# Usage Based Clustering of Customers for Mobile Service Packaging

BY: DEMELASH BIRU

ADVISOR: Dr. BENEYAM BERHANU (PHD)

A Thesis submitted to School of Electrical and Computer Engineering

Addis Ababa Institute of Technology

in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecommunication Engineering.


Addis Ababa University

Addis Ababa, Ethiopia

December ,2019

## Declaration

I, undersigned, declare that this MSc thesis comprises my work and it is done in compliance with internationally accepted practices. I acknowledged and referred to all materials used in this thesis work.

| Demelash Biru | _____ | _____ |
|---|---|---|
| Name | Signature | Date |

This thesis has presented for examination with my approval as a university advisor.

| Dr. Beneyam Berhanu | _____ | _____ |
|---|---|---|
| Advisor | Signature | Date |

# EXAMINERS


## Addis Ababa University

## Addis Ababa Institute of Technology

## School of Electrical and Computer Engineering

This is to certify that this paper is prepared by, **Demelash Biru** on the title **Usage-Based Clustering of Customers for Mobile Service Packaging** and submitted in partial fulfillment of the requirements for degree of Master of Science in Telecommunication Engineering. It complies with the regulation of the university and meets the accepted standards with respect to originality and quality.

Examiner Dr. Surafel Lemma Signature _____ Date _____

Examiner Dr. Yihenew Wondie Signature _____ Date _____

Advisor: Beneyam Berhanu (PhD) Signature _____ Date _____

Director of Post Graduate Program: _____ Signature_____ Date _____

---

*Dean, School of Electrical and Computer Engineering.*

# DEDICATION

This paper is dedicated to my family and friends.

# ABSTRACT

Satisfaction of customers is the most important factor for mobile operators to be successful. This needs effective customer segmentation and segment targeted mobile service packaging and delivery. Segmentation differentiates customers into multiple groups that manifest different service needs and preferences, thus different service packages. It has been traditionally performed using demographic and value-based segmentation methods based on customer survey data. For improved efficiency, advanced clustering techniques that exploit existing historical customer data from network management system have been applied. Instead of using a single dimension of value-based segmentation, the historical data set with many features was applied to assess the customer service usage behavior from different dimensions. For a dataset with many attributes, such advanced clustering techniques have not been investigated in the Ethiopian context.

The thesis work investigates and compares the performance of K-means and expectation-maximization algorithms for usage-based clustering using voice, SMS and internet service usage call detail record data of mobile customers. The performance was compared using metrics such as cluster size or ratio, cluster cohesion or compactness and separation between centroid values. These metrics were used to evaluate the quality of the clustering result of the algorithms in identifying distinguished customer segments from each service usage dataset for mobile service packaging purposes. Optimal cluster size per dataset was determined using elbow method. In the study, data processing and algorithm implementations were performed using WEKA data mining tool.

Achieved results indicate that for all the datasets the EM algorithm formed compact clusters with low level of within cluster variance. On the other hand, K-means clustering has a better quality in assigning instances to each cluster fairly. In general, the study identified important additional attributes from the CDR dataset to differentiate customers for mobile service packaging purpose. These additional features enhance the insight on customers to provide well differentiated mobile service packages.

**Key words**:

*Customer Segmentation, Usage based, K-means, EM, Mobile Service Packaging, CDR*

# ACKNOWLEDGMENT

*This MSc thesis completes my study in Telecommunication Engineering at AAiT. Writing this thesis document has been an excellent learning process and a long journey. It is a pleasure to thank all the people involved in the journey and made the thesis possible.*

*First and foremost, sincere thanks to my advisor Dr. Beneyam Berhanu, for the insightful discussion and guidance throughout the study. I am indebted to ethiotelecom and AAiT for allowing me to attend the MSc program. I would like to express my appreciation to everyone who has helped me with this work. This includes the IS and marketing division staff of ethiotelecom, who helped in responding to my request through the provisioning of various materials and documents.*

*Finally, I would like to express my deepest gratitude to my family and friends for their patience and essential supports.*

*Addis Ababa, December 2019*

*Demelash Biru*

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **ARFF** | Attribute Relationship File Format |
| **ARPU** | Average Revenue Per Unit |
| **BICP** | Business Intelligence Communication Platform |
| **CDR** | Call Detail Record |
| **CRISP** | Cross Industry Standard Process |
| **CRM** | Customer Relationship Management |
| **DM** | Data Mining |
| **EDA** | Exploratory Data Analysis |
| **EM** | Expectation Maximization |
| **GUI** | Graphical User Interface |
| **KDD** | Knowledge Discovery Database |
| **KPI** | Key Performance Indicators |
| **MB** | Megabyte |
| **PAG** | Pay As you Go |
| **RFM** | Recency, Frequency, Monetary |
| **SMS** | Short Messaging Service |
| **SOM** | Self-Organizing Map |
| **SSE** | Sum of Squared Error |
| **WEKA** | Waikato Environment for Knowledge Analysis |

# 1 Introduction

## 1.1 Background and Motivation

Customers are usually stated as key assets of any customer-oriented enterprises and the success of an organization is highly correlated with the strength of relationship with customer. Hence identifying the level of heterogeneity in the customer base enables to enhance the insight on the customer and strength the relationship with them by effectively identifying their needs to introduce various retention mechanisms. As discussed in [19] organizations have been employing Customer Relationship Management (CRM) to easily identify the customer needs of established customers.

In this regard customer segmentation has a vital role in the CRM system to distinguish customers into different groups. As discussed in [45,46] modern marketing is moving from mass-marketing to target-marketing and hence customer segmentation becomes a compulsory task to propose a customer-oriented marketing strategy. According to [30,44] segmentation process divides the customer base into well separated & at the same time internally homogenous groups to develop segment targeted marketing actions. Thus, it helps to minimize the number of individuals to deal with to limited number of groups.

The mobile customer base of ethiotelecom has reached around 40 million customers and majority of them are prepaid customers [55]. These customers recharge their account in advance to use the service. To influence service usage behavior and profitability of the prepaid customer base, the mobile operator has been offering various mobile service package options as an alternative with different volume for voice, SMS and Internet services. The mobile service packages are provided for a discount price with a fixed validity period and the goal of service packaging is to influence service usage and thereby enhance the company revenue by encouraging service usage intensity.

In the mobile service package development process, traditional customer segmentation technique has been applied on survey data to build the customer segments. But as Figure 1 shows, these packages have low level of demand from customers (indicated by low & inconsistent purchase frequency). As a result, it has a little impact on influencing the usage behavior of subscribers. It designates that the current mobile service package options and its development process are far from being customer centric to attract customers' need. The

attractiveness of mobile service package alternatives depends on the development process followed to propose these products and the level of insight about customer service usage behavior. Hence an effective implementation of customer segmentation helps to identify the needs and preferences of customer base to provide segment targeted mobile service packages.

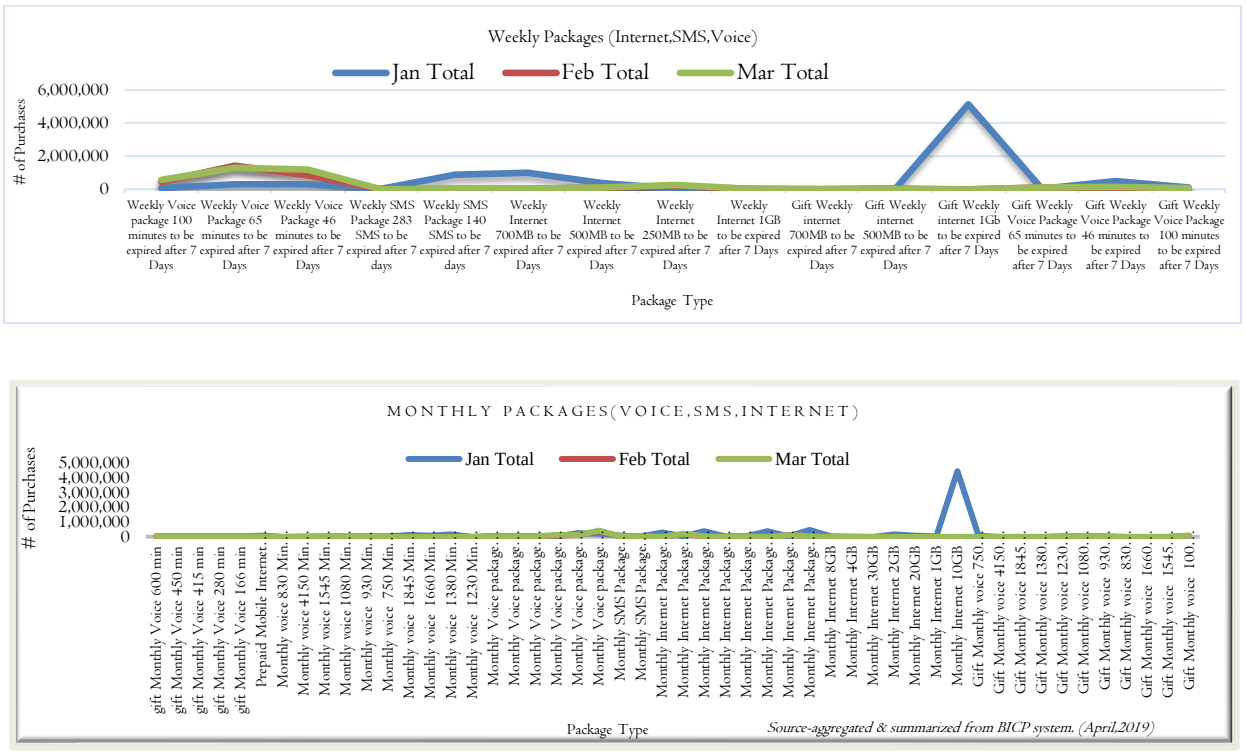



Figure 1 : Mobile service package purchase frequency.

To influence customer usage behavior and enhance the Average Revenue Per Unit (ARPU) of the prepaid mobile customer base, different marketing actions have been implemented such as provisioning of mobile service package for voice, SMS and internet services in various options. Though with these efforts, the package purchase frequency has little impact on influencing the usage behavior of subscribers. According [55] only 21% of customers are package users and the rest 79% still uses the Pay as you go (PAG) service usage options, which reflects the gap of the current mobile service package option in attracting the subscriber's attention.

ethio telecom serve large number of customers with diverse needs, usage behaviors and preferences. The operator preserves various data about subscriber such as CDR data, which reflect the customer behavior in terms of usage amount, usage time, usage day, usage day, and spending potential. But CDR set is large in

volume with many features and needs the application of advanced data mining technique to analyze the customer behavior and extract full insight about the customer.

Data mining techniques such as clustering, helps to build data driven segments by analyzing behavioral or service usage data to establish natural groupings of customers [19]. Customers differ in terms of behavior, needs, wants & characteristics and the goal of clustering is to identify differentiated customer types and segment the customer base into clusters for segment targeted marketing actions. As discussed in [44] clustering technique reveals internally homogeneous and externally heterogeneous groups. This research is motivated by the drawbacks of the existing customer segmentation approach in its process of differentiating the customer base for the provisioning of mobile service packages.

## 1.2 Statement of the Problem

The level of insight on customers' service usage behavior affects the type of products to propose and the development process to be followed. Which in turn, affects the desirability of the product by the customer. In this regard, customer segmentation has a vital role in enhancing the insight on the customer by distinguishing or identifying the needs and preferences of customers to provide tailored service closed to each segment needs and preferences.

Currently, the Ethiopian sole and incumbent mobile operator applies demographic and value-based segmentation to distinguish the customer base and offer segment targeted mobile service packages. Statistical technique has been applied to build customer segments using customer survey data. The downside of this segmentation approach is the unreliability of results due to a small number of respondents to surveys, expensiveness & time-consuming nature of survey data collection method, and survey data is less reliable in reflecting the actual behavior of customers due to customers integrity problem to responses. Moreover, value-based and demographic segmentation approaches as well as input datasets are inadequate for mobile service packaging as the segmentation features fail to reveal detail subscriber service usage patterns.

In value-based segmentation, customers are simply binned into different groups based on an aggregate monthly spending amount attribute. A single attribute is used to differentiate the heterogeneous customer base. This feature is insufficient to fully explain the level of differentiability in the mobile customer base as it neglects to include features such as preferred service type of customers, service usage day and time, and weekly consumption ratio of customers. Thus, the segmentation approach missing these relevant features is less important in effectively identifying the customers into different groups to build a suitable mobile service

package alternative. According to [45] value-based segmentation technique is less important for product development purposes.

In the same context, demographic segmentation has been applied for multiple purposes ranging from mobile service package development to distribution channel selection. This segmentation is mainly based on survey data sources collected from customers. Attributes used for demographic segmentation are age, sex, and religion. These attributes are less important to objectively differentiate customers for mobile service packaging. As discussed in [30] though it is widely applicable, it is criticized for being untrustworthy as people with the same demographic value might have different attitudes. According to [45] customer segmentation for product development purposes demands behavioral attributes that reflect the service usage behavior of subscribers. Hence behavioral attributes are more relevant for the segmentation of customers for mobile service packaging purposes.

Clustering is one of the data mining techniques widely deployed for customer segmentation and in this thesis, clustering techniques are compared to build usage-based customer segments by using CDR data of customers.

## 1.3 Objectives

### 1.3.1 General Objective

The main objective of the thesis is to compare the performances of clustering algorithms to build differentiated customer segments for mobile service packaging by using voice, SMS and internet service CDR data of Ethiopian mobile customers.

### 1.3.2 Specific Objectives

The specific objectives of the thesis are:

- To construct additional features relevant for usage-based segmentation.
- To preprocess service usage CDR datasets for usage-based clustering.
- To evaluate the clustering results based on the cluster evaluation techniques.
- To compare and identify a clustering algorithm suitable for segmentation of mobile customers.
- To profile and interpret the clustering results of the best performing algorithm.

## 1.4 Methodology

In this thesis, clustering techniques were applied to CDR data of mobile subscribers to build customer segments. Expectation-Maximization (EM) and simple K-means clustering algorithms were compared based on the quality of the clustering result for each service usage dataset. Fig. 2 shows the applied methodology to undertake the comparison of EM & K-means algorithms for customer segmentation. The description of its parts follows.

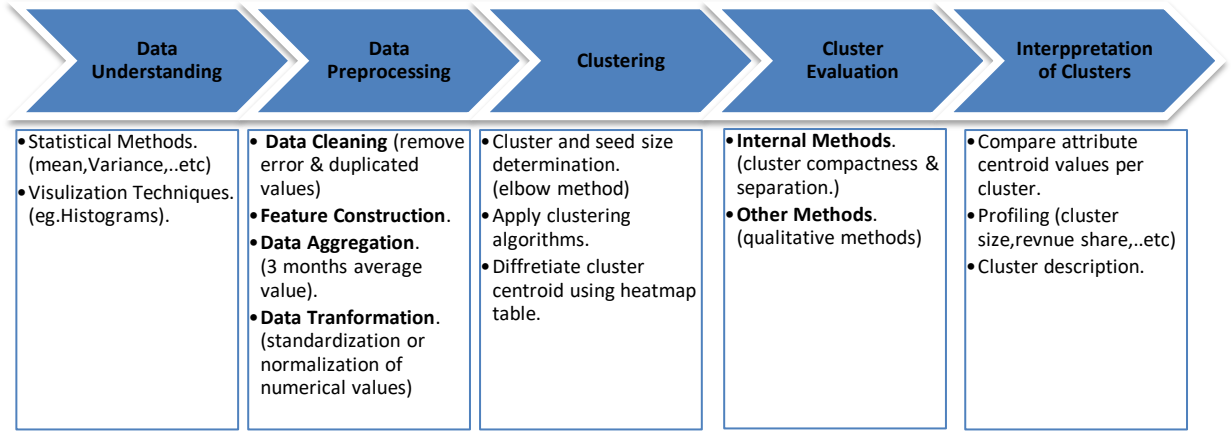



Figure 2 : Research framework.

### 1. Literature Review

Research works and journal papers in the area of data mining and customer segmentation were reviewed. Besides, marketing documents and reports were analyzed for a better understanding of the marketing domain as well as mobile service package development process.

### 2. Data Collection and Understanding

Mobile customers CDR data was collected from the BICP platform of the CRM system. The data required for this study include CDR data of each service category such as voice, SMS and internet. For the data understanding, visualization and statistical techniques such as mean, variance and histogram were used. It gives an insight on the distribution of instance for each attribute and identify the importance of an attribute in differentiating customers.

### 3. Data Preprocessing

The data preprocessing task convert the original dataset to a format suitable for clustering purpose. The detail service usage transaction of the subscriber was aggregated, and monthly average value was utilized to reflect subscriber usage behavior. Additional fields were also constructed to enhance the data quality and capture essential behaviors of customers. The detail of the data preprocessing task is presented in Section 4.3 and the major tasks include data cleaning, construction of additional features, data aggregation and selection of clustering features. The tools used for data preprocessing are MS excel, IBM SPSS and Weka.

### 4. Clustering Methods

Clustering tasks were implemented on each service usage dataset using an open-source data mining tool called, "Weka 3.8.3", to build customer segments. The cluster size with additional input parameters of the tool was applied to the prepared dataset & the major tasks in this stage are determination of the optimal seed and cluster size combination, and comparison of clustering results. For cluster size determination, the auto clustering option was compared with the elbow method to identify the optimal cluster size.

### 5. Cluster Result Evaluation

The quality of the clustering result derived from each clustering algorithm (K-means & EM) was compared and evaluated based on the cluster evaluation techniques mentioned in Section 5.8. The major tasks in this stage include evaluation of clustering algorithms in terms of cluster cohesion, cluster separation, and distribution of instances per cluster.

### 6. Cluster Profiling & Interpretation

The clustering process was concluded by an interpretation of clustering results using cluster profiling techniques. Therefore, the characteristics of each segment was described using visualization techniques, in which clusters are compared based on standardized attribute values. Moreover, clusters were labeled based on nomenclatures that uniquely signify segment characteristics to easily identify segments instead of referring cluster index.

## 1.5 Scope and Limitations

### 1.5.1 Scope of the Study

The study focusses on the clustering of mobile customers using CDR data of prepaid mobile customer base. Randomly selected CDR data of mobile customers were targeted as a case study and the clustering results of K-means and EM clustering algorithms were compared for each service usage dataset.

### 1.5.2 Limitations of the Study

The clustering task comprises CDR data of prepaid mobile customers, which account for most of the mobile subscriber customer base. Besides this, in the Ethiopian context the spending potential as well as usage behavior of the postpaid customer is distinct from the prepaid users and hence the study is limited to the clustering of prepaid mobile customers. Moreover, due to the unavailability of detail call records for long periods (above three months) and complexity to preprocess large volume data, the study is limited to three months of CDR data of mobile customers.

## 1.6 Contribution of the Study

The study and its results will help ethiotelecom to identify an algorithm suitable for clustering of customers from each service usage datasets and identify additional attributes besides the monthly spending amount to differentiate customers. This enables the operator to properly recognize the level of differentiability in the mobile customer base for the provisioning of targeted service package options.

Since the CDR dataset is accessible, customer segmentation based on these data source can be easily implemented. Besides, the dataset reflects the actual behavior of the customer and hence mobile service packaging based on such dataset appears to be more customer centric. Moreover, the clustering result will be used as a preprocessing step for further analysis.

## 1.7 Related Works

Various research works conducted in the area of customer segmentation are discussed here. Most of the literatures applied clustering techniques to build customer segments and among the various clustering algorithms K-means, two-step, and SOM clustering algorithms are mainly applied to the dataset to build customer segments. Pertinent works of literature in the marketing domain of the telecom sector are reviewed as follows.

In [31] the study used CDR data of customers to build customer segments for product development and business planning conducted using nine days CDR data of 5,000 subscribers based on service usage and revenue share attributes. Data preprocessing techniques such as correlation analysis was applied to eliminate highly correlated attribute values & two-step clustering algorithm applied to the prepared dataset. The algorithm helps to determine the cluster size automatically. In the paper, the silhouette measure of cluster quality was used to evaluate the cluster quality and the resulting clusters are profiled in terms of cluster size and revenue share. And the cluster index was labeled based on unique names to reflect the group behavior. Based on the cluster results, marketing actions were defined according to each segment's unique behavior. However, the study targeted a small number of customers with a few days of usage information. Besides this, the feasibility of the cluster size was not properly assessed from the business perspective as it is automatically determined by the algorithm.

In [6] the study assessed the call behaviors of customers using three-dimensions called RFM techniques. The objective of the study was to identify the profitability of telecom customers for the provisioning of personalized services. The study applied K-means clustering for its simplicity and ability to handle large datasets. RFM model was used as a dimension reduction technique to minimize the dimension of the input dataset into three attributes named R, F & M and capture different behavior of the customers. In the study, the RFM model was preferred over simply using a single attribute i.e. revenue to determine customer profitability and additional two features were used to identify the customer profitability. For optimal cluster size determination, the elbow method or distortion curve technique was applied. The clusters were compared based on weighted RFM variables. Accordingly, the K-means clustering identified four clusters of different profitability levels helpful for designing of suitable marketing strategies.

In [21] the study used the SOM algorithm on the customer service usage behavior dataset to define targeted marketing strategies and enhance customer satisfaction. Six months period usage behavior data of randomly selected subscribers were used in the study. Ten attributes were used to represent subscriber service usage behavior and the average value is used. Data preprocessing techniques like normalization was used to avoid biases towards large values. Visualization techniques available in SOM clustering such as D-matrix and component planes are applied to easily analyze the clustering results. The study used the auto cluster size determination of the SOM algorithm and K-means clustering technique applied to the component plane to identify the cluster boundary easily. Accordingly, six clusters were identified, and analyzed using the key values of each segment. Besides this, each segment was profiled based on the segment size and the value of

each segment to the operator and compares clusters based on the loyalty rate. The study also highlighted the importance of clustering knowledge to build classification models.

## 1.8 Thesis Organization

The remaining part of the research paper consists of six chapters. *Chapter 2* presents an overview, type and benefits of customer segmentation and its application in the telecom and other sectors. *Chapter 3* introduces data mining techniques, different clustering methods and the detail on the working principle of clustering techniques or algorithms applied to this research. *Chapter 4* presents the detail on problem domain understanding, dataset description and preprocessing of CDR datasets. *Chapter 5* is the core part of the thesis and mainly focuses on cluster size determination, and implementation of clustering algorithms. *Chapter 6* focuses on the result, evaluation and interpretation of clustering results and *Chapter 7* covers conclusion and future work.

# 2 Overview on Customer Segmentation

This chapter discusses the basics of customer segmentation, types of customer segmentation and its importance in business analytics.

## 2.1 Customer Segmentation

Customer segmentation technique is widely used to partition the customer base of the company into different groups and customers within a group should be cohesive but well separated from the other groups [31]. As discussed in [13] segmenting means putting the population into segments based on shared characteristics and the process of segmentation illustrates the group(cluster) characteristic within the data. It is usually used as a data preprocessing step for supplementary study and occasionally as a standalone technique to establish targeted relationship with customers.

## 2.2 Benefit of Customer Segmentation

According to [30,45] customer segmentation is used in many business firms to propose important marketing actions by distinguishing the heterogenous customer base into few groups for targeted service provisioning. As shown in [13] customer segmentation is used to differentiate customers for a better management by designing of tailored marketing actions. It is being used as a decision support tool for the provisioning of new or customized products. Moreover, it is also serving as an analytical tool to establish an effective relationship with various customer groups through a better understanding of customers value to the business and enhancing the insight on behavior. Therefore, customer segmentation helps to understand customer preferences and establish strong relationship with them.

In general, as discussed in [30] the benefits of customer segmentation for business enterprise can be summarized as: to better understand customer, to identify the attractive customer segments, efficiently prioritize resources, form personalized campaign and select the best performing distribution channels.

## 2.3 Types of Customer Segmentation

Customers can be segmented into different groups based on their buying behavior, usage frequency, demography. Hence the basis for segmentation vary with its purpose and application. According to [30] each type of segmentation has its own goals. The most common and well-known segmentation types are summarized in the Table 1:

Table 1 : Types of segementation adapted from [30]

| Segmentation Type | Description | Attributes |
|---|---|---|
| *Behavioral Segmentation* | Grouping based on usage, attitude & behavior regarding a product or promotion. | product ownership, type & frequency of transactions, revenue, payments, and utilization |
| *Loyalty segmentation* | Grouping of customers based on degrees of loyalty to the company or brand. | Loyalty score, frequency of purchases, number of complains, new or old customer. |
| *Socio Demographic* | Demographic or social characteristics. | Age, gender, income, marital status, education & other personal details. |
| *Geographic segmentation* | Group customers based on geographic factors. | Overlap with socio demographic segmentation |
| *Psychographic segmentation* | Group customers according to different degrees of lifestyle, social behavior and personality. | preferences, interests, values & consumers believe, social class, political orientation and personality |

3

# 3 Data Mining Techniques for Customer Segmentation

This chapter mainly discuss about data mining techniques, data mining process, clustering techniques and various cluster evaluation techniques used to assess the quality of the derived clusters. Furthermore, the working principle of the clustering algorithms used in the study are also presented here.

As a data mining task, clustering technique was applied to the CDR data set of mobile customers to build customer segments based on service usage behavior for mobile service packaging purpose. According to [2] CDR is a record that contains detail information about a telecom transaction, such as call start and end time, duration, call parties, cell ID, requested web sites and other related information when a call is placed on a telecommunication network. CDR data is generated when usage event occurs and updated accordingly by retrieving the information from different network elements. CDR data has no class label attributes and hence the clustering technique is used to build cluster labels according to the level of similarity in the dataset. The attributes of CDR data reflect the actual behavior of subscribers and hence segmentation based on such data source are reliable and can be implemented using an easily accessible data source. But the challenge with these datasets are lack of skills and appropriate technologies to analyze voluminous data.

## 3.1 Data Mining Process

Data mining is the process of extraction of important information, rules and patterns from large databases. It helps companies to predict customers' future trends and behaviors using the available data, allows to assess the existing situation and make an informed decision. Data mining tools can solve complex and time-consuming task easily and promptly [47]. As discussed in [13,45] data mining tasks are categorized into classification, clustering, estimation, prediction, defining association rules and data visualization. An informed decision-making process based on data mining is shown in Fig 3.

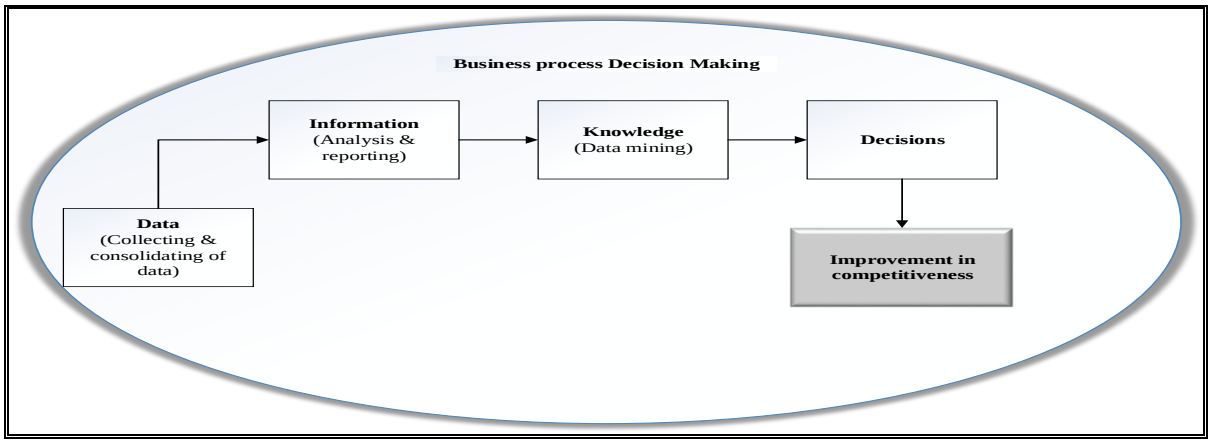



Figure 3 : Business decision making process [Olszak].

## Data Mining Application

In (Mitra, Pal, & Mitra, 2002) data mining is explained as a stage in Knowledge Discovery in Databases (KDD), involving the application of specific algorithms for pattern extraction. It is applied in different areas such as banking and finance, insurance and telecom marketing domains. Data mining brings various techniques together to discover patterns or rules and to construct models from databases. As discussed in (Rygielski et al., 2002), it is a part of CRM and business intelligence to support companies in their effort to become more customer centric. It is also widely applied in the telecommunication sector for the following purpose:

- **Call Detail Record Analysis**: identify customer segments with similar usage pattern by analyzing CDR data of subscribers. It is used to develop attractive pricing and targeted promotions.
- **Customer Loyalty**: in a competitive telecom market, it helps to identify the characteristics of loyal customers and switching customers and their profitability.

## Models for Data mining

The model refers the steps to be followed to meet of the objective of the data mining task. Most of the data mining models corresponds on the common tasks and put the data mining steps as data gathering, data analysis, implementation of results. One of the well-known models is, CRISP (Cross-Industry Standard Process for data mining) was proposed in the mid-1990s by a European consortium of companies to serve as a non-proprietary standard process model for data mining. The general approach of the model is shown in Figure 4.

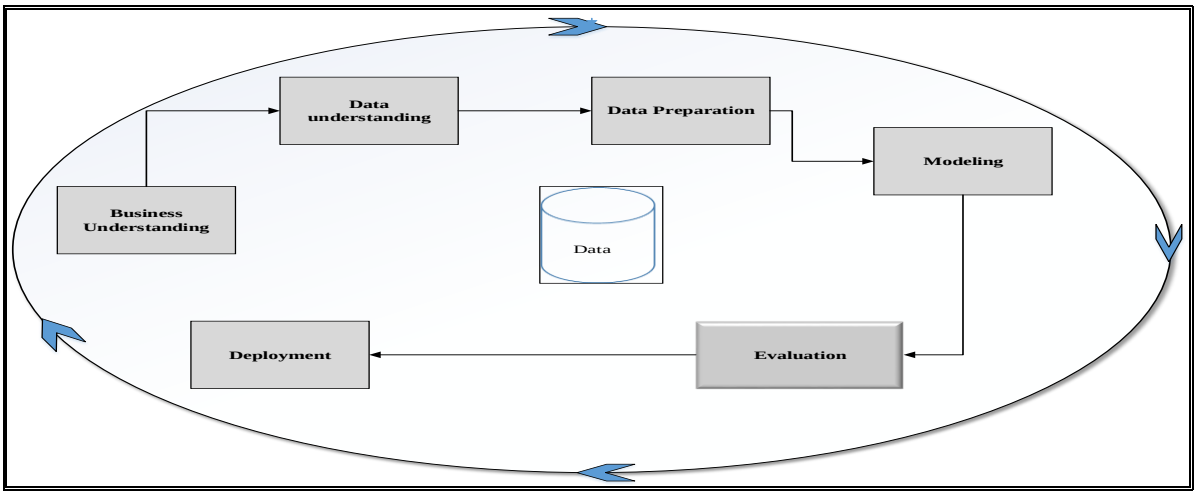



Figure 4: CRISP model adapted from, Larose,2006.

According to Fayyad et al,1997, the other data mining model is the KDD, follows steps such as data collection, data preprocessing, data transformation, data mining, evaluation and knowledge extraction. Both data mining models follow almost the same procedure in terms extracting essential knowledge from dataset. Specially the data preprocessing and evaluation of modeling results are the common tasks to extract knowledge. The KDD data mining model is depicted in Fig 5.

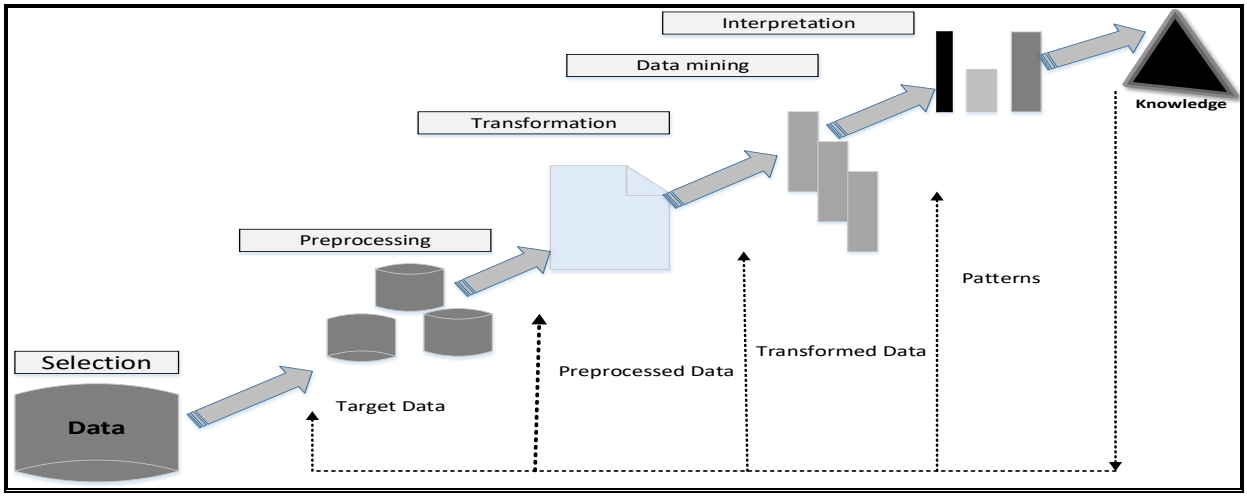



Figure 5 : KDD steps in data mining process (Fayyad et al.19970).

## 3.2 Clustering as a Data Mining Technique

Customer segmentation helps to distinguish the heterogenous customer base into a few manageable groups based on their similarity to arrange targeted communication with customers. According to [4] clustering is

known as unsupervised learning because the class label information is unavailable, and it is widely used for customer segmentation purposes. It categorizes a set of instances into various groups or clusters and objects within a cluster are highly cohesive and dissimilar with the other clusters. Usually, the similarity or dissimilarity is measured based on the distance between attribute values per cluster. As discussed in [4] clustering also called data segmentation used to facilitate the development of marketing actions to improve the relationship with the customer. As a data mining function, cluster analysis can be used as a standalone tool to extract patterns in the dataset and distinguish customers based on their characteristics. On the other hand, it can also serve as a preprocessing step for other tasks such as classification, attribute selection, and characterization.

There are different types of clustering techniques and the next section discusses the type of clustering techniques and the working principle of the clustering algorithms used in the study.

### 3.2.1 Types of Clustering Techniques

Clustering algorithms mainly concentrate on identification of clusters or class labels based on the similarity exist in the dataset. According to [4, p.448] clustering methods are classified into the following categories and discussed as follows:

#### 3.2.1.1 Partitioning Methods

Given a set of $n$ objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$ which means, it splits the data into k groups such that each groups or clusters at least have an instance as a member. This method naturally builds mutually exclusive clusters, or each object must belong to exactly one group. In this method, the criteria of the cluster quality are closeness to the cluster centroid and separation between centroid values for different clusters. The well-known partitioning methods are k-means and k-medoids.

#### 3.2.1.2 Hierarchical Methods

This method builds a hierarchical breakdown of the datasets or instances in a dataset. It can be classified as agglomerative clustering or divisive clustering based on the decomposition type formed. In agglomerative clustering it starts with every instance building a distinct group and successively merge the instances or clusters close to one another to the top level till the termination condition satisfied. However, the divisive

clustering begins with all the objects in the similar cluster then successively in each iteration, a cluster is split into smaller groups, till every instance are assigned to a cluster or a termination condition satisfied.

#### 3.2.1.3 Density-Based Methods

Partitioning methods build clusters based on the distance among objects. It works well when a spherical shaped cluster formed and challenged to discover clusters of arbitrary shapes. But in density-based clustering, it divided instances into mutually exclusive clusters or hierarchy of clusters. The clusters continue to grow in a given cluster if the density (number of objects or data points) near to a given radius exceeds some threshold or minimum number of points. This method is effective in identifying outlier values and clusters of arbitrary shape. Typically, density-based methods consider exclusive clusters only, and do not consider fuzzy clusters. DBSCAN, OPTICS.

Table 2 : Overview on clustering methods adapted from [4].

| Method | General Characteristics |
|---|---|
| Partitioning Methods | • Find mutually exclusive clusters of spherical shape.<br>• Distance based.<br>• May use mean or medoid to represent cluster centers.<br>• Effective for small-medium sized datasets. |
| Hierarchical Methods | • Clustering is a hierarchical decomposition (i.e. multiple levels).<br>• Cannot correct erroneous merges or splits.<br>• May include techniques like micro clustering or consider object linkages. |
| Density-Based Methods | • Can find arbitrary shaped clusters.<br>• Clusters are dense regions of objects in space that are separated by low density regions.<br>• Cluster density: each point must have a minimum number of points within its neighborhood.<br>• May filter out outliers. |

### 3.2.2 Clustering Algorithms Used in The Study

In this thesis partitioning based clustering methods were applied on each dataset to build mutually exclusive assignment of instances to each cluster. This method is efficient in handling small-medium size datasets and as well as numerical attribute values [4]. Because of these and other merits, partitioning based clustering algorithms were used in the study. Among the variants of this clustering method, the performance of K-means and EM algorithms were compared based on the quality clustering results for each dataset. Both algorithms use centroids as representative of the cluster, and this is suitable to compare the clustering results. According to [45] a cluster centroid is derived by the averaging value of the input fields over the number of instances per cluster and it is considered as a prototype or representative of the cluster. Other clustering algorithms such as the hierarchical algorithm look unsuitable for large dataset and lacks flexibility in determining the cluster size. Thus, the cluster comparison was implemented only using the two clustering algorithms. The detail on the working principle of K-means and expectation maximization (EM) clustering algorithms are illustrated in the next section.

#### 3.2.2.1 K-means algorithm

As discussed in [4] K-means is a centroid based method. For example, for a data set, *D*, containing *n* objects in Euclidean space. Partitioning methods distribute the objects in *D* into *k* clusters, *C*1, … ,*Ck*, that is, $Ci \subset D$ and $Ci \cap Cj = \varnothing$ for $(1 \leq i, j \leq k)$. The objective function is to maximize within cluster similarity and minimize between cluster similarity. As a partitioning method, K-means uses the *centroid* of a cluster, *Ci*, to represent that cluster. The difference between an object $p \in Ci$ and *ci*, the representative of the cluster, is measured by *dist* (*p*, *ci*), where *dist*(*x*,*y*) is the Euclidean distance between two points *x* and *y*. The quality of cluster *Ci* can be measured by the within cluster variation, which is the sum of *squared error* between all objects in *Ci* and the centroid *ci*, defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} dist(p, ci)^2 \qquad (3.1)$$

Where, E is the sum of the squared error for all objects in the data set, p is the point in space representing a given object, and ci is the centroid of cluster *ci* (both p and ci are multidimensional). Instead, for each object in each cluster, the distance of an object from cluster centroid is squared, and the distances are summed. The objective function tries to make the resulting k clusters as compact and as separate as possible. The working principle of *k* -means algorithm is depicted in Fig 6.

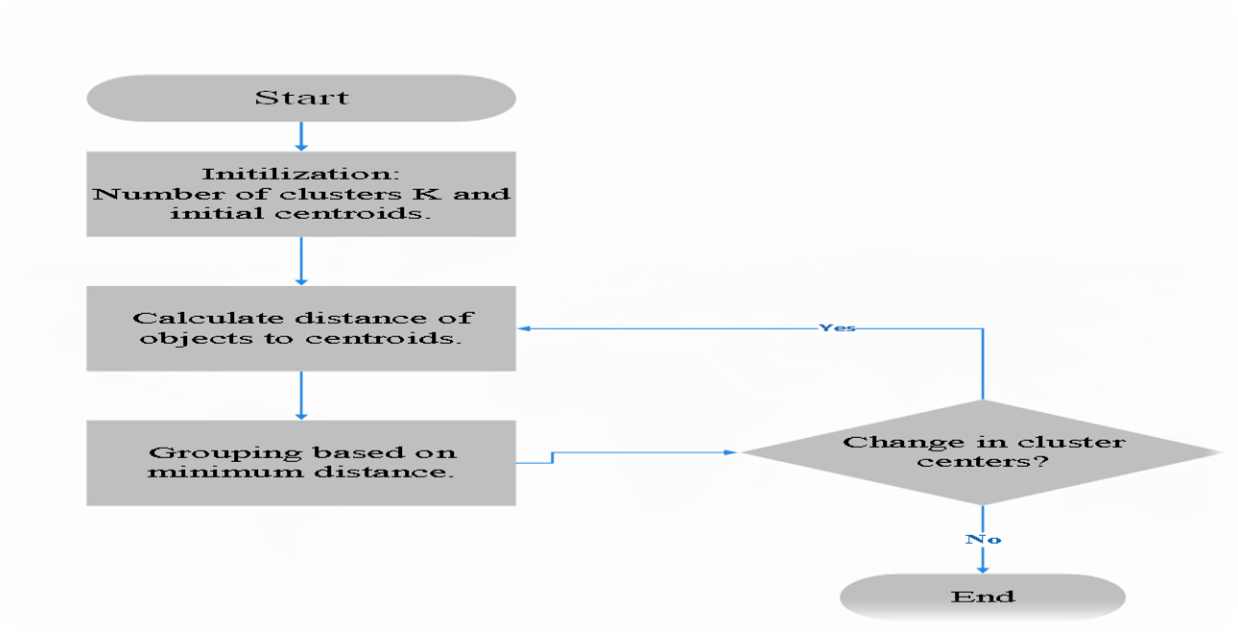



Figure 6 : K-means clustering algorithm, adapted from [4].

### 3.2.2.2 Expectation-Maximization (EM) Algorithm.

According to [12] EM algorithm is an iterative procedure to compute the maximum likelihood (ML) estimate of data distribution usually for incomplete or missing datasets. In estimating the maximum likelihood, the objective is to estimate the model parameter for which the data distribution is most likely. As presented in [12,34], each iteration of EM algorithm has two steps. i.e. the E-step and the M-step. The algorithm begins with estimating an initial parameter. Then, an *expectation* step is applied where the known data values are used to compute the expected values of the unknown data. While in the *maximization* step where the known and expected values of the data are used to generate a new estimate of the parameters. The expectation and maximization steps are run iteratively till convergence. The detail working process is described as follows:

**E (Expectation step) step** - is responsible to estimate the probability of each element belong to each cluster $P(C_J \mid x_k)$ . Each element is represented by an attribute vector $x_k$. The relevance degree of the points of each cluster is given by the likelihood of each element attribute in comparison with the attributes of the other elements of cluster $C_J$.

$$P(C_J \mid x_k) = \frac{|\sum_j(t)|^{-\frac{1}{2}} \exp^{nj} P_j(t)}{\sum_{k=1}^{M} |\sum_j(t)|^{-\frac{1}{2}} \exp^{nj} P_k(t)} \quad (3.2)$$

*Where:*

X is the input dataset.

M is the total number of clusters.

t is an instance and initial instance is zero.

**M (maximization) step –** in this step the parameters of the probability distribution of each group estimated for the next step. First it computes the mean (μj) of class j obtained through the mean of all points in function of the relevance degree of each point. The covariance matrix for each iteration is computed by using Bayes theorem. The probability of occurrence of each class is computed through the mean of probabilities (C_j) in function of the relevance degree of each point from the class. The mathematical formula for the EM algorithm is:

$$P_j(t+1) = \frac{1}{N} \sum_{k=1}^{N} P(C_j \mid x_k) \quad (3.3)$$

Where:

X is input dataset

N is the total number of clusters

t is an initial instance and initial instance is zero.

The general flow of the expectation maximization algorithm is diagrammatically represented as follows:

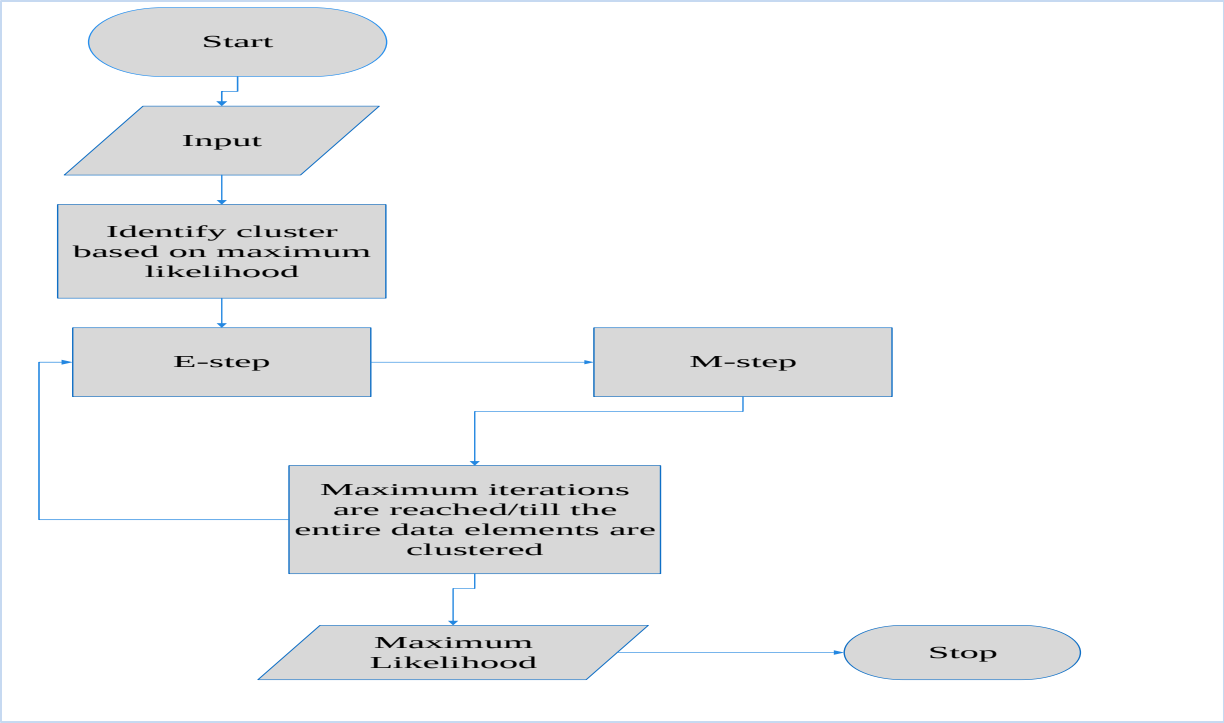



Figure 7 : EM clustering algorithm, adapted from [34].

## 3.3 Cluster Interpretation

In [45] cluster interpretation is used to understand the resulting clusters and it comprises tasks such as identification of the differentiating characteristics of each cluster through profiling using descriptive statistics and charts. In general, cluster profiling mainly includes tasks like:

- *Compare clusters using clustering fields* – compare clusters based on centroid value to identify attributes that differentiate each cluster.
- *Comparison of clusters with respect to other KPIs* – the cluster can be profiled using external fields not used for clustering such as KPIs and demographic variables.
- *Visual exploration* – visualization techniques such as charts and scatter plots are important to easily identify the most differentiating fields.
- *Labeling the segments* - profiling and interpretation usually conclude by labeling of the clusters with an informative nomenclature that reflect the unique characteristics of the segment.

## 3.4 Cluster Result Validation or Evaluation Techniques

As shown in [51], the evaluation of clustering results is not a developed process because clustering is usually used as a data preprocessing step for other data mining tasks. Moreover, each clustering algorithm define its clusters and there is a lack of commonly applied evaluation techniques. According to [2] cluster validation refers to evaluating the appropriateness of the clusters derived from the dataset and conduct various experiments by altering parameters for a better cluster solution. Additionally, cluster validation also indicates determining the appropriate number of clusters. As shown in [4,51] cluster evaluation measures are categorized as follows:

1. **Unsupervised –** measures a clustering structure with respect to internal information like SSE.
    - ***Cluster Cohesion*** - measures cluster compactness or tightness and determine how closely related the objects in the cluster are.
    - ***Cluster Separation*** – measures the distance or separation between clusters.
2. **Supervised** - measures the extent of similarity between the cluster result of an algorithm and the external structure. It evaluates the level of similarity between the cluster label with externally supplied class labels. It uses information outside of the dataset or out of the clustering attributes.
3. **Other methods**- methods used to assess the cluster result using qualitative cluster evaluation techniques. According [45] cluster evaluation should also be supported by qualitative cluster evaluation criteria. Such as:
    - *Measurability*: the characteristics of the segments can be measured and identified.
    - *Substantiality*: the segments should be large and profitable enough, worth to invest and feasible to reach with a tailored marketing program.
    - *Accessibility*: the segments should be meaningful to the company to be effectively reached and served.
    - *Differentiability*: each segment needs to be distinguishable from others and to respond differently to a marketing program.
    - *Actionability*: effective marketing programs can be carried out for the selected segments by considering the objectives and resources of the company.
    - Stability: is a key factor to be considered while evaluating the appropriateness of a segmentation model.

# 4 Business Domain Understanding and Data Preprocessing

Data mining commonly starts with an understanding of the problem domain and assessment of the gap in the problem area to be addressed by the data mining task [45]. The key tasks accomplished for domain understanding embrace document analysis, analysis of mobile service package development procedures or templates and discussion with marketing experts. Then it is followed by identification or acquisition of datasets and attributes essential for usage-based segmentation. The major tasks in the chapter are presented in the following subsections.

## 4.1 Understanding of Business Domain

Understanding of the problem domain starts by doing critical activities like defining the business objectives, selecting the segmentation criteria and determining the segmentation population [45]. Telecom marketing assumes key responsibilities such as influencing customer usage behavior by offering mobile service packages. The objective of offering these packages is for efficient utilization of network resources, improve customer satisfaction and enhance the company revenue. In order to meet these objectives, the operator should apply an efficient technique of identifying the level of differentiability in the customer base to enhance the insight on the customer and propose segment targeted services to customers. However, the Ethiopian sole mobile operator still concentrates on mass marketing strategy with little implementation of targeted marketing. Identifying customers with a single dimension or attribute in the current segmentation approach has limitations in fully extracting insight about customers' needs, preferences, and usage behaviors. However, mobile service package provisioning demands an effective assessment of the usage behavior of subscribers in terms of preferred usage time, service preference and spending patterns of customers.

In this regard, CDR datasets comprise important features that reflect the actual usage behavior of subscribers and hence the application of advanced clustering techniques is appropriate for handling large datasets with many features and build the customer segments. The goal of segmentation is to group customers into a few and manageable segments in order to design feasible marketing actions. The overall mobile service package development process and its purpose are shown in Fig 8.

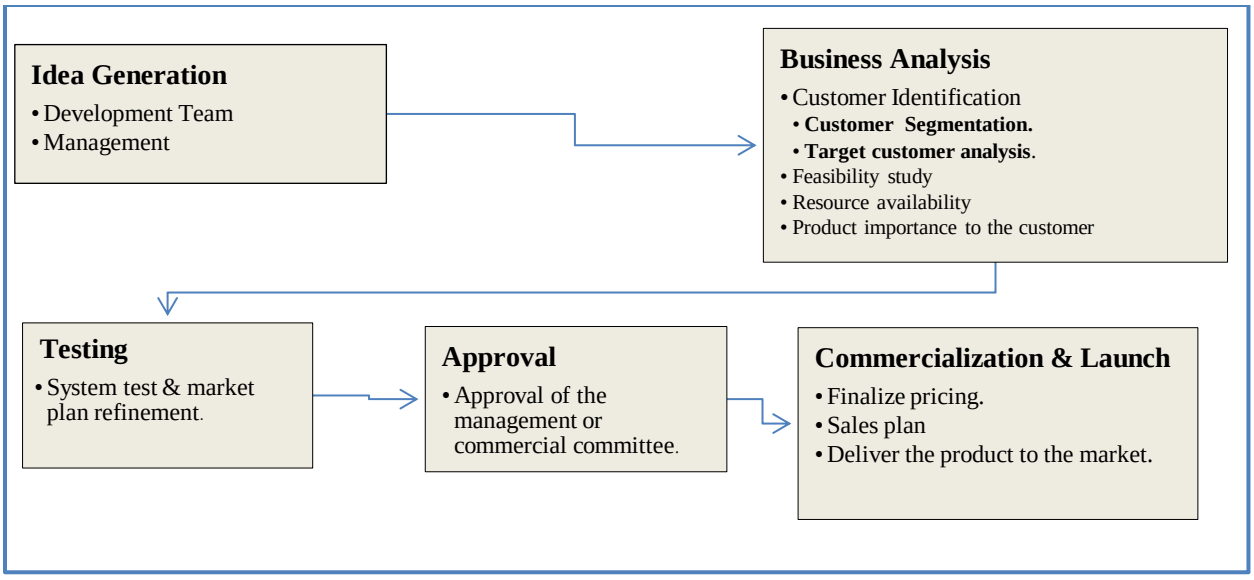



Figure 8 : ethiotelecom marketing service package development process.

As presented in Fig 8, customer segmentation has been applied in the service packaging process to identify different customer groups based on the similarity in behavior. The data sources of the segmentation were mainly from survey data and summarized subscriber database data sources such as spending amount.

### 4.1.1 Purpose of Customer Segmentation in ethiotelecom

Different segmentation techniques have been applied for various purposes such as enhancing the customer insight and implement various marketing actions on the key marketing pillars of (Product, Pricing, Places/sales channels, Promotion also called 4Ps). Though it is traditional and generic, customer segmentation has been predominantly applied to distinguish customers into different segments for the purpose of developing new products and customization of existing products.

### 4.1.2 Gaps in the Existing Segmentation Method.

The gaps in the existing customer segmentation approach was assessed from different perspectives such as data sources used, segmentation techniques followed and flexibility or simplicity of the approach. It is summarized as follows:

- Customer segmentation has been based on the easily inaccessible dataset and hence it is difficult to conduct segmentation easily to assess the impact of new offers (mobile service packages) on the customer service usage behavior. In addition, survey data collection is time-consuming and expensive.

- The dataset and features used for customer segmentation are insufficient to reflect the actual usage behavior of the customer.
- The segmentation type applied for the mobile service packaging process are less applicable. According to [45] behavioral or usage-based segmentation is usually preferable for product development since data sources for such type of segmentation can be directly accessible from the company data warehouse or data store.

## 4.2 Initial Data Understanding

In this study, mobile subscribers CDR data was used to build usage-based customer segments. CDR data reflects the actual subscribers' usage behavior and mobile service packaging demand understanding of the customer usage behavior, preference and potential. Attributes used as an in input to the clustering algorithm were related to subscribers' service usage amount, usage frequency, usage day and time, spending amount, week to week usage plan, ...etc. In the study, attribute values were aggregated or summarized in a record to indicate the usage behavior of subscribers. Hence information about each subscriber was summarized using sum, average, frequency, and ratio. The main tasks in this subsection are discussion on the data acquisition, data set description and data quality verification.

### 4.2.1 Data Acquisition

CDR data was extracted from the BICP platform of the CRM system. The database consists of a different table on customer profile, customer type, usage time, usage amount, charge amount and other information of the subscribers. The customer profile was aggregated with CDR information of each subscriber during extraction and the service number of the subscriber was used to aggregate detail customer usage information and uniquely identify the customers.

For this study, initially around 39,000 unique subscribers CDR data for a month period was collected. But acquiring detail data on many subscribers for above a 1-month period was impossible and as a result, CDR data of samples customers were selected randomly. The sample was selected carefully to include customers from each group based on spending amount. Then, using Weka resampling technique random sample of 7,502 subscribers were chosen and finally three months CDR data for the sample size was collected based on the attributes prepared for the data request template. The data was saved from the source in a CSV file format and converted to an excel file format for preliminary data preparation.

### 4.2.2 Data Set Description

Data understanding is a critical step in the data mining process to grasp broad information about the dataset. Visualization and statistical techniques were applied to understand the behavior of the dataset and the distribution of customers. Weka has visualization options to understand the data using the histogram of each attribute. Statistical techniques like minimum, maximum, mean, variance and standard deviation of attribute values were also helpful in illustrating the dataset at least at a conceptual level. Moreover, the researcher's experience in the domain was advantageous in identifying datasets and attributes relevant to usage-based customer segmentation. Fig 9 shows the distribution of customers for a sample of voice dataset attribute value.

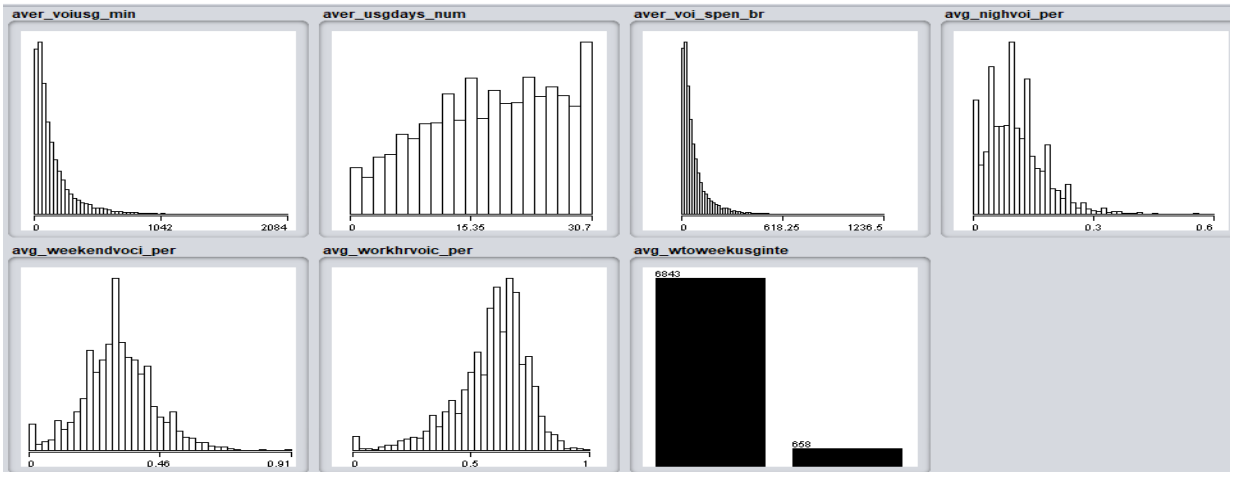



Figure 9 : Histogram of distribution of instances for each attribute -voice dataset.

Figure 9 demonstrates the frequency of instance for each clustering attribute value and the differentiating capability of variables to be used for clustering.

The dataset was extracted and prepared separately for each service type of voice, SMS and internet usage. Since the CDR data consist of detail usage information, the data was aggregated based on service number and additional attributes were constructed from the raw data to enrich the data quality. Therefore, the final attributes used as an input for clustering purpose were briefly described as follows:

#### Voice Usage Dataset

The voice dataset comprises attributes that indicate the voice service usage behaviors of subscribers. Attributes used for clustering of customers based on voice usage were *voiusg_min, avervoic_usgdays,*

*aver_voicspend_br, nightvocie_perc, workinghourvoic_perc, ave_weekendvoicusg_perc and avg_wtoweekusgint*. These attributes with other clustering input parameters such as cluster size and seed value were provided to distinguish customers from the voice service usage dataset. The cluster size was determined according to the cluster size determination step mentioned in Section 5.6. The description of attributes used in the voice usage datasets are:

Table 3 : Description of voice service usage dataset.

| Attribute | Data Type | Description | Unit | Purpose |
|---|---|---|---|---|
| **voiusg_min (VUM)** | Numeric | Monthly average voice service usage amount. | Minutes | Usage amount in minute. |
| **usgdays_num (Udys)** | Numeric | The average # of days of voice service usage per month. | Number | Usage Frequency |
| **voi_spen_br (VSpBr)** | Numeric | The average monthly spending on voice service. | Br. | Spending on voice. |
| **nighvoi_per (NvP)** | % | The average of percentage of night voice calls per month. Usage during night hour. (0:00 – 7:00) | % | The proportion of night (0:00-7:00) voice call from their total minute. |
| **weekendvoci_per (WvP)** | % | The percentage of weekend voice calls per month. Usage during weekend days. (Saturday & Sunday) | % | The proportion of weekend (Sat. & Sun.) voice call from their total minute. |

| | | | | |
|---|---|---|---|---|
| **workhrvoic_per (WoVp)** | % | The average of percentage of calls made during working hours of the day. Usage during working hour. (9:00– 18:00) | % | The proportion of working hour (9 -18) voice call from their total minute. |
| **wtoweekusginte_voi (avg_wtoweekugint)** | String | The average week to week voice service consumption plan from the total month usage. | Yes/No | The proportion of usage in a given week from the monthly usage. |

SMS Usage Dataset

The SMS dataset contains attributes insightful of the SMS usage behaviors of subscribers. The attributes used for the clustering of customers based on SMS service usage were *tot_smsusg_num, sms_usgdays, tot_sms_spend_br, nightsms_perc, weekendsms_per, workinghrsms_per*. These attributes with other clustering input parameters such as cluster size and seed value are provided to establish customer segments based on SMS service usage behavior. The attributes values of SMS usage dataset are described as follows:

Table 4 : Description of SMS service usage dataset.

| Attribute | Data Type | Description | Unit | Purpose |
|---|---|---|---|---|
| **tot_smsusg_num (SMSUN)** | Numeric | Monthly average SMS service usage amount. | Number | SMS Usage amount in number. |
| **sms_usgdays (SMSUdys)** | Numeric | The average # of days of SMS service usage per month. | Number | Usage Frequency |
| **tot_sms_spend_br (SMSSpBr)** | Numeric | The average monthly spending on SMS service. | Br. | Spending on SMS. |
| **nightsms_perc** | % | The average of percentage of night SMS calls per month. | % | The proportion of night (0:00-7:00) |

| | | | | |
|---|---|---|---|---|
| (NSMSP) | | | | SMS call from their total number. |
| **weekendsms_per (WSMSP)** | % | The average of percentage of weekend SMS calls per month. | % | The proportion of weekend (Sat. & Sun.) SMS call from their total number. |
| **workinghrsms_per (WoSMSP)** | % | The average of percentage of SMS sent during working hours of the day. | % | The proportion of working hour (9 -18) SMS call from their total number. |
| **wtwusginten_sms** | String | The average week to week voice service consumption plan. | Yes/No | the proportion of usage in a given week from the monthly usage. |

### Internet Usage Dataset

This dataset contains attributes that reflect the internet usage behaviors of subscribers. Attributes used clustering of customers based on internet usage are *aveintusg_MB, aver_Inter_usgdays, aver_intspend_br, ave_weekendintusg_perc, avg_wtoweekusgint*. These attributes with other clustering input parameters such as cluster size, seed size was provided to establish customer segments based on internet usage behavior. The abbreviation of the values of the attributes of internet usage dataset are described as follows:

Table 5 : Description of internet service usage dataset.

| Attribute | Data Type | Description | Unit | Purpose |
|---|---|---|---|---|
| **Intusg_MB (IUMB)** | Numeric | Monthly average internet service usage amount. | MB | Internet Usage amount in MB. |
| **Inter_usgdays (Udys)** | Numeric | The average # of days of internet service usage per month. | Number | Usage Frequency |

| | | | | |
|---|---|---|---|---|
| **Intspend_br (ISpBr)** | Numeric | The average monthly spending on internet service. | Br. | Spending on Internet. |
| **weekendintusg_perc (WIntP)** | % | The average of percentage of weekend internet usage per month. | % | The proportion of weekend (Sat. & Sun.) Internet usage from their total usage. |
| **wtoweekusg_int (avg_wtoweekugint)** | String | The average week to week internet service consumption plan. | | the proportion of usage in a given week from the monthly usage. |

### 4.2.3 Data Quality Verification

The CDR data was extracted from the BICP system as per the data request template prepared by the researcher. However, minor computational errors and inconsistency on some attribute values were addressed using manual computation by multiplying the usage amount with the tariff amount. Besides, attributes irrelevant to the study were manually excluded from the dataset and instances with incomplete or missing attribute value were excluded from the sample. The data quality was further enhanced by constructing additional attributes to the original datasets.

## 4.3 Data Preparation

Data gathering methods are loosely controlled and might result in out of range values, wrong combinations, missing values, and computational errors. Data quality highly affects the quality of the derived results and hence the role of data preparation is vital in determining the quality of the results [33]. For data preprocessing and visualization of clustering results tools such as Ms Excel, IBM SPSS and Weka were used. The explorer environment of Weka comprises various features supportive of the data preprocessing tasks. The key data preprocessing tasks supported by the filtering option of the tool includes resampling, feature selection, dimension reduction, supervised and unsupervised standardization, discretization, correlation analysis, …etc.

A considerable amount of time was spent for data processing and the main data preprocessing tasks applied to prepare the data were presented as follows:

### 4.3.1 Data Cleaning

Data can be incomplete or lacking attribute values or containing errors, outlier and inconsistent values [33,45]. Though it is possible to extract meaningful patterns from outlier values however, in this case, the number of subscribers with extremely high attribute values was insignificant and hence these values were manually excluded from the dataset. Thus, the data cleaning task was applied to the original dataset before constructing additional features. The tasks in this stage include removing incomplete, erroneous, unnecessary and redundant values manually.

### 4.3.2 Feature Construction

The original CDR data holds a few attributes directly used for clustering and hence additional fields were constructed to further explain the customer behavior and improve data quality. It was constructed by using sum, average, and percentage or ratio on the original dataset and additional attributes such as monthly spending, usage amount, usage percentage were constructed.

### 4.3.3 Data Integration

The dataset used for customer segmentation in this study was extracted from a single source and the data integration process was accomplished during the data extraction stage. The database expert integrated data from the different tables such as customer base, customer profile, usage and charging system to generate the requested data using the service number as a unique identifier.

### 4.3.4 Data Aggregation

To extract meaningful patterns, data aggregation task was applied to summarize the average value of each attribute per subscriber in a single record. The value for each attribute indicates the average of the three months value and hence every usage information about a subscriber is completed in a record.

### 4.3.5 Data Transformation

Data transformation means transforming the data in a way suitable for the data mining process. According to [33] the key data transformation tasks include normalization, standardization, binning, data aggregation, … etc. Standardization of values helps to adjust large difference in measurement values to be within a small

range and minimize the effect of extremely different values on an algorithm [2]. Standardized values are utilized to compare clusters using visualization techniques and identify the differentiating attributes. The attribute values were standardized or normalized using Weka preprocessing standardization technique, which was applied to numerical attribute values using the z-score standardization technique. It is mathematically described as follow:

$$\frac{x-\mu}{\sigma} \qquad (4.1)$$

*Where*:

$x$ – The attribute value of an instance in the dataset.

$\mu$ - The mean value of a given attribute.

$\sigma$ - Standard Deviation

Moreover, cluster centroid values were normalized to visualize separation between cluster centroid values for each attribute simultaneously or to plot separation between attribute values per cluster in a single figure.

### 4.3.6 Attribute Selection

CDR data has no predefined class label to apply supervised attribute selection filter in Weka and hence the attribute selection process was conducted during the data acquisition process to include essential attributes for usage-based clustering. But EDA or visualization technique was applied to differentiate the customer based on each attribute value. As Fig 9 shows, visualization technique like histogram slightly displays the distribution of instances per attribute and it helps to consider attributes that can be used to distinguish customers. Weka has a visualization option that indicates the difference in the usage behavior of the subscriber.

### 4.3.7 Data Formatting

Data formatting was used to convert the file format of the dataset into a format suitable for the data mining tool. Weka 3.8.3 is the main data mining tool used for clustering implementation. The tool supports data of different formats such as CSV or ARFF file format. Nevertheless, Weka prefers the ARFF file format. Thus, the CSV file format is converted into ARFF using the tool menu of Weka.

# 5 Usage Based Clustering Method and Implementation

In this study, K-means and EM algorithms were implemented to each service usage dataset to build usage-based customer segments. Visualization and cluster profiling techniques were utilized to interpret the cluster results. For the purpose of evaluation of the clustering results and to compare the performance of clustering algorithms, different techniques were applied such as visualization, distribution of instances per cluster and separation between cluster centroid. The major tasks in this chapter are summarized as cluster size determination, comparison of the results of clustering algorithms, identifying cluster evaluation metrics and cluster labeling.

## 5.1 Data Mining Tool Utilized in the Study

In this study Weka 3.8.3 is used for data preprocessing as well as implementing the clustering techniques to CDR dataset. It supports various data mining tasks such as preprocessing, clustering, classification, association rule, and visualization. The tool has suitable filtering options for data preprocessing purposes with both a GUI and a command-line interface to support various clustering tasks. Besides this additional packages or algorithms can be installed from the online repository. The tool was selected for clustering implementation due to compatibility with the existing operating system and support for a wide variety of clustering algorithms & data preprocessing capability.

## 5.2 Implementation Conditions

The clustering experiment was conducted on a separate voice, SMS, and internet usage datasets to build customer segments. Attributes used for the clustering purpose were related to service usage amount, usage time, usage day, usage frequency, spending amount and week to week usage plan of subscribers. The conditions to be satisfied for the implementation of clustering algorithms to each service usage CDR dataset are:

- The cluster and seed size are used as an input parameter for each algorithm to build the clusters and different experiments should run for each dataset to identify the optimal cluster and seed size combination. Moreover, the same cluster size should be used in both K-means and EM algorithms.
- The performance of clustering algorithms should be compared for the same input parameters such as cluster size, seed value, and other important parameters. The cluster quality is evaluated based on the cluster evaluation techniques discussed in Section 5.8.
- Weka 3.8.3 is used for clustering implementation as well as data preparation with Ms-excel, and IBM SPSS for data preparation. The installation of the data mining tool was done on a laptop with the specification of Window 10,64-bit OS, core i7 with CPU of 1.9 GHz and 8 GB RAM.

## 5.3 Clustering Methods

Weka supports various clustering algorithms. In this study, the performance of K-means and EM clustering algorithms were compared in terms of the quality of clustering results in differentiating customers based on service usage behavior. These algorithms were selected because of the similarity or flexibility in determining the required input parameters such as cluster size. Both algorithms use centroids as a cluster representative and have lower execution time as well as memory requirements. Besides, both algorithms can handle a combination of numeric and categorical attribute values. The detail on the working guides of K-means and EM algorithms is discussed in Section 3.2.2.

## 5.4 Assumptions and Input Parameters

- The customer segments or the clusters should have well-separated centroid values for most of the attributes and clusters with similar centroid values for most of the attributes indicate an overlapping cluster. Moreover, customer class labels should be unique or non-overlapping.
- Customers in a segment or a cluster are assumed as homogenous to be treated under a common service package offering regardless of variation within a cluster.
- The distribution of instances per cluster should be large and feasible enough to propose viable service packages and hence clusters with very small or large instances are not feasible for the study. As discussed in [53] the ratio of largest to the smallest cluster should be small to be good.
- Finally, the principal inputs for the clustering algorithm are preprocessed data set, clustering attributes, seed value, and cluster size.

## 5.5 Cluster Class Labeling

Cluster labeling aims to represent each cluster uniquely by a name representative of each segment behavior. The class labels were allocated to each cluster based on a unique feature and it helps to easily identify the derived cluster, compare cluster assignments and to make cluster analysis easily [45]. So, identifying customers by a class label is easier than using the clustered index.

The usage amount attribute together with an additional unique feature was used to uniquely identify the cluster through cluster labeling. Standardized values of each attribute were used to build graphs and label clusters based on the unique features that differentiate the cluster. Attribute values were compared based on the distance from the total dataset mean value. Hence, a value above the mean of the total dataset indicates a very high or high value based on the bar length. While values below the mean of the total indicate very low or low values. So, the length of the bar indicates the magnitude or deviation from the total mean value.

## 5.6 Cluster Size Determination.

Identifying the optimal cluster size is a critical task. One of the key input parameters in the clustering process for the algorithms were the cluster size, which determines the number of customer segments to derive from the dataset. Since the mobile service packages are offered based on the differentiability of the clusters, the cluster size should not be too small or too large. At the same time, the distribution of instances per cluster should be substantial to design a package. As discussed in [22,53] the distribution of instance per cluster should be substantial or large and profitable enough to invest and hence the cluster ratio (largest to the smallest cluster) recommended to be in 2:1.

K-means clustering algorithm requires the cluster size and seed value (data points) as a compulsory input to derive clusters and hence identifying the optimal K needs conducting different experiments. For different values of K, the clustering algorithm resulted in clusters with distinct instance distribution, iteration, execution time and within-cluster SSE values. On the other hand, the EM clustering algorithm has both manual and auto clustering options. The auto clustering option was used to decide on the maximum number of clusters that could exist in the dataset. In the auto cluster size determination option, the cluster size was determined by the clustering algorithm while for the manual option, it needs priori information on cluster size to insert manually.

In this study, the elbow method or distortion curve technique was used to identify the optimal K value in the dataset, which depicts within-cluster SSE for each cluster. SSE value shows the variance within a cluster and a cluster size with minimum within-cluster SSE value is preferable. As presented in [2] elbow criterion is a rough rule of thumb to determine the number of clusters to be chosen. Depending on this method, increasing cluster size beyond a certain value has little impact on reducing the SSE value and the point where the marginal gain is low is called the optimal point. But in the service usage CDR dataset, the method was not free from doubt to identify the elbow point easily. So, the cluster size derived from the auto clustering option was used to set a threshold K value and it receives the judgment of the researcher to decide the elbow point by considering the clustering purpose.

## 5.7 Clustering of Customers Based on Service Usage Dataset

The clustering algorithms were applied to each service of voice, SMS and internet usage datasets to analyze the differentiability in service usage behaviors of customers. The sample size and number of attributes for each service usage dataset were described as follows:

- Voice dataset has 7,501 unique instances and seven attributes.
- SMS dataset has 7,501 unique instances and seven attributes.
- Internet dataset has 7,501 unique instances and five attributes.

### 5.7.1 Clustering of Voice Usage Dataset

#### Cluster Size Determination on Voice Dataset

Based on the cluster size determination technique mentioned in Section 5.6, from different trials of K-means algorithm to voice usage dataset, the cluster size K=5 and seed value of 1000 provides in minimum within-cluster SSE. Therefore, it is identified as the optimal cluster size for voice usage dataset. Figure 10 shows results of the elbow criterion for voice usage dataset.

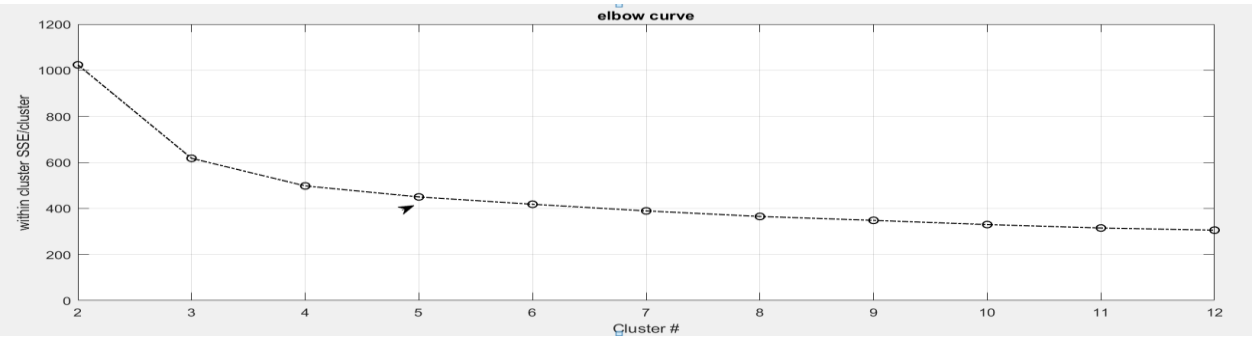



Figure 10 : Elbow curve of voice dataset

The auto cluster size determination option of the EM algorithm was applied to the same dataset and identified about 13 overlapping clusters indicated by very close centroid values for most of the attributes. Based on the qualitative cluster evaluation techniques mentioned in Section 5.8, the cluster size was vast and inefficient to design differentiated voice service packages. Therefore, the auto cluster size result of the EM algorithm was abandoned, and the cluster size determined from the elbow curve as K=5 will be used as an optimal cluster size for voice dataset. As Figure 10 indicates, increasing cluster size beyond K=5 has little impact in reducing the marginal within cluster SSE and it results in overlapping clusters indicated by similar centroid values. Therefore, highly cohesive (indicated by low SSE) and separated clusters can be derived from the voice usage dataset when K=5.

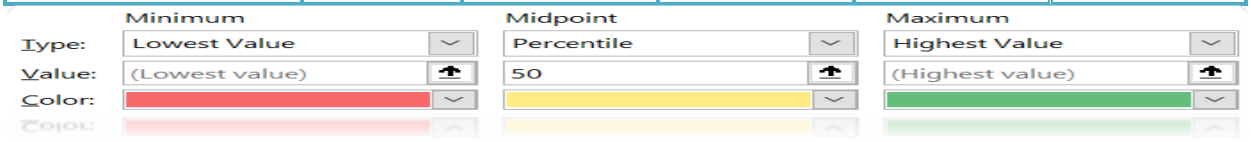
To build the customer segments, the clustering attributes with cluster size and seed values were provided to each clustering algorithms. The detail on the working principles of both algorithms were addressed in Section 3.2.2.

### K-means Clustering on Voice Dataset.

Based on the cluster size, seed value and other input parameters the clustering results of the K-means algorithm are presented in Table 6, with a heat map table comprising centroid values. A heat map is a table in excel used to show the difference between values using colors. The green color indicates "Highest" values and red color indicates the "lowest" values. The clustering result consists of numeric centroid value for each attribute per cluster, distribution of instances per cluster, time taken to build the model, the number of iterations done to complete the execution and within-cluster SSE. But in this study, the most important values to compare the cluster solution were the distribution of instances, within-cluster SSE and attribute centroid value. These metrics used to compare and evaluate the quality of the clustering results.

Table 6 : Behavioral distribution of voice service usage - K-means Clustering.

| | Cluster # | | | | |
|---|---|---|---|---|---|
| | 0 | I | 2 | 3 | 4 |
| Attributes | 658(9%) | 1867(25%) | 1206(16%) | 2273(30%) | 1497(20%) |
| VUM | 92.4 | 342.4 | 53 | 128.9 | 56.9 |
| Udys | 12 | 27.8 | 6.9 | 20.9 | 11.9 |
| VSpBr | 44.7 | 166.7 | 25.9 | 62.6 | 27.6 |
| NvP | 0.11 | 0.12 | 0.06 | 0.11 | 0.1 |
| WvP | 0.33 | 0.3 | 0.18 | 0.33 | 0.37 |
| WoVp | 0.63 | 0.61 | 0.36 | 0.62 | 0.65 |
| avg_wtoweekugint | Yes | No | No | No | No |




As Table 6 demonstrates, from the total instances 30% of customers were assigned to C3 and 25% were assigned to C2 while 9% assigned to C0. The separation between clusters is indicated by colors for each clustering field (green color-highest values & red- lowest values). According to the heat map table, customers in Cluster1 had the highest VUM, NvP & Udys values. On the other hand, customers in Cluster2 had the lowest VUM, Udys & other values. Customers in C0 consume above 50% of their voice service in a given week of the month.

### Expectation-Maximization (EM) Algorithm

To compare the performance of the clustering algorithm, the same datasets and input parameters used for K-means were applied to the EM algorithm to build customer segments and Table 7 shows the clustering result of the EM algorithm.

Table 7 : Behavioral distribution of voice service usage - EM clustering.

| | Cluster # | | | | |
|---|---|---|---|---|---|
| Attributes | 0 | 1 | 2 | 3 | 4 |
| | 2052(27%) | 1471(20%) | 1402(19%) | 856(11%) | 1720(23%) |
| VUM | 56.3 | 225.6 | 17.6 | 536 | 117 |
| Udys | 15.3 | 23.9 | 6.2 | 26.8 | 20.4 |
| VSpBr | 27.3 | 109.6 | 8.5 | 261.4 | 56.8 |
| NvP | 0.11 | 0.11 | 0.08 | 0.11 | 0.11 |
| WvP | 0.32 | 0.31 | 0.26 | 0.29 | 0.32 |
| WoVp | 0.6 | 0.6 | 0.5 | 0.59 | 0.61 |
| avg_wtoweekugint | No | No | No | No | No |

As Table 7 demonstrates, 27% of customers were assigned to C0 and 11% were assigned to C3. All clusters had a proportionate week to week usage plan. According to the heat map table, customers in Cluster3 had the highest VUM, Udys & NvP values. On the other hand, customers in Cluster2 had the lowest VUM & Udys.

### 5.7.2 Clustering of Internet Usage Dataset

#### Cluster Size Determination on Internet Dataset

Based on the cluster size determination technique mentioned in Section 5.6, from the different trials of the K-means algorithm to the Internet usage dataset, the cluster size K=6 and seed value of 100 resulted in minimum within-cluster SSE and hence it is identified as the optimal cluster size for internet usage dataset. Fig. 11 shows the results of the elbow criteria for internet usage dataset.

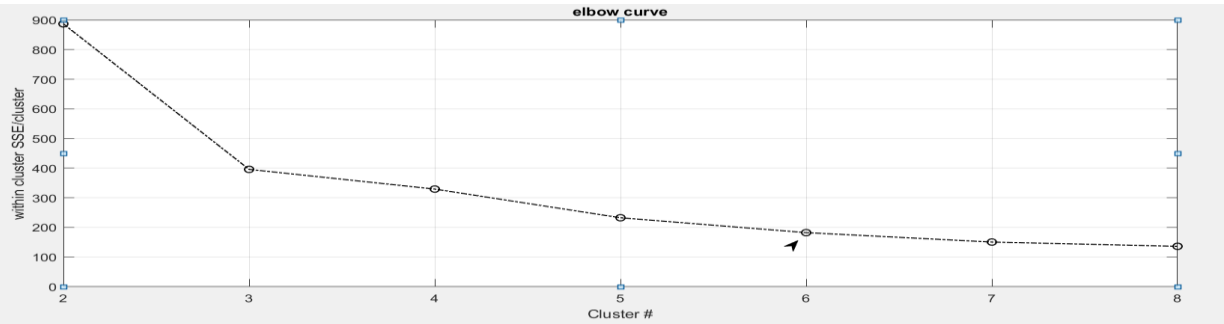



Figure 11 : Elbow curve of internet dataset.

The auto cluster size determination option of the EM algorithm was also applied to the same dataset and identified about 8 clusters of overlapping centroid values. It is closed to the results obtained from the elbow method. But for a better cluster quality with dissimilar centroid values, the auto cluster size of the EM was rejected, and the cluster size determined from the elbow curve at K=6 used as an optimal cluster size for the internet dataset. Fig 11 shows within-cluster SSE per cluster and according to the figure, increasing the cluster size beyond 6 has little impact in reducing the marginal within-cluster SSE and it results in overlapping clusters of similar centroid values. Hence compact (indicated by low SSE) and separated clusters can be derived from the internet usage dataset at K=6. Thus, the clustering attributes with cluster size and seed values were provided to the clustering algorithms to build the customer segments. The clustering result of K-means and EM algorithm for Internet usage dataset are presented here.

### K-means Clustering

Based on the input parameters such as cluster size, seed value and other input parameters, Table 8 presents the clustering result of K-means clustering algorithm. It consists of numeric centroid values per attribute for each cluster, distribution of instances per cluster, time taken to build the model, the number of iterations done to complete the execution & within-cluster SSE. But the most important values to compare the cluster solution was the distribution of instances, within-cluster SSE and attribute centroid value. These values were useful in evaluating the feasibility of the clustering results and the separation between customer segments.

Table 8 : Behavioral distribution of internet service usage - K-means clustering.

| | Cluster # | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| Attributes | 1924 (26%) | 647 (9%) | 1426 (19%) | 1350 (18%) | 852 (11%) | 1302 (17%) |
| IUMB | 28.3 | 424.2 | 825.1 | 44 | 34.4 | 233.3 |
| IUdys | 3.3 | 18.2 | 26.5 | 6.4 | 5 | 14.8 |
| ISpBr | 5.6 | 85 | 165.1 | 8.8 | 6.9 | 46.6 |
| WIntP | 0.03 | 0.29 | 0.29 | 0.29 | 0.39 | 0.24 |
| avg_wtoweekugint | No | Yes | No | Yes | No | No |

As Table 8 shows, 26% of customers were assigned to C0 and 19% are assigned to C2 while 9% were assigned to C1. According to the heat map table, customers in Cluster2 had highest IUMB & IUdys values. On the other hand, customers in Cluster0 had lowest IUMB & IUdys values. Customers in C0, C3 & C4 had very close or similar internet usage amount but vary on at least one of the remaining attributes. Customers in C1 & C3 consumed above 50% of their internet service in a given week of the month.

### Expectation Maximization (EM)

To compare the performance of the clustering algorithm, the same datasets and input parameters used for K-means are used the EM algorithm to build customer segments and Table 9 shows the clustering results of EM algorithm.

Table 9 : Behavioral distribution of internet service usage - EM clustering.

| | Cluster # | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| Attributes | 2141 (29%) | 1065 (14%) | 2127 (28%) | 1020 (14%) | 923 (12%) | 225 (3%) |
| IUMB | 7.6 | 4.5 | 125.8 | 749 | 0.86 | 3972.5 |
| IUdys | 8.6 | 7.2 | 14.8 | 23.6 | 1.6 | 25.3 |
| ISpBr | 1.5 | 0.9 | 25.1 | 149.8 | 0.16 | 795.5 |
| WIntP | 0.23 | 0.31 | 0.25 | 0.27 | 0.0001 | 0.29 |
| avg_wtoweekugint | No | Yes | Yes | No | No | No |

According to Table 9, from the total instances, 29% of customers were assigned to C0 and 3% were assigned to C5. Customers in C1 & C2 had unproportionally week to week usage intensity per month. According to the heatmap table, customers in C5 had high IUMB, WIntP & IUdys values. On the other hand, customers in Cluster4 had the lowest IUMB & IUdys values. The algorithm resulted in clusters of closed centroid value for most of the attributes such as C0 & C1.

### 5.7.3 Clustering of SMS usage dataset

### Cluster Size Determination on SMS Usage Dataset

Based on the cluster size determination technique mentioned in Section 5.6, from the different trials of K-means algorithm to SMS service usage dataset, the cluster size K=6 and seed value of 10 resulted in minimum within cluster SSE and hence it is identified as an optimal cluster size for SMS usage dataset. Fig. 12 shows the results of the elbow criterion for SMS usage dataset.

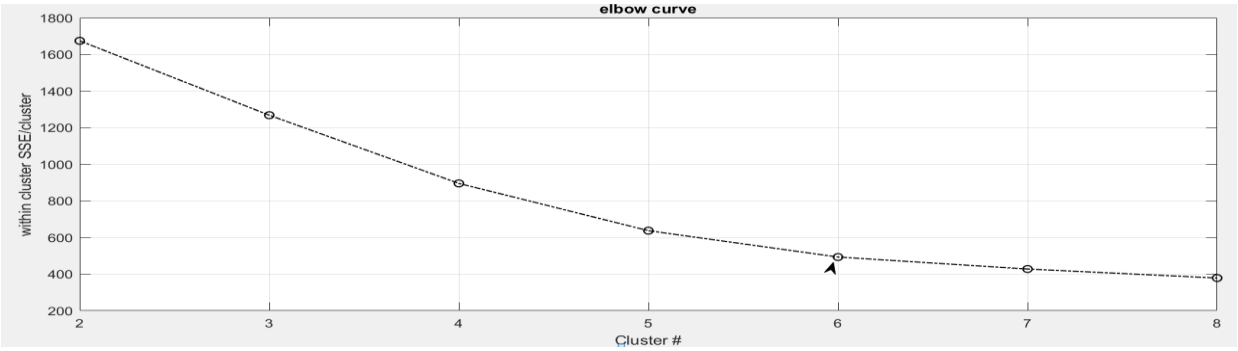



Figure 12 : Elbow curve of SMS dataset.

The auto cluster size determination option of the EM algorithm is tested on the same SMS usage dataset and identified six customer segments. The auto clustering result of the EM exhibit similar results with the elbow method and hence the result identified by the elbow method can be used as a cluster size input parameter for clustering. Fig 12 shows within-cluster SSE per cluster and according to the figure, increasing the cluster size beyond 6 has little impact in reducing the marginal within-cluster SSE and it can also impact the cluster quality by identifying overlapping clusters of similar centroid values. Hence compact and well separated clusters can be derived from the SMS usage dataset at K=6. The clustering results of K-means and EM algorithm on the SMS usage dataset are presented here.

### K-means Clustering.

Based on the input parameters such as cluster size, seed value and other input parameters Table 10 displays the clustering results of K-means clustering. The clustering results consists numeric centroid value per attribute for each cluster, distribution of instances per cluster, time taken to build the model, the number of iterations done to complete the execution & within cluster SSE. But in this research the most important values to compare the cluster solution were the distribution of instances, within cluster SSE and attribute centroid value. These values were helpful to evaluate the feasibility of the clustering results and the separation between customer segments.

Table 10 : Behavioral distribution of SMS service usage - K-means Clustering.

| Attributes | Cluster # | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| | 1112(15%) | 844(11%) | 3572(48%) | 584(8%) | 858(11%) | 531(7%) |
| SMSUN | 14 | 39.3 | 0.38 | 15.5 | 7.7 | 18.6 |
| SMSUdys | 3.4 | 8.9 | 0.18 | 3.6 | 2.6 | 3.5 |
| SMSSpBr | 2.2 | 5.3 | 0.07 | 2.2 | 1.3 | 2.6 |
| NSMSP | 0.07 | 0.04 | 0.003 | 0.01 | 0.009 | 0.14 |
| WSMSP | 0.08 | 0.33 | 0.008 | 0.81 | 0.06 | 0.84 |
| WoSMSP | 0.13 | 0.55 | 0.002 | 0.85 | 0.9 | 0.12 |
| WtWint | Yes | No | No | Yes | Yes | Yes |

According Table 10, from the total instances 48% of customers were assigned to C2 and 7% were assigned to C5. According to the cluster result heatmap table, customers in C1 had highest SMSUN & SMSUdys values. On the other hand, customers in Cluster2 had lowest SMSUN & SMSUdys values. Customers in C0, C3 & C4 consumed above 50% of their SMS service in a given week of the month which means unproportionate weekly usage plan.

### Expectation-Maximization (EM) Algorithm

To compare the performance of the clustering algorithm, the same datasets and input parameters used for K-means are used by the EM algorithm to build customer segments and Table 11 displays the clustering results of EM algorithm.

Table 11 : Behavioral distribution of SMS service usage - EM Clustering.

| Attributes | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| | 3862(51%) | 514(7%) | 705(9%) | 795(11%) | 1028(14%) | 597(8%) |
| SMSUN | 0.21 | 14.1 | 14 | 12.1 | 4.5 | 51.1 |
| SMSUdys | 0.15 | 3.1 | 4.5 | 3.5 | 2.2 | 9.7 |
| SMSSpBr | 0.04 | 2.1 | 2.6 | 1.8 | 0.9 | 6.7 |
| NSMSP | 0 | 0.15 | 0.14 | 0.006 | 0.002 | 0.03 |
| WSMSP | 0.0001 | 0.83 | 0.14 | 0.78 | 0.04 | 0.33 |
| WoSMSP | 0 | 0.12 | 0.21 | 0.82 | 0.89 | 0.47 |
| WtWint | No | Yes | Yes | Yes | Yes | No |

According to Table 11, from the total instances 51% of customers were assigned to C0 and 7% were assigned to C1. Customers in C1, C2, C3 & C4 had unproportionally week to week usage intensity per month. According to the heatmap table, customers in Cluster5 had highest SMSUN & SMSUdys values. On the other hand, customers in Cluster0 had lowest SMSUN & SMSUdys values. The algorithm resulted in clusters of closed centroid value for most of the attributes such as C0 & C4.

## 5.8 Cluster Results Evaluation Methods

The goal of clustering is to build customer segments with differentiated service usage behavior. According to [51] a good clustering result is highly cohesive internally and well separated from the other clusters. Within cluster variance is used to evaluate the internal quality of the cluster and distance between centroid value reflects the separation of the resulting clusters. In this thesis, different cluster evaluation techniques and metrics are applied to validate cluster results as well as compare the performance of clustering algorithms for each service usage dataset. The purpose of cluster evaluation is to assess the suitability of the clustering results and compare clustering algorithms' performance in building quality customer segments. The detail of cluster evaluation techniques is mentioned earlier in Section 3.4. The CDR data has no class label feature to be used as a ground truth in evaluating the cluster quality using extrinsic or supervised techniques. Hence, internal methods of cluster quality evaluation techniques will be applied to the study. The cluster evaluation techniques applied in this study to evaluate the clustering results are summarized as follows:

### 5.8.I Internal Methods of Cluster Evaluation

#### Within Cluster Sum of Squared Error (SSE)

It shows the variance in a cluster for each attribute value and small variance is expected for compact clusters. According to [51] one of the internal measures of cluster quality is Within cluster Sum of Squared Error. It was applied before to determine the optimal cluster size in the dataset in the cluster size determination step.

#### Cluster Distribution

It shows the number of instances assigned per cluster and cluster labels are used to identify the distribution of instances per cluster instead of the cluster index. Mobile service packages are offered per segments and hence the cluster size should be large enough and feasible. Therefore, clusters should not be extremely small or too large. According to [22] qualitative cluster evaluation techniques are used to evaluate the cluster distribution as well as applicability of the clustering results to propose an applicable marketing action. The qualitative criteria are:

- *Measurable:* the attributes used to distinguish customer are measurable as the most important attributes used to build the cluster are related to usage amount, usage time, spending amount.
- *Substantial:* the distribution of instances per cluster should be substantial and feasible enough to invest in.
- *Accessible:* the customer segments are built using easily accessible CDR data that indicates substantial usage behavior and hence the derived results applicable for offering customer centric mobile packages.
- *Differentiable:* as indicated in the cluster result table, customers in each segment have differentiated centroid value for most clustering attributes and hence the segments are distinguishable. It is indicated by different colors in the heatmap table for each attribute value per clustering algorithm.
- *Actionable*: the number of clusters in this study is less than 6 and this segment size is manageable and easily actionable.

### Separation of Attribute Values

The separation between cluster is indicated by distance between centroid values. According to [2] the attribute values per cluster should be well separated. Very close or similar centroid values are an indicator of poor clustering results or overlapping clusters.

# 6 Results, Evaluation and Interpretation

This chapter discuss and compare the clustering results of K-means and EM algorithm. Clustering algorithms were compared based on the cluster evaluation techniques presented in Section 5.8 to identify an algorithm suitable for the clustering of customers. Moreover, the cluster results were interpreted for a better understanding of the results using descriptive methods and visualization techniques.

## 6.1 Evaluation of Clustering Results

The clustering results of K-means and EM clustering algorithm were compared for each service usage dataset based on the key metrics such as distribution of instances, within cluster cohesion, & separation between cluster centroid values per attribute.

### 6.1.1 Comparison of Clustering Results for Voice Dataset

Different cluster evaluation techniques were used to compare the algorithm's result for the voice usage dataset. One of the techniques to evaluate the clustering result was based on the distribution of instances in each cluster or cluster ratio. To make the cluster comparison identifiable, cluster labels were used instead of the cluster index. Because the cluster labels indicate the type the customer segment, the cluster labeling was mainly based on the usage amount level and other unique features. Based on the cluster centroid of Table 6 and 7, Figure 13 shows the distribution of instances.

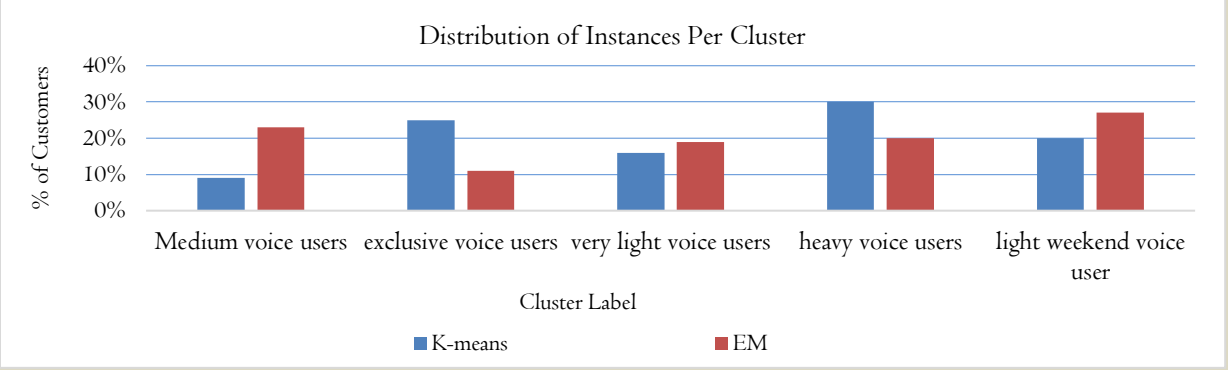



Figure 13 : Distribution of instances per cluster -voice dataset.

According to Figure 13, five clusters of unique class labels were derived from the voice dataset by both clustering algorithms. Based on the distribution of instances per cluster the smallest cluster size for K-means was 9% and 11% for EM clustering. The cluster ratio of K-means was 3.3:1 while for EM, it was 2.4:1. Hence the cluster ratio of EM algorithm was closed to the recommended value mentioned in Section 5.8. The clusters were also compared based on the compactness of the cluster indicated by low variance for clustering attributes as the box plot or Fig 14 shows. Mobile services are designed for segments and it should be close to instances in a cluster to meet the needs of customers in a segment.

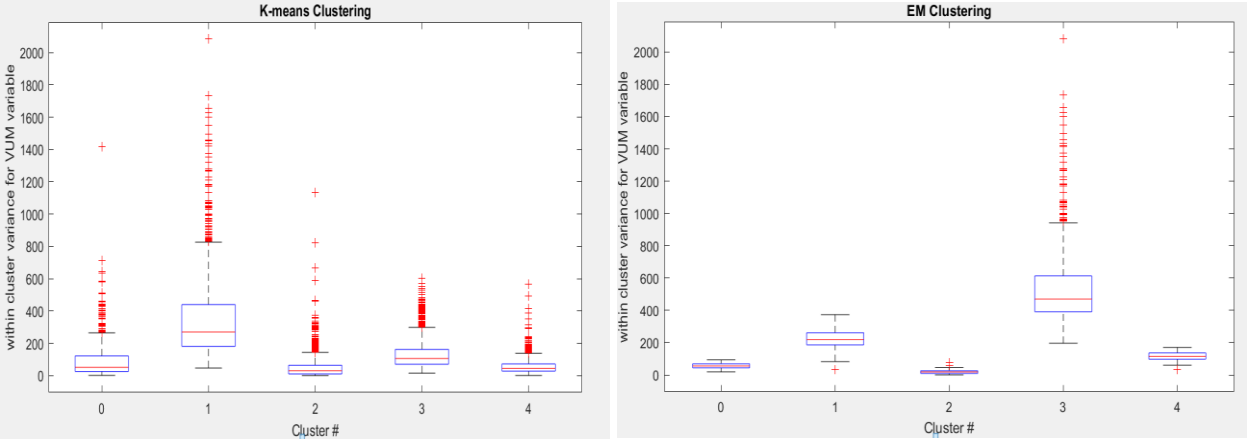



Figure 14 : Within cluster variance per cluster for VUM attribute.

According to the Fig 14, based on the VUM attribute the clustering results of K-means clustering algorithm had high level of within-cluster variance (deviation from cluster centroid value) for all customer segments but the EM built compact clusters except to cluster 3.Thefore, EM algorithm established compact clusters for the voice usage dataset.

The other technique was the separation between clusters. According to [45] well-derived clusters should have well-separated centroid value for most attributes per cluster. For this purpose, visualization technique was used to compare the separation between clusters and normalized numeric attributes values were used as input for clustering. Weka has a filtering option to normalize the numeric attribute value. According to [2] normalization of used to minimize the difference between very large and small values in a small range i.e. between [0,1]. This helps to easily visualize large and small values in a single figure to make comparison simpler. Figure 15 shows the separation between attribute values.

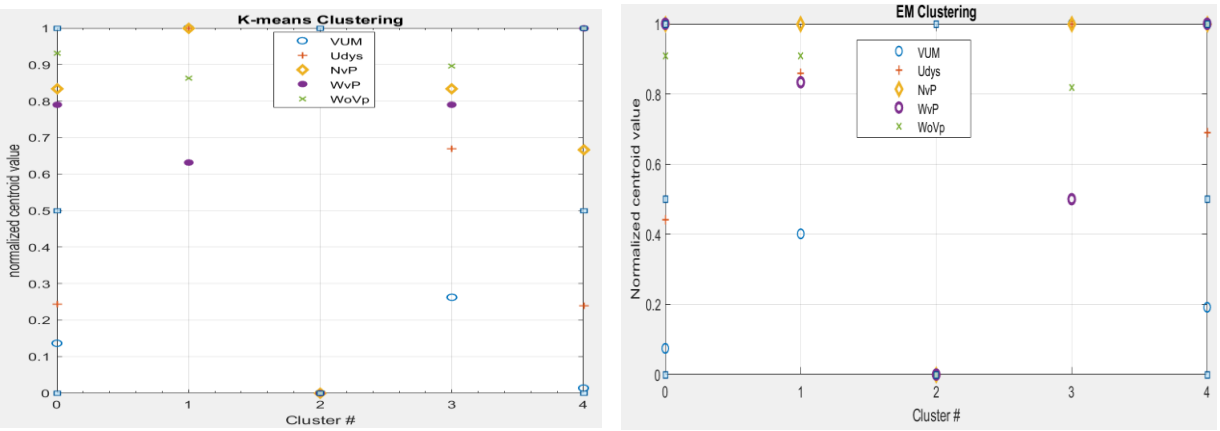



Figure 15 : Separation of cluster centroid value per attribute -voice dataset.

Both clustering algorithms (K-means & EM) build well separated customer segments for most of the centroid values of voice dataset. As Fig 15 indicates, based on VUM attribute K-means derived two clusters of closed centroid values such as C2 & C4. But for the same attribute, EM identified relatively well-separated clusters. In the figure, the value of WoVp was very close for most of the clusters in both algorithms and hence it shows that the variable is a weak predictor to differentiate customers based on their service usage as majority of customers made most of their voice calls during working hours.

The quality of results of both clustering algorithms on the voice service usage dataset are compared based on the evaluation metrics summarized in Table 12.

Table 12 : Summary of performance of clustering algorithms - voice dataset.

| Evaluation Metrics | K-means | EM | Better Values [2, 3,53] |
|---|---|---|---|
| *Cluster Ratio* *(largest to smallest)* | 3.3:1 | 2.4:1 | 2.0:1 |
| *Cluster Cohesion* *(Variance)* | High | Low | Low |
| *Cluster Separation* | Good | Good | well separated |

In general, for most of the cluster evaluation techniques, the EM clustering algorithm exhibited better performances in establishing quality clusters for the voice usage dataset than the K-means clustering. The clustering result of the EM algorithm is discussed in the next subsection.

### 6.1.2 Comparison of Clustering Results for Internet Dataset

Clustering results for internet dataset was evaluated using various metrics. One of the techniques to evaluate the clustering result was based on the distribution of instances in each cluster. Just like the voice dataset cluster labels were used instead of the cluster index. Based on Tables 8 and 9 of the clustering results of Internet dataset, Figure 16 shows the distribution of instances.

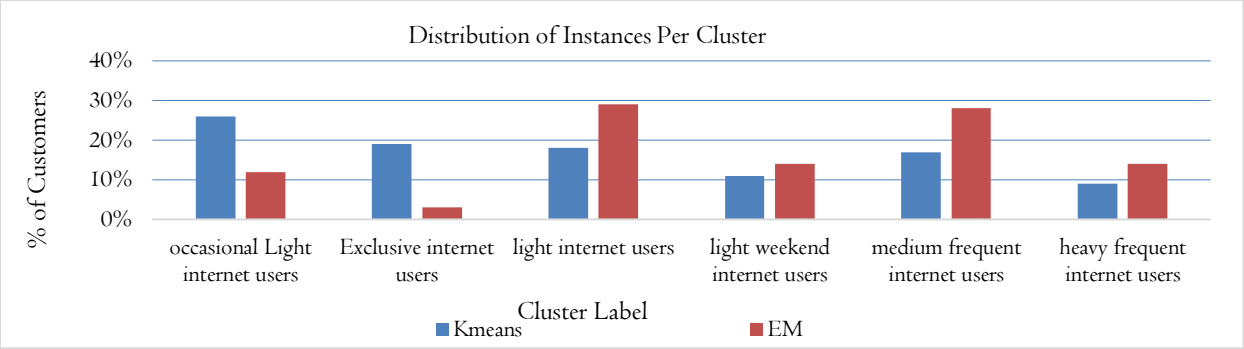



Figure 16 : Distribution of instances per cluster - Internet dataset.

According to Fig 16, both clustering algorithms established six clusters with six unique class labels. Based on the distribution of instances per cluster the smallest cluster size for K-means was 9% and 3% for EM clustering. The cluster ratio of K-means was 2.8:1 while for EM, it was 9.6:1. Hence the cluster ratio of K-means was closed to the recommended value mentioned in Section 5.8. Based on this metric, the clustering results of the K-means algorithm is suitable for mobile service packaging. Furthermore, clusters should be internally cohesive indicated by low variance on clustering attributes as Figure 17 demonstrates with a box plot.

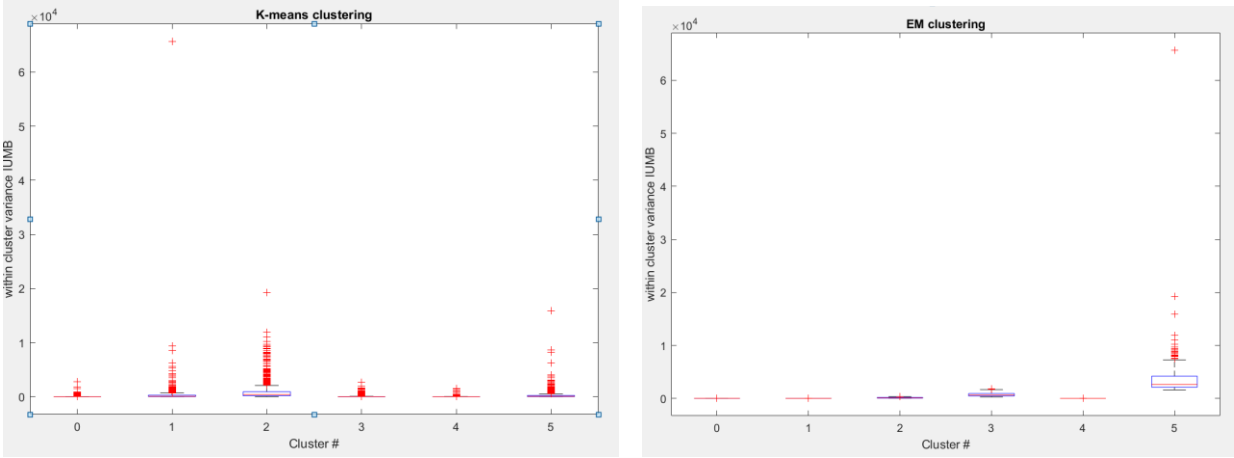



Figure 17 : Within cluster variance per cluster for IUMB attribute.

According to Figure 17 based on the IUMB attribute, the clustering results of the K-means clustering algorithm had a high level of within-cluster variance (deviation from cluster centroid value) for all the clusters. However, the EM algorithm resulted highly cohesive clusters except to cluster 5. Therefore, EM established compact clusters for the internet usage dataset. The other metric to compare the derived cluster was using separation between clusters. It is based on the distance between clustering attributes centroid value. According to [45] clusters are expected to have well-separated centroid value for most of the attributes per cluster. Visualization technique was used to compare the separation of clusters and for this purpose, normalized numeric attribute values were used as input for clustering. Figure 18 shows the separation between attribute values.

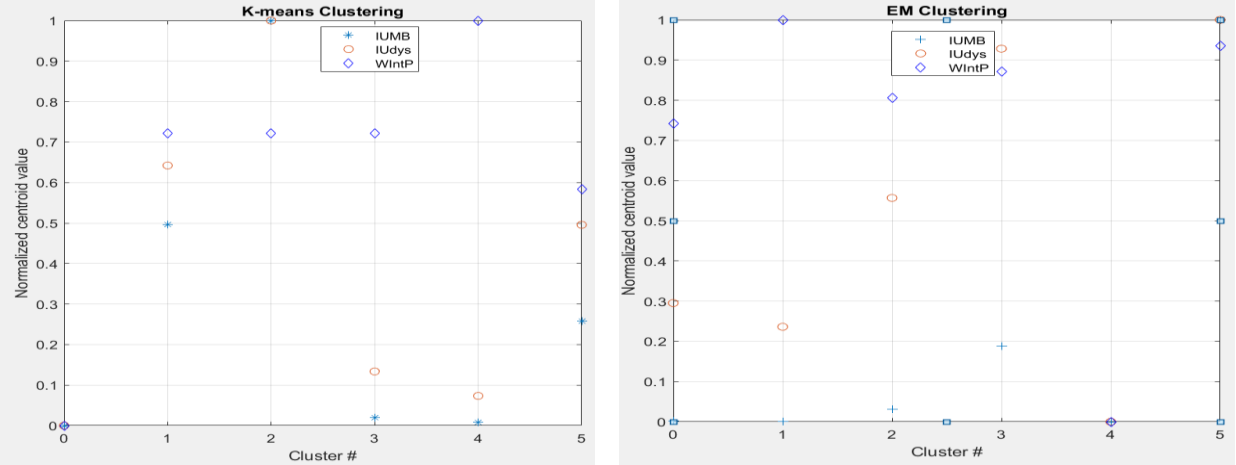



Figure 18 : Separation of cluster centroid value per attribute - Internet dataset.

Both clustering algorithms (K-means & EM) built customer segments highly separated by most of the centroid values from internet usage dataset. But the clustering result of the EM algorithm were relatively well-separated. According to Figure 18 based on the IUMB attribute, both algorithms identified three clusters of closed centroid values such as C0, C3 & C4 for K-means and C0, C1 & C4 for EM algorithm. But they were differentiable for the rest of the attributes. This indicates different usage behavior regardless of similarity in usage amount. In both algorithms, the IUdys was well separated and hence it was a good predictor to differentiate customers based on internet usage behavior. In both algorithms, the usage amount attribute was a weak variable to differentiate customers.

Table 13 presents the summary of the clustering algorithms performance compared based on the evaluation metrics.

Table 13 : Summary of performance of clustering algorithms - Internet dataset.

| Evaluation Metrics | K-means | EM | Better Values [2, 3,53] |
|---|---|---|---|
| Cluster Ratio (largest to smallest) | 2.8:1 | 9.6:1 | 2.0:1 |
| Cluster Cohesion (Variance) | High | Low | Low |
| Cluster Separation | Good | Good | Well Separated |

In general, both algorithms build clusters of separated centroid values. But K-means builds clusters with fair distribution of instances per cluster. On the other hand, EM algorithm establish compact clusters with low level of within cluster variance. Since cluster ratio is an essential metrics to propose feasible service package offers, the results of K-means algorithm preferable than the EM algorithm for the internet dataset. Therefore, the clustering result of the K-means algorithm was further discussed in the next section.

### 6.1.3 Comparison of Clustering Results for SMS service usage Dataset

Clustering results for SMS dataset was evaluated using various metrics. One of the techniques to evaluate the clustering result was based on the distribution of instances in each cluster. Just like the voice and internet dataset cluster labels were used instead of the cluster index for SMS dataset. Based on Table 10 and 11 of the clustering results of the SMS dataset, Figure 19 shows the distribution of instances.

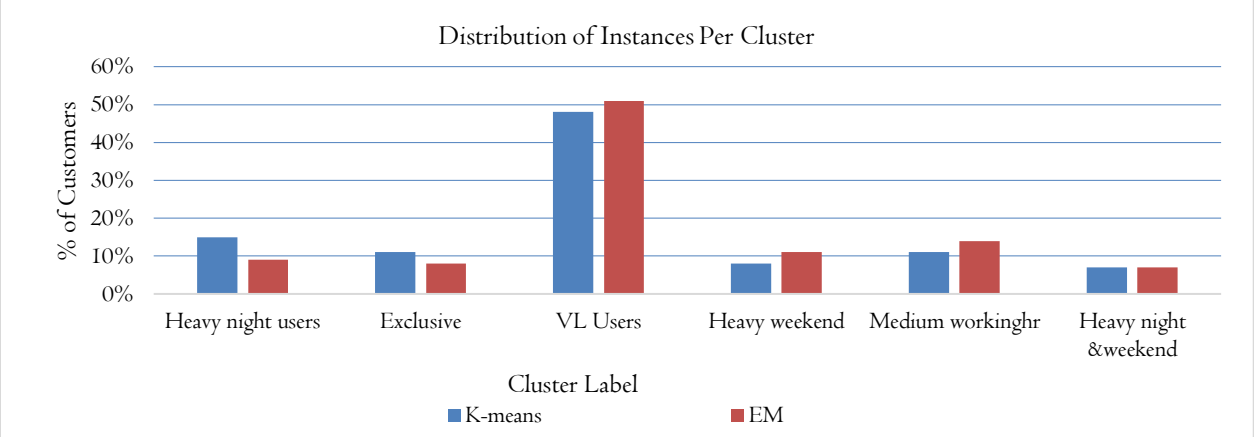



Figure 19:Distribution of instances per cluster - SMS dataset.

According to Figure 19, both clustering algorithms established six clusters of unique class labels. Based on the distribution of instances per cluster the smallest cluster size for both clustering algorithms was 7% but the cluster ratio of K-means was 6.8:1 while for EM, it was 7.2:1. Hence the cluster ratio of K-means is closed to the recommended value mentioned in Section 5.8. Hence, based on this qualitative metric, the clustering results of K-means algorithm is suitable for mobile service packaging.

Furthermore, the compactness of clusters was indicated the variance within a cluster for each clustering filed using box plot of Figure 20. Mobile service packages designed for each segment should address the needs of most of the instances in the segment.

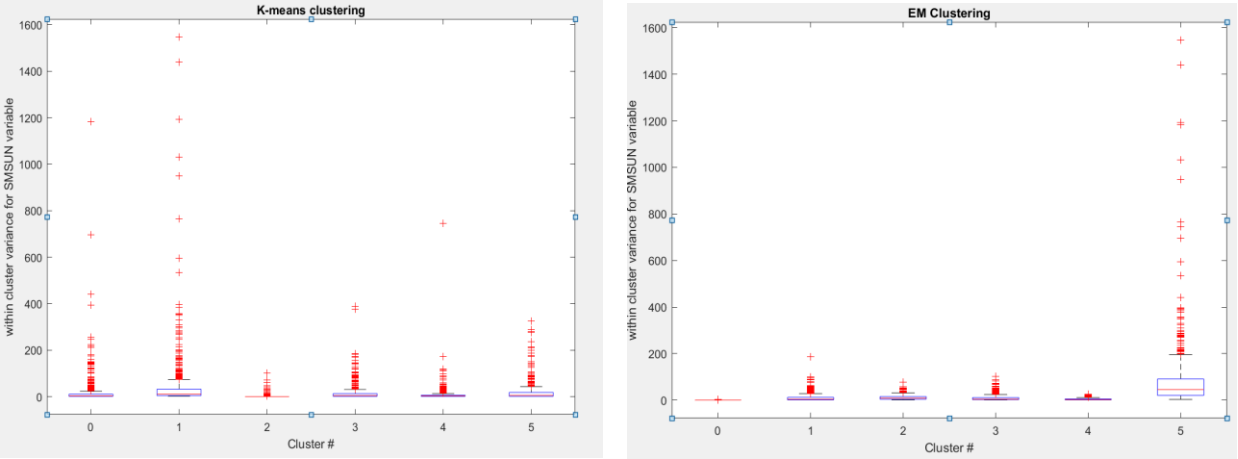



Figure 20 : Within cluster variance per cluster for SMSUN variable - SMS dataset.

According to Figure 20, based on the SMSUN attribute both algorithms identified clusters with a high level of variance. Relatively the clustering results of the EM clustering algorithm had compact clusters with minimum within-cluster variance for most of the customer segments except to cluster 5 but the clustering results of K-means had a high level of within-cluster variance for most of the clusters. Therefore, EM established highly cohesive clusters for the SMS usage dataset.

The other technique to compare the derived cluster is a separation of customer segments. It is based on the distance between clustering attributes centroid value. Figure 21 shows the comparison of the attribute values.

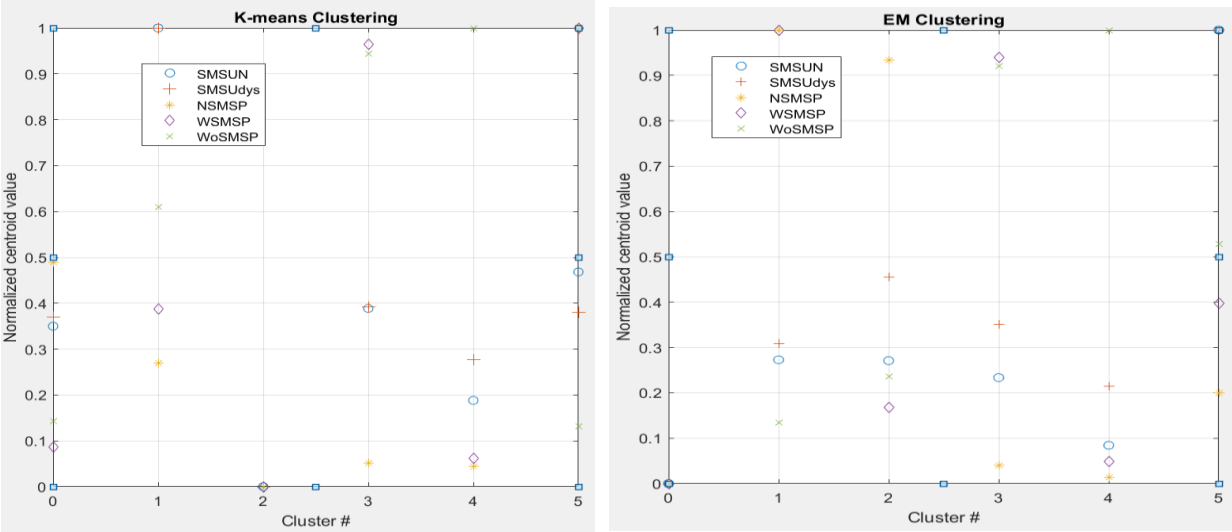



Figure 21 : Separation of cluster centroid value per attribute - SMS dataset.

The application of clustering algorithms (K-means & EM) on the SMS usage dataset established customer segments highly separated by most of the centroid values. According to Figure 21, based on SMSUN attribute K-means algorithms identified differentiated clusters with separated centroid value. But for the same attribute EM clustering identified clusters with very close value such as C1, C2 & C3 though they were differentiable on the remaining attributes. In both algorithms, WoSMSP & WSMSP values were well separated and hence these attributes are a good predictor to differentiate customers based on SMS usage behavior. In both algorithms, the usage amount attribute was a weak variable to differentiate customers.

The clustering algorithms were compared based on the evaluation metrics and summarized in Table 14.

Table 14 : Summary of performances of clustering algorithms on SMS dataset.

| Evaluation Metrics | K-means | EM | Better Values [2, 3,53] |
|---|---|---|---|
| **Cluster ratio** (largest to smallest) | 6.8:1 | 7.2:1 | 2.0:1 |
| **Cluster cohesion** (Variance) | High | Low | Low |
| **Cluster separation** | Good | Good | well separated |

In general, both algorithms build clusters of separated centroid values. But K-means builds clusters with fair distribution of instances per cluster. On the other hand, EM algorithm establish compact clusters with low

level of within cluster variance. Since cluster ratio is an essential metrics to propose feasible service package offers, the results of K-means algorithm preferable than the EM algorithm for the SMS dataset. Therefore, the clustering result of the K-means algorithm was further discussed in the next section.

## 6.2 Interpretation of Clustering Results.

The cluster profiling technique was used to interpret and understand the clustering results of each algorithm. Thus, cluster attribute values were standardized to minimize the large difference between values and visualize in a graph. Weka's filtering options support the standardization of numeric attribute values. The characteristics of the customer in each segment and differentiating attributes were represented using bar charts. The value in the Y-axis shows standardized attribute value, the x-axis shows the cluster index and the horizontal line with zero value shows the total population mean value. Hence any value above the total mean value shows the highest value and below it indicates lower values based on the deviation of the bar from the horizontal line. The length of the bar indicates the deviation from the total mean value. As discussed in Section 5.5, the cluster labeling was done based on the deviation of the cluster from the total mean value for each attribute. This was helpful to uniquely identify the characteristics of each cluster and attributes to efficiently differentiate each cluster.

### 6.2.1 Cluster Profiling and Interpretation-Voice Dataset

Based on the cluster evaluation techniques applied to the voice usage dataset, the EM clustering algorithm derived well separated, and highly cohesive clusters and it was described using cluster profiling technique. Figure 22 illustartes the behavioral profile of the EM clustering algorithm for voice service usage dataset, which describe the difference in usage behavior between cluster.

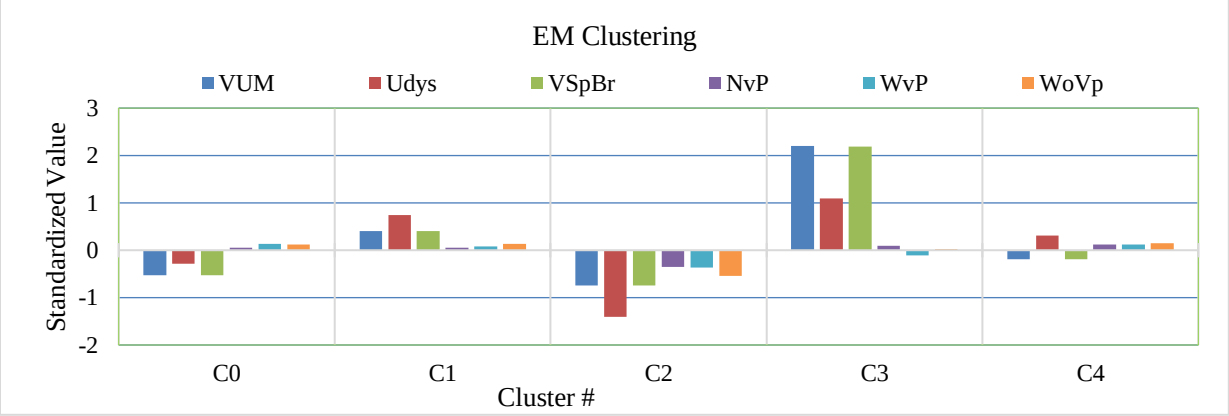



Figure 22 : Segment characteristics of EM clustering-voice dataset.

The segment revenue contribution, cluster size and other characteristics of the clustering result of the EM algorithm were described and the preferred service package options are forwarded after analyzing the customer usage behavior in each segment. The behavioral profile of the derived customer segments of the EM clustering algorithm on voice usage dataset are described as follows:

### Cluster_0: light weekend voice user

**Contribution to voice revenue – 10%**

According to Table 7, it is the largest cluster with 27% of customers assigned to this segment and they have a proportionate week to week consumption plan. Figure 22 shows that customers in this segment have a low voice usage amount with a few days per month usage frequency. Customers have very high weekend voice calls and high working hours and night voice calls. VUM in negative and WvP in a positive direction are most differentiating attributes.

### Cluster_1: frequent heavy voice users

**Contribution to voice revenue – 29%**

Table 7 shows that 20% of customers are assigned to this segment and they had a proportionate week to week usage plan. Figure 22 shows that customers have high VUM and very high Udys frequency per month. A medium volume monthly voice package is the preferred marketing approach. Udys is the most differentiating attribute.

### Cluster_2: very light voice user

**Contribution to voice revenue – 2%**

Table 11-shows that 19% of customers are assigned to this segment and proportionate week to week usage plan. According to Figure 22, customers have a very low VUM and Udys frequency and very low WvP, NvP & WoVp call usage. Udys variable is the most differentiating in a negative direction. Low volume voice packages, especially on working hours, is a suitable approach.

### Cluster_3: Exclusive voice users

**Contribution to voice revenue – 41%**

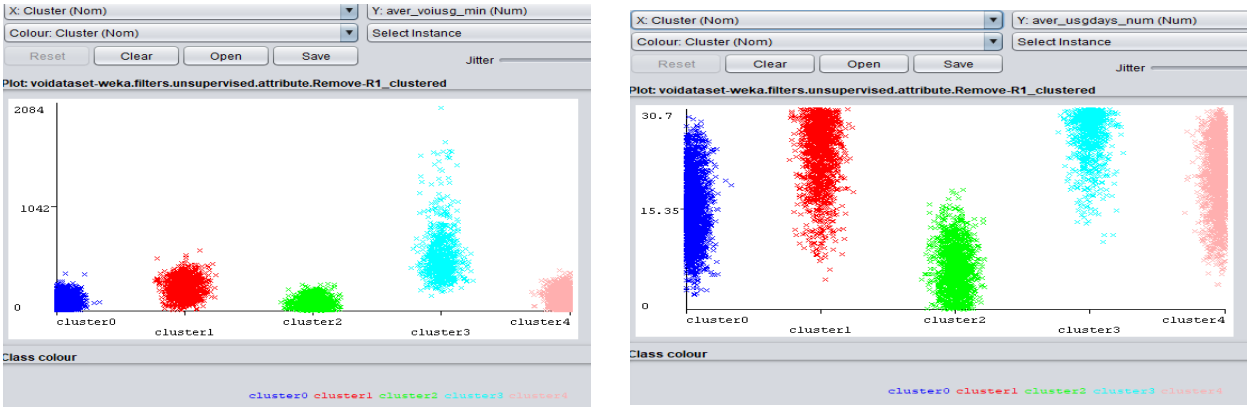
Table 7 shows that 11% of customers are assigned to this segment and they had a proportionate week to week voice usage plan. Customers had high VUM, Udys, and VSpBr. The WvP voice call ratio is relatively low. Premium monthly voice packages are preferable. VUM in positive and WvP in negative direction were the most differentiating attributes.

### Cluster_4: medium frequent voice users

**Contribution to voice revenue – 18%**

According to Table, 7,23% of customers are assigned to this segment and they had a proportionate week to week voice usage plan. Customers in this segment had low VUM with high Udys frequency per month. Customers in this segment have relatively high WvP. VUM and Udys frequency are the most differentiating variables in positive and negative direction respectively. Offering weekend and working hour voice packages are more suitable.

In general, the VUM or VSpBr & Udys are the most important predictors to differentiate customers from the voice dataset. Figure 23 shows the inter-cluster difference of clustering results of the EM algorithm based on the two attributes or predictors (VUM & Udys). It shows the distribution of instances based on the monthly voice usage amount and usage day frequency of customers.




(a) (b)

Figure 23 : Distribution of instances per cluster for VUM(a) & Udys (b) attributes-EM.

Based on Figure 23, Customers in C3 have high VUM with very high usage days frequency per month. On the other hand, customers in C2 have very low voice usage with a few days per month usage frequency.

### 6.2.2 Cluster Profiling and Interpretation-Internet Dataset

Based on the cluster evaluation techniques applied, the K-means clustering algorithm established well separated, substantial and relatively stable customer segments than EM for internet datasets. The segments were described using the cluster profiling technique. To profile the cluster, the standardized attribute values of the internet usage dataset were provided for the K-means clustering algorithm. Figure 24 illustrates the

clustering results of the K-means clustering algorithm, which describe the difference in usage behavior between cluster.

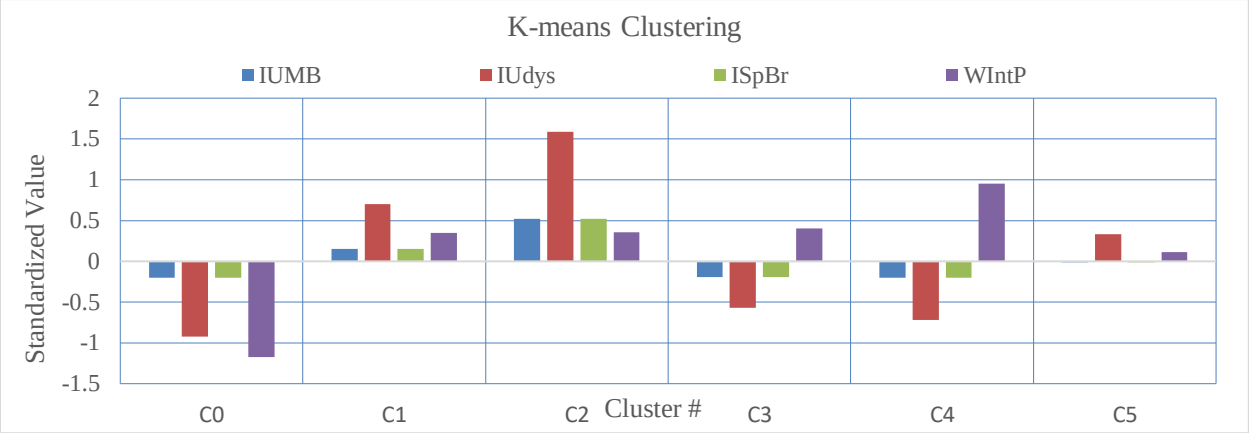



Figure 24 : Segment characteristics of K-means clustering -Internet dataset.

The segment revenue contribution, cluster size and other characteristics of the derived customer segments of the K-means algorithm was described and the service package options are forwarded after analyzing the customer usage behavior in each segment. The behavioral profile of the derived customer segments of the K-means clustering algorithm on internet usage dataset were described as follows:

**Cluster_0: occasional Light internet users**

***Contribution to internet revenue – 3%***

According to Table 8, It is the largest cluster with 26% of customers are assigned to this segment and they have a proportionate week to week consumption plan. Figure 24 indicates that customers in this segment have low IUMB with very few days IUdys per moth usage frequency. Moreover, customers in this segment have very low WIntP usage. IUdys frequency and WIntP are the differentiating attributes on the negative side. Low volume daily internet packages are suitable.

**Cluster_1: heavy frequent internet users**

***Contribution to internet revenue – 14%***

According to Table 8, the smallest cluster with 9% of customers are assigned to the cluster. Customers in this segment have high IUMB with very high IUdys frequency. They also have high WIntP. IUdys is differentiating variable in a positive direction. Consume >50% of total internet usage in a given week of the month. Medium volume fortnight internet packages are preferable.

### Cluster_2: Exclusive internet users

*Contribution to internet revenue – 62%*

According to Table, 8,19% of customers are assigned to this cluster. Based on Figure 24. Customers have high IUMB, IUdys & WIntP and proportionate week to week consumption plans. IUdys frequency is a differentiating variable in a positive direction. large volume monthly internet packages are preferable.

### Cluster_3: light internet users

*Contribution to internet revenue – 3%*

According to Table 8,18% of customers are assigned to this segment Consume >50% total internet usage in a given week of the month. Based on Figure 24 customers in this segment have low IUdys frequency and high WIntP. These variables are the cluster differentiating in the negative and positive direction respectively. Small volume of weekly internet packages is preferable.

### Cluster_4: light weekend internet users

*Contribution to internet revenue – 2%*

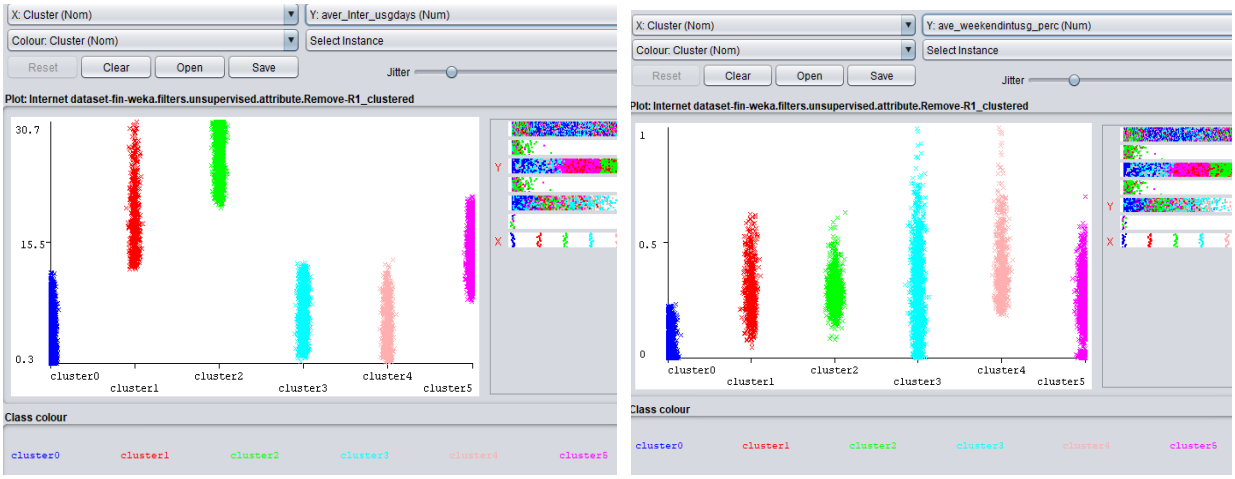
According to Table, 8,11% of customers are assigned to this segment and they have a proportionate week to week internet usage plan. Based on Figure 24 customers in this segment have low IUMB with few days per month IUdys and they prefer internet usage on weekends. WIntP is the differentiating variable of the cluster. Small volume weekend internet packages are suitable.

### Cluster_5: medium frequent internet users

*Contribution to internet revenue – 16%*

According to Table 8, 17% of customers are assigned ton this segment and they have a proportionate week to week internet usage plan. Based on Figure 24 they have medium IUMB with high IUdys frequency per month. IUdys variable differentiates the cluster. A small volume of monthly internet packages week is suitable.

In general, IUdys and WIntP are the most important predictors to differentiate customers from internet usage dataset. Figure 25 shows the difference in usage behavior of the clustering results of the K-means clustering algorithm based on these attributes. It shows the distribution of instances based on the monthly internet usage amount and usage days frequency of customers' attributes.




(a) (b)

Figure 25 : Distribution of instances per cluster for IUdys(a) & WIntP(b) attributes-K-means.

Based on Figure 25 Customers in Cluster2 have high usage days frequency per month. On the other hand, customers in C0 have very low internet usage on weekends with few days per month usage frequency.

### 6.2.3 Cluster Profiling and Interpretation-SMS Dataset

Based on the cluster evaluation techniques applied to the SMS usage dataset, the K-means clustering algorithm established well separated, substantial, relatively stable customer segments than EM. The resulting segments were described using the cluster profiling technique. To profile the cluster the standardized attribute values of the internet usage dataset were provided for the K-means clustering algorithm. Fig 26 illustrates the clustering results of the K-means clustering algorithm, which describe the difference in usage behavior between cluster.

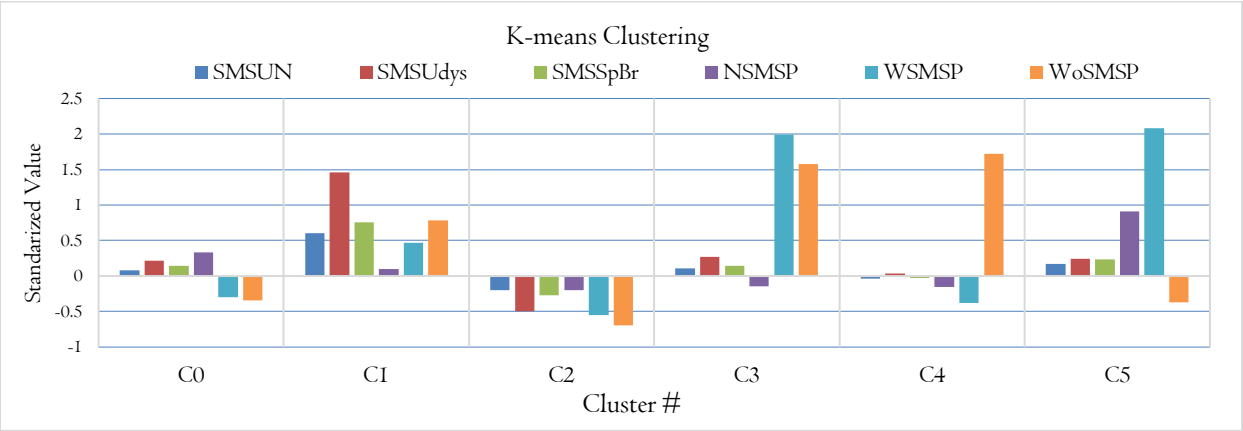



Figure 26 : Segment characteristics of K-means clustering - SMS dataset.

The segment revenue contribution, cluster size and other characteristics of the derived customer segments of the K-means algorithm was described and the preferred service package options were forwarded after analyzing the customer usage behavior in each segment. The behavioral profile of the derived customer segments of the K-means clustering algorithm on the SMS usage dataset was described as follows.

### Cluster_0: Heavy night SMS users

*Contribution to SMS revenue – 22%*

According to Table 10,15% of customers are assigned to this segment and they Consume >50% total SMS usage in a given week of the month. And based on Figure 26 customers in this segment have high SMSUN, SMSUdys, and NSMSP. Weekly SMS packages are suitable since their usage frequency is relatively high.

### Cluster_1: Exclusive SMS users

*Contribution to SMS revenue – 40%*

According to Table 10, 11% of customers are assigned to this segment and they have a proportionate week to week usage per month. Customers in this segment have relatively very high SMSUN and SMSUdys frequency per month. Customers have low NSMSP. Medium usage on weekends and working hours. Offering monthly SMS packages.

### Cluster_2: Very light SMS users

*Contribution to SMS revenue – 3%*

According to Table 10,48% of customers are assigned to this segment and they have a proportionate week to week usage. Based on Figure 26 customers have very a low value for almost all attributes and customer in this segment has low preference for the SMS service and it is the largest cluster. Subsidized daily SMS packages are preferable.

### Cluster_3: Heavy Weekend SMS users

*Contribution to SMS revenue – 12%*

According to Table 10,8% of customers are assigned to this segment and they Consume >50% of total SMS usage in a given week of the month. Based on Figure 26 customers have high SMSUN and high SMSUdys. Customers in this segment have very high WSMSP usage and WoSMSP. Offering high volume working hour packages.

### Cluster_4: Medium working hour SMS users

*Contribution to SMS revenue – 10%*

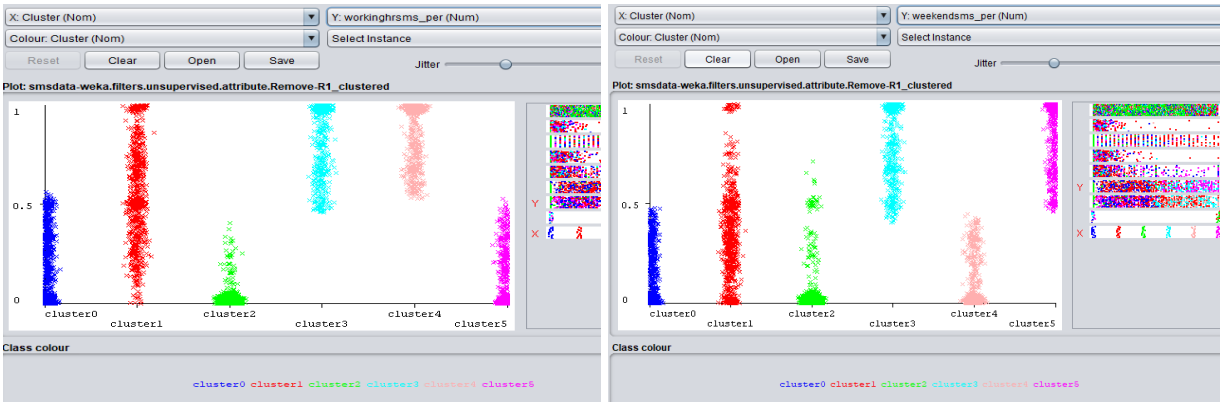
According to Table 10, 11% of customers are assigned to this segment and they Consume >50% total SMS usage in a given week of the month. Customers have medium SMSUN and Low WSMSP, but they have very high WoSMSP. Offering medium volume weekly working hour packages.

### Cluster_5: Heavy weekend & night SMS users

*Contribution to SMS revenue – 13%*

According to Table 10, 7% of customers are assigned to this segment and they Customers Consume >50% total SMS usage in a given week of the month. Customers in this segment have high SMSUN and high SMSUdys frequency per month. Customers have very high WSMSP and high NSMSP. Low WoSMSP and it is the smallest cluster. Weekend packages preferable.

In general, WoSMSP & WSMSP are the most important predictors to differentiate customers based on SMS service usage behavior. The difference in usage behavior of the clustering results of the K-means clustering algorithm based on these attributes is shown in Figure 27. It shows the distribution of instances based on the two attributes of WoSMSP & WSMSP.




(a) (b)

Figure 27 : Distribution of instances per cluster for WoSMSP(a) & WSMSP(b) attributes - K-means.

Figure 27 shows the distribution of instances based on working hours and weekend SMS service usage ratio. Customers in C3 & C4 have relatively high SMS usage during working hours. On the other hand, customers in C3 & C5 have very high SMS usage during weekend days.

# 7

# 7.0 Conclusion and Future work

## 7.1 Conclusion

Currently, mobile customers have diverse needs and preferences for services and effectively identifying their needs, preferences, and potentials have very essential to offer customer-centric mobile service packages. In this regard, customer segmentation has a crucial role in distinguishing customers according to their similarity and differentiability in service usage behavior. Hence it can be used as a tool to identify the level of heterogeneity in the customer base and categorize them into a few manageable groups for segment-based service packaging.

In this thesis, the advanced clustering technique was applied to the CDR dataset of mobile customers to build customer segments. The performance of the clustering results of two clustering algorithms (K-means and EM) were compared based on the cluster quality evaluation metrics. The clustering technique was applied to mobile subscribers CDR data that can reflect the actual usage behavior of the mobile customer and the cluster size was identified only based on the differentiability that should exist in the dataset. The study is unique as it is applied to real CDR data of mobile subscribers.

In the study, data preprocessing tasks were applied to preprocess the original data for clustering purposes. One of the tasks was construction of additional fields (such as attributes that indicate service usage day & hour) to the original dataset to include features that can reflect customer behavior. Besides this, numeric attribute values were standardized or normalized to easily visualize the distribution of instances for each attribute using a single figure. The refined dataset was provided to the clustering algorithms to build customer segments. The primary task to build customer segment was cluster size determination and the elbow method or distortion curve technique was applied and compared with the auto clustering of the EM algorithm. Besides this qualitative cluster evaluation techniques were used to decide on the optimal cluster size. Consequently, the clustering algorithms were applied on each dataset and the derived clusters were compared based cluster evaluation metrics such as cluster ratio, cohesion, and cluster separation.

Based on the cluster evaluation metrics such as cluster distribution and within-cluster variance, the EM algorithm established quality clusters than K-means clustering for voice usage dataset. On the other hand,

for SMS and internet usage datasets, the K-means clustering algorithm derived quality cluster indicated by highly cohesive and well separated segments than the EM algorithm. In general, the EM algorithm is efficient in establishing cohesive or compact clusters with a low level of within-cluster variance for all the datasets. On the other hand, K-means clustering build clusters with fair assignment of instances in each cluster.

Finally, the results of the study help offer customer centric mobile service packages and build a strong relationship with customers by enhancing the insight on the customer. In the existing value-based segmentation, monthly spending amount attribute is mainly used to build the customer segments and with this approach it is difficult to fully capture the behavior of customers and propose differentiated and customer centric mobile service packages. However, this study identified additional attributes to enhance insight on customer service usage behavior for a better segmentation purpose. The most important predictors or attributes to distinguish the cluster are VUM & Udys for voice dataset. The IUdys and WIntP for internet dataset and WoSMSP & WSMSP for the SMS dataset.

## 7.2 Future Work

The thesis was conducted using a small sample of subscribers CDR for a short period (3 moths detail data). But the results of the study can be more strengthened by increasing the sample size targeted for the study and the CDR period by including longer periods of above 3 months. In addition, combined CDR attributes of each service type can be utilized to analyze the combined service usage behavior of customers to propose combined service packages.

# References


[1] F. Abdi and S. Abolmakarem, "Customer Behavior Mining Framework (CBMF) using clustering and classification techniques," *J. Ind. Eng. Int.*, vol. 3, 2018.

[2] S. Aheleroff, "Customer segmentation for a mobile telecommunications company based on service usage behavior," *Proc. - 3rd Int. Conf. Data Min. Intell. Inf. Technol. Appl. ICMIA 2011*, pp. 308–313, 2011.

[3] C. Analysis, "Chapter 6 Cluster analysis," *Process Metall.*, vol. 12, no. C, pp. 199–227, 2002.

[4] C. Analysis and B. Concepts, "Cluster Analysis: Basic Concepts and Methods," pp. 443–495, 2012.

[5] A. Arora and R. Vohra, "Segmentation of Mobile Customers for Improving Profitability Using Data Mining Techniques," vol. 5, no. 4, pp. 5241–5244, 2014.

[6] H. I. Arumawadu, R. M. K. T. Rathnayaka, and S. K. Illangarathne, "Mining Profitability of Telecommunication Customers Using K-Means Clustering," *J. Data Anal. Inf. Process.*, vol. 03, no. 03, pp. 63–71, 2015.

[7] W. Atkinson, S. Roberts, and M. Savage, "Class inequality in austerity Britain: Power, difference and suffering," *Cl. Inequal. Austerity Britain Power, Differ. Suff.*, vol. 4, no. 9, pp. 1–195, 2012.

[8] M.-F. Băcilă and A. Rădulescu, "Consumption-based segmentation : An analysis of a telecom company ' s customers," no. April 2015, pp. 48–59, 2011.

[9] J. Bayer, "Customer segmentation in the telecommunications industry," *J. Database Mark. Cust. Strateg. Manag.*, vol. 17, no. 3–4, pp. 247–256, 2010.

[10] B. B. Bezabeh, "the Application of Data Mining Techniques To Support Customer Relationship Management: the Case of Ethiopian Revenue and Customs Authority," *Int. J. Adv. Stud. Comput. Sci. Eng.*, vol. 6, no. 6, pp. 35–41, 2017.

[11] D. Birant, "Data Mining Using RFM Analysis," *Knowledge-Oriented Appl. Data Min.*, no. iii, 2011.

[12] S. Borman, "(Tutorial) The Expectation Maximization Algorithm," *Submitt. Publ.*, vol. 25, no. x, pp. 1–9, 2009.

[13] C. Bounsaythip and E. Rinta-runsala, "10.1.1.22.3279.Pdf," 2001.

[14] J. Chen, "Leveraging Purchase History and Customer Feedback for CRM: a Case Study on eBay's "Buy It Now"," 2013.

[15] V. Components, C. Study, and A. P. Bank, "JIENG233281321907400.pdf," pp. 79–93, 2011.

[16] S. Daniel and E. Cayirci, "Predictive modeling of trust to social media content," 2014.

[17] C. Dullaghan and E. Rozaki, "Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers," *Int. J. Data Min. Knowl. Manag. Process*,

vol. 7, no. 1, pp. 13–24, 2017.

[18] T. Engineering, "Performance Evaluation of Unsuper- vised Learning Techniques for Enter- prise Toll Fraud Detection," 2018.

[19] B. Farhangian, M. Shamsi, and R. Ahsan, "Identification of Customers in the CRM system using Data Mining and Fuzzy AHP Method Introduction :," vol. 2, no. 12, pp. 37–53, 2015.

[20] R. Florez-Lopez and J. M. Ramon-Jeronimo, "Marketing segmentation through machine learning models: An approach based on customer relationship management and customer profitability accounting," *Soc. Sci. Comput. Rev.*, vol. 27, no. 1, pp. 96–117, 2009.

[21] R. Ghnemat and E. Jaser, "Classification of mobile customers behavior and usage patterns using self-organizing neural networks," *Int. J. Interact. Mob. Technol.*, vol. 9, no. 4, pp. 4–11, 2015.

[22] R. Ghnemat *et al.*, "Visual Customer Segmentation and Behavior Analysis A SOM-Based Approach," *Expert Syst. Appl.*, vol. 7, no. 1, pp. 33–42, 2018.

[23] Y. Gopi, V. Sumalatha, and A. Professor, "Tele Comm. Customer Data Analysis using Multi-Layer Clustering Model," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 1, no. 1, pp. 1443–1448, 2018.

[24] B. Farhangian, M. Shamsi, and R. Ahsan, "Identification of Customers in the CRM system using Data Mining and Fuzzy AHP Method Introduction :," vol. 2, no. 12, pp. 37–53, 2015.

[25] P. Hanafizadeh and M. Mirzazadeh, "Visualizing market segmentation using self-organizing maps and Fuzzy Delphi method - ADSL market of a telecommunication company," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 198–205, 2011.

[26] K. R. Kashwan and C. M. Velu, "Customer Segmentation Using Clustering and Data Mining Techniques," *Int. J. Comput. Theory Eng.*, vol. 5, no. 6, pp. 856–861, 2013.

[27] S. Y. Kim, T. S. Jung, E. H. Suh, and H. S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert Syst. Appl.*, vol. 31, no. 1, pp. 101–107, 2006.

[28] R. Kohavi and R. Parekh, "Visualizing RFM segmentation," *SIAM Proc. Ser.*, pp. 391–399, 2004.

[29] J. K. Kumar, E. Karunakaran, and K. M. Sabarivelan, "Cluster Based Data Mining Technique for Identification of User Behavior," *Int. J. Eng. Comput. Sci.*, vol. 6, no. 3, pp. 20801–20805, 2017.

[30] N. Lindqvist and D. Thesis, "Green segmentation," *Int. Bus.*, 2010.

[31] S. Masood, M. Ali, F. Arshad, A. M. Qamar, A. Kamal, and A. Rehman, "Customer segmentation and analysis of a mobile telecommunication company of Pakistan using two phase clustering algorithm," *8th Int. Conf. Digit. Inf. Manag. ICDIM 2013*, no. September, pp. 137–142, 2013.

[32] E. Mattila, "Behavioral Segmentation of Telecommunication Customers Master of Science Thesis Behavioral Segmentation of Telecommunication Customers," *R. Inst. Technol. Sch. Comput. Sci. Commun.*, 2008.

[33] D. P. Methods, "Data Preprocessing Techniques for Data Mining."

[34] T. R. Prajwala and V. I. Sangeeta, "Comparative Analysis of EM Clustering Algorithm and Density Based Clustering Algorithm Using WEKA tool .," *Int. J. Eng. Res. Dev.*, vol. 9, no. 8, pp. 19–24, 2014.

[35] M. T. Report, "Master Thesis Report Analysis of free-cooling system for telecom data centres ( Base Transceiver Stations ) - big data analytics and pattern detection model," no. June, 2018.

[36] H. Roshan and M. Afsharinezhad, "The New Approach in Market Segmentation by Using RFM Model," *J. Appl. Res. Ind. Eng.*, vol. 4, no. 4, pp. 259–267, 2017.

[37] S. A. Said, "Telecommunication Engineering Graduate Program Enhancing Mobile Banking Service Availability Using Machine Learning," no. October, 2018.

[38] R. Sankar, "C Ustomer Data Clustering Using D Ata," *Int. J. Database Manag. Syst.*, vol. 3, no. 4, pp. 1–11, 2011.

[39] D. Şchiopu, "Applying TwoStep cluster analysis for identifying bank customers' profile," *Ştiinţe Econ.*, vol. LXII, no. 3, pp. 66–75, 2010.

[40] A. Sciences, "Customer Segmentation By Using Rfm Model and Clustering Methods : a Case," no. July, pp. 1–19, 2018.

[41] J. J. Shen, P. H. Lee, J. J. A. Holden, and H. Shatkay, "Using cluster ensemble and validation to identify subtypes of pervasive developmental disorders.," *AMIA Annu. Symp. Proc.*, no. November, pp. 666–670, 2007.

[42] S. SIMIĆ, "Business Customers Segmentation With the Use of K-Means and Self-Organizing Maps: an Exploratory Study in the Case of a Slovenian Bank," no. December, 2015.

[43] R. Soudagar, "MASTER ' S THESIS Customer Segmentation and Strategy Definition in Segments Customer Segmentation and Strategy."

[44] S. Tripathi, A. Bhardwaj, and P. E, "Approaches to Clustering in Customer Segmentation," *Int. J. Eng. Technol.*, vol. 7, no. 3.12, p. 802, 2018.

[45] K. Tsiptsis and A. Chorianopoulos, *Data Mining Techniques in CRM: Inside Customer Segmentation*. 2010.

[46] A. C. Tynan and J. Drayton, "Market segmentation," *J. Mark. Manag.*, vol. 2, no. 3, pp. 301–335, 1987.

[47] P. T. Upadhyay, "Customer Profiling and Segmentation using Data Mining Techniques," *Int. J. Comput. Sci. Commun.*, vol. 7, no. 2, pp. 65–67, 2016.

[48] L. Vera, "Management in Retailing Supported by Data Mining Techniques," 2012.

[49] Z. T. Victor, "Telecom Customer Segmentation and Precise Package Design by Using Data Mining by using Data Mining," no. October, 2018.

[50] S. W., E. A., and E. Ahishakiye, "Consumer Segmentation and Profiling using Demographic Data and Spending Habits Obtained through Daily Mobile Conversations," *Int. J. Comput. Appl.*, vol. 181, no. 9, pp. 33–42, 2018.

[51] D. S. Wilks, "Cluster Analysis," *Int. Geophys.*, vol. 100, pp. 603–616, 2011.

[52] Z. Yao and T. Dissertations, *Visual Customer Segmentation and Behavior Analysis A SOM-Based Approach*, no. 163. 2013.

[53] AAU School of Commerce,"Market Segmentation Study for ethiotelecom", 2015.

[54] Birhanu B., "Super Unlimited Offer", Dept.Mkting, ethiotelecom,Addis Ababa,Marketing Rep. Feb,2019.

[55] Birhanu B., "Mobile Internet Tariff Revision", Dept.Mkting, ethiotelecom, Addis Ababa, Marketing Rep. Aug,2018.

# APPENDIX A

Table A.I 1: Sample CDR data.

| | B | C | D | F | G | I | J | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SERV_NO | OTHER_NO | CUST_TYPE_NAM | STATUS_NAME | SGMT_NAME | START_TIME | END_TIME | Usage (Minute,#,M | TOTAL_FEE_ETB | voice | CELL_ID |
| 2 | 2519\ | 9622 | 251953 | 3952 | Individual | Active | Prepaid | 20190201002132 | 20190201002145 | 0.33 | 0.115 | voice | 636012412044331 |
| 3 | 251944( | 1505 | 251926 | 6092 | Individual | Active | Prepaid | 20190201004604 | 20190201004621 | 0.33 | 0.115 | voice | 636011802211413 |
| 4 | 251915( | 1062 | 251915 | 5001 | Individual | Active | Prepaid | 20190201012733 | 20190201013323 | 5.83 | 2.0125 | voice | 636011500610033 |
| 5 | 25198/ | .7639 | 251910 | 5943 | Individual | Active | Prepaid | 20190201022116 | 20190201022143 | 0.50 | 0.1725 | voice | 636011100612681 |
| 6 | 25191 | 58255 | 251973 | 4324 | Individual | Active | Prepaid | 20190201034515 | 20190201034531 | 0.33 | 0.115 | voice | 636010150130929 |
| 7 | 25191 | 58255 | 251909 | 0059 | Individual | Active | Prepaid | 20190201050255 | 20190201050306 | 0.33 | 0.115 | voice | 636010150130929 |
| 8 | 25194 | 55496 | 251961 | 0204 | Individual | Active | Prepaid | 20190201054807 | 20190201054830 | 0.50 | 0.1725 | voice | 636011100511807 |
| 9 | 25191 | 58255 | 251973 | 4324 | Individual | Active | Prepaid | 20190201062910 | 20190201062923 | 0.33 | 0.115 | voice | 636010150130929 |
| 0 | 25193 | 28128 | 251949. | 882 | Individual | Active | Prepaid | 20190201062928 | 20190201063019 | 1.00 | 0.345 | voice | 636011200918762 |
| 1 | 25193 | 28128 | 2519442 | 784 | Individual | Active | Prepaid | 20190201064241 | 20190201064359 | 1.33 | 0.46 | voice | 636011200918762 |
| 2 | 25191 | 90471 | 2519626 | 125 | Individual | Active | Prepaid | 20190201065022 | 20190201065041 | 0.33 | 0.115 | voice | 636011200513581 |
| 3 | 25194 | 74756 | 25191680 | 582 | Individual | Active | Prepaid | 20190201065801 | 20190201070034 | 2.67 | 1.0235 | voice | 636012214420091 |
| 4 | 25196 | 40306 | 2519669 | 307 | Individual | Active | Prepaid | 20190201071106 | 20190201071120 | 0.33 | 0.1668 | voice | 636011700820836 |
| 5 | 25191 | 52546 | 251940( | 646 | Individual | Active | Prepaid | 20190201071126 | 20190201071336 | 2.17 | 1.0839 | voice | 636011200715973 |
| 6 | 25190 | 19622 | 251910 | 4527 | Individual | Active | Prepaid | 20190201071614 | 20190201071659 | 0.83 | 0.4169 | voice | 636011400210285 |
| 7 | 25198 | 24015 | 251915 | 4076 | Individual | Active | Prepaid | 20190201071616 | 20190201071735 | 1.33 | 0.667 | voice | 636012113514331 |
| 8 | 25191 | 55670 | 251914 | 5395 | Individual | Active | Prepaid | 20190201071710 | 20190201071821 | 1.33 | 0.667 | voice | 636010120132831 |
| 9 | 25191 | 51064 | 25191/ | 1056 | Individual | Active | Prepaid | 20190201072033 | 20190201072126 | 1.00 | 0.5003 | voice | 636010130132365 |
| 0 | 25191 | 27545 | 2519! | 17369 | Individual | Active | Prepaid | 20190201072209 | 20190201072252 | 0.83 | 0.4169 | voice | 636011700713536 |
| 1 | 25198 | 32398 | 2519: | 11912 | Individual | Active | Prepaid | 20190201072251 | 20190201072335 | 0.83 | 0.4169 | voice | 636011102710237 |
| 2 | 25194 | 30331 | 25192 | 66055 | Individual | Active | Prepaid | 20190201072510 | 20190201072535 | 0.50 | 0.2501 | voice | 636011102016415 |

# APPENDIX B

## Clustering Implementation Results

Table B 1: Clustering result of EM algorithm for voice usage dataset.

```
Number of clusters: 5
Number of iterations performed: 63

                          Cluster
Attribute                      0          1          2          3          4
                          (0.27)     (0.19)     (0.19)     (0.12)     (0.23)
=============================================================================
aver_voiusg_min
  mean                    56.3003   225.6259    17.6972   536.0336   117.0049
  std. dev.               15.5865    50.0821    10.8442   235.7704    23.7254

aver_usgdays_num
  mean                     15.324    23.9885     6.2616     26.841    20.4047
  std. dev.                5.0141     5.1466      3.755     3.9131     5.4463

aver_voi_spen_br
  mean                    27.3057   109.6174     8.5691   261.4821    56.8254
  std. dev.                7.5723      24.11       5.27   117.9178    11.4217

avg_nighvoi_per
  mean                     0.1103       0.11     0.0804     0.1137     0.1146
  std. dev.                0.0711     0.0594     0.0913     0.0565     0.0672

avg_weekendvoci_per
  mean                      0.327     0.3193     0.2632     0.2966     0.3257
  std. dev.                 0.122     0.0919     0.1741     0.0723     0.1062

avg_workhrvoic_per
  mean                     0.6053     0.6097     0.5058     0.5921     0.6118
  std. dev.                0.1323     0.1131     0.2205     0.1121     0.1227

avg_wtoweekusginte
  No                   1861.3751 1378.2902 1163.4541   858.2998 1586.5807
  Yes                     189.67    78.9377    235.683    26.5831   132.1262
  [total]              2051.0451 1457.2279 1399.1372   884.8829 1718.7069

Time taken to build model (full training data) : 2.68 seconds
```

Table B 2 : Clustering result of K-means algorithm for internet usage dataset.

```
Final cluster centroids:
                                Cluster#
Attribute                Full Data          0          1          2          3          4          5
                          (7501.0)   (1924.0)    (647.0)   (1426.0)   (1350.0)    (852.0)   (1302.0)
=============================================================================================================
aveintusg_MB             253.0808    28.3154   424.2448   825.1888    44.0711    34.4415   233.3595
aver_Inter_usgdays        11.7886     3.3043    18.2983    26.5137     6.4498     5.081    14.8889
aver_intspend_br          50.6509     5.6617    85.0329   165.1157     8.8341     6.9185    46.6571
ave_weekendintusg_perc     0.2304     0.0327     0.2904     0.292      0.299     0.3906     0.2491
avg_wtoweekusgint             No         No        Yes         No        Yes         No         No



Time taken to build model (full training data) : 0.09 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1924 ( 26%)
1       647 (  9%)
2      1426 ( 19%)
3      1350 ( 18%)
4       852 ( 11%)
5      1302 ( 17%)
```

Table B 3 : Clustering result of K-means algorithm for SMS usage dataset.

```
Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4          5
                         (7501.0)   (1112.0)    (844.0)   (3572.0)    (584.0)    (858.0)    (531.0)
=============================================================================================================
tot_smsusg_num            10.1145    14.0854    39.3436     0.3877    15.5154     7.7249    18.693
sms_usgdays                2.4426     3.4577     8.9336     0.1837     3.6627     2.6469     3.5235
tot_sms_spend_br           1.4904     2.2392     5.3621     0.0791     2.2429     1.3337     2.6879
nightsms_perc               0.029     0.0704     0.0415     0.0039     0.0113     0.0093     0.1427
weekendsms_per             0.1849     0.0879     0.3339     0.0088     0.8137     0.0633     0.8406
workinghrsms_per           0.2615     0.1338     0.5525     0.0029      0.852     0.9017     0.1223
wtwusginten                    No        Yes         No         No        Yes        Yes        Yes



Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       1112 ( 15%)
1        844 ( 11%)
2       3572 ( 48%)
3        584 (  8%)
4        858 ( 11%)
5        531 (  7%)
```