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. Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern (Frenkel et al., 2013)

Classification approaches normally use a training set where all objects are already associated with known class labels. Then the classification algorithm learns from the training set and builds a model. The model is used to classify new objects. In other words, classification is a two-step process, first a classification model is built based on training data set and then the model is applied to new data for classification. 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.

Sheu et al. (2009) adopted decision tree to explore the potential relationship between important influential factors and customer loyalty. The findings of these studies inspire us to adopt the decision tree to explore the relationship between customers' purchase amounts and customers' demographic and behavioral characteristics, with special attention to the characteristics of high-and low-spending customers.

7.0 Clustering techniques for customer segmentation

Clustering can be defined as the process of grouping a set of physical or abstract objects into classes of similar objects. Clustering is also called unsupervised classification, because the classification is not dictated ordered by given class labels. There are many clustering approaches, all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity). Clustering is similar to classification, but classes are not predefined and it is up to the clustering algorithm to discover acceptable classes (Tsiptsis & Chorianopoulos, 2011). Often, it is necessary to modify the clustering by excluding variables that have been employed to group instances because, upon examination, the user identifies them as irrelevant or not meaningful. After clusters are found that reasonably segment the database, these clusters are then used to classify new data. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means. Clustering is different from segmentation. Segmentation refers to the general problem of identifying groups that have common characteristics whereas clustering is a way to segment data into groups that are not previously defined. Clustering is useful for finding natural groups of data which are called clusters. A cluster is a collection of data that are similar to one another. Clustering can be used to group customers with similar behavior and to make business decisions in industry to as unsupervised learning (Berkhin, 2006). 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 cluster are more similar to each other than points in different clusters. Berkhin (2006) use clustering technique to segment customers and markets. The K-means clustering algorithm and the Kohonen self-organizing map are the two most popular clustering techniques.

Samira et al. (2007) have 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, in order 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".

Pramod et al. (2011) 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 in order to provide personalized and efficient customer service. Huang et al. (2009) 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

8.0 Research Design

The Model of this research has been implemented using R studio software. K-means clustering algorithm is employed in order to generate customer segments. Customers' behavioral variables were taken as an input of k-means algorithm. Clustering results are evaluated and the most profitable segment was determined using completely

randomized design (CRD). The proposed approach is a two phased model. In first phase, data was extracted from the https://archive.ics.uci.edu/ml/datasets.html and target population and then data cleansing was done. It involved removing the noisy, irrelevant data and then formatted according to the required standard. In second phase, clustering phase, I generated the clusters using K-means algorithm, then clustering results are evaluated and the most profitable segment was determined using completely randomized design (CRD). Finally, presentation of the results of data analysis.

9.0 Principal Components Analysis - PCA)

Principal Component Analysis (PCA) is a statistical procedure that transforms and converts a data set into a new data set containing linearly uncorrelated variables, known as principal components. The basic idea is that the data set is transformed into a set of components where each one attempts to capture as much of the variance (information) in data as possible.

Experiment Results for PCA

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.3611	1.1777	1.0420	0.69264	0.4416
Proportion of Variance	0.3705	0.2774	0.217	2 0.09595	0.0390
Cumulative Proportion	0.3705	0.6479	0.865	0 0.96100	1.0000

The above results shows that pc1 contributes the maximum variability and pc5 the least variability. This can be elaborated below;
PC1 contributes approximately 37.05%
PC2 contributes approximately 27.74%
PC3 contributes approximately 21.72%
PC4 contributes approximately 9.59%
PC5 contributes approximately 3.9%
The standard deviations of the above componenets results i.e. 1.3611 ,1 .1777 ,1 .0420 ,0.69264, 0.4416 indicates that PC1,PC2,PC3 are the Eigen Values since their standard deviation is greater than one. These components or variables with Eigen value are retained for further analysis since they account most of the variation. That is why PC1, PC2, PC3 was capable to account for 86.5% of the variability hence reduction of dimension was successfully. PC4 and PC5 neglected since they are not Eigen values.
The Summary (pca_out) Shows that the cumulative Variance explained by the pc i.e Cumulative Proportion 0.3705 ,0.6479, 0.8650.

10.0 Conclusion

The mobile telecommunication marketplace is highly competitive. The operators often need to design distinguishable marketing strategy based on different behavior of their mobile subscribers in order to improve their marketing result and revenue. Call Detail Records describe customer utilization behavior. They have more information to describe customer behavior than billing system data. Clustering analysis based on call detail records can give more information than other clustering analysis for marketing management. We suggested a customer life cycle model considering the past contribution, potential value, and churn probability at the same time. The model used for customer segmentation. Three perspectives on customer value (current value, potential value, and customer loyalty) assist marketing managers in identifying customer's segmentation with more balanced viewpoints. After identification of subscriber's behavior and identification of loyal groups, it's feasible and possible to put mobile customer clusters in place and make an applicable strategic plan for each group to achieve customer satisfaction

References

1) AAggarwal, C. C., & Reddy, C. K. (2013). *Data clustering: algorithms and applications*. CRC press.
2) Ahn, H., Ahn, J. J., Oh, K. J., & Kim, D. H. (2011). Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques. *Expert Systems with Applications*, *38*(5), 5005-5012.
3) Arasa, R., & Githinji, L. (2014). The relationship between competitive strategies and firm performance: A case of mobile telecommunication companies in Kenya. *International Journal of Economics, Commerce and Management*, *2*(9), 1-15.
4) Arora, D., & Malik, P. (2015). Analytics: Key to go from generating big data to deriving business value. In *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on* (pp. 446-452). IEEE.
5) Bahari, T. F., & Elayidom, M. S. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, *46*, 725-731.
6) Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping multidimensional data*, *25*, 71.

7) Borg, W. & Gall, M. (2006). *Educational Research: An introduction*. New York: Longman Inc.
8) Bose, I., & Chen, X. (2010). Exploring business opportunities from mobile services data of customers: An inter-cluster analysis approach. *Electronic Commerce Research and Applications*, *9*(3), 197-208.
9) Braha, D. (2013). *Data mining for design and manufacturing: methods and applications* (Vol. 3). Springer Science & Business Media.
10) Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for segmentation prediction in telco. *Expert Systems with Applications: An International Journal*, *85*(C), 204-220.
11) Chen, K., Hu, Y. H., & Hsieh, Y. C. (2015). Predicting customer segmentation from valuable B2B customers in the logistics industry: a case study. *Information Systems and e-Business Management*, *13*(3), 475-494.
12) Chen, Y. L., Kuo, M. H., Wu, S. Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications*, *8*(5), 241-251.
13) Crespo, F., & Weber, R. (2015). A methodology for dynamic data mining based on fuzzy clustering. *Fuzzy Sets and Systems*, *150*(2), 267-284.
14) Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Segmentation prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing*, *19*, 31-40. Gerpott, T. J., Ahmadi, N., & Weimar, D. (2015). Who is (not) convinced to withdraw a contract termination announcement?–A discriminant analysis of mobile communications customers in Germany. *Telecommunications Policy*, *39*(1), 38-52.
15) Floh, A., Zauner, A., Koller, M., & Rusch, T. (2014). Customer segmentation using unobserved heterogeneity in the perceived-value–loyalty–intentions link. *Journal of Business Research*, *67*(5), 974-982.
16) Frenkel, D., Wilkinson, K., Zhao, L., Hickman, S. E., Means, T. K., Puckett, L., ... & El Khoury, J. (2013). Scara1 deficiency impairs clearance of soluble amyloid-β by mononuclear phagocytes and accelerates Alzheimer's-like disease progression. *Nature communications*, *4*.
17) Guha, S., & Mishra, N. (2016). Clustering data streams. In *Data Stream Management* (pp. 169-187). Springer Berlin Heidelberg.
18) Gupta, G. K. (2014). *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd..
19) Hamka, F., Bouwman, H., De Reuver, M., & Kroesen, M. (2014). Mobile customer segmentation based on smartphone measurement. *Telematics and Informatics*, *31*(2), 220-227.
20) Hickman, S. E., Kingery, N. D., Ohsumi, T. K., Borowsky, M. L., Wang, L. C., Means, T. K., & El Khoury, J. (2013). The microglial sensome revealed by direct RNA sequencing. *Nature neuroscience*, *16*(12), 1896-1905.
21) Hoegele, D., Schmidt, S. L., & Torgler, B. (2016). The importance of key celebrity characteristics for customer segmentation by age and gender: Does beauty matter in professional football?. *Review of Managerial Science*, *10*(3), 601-627.
22) Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved segmentation prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, *24*, 994-1012.
23) Keter, A. (2015). *Challenges of Strategy Implementation in the Telecommunication Industry in Kenya: A Case of Safaricom Limited* (Doctoral dissertation, United States International University-Africa)..
24) Kothari, C. R. (2009). *Research methodology: Methods and techniques*. New Age International
25) Kwach, J., Flora, J., & Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, *5*(1), 420-430.
26) Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.
27) 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.
28) Mahajan, V., Misra, R., & Mahajan, R. (2017). Review on factors affecting customer segmentation in telecom sector. *International Journal of Data Analysis Techniques and Strategies*, *9*(2), 122-144.
29) Maheshwari, A. (2014). *Business Intelligence and Data Mining*. Business Expert Press.
30) Malhotra, A., & Kubowicz Malhotra, C. (2013). Exploring switching behavior of US mobile

service customers. *Journal of Services Marketing*, *27*(1), 13-24.

Manzano, B. L., Means, B. K., Begley, C. T., & Zechini, M. (2015). Using Digital 3D Scanning to Create "Artifictions" of the Passenger Pigeon and Harelip Sucker, Two Extinct Species in Eastern North America: The Future Examines the Past. *Ethnobiology Letters*, *6*(2), 232-241.

Mehta, S., Cronkite, D. A., Basavappa, M., Saunders, T. L., Adiliaghdam, F., Amatullah, H., ... & Lauer, G. M. (2017). Maintenance of macrophage transcriptional programs and intestinal homeostasis by epigenetic reader SP140. *Science immunology*, *2*(9).

Mugenda & Mugenda, G.A. (2003). *Research Methods, Qualitative and Quantitative Approaches*, Kenya: *ACTS press.*