Market Segmentation of Mobile Internet Customers Using Clustering Algorithms: The Case of Ethio Telecom

BY: BISHAW YADESSA
SUPERVISOR: EPHREM TESHALE (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
(Telecommunication Engineering)

Addis Ababa University
Addis Ababa, Ethiopia
February 25, 2020

Declaration

I, the undersigned, declare that the thesis comprises my own work in compliance with internationally accepted practices; I have fully acknowledged and referred all materials used in this thesis work.

Bishaw Yadessa

Name

Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Bishaw Yadessa**, entitled *Market Segmentation of Mobile Internet Customers Using Clustering Algorithms: The Case of Ethio Telecom* and submitted in partial fulfillment of the requirements for the degree of Master of Science (Telecommunication Engineering) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner 1	______________	Signature	________	Date	______
Examiner 2	______________	Signature	________	Date	______
Supervisor	Ephrem Teshale (PhD)	Signature	________	Date	______

____

Dean, School of Electrical and Computer Engineering

ABSTRACT

Telecom companies utilize data mining algorithms and tools to understand the behaviors of their customers. Cluster analysis is one of the techniques used to identify homogenous groups of customers from a heterogeneous group based on the customers' service usage records. Clustering algorithms aim to find natural groupings of subscribers and are widely applied for customer profiling and market segmentation.

As telecom customers use services like cellular Internet, huge amount of data is generated in the form of call detail record (CDR) primarily for billing purpose. This data can also be a source of information for marketing and network management tasks. Previously, different algorithms have been studied and implemented by researchers to understand how, why, when and where people access mobile Internet using already available data from the telecom systems.

In this thesis, two clustering algorithms namely K-means and Two-Step were implemented on real CDR and CRM datasets of a telecom company in Ethiopia to segment mobile Internet users. Behavioral segmentation was performed using aggregated data of sample number of customers based on derived features. Moreover, the insights obtained from each segment were analyzed before suggesting marketing strategies for personalized services and targeted campaigns.

K-means was applied on CDR dataset and meaningful clusters showing characteristics of users were obtained and explained. Two-Step clustering was found to be more suitable for segmenting user groups and has a better silhouette score than K-means. The results of the analysis show the possibility of using data stored by mobile operators for market segmentation purposes. It was examined from this research that clustering algorithms like K-means and Two-Step are applicable to observe patterns among ethio telecom customers based on service usage and demographic information of subscribers from database systems .

Keywords: Mobile Internet, Market segmentation, Cluster analysis, CDR, K-means, Two-Step.

ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to my advisor, Dr. Ephrem Teshale, for his time, continued support and excellent comments. He always listened and encouraged me. The role of the thesis committee members was also great. Their wise counsel during the seminars shaped the way I work on this project.

My colleagues at ethio telecom provided computational resources, data and firsthand knowledge. My fellow students especially Derebe and Getachew have been instrumental to my success. I couldn't have written this paper without the help of these people. Moreover, I would like to thank ethio telecom, my employer, for giving me a chance to pursue my masters degree and for financing this research.

I am grateful to my family for their unreserved support throughout the course work and during this project. Loza Abayneh, my wife, is a wonderful person. She has been on my side. Thanks also to our beloved daughter, Eva, for making our home joyful. My friend Tekeste deserves my appreciation for his frequent encouragement and for the invaluable feedback.

Finally, I would like to say praise and thanks be to the Almighty God, who is the source of all good things. And to my savior Jesus Christ for his eternal love and mercy.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

3G	Third Generation
CBS	Convergent Billing System
CDMA	Code Division Multiple Access
CDR	Call Detail Record
CRM	Customer Relationship Management
CSV	Comma Separated Values
DCNO	Data Collected by the Network Operator
DCSP	Data Collected by the Sensors on the Phone
EDGE	Enhanced Data Rates for GSM Evolution
ETC	Ethiopian Telecommunications Corporation
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HSDPA	High Speed Downlink Packet Access
LTE	Long Term Evolution
M2M	Machine to Machine
MBB	Mobile Broadband
MSISDN	Mobile Station International Subscriber Directory Number
NB	Narrowband
SMS	Short Message Service
SOM	Self Organizing Map
UMTS	Universal Mobile Telecommunications System
WCDMA	Wideband Code Division Multiple Access

1

INTRODUCTION

Mobile phones are now the primary tools for connecting to the Internet. Previously, personal computers were commonly used devices for people's digital presence. The innovation of smartphones, improvements in cellular network coverage, affordability, literacy rate and content productions have resulted in rapid growth in mobile Internet adoption. According to the latest GSMA report, by 2025, 5 billion people across the globe (more than 60% of the population) will be mobile Internet subscribers [1]. That means three out of five people in the world will access the Internet through their smartphones. In Ethiopia, as of end of 2018, the total number of Internet users was 18.62 million and about 96% were mobile data users [2].

Understanding the characteristics of customers of specific services is essential for mobile operators such as ethio telecom. In this regard, the rapidly growing cellular data users deserve particular attention. Analysis of customers' behavior benefits service providers and users alike. This provides the telecom companies with important information that can be used for corporate strategies, network planning, service design, marketing promotion and so on. The customers on the other hand will benefit from targeted offers and quality services to mention a few.

Cluster analysis is one of the most widely used data mining techniques to categorize subscribers into distinct groups, and one of the major areas of its application in telecom is market segmentation [3][4]. Market segmentation helps service providers give personalized offers, better price plans and quality services for their customers.

Various researchers have used clustering algorithms like K-means, Self Organizing Map (SOM) and Two-Step on different datasets obtained from the telecom industry to segment customers for marketing purposes [3–9]. In this study, real CDR dataset is used to perform cluster analysis of mobile Internet customers. Data service subscribers differ in their frequency of use, consumption amount, service choices and price plans. Therefore, behavioral based segmentation was applied using two clustering algorithms namely K-means and Two- Step. To the best of the author's knowledge, these algorithms were not used together for market segmentation to date. There is also a need to study ethio telecom mobile Internet users' behavior based on CDR. Hence, this study will attempt to fill these gaps.

1.1 STATEMENT OF THE PROBLEM

Advances in tools and algorithms of data analytics enabled companies such as telecom operators to make the best use of the data they store. In the past, telecom companies lacked the different computational resources to extract meanings out of the different types of data they stored for decades. Consequently, market segmentation and other business decisions were done based on small samples. Surveys are still the preferred method for marketing studies at ethio telecom, and the maximum sample size taken in so far was around 1,500 [10]. The main problem with survey method is that the results are biased by the subjectivity of the participants' answers [11].

Call detail records, rather, enable access to the information of millions of people at a time that are real observations. And as described above, researchers have employed cluster analysis techniques on telecom datasets like CDRs for market segmentation and other purposes. Studying consumption behaviors of customers using real billing CDR dataset and possibly combining it with their demographic information from CRM can be a better alternative to surveys.

This thesis presents segmentation of mobile Internet customers based on their service utilization behavior using unsupervised machine learning algorithms. The following two major metrics are taken as criteria for usage analysis: the preferred time period customers want to use the service (time preferences) and the amount they consume (data volume). K-means and Two-Step algorithms will be evaluated for segmenting users according to these metrics. In order to achieve such segmentation of customers using the algorithms and metrices, it is necessary to answer the following two research questions:

1. What CDR features are suitable to describe cellular data users for market segmentation?
2. Which clustering algorithm better segments mobile Internet customers with CDR data ?

1.2 OBJECTIVES

The main objective of this study is to analyze mobile Internet customers with clustering algorithms and find segments for marketing based on real CDR dataset. The specific objectives are as follows:

- To analyze the behaviors of mobile Internet users and select appropriate CDR features for market segmentation.
- To evaluate clustering algorithms based on actual usage data.

1.3 SCOPE AND LIMITATIONS

The research is restricted to analyzing mobile Internet users in Ethiopia with data from ethio telecom, the sole telecom service provider in the country. This is done by performing cluster analysis on actual usage data of customers. The final goal of the research is to segment and profile users according to their behavior. Further study is required to propose working technical and business deployment models in the future.

1.4 CONTRIBUTIONS OF THE RESEARCH

Previous studies conducted on customer segmentation in ethio telecom were based on primary data collected through surveys. However, in this thesis, the researcher has tried to use already available data or the CDR dataset that the network operator stores for billing purpose. The finding of the research can be used by marketers for designing new services and offers. The use of clustering algorithms for segmentation purposes in telecommunications companies supports the need for a more efficient and practical approach of applying clustering techniques to specific types of services.

1.5 LITERATURE REVIEW

Review of the previous body of work show that different variables from real data were used to segment telecom customers. Through combining beneficial and behavioral features, Namvar, et al. [3] used CDR to build a customer segmentation framework for marketing. In [4], subscribers were segmented based on their call, SMS and revenue variables. Luo Ye, et al.

[5] included value information of customers for showing characteristics of different groups with figure. Customers were also clustered into 7 groups based on historical data [7]. Table 1.1 summarizes these four works.

Table 1.1: Summary of Literature on Customer Segmenation in Telecom

Ref.	Methodology	Key Findings	Limitations
[3]	Data : CDR Algorithm(s): K-means	Using beneficial & behavioral features separately is better than considering all features in one phase.	Marketing strategies did not consider categorical values.
[4]	Data : call usage, sms usage and revenue generation Algorithm: Two-Step clustering algorithm	VAS usage was greater than the SMS usage for all customers' segments.	Analysis of each segment based on usage records of different services is not performed.
[5]	Data: basic, value and behavioral information of customers Algorithm: K-means	Consumption characteristics of groups can be showed in figure.	Small number of customers were selected
[7]	Data: historical data Algorithm: SOM	7 groups of customers were identified	Internet users were not included

The use of clustering algorithms has interested researchers in recent years. In addition to the studies listed above, for example, authors in [6] & [9] used Two-Step and K-means algorithms respectively for analyzing behaviors of customers for various segmentation reasons. Furthermore, with millions of customers, analyzing each individual customer's usage pattern becomes a challenging task and sophisticated data mining techniques need to be applied [8]. Segmenting customer employing data mining techniques is a challenging task which is confronted with two types of problems: 1) selecting or summarizing suitable features from a huge amount of raw transactional data. 2) Proposing an algorithm that has the best performance on huge dataset [3].

Ethio telecom segments its customers using different variables. However, for market researches and business strategy decisions, it uses primary data collected through questionnaires. Primary data collected from sample population has two major limitations. One, for different reasons including cultural and personal attitude, respondents may not speak the truth. Hence, surveys may provide us with inaccurate responses. Second, the samples may not be representative enough to show us patterns in usage and different behaviors in the same way analysis of big data would.

In this thesis, however, service usage information which is stored for billing purpose will be utilized as the main source of clustering. In addition, so far all the three major mobile services namely voice, SMS and Internet were all included when marketing studies were conducted. The scope of this thesis is however, exclusively limited to market segmentation of mobile data service customers.

1.6 THESIS ORGANIZATION

The rest of the thesis is organized as follows. Chapter 2 gives the background. It discusses market segmentation concepts and customer behavior analysis in telecom. An overview of mobile networks and services is also included. Chapter 3 provides theoretical knowledge to clustering algorithms giving emphasis to the algorithms used in the research. The data collection, preparation and analysis methods are then described in chapter 4. The results are presented and discussed in chapter 5. Finally, conclusions and recommendations are provided in chapter 6.

2

MARKET SEGMENTATION IN TELECOM

This chapter provides a discussion on the basic concept of market segmentation and presents literature on analysis of mobile Internet customer behavior. Overview of cellular networks and datasets available in telecom industry are also discussed.

2.1 THE CONCEPT OF MARKET SEGMENTATION

Throughout this thesis, customer segmentation and market segmentation are used interchangeably, unless indicated otherwise. However, it is important to note that market segmentation concerns with high level strategy while customer segmentation is about gaining insights on customer habits. Thus, in this thesis, market segmentation is all about customers not others like business partners and prospects.

Market segmentation is dividing markets into groups (segments) of consumers for the purpose of successfully matching of products and services to the needs of customers. The concept was first defined by Smith in 1956. According to him, market segmentation involves viewing a heterogeneous market as a number of smaller homogenous markets, in response to differing preferences, attributable to the desires of consumers for more precise satisfaction of their varying needs [12].

Market segmentation is one of the most fundamental marketing activities. Companies need to divide markets into groups (segments) of consumers and customers for providing customized products, targeted advertisements, and pricing strategies. Tools like segmentation have the greatest impact on marketing decisions, and proved to be the key to success as the most successful companies are those who carefully segment their markets [13].

In short, segmentation is the process of dividing an entire customer base into groups of similar customers depending on their characteristics so that targeting each group is possible. Customers in a well segmented single group share useful similarities in demographic, behavioral and other variables. For example, a telecom service provider can offer a special price or gifts to a certain group.

There is no single way to segment a market as a marketer needs to try different segmentation variables, separately and in combination, to find the best way of viewing the market structure [14]. Generally, there are four major segmentation variables; namely geographic, demographic, psychographic, and behavioral variables.

- *Geographic segmentation*: geographical unites such as nations, regions, states, cities, or even neighborhoods are used to segment the market.
- *Demographic segmentation*: is based on variables such as age, life-cycle stage, gender, income, occupation, education, religion and ethnicity.
- *Psychographic segmentation*: divides buyers into different segments based on social class, lifestyle, or personality characteristics.
- *Behavioral Segmentation*: groups customers based on their knowledge, attitudes, uses, or responses concerning a product. Many marketers believe that behavior variables are the best starting point for building market segments.

2.2 ANALYSIS OF CUSTOMER BEHAVIOR IN TELECOM

The convenience of smartphones for easy Internet browsing has contributed to the rapid increase in the number of their users. With the ever increasing popularity of mobile Internet, telecom companies have also come to realize the need to deploy various technologies that can give them insight into what their customers want a priori. This notion can perhaps be best described by the words of Steve Jobs, the founder of Apple company who noted "Get closer than ever to your customers. So close that you tell them what they need well before they realize it themselves [15]."

Understanding customers' mobile usage behaviors is essential for successful service design, product marketing and network planning. Some of the business decisions telecom companies can base on the results of mobile customers' usage behavior analysis include: customer identification for customized product offers, building referral-based promotions from a user's social network, recognizing leaders or influential users to use them as brand ambassadors, and offering location aware services.

Studies have been conducted on characteristics of telecom service subscribers to answer questions related to usage pattern analysis, customer retention, product differentiations, resource management and so on. Some used simple statistical analysis [16] and others applied advanced data mining techniques [17] to observe patterns among telecom customers. Literature related to customers behavior analysis in telecom industry are listed below.

Jiang, et al. [17] observed 4G users behavior from mobility and application usage perspectives. Using K-means to determine the best clustering choice for users' occurrence and application usage, they found out that mobility pattern was highly correlated with application usage.

The study by Liao, et al. [18] also analyzed heavy users of online mobile video services using statistical tests. Their observation of user behaviors of several competing online video apps using network connection records of mobile and broadband customers indicated that female and the youth account for higher proportions, travel more, have more contacts to call and download more data.

Jie Yang, et al. [19] studied mobile data usage, mobility pattern and application usage behavior by collecting mobile traffic data at core metropolitan 2G and 3G networks over a week. They used HTTP (Hypertext Transfer Protocol) flows of mobile network traffic data containing phone no (anonymized user identifier), timestamp, total number of packets & bytes for this flow, URI, hostname & server IP. They concluded that both the data usage and the mobility pattern are closely related to the application access behavior of the users. Data usage among mobile Internet users was found to be highly uneven.

Authors in [20] analyzed the behavioral differences of two types of users(i.e. teachers and students) in making calls, sending messages and surfing the Internet, and identified the users by machine learning algorithm according to the characteristics of users. Teachers were found to have large social circles than students (social circle of teacher is more diversified) and students' social circle can be assumed as confined among students themselves. The streaming data usage between the two user groups is similar.

Similarly, Yalcin [21] presented probabilistic model to predict the user profile of the customer, namely gender and age, based on usage behaviors of mobile applications, phone brand and model from a statistical characteristic profiled from dataset of TalkingData company. This can be used to track user tendencies important to deliver personalized services like targeted advertisement.

Mobile Internet allows high-speed data connectivity for people with smartphones, regardless of their location and time. As the focus of this study is to distinguish between different user groups in mobile Internet, it is necessary to discuss about mobile technology and services.

2.3 OVERVIEW OF MOBILE INTERNET TECHNOLOGY

Telecommunications industry has advanced in many aspects. In terms of technology, analog transmission and switching systems are obsolete. Now digital technologies are available everywhere. This section presents how mobile networks evolved from 1G to 5G shortly after.

Services offered by telecom companies have also changed. Voice and SMS are no longer the only services. Innovations in mobile phone technologies resulted the adoption of Internet access through mobile handsets. Now people with smartphones can use online services, transmit video calls and use social media conveniently almost anywhere and anytime.

Subscribers are purchasing multipurpose smartphones that handle Internet access and apps in addition to voice. In the face of declining voice revenue, carriers are seeking to increase revenue and profits by promoting data plans to customers and upgrading networks for additional capacity [22].

2.3.1 *Mobile Networks and Services*

The first cellular systems or 1G (1st Generation) use analog technology and support voice only calls. In analog systems, information is represented as a continuous electromagnetic waveform. Whereas digital communications involve modulating (i.e., changing) the analog waveform in order to represent information in binary form (1 s and 0 s). Digital systems provide improved error performance, better bandwidth utilization through compression, enhanced security through encryption and support to data communications.

Digital cellular network in Europe was launched for the first time based on Global System for Mobile Communications (GSM) standard in 1991. GSM operates in the 800 - MHz and 900 - MHz frequency bands. It became the primary international standard for mobile networks. GSM also called 2G or the second-generation cellular technology started offering all digital data communications advantages enjoyed in the wired world.

General Packet Radio Service (GPRS) is a 2.5G cellular technology designed to enhance the data service of GSM. GPRS is a packet - switched service that enables high - speed mobile data usage with a theoretical transmission rate as high as 171.2 kbps. An extension development of the GPRS known as Enhanced Data Rates for GSM Evolution (EDGE) supports data rates as high as 473.6 kbps. GPRS support simultaneous voice and data communications over the same wireless link, with voice taking precedence [22].

Universal Mobile Telecommunications System (UMTS) which uses Wideband Code Division Multiple Access (WCDMA) technology is a 3G technology that is seen as a logical upgrade to GSM, although the two are not compatible. UMTS specifications include 128 kbps for high - mobility applications, 384 kbps for pedestrian speed applications, and 2 Mbps (1.920 Mbps) for fixed in - building applications.

UMTS networks are currently being upgraded with High - Speed Downlink Packet Access (HSDPA), which sometimes is characterized as a 3.5G technology. (HSDPA) promises to increase theoretical downlink data rates to 14.4 Mbps, although current implementations support speeds more typically in the range of 400 – 700 kbps, bursting up to 3.6 Mbps for short periods of time using an adaptive modulation technique to throttle bit rates up and down as the link permits [22].

LTE (Long Term Evolution) is the fourth generation (4G) mobile network technology that supports services such as gaming, video conferencing, home Internet access, HD mobile TV and other features in addition to mobile web access supported by 3G. These services require high speed connection. The maximum speed of a 4G network for moving device is 100Mbps. A speed up to 1Gbps can be achieved for a stationary or walking user. Most current cell phone models support both 4G and 3G technologies.

5G promises significantly faster data rates (speeds), higher connection density, much lower latency, and energy savings, among other improvements from the previous generation. The expected theoretical speed of 5G connections is up to 20 Gbps. These will be achieved because of advancement in radio technology and availability of more spectrum for 5G. Table 2.1 shows the difference between mobile network technologies avaliable in most part of the world including in our country.

GSM and CDMA are the most commonly deployed mobile phone networks. They use different algorithms which allow multiple mobile phone users to share the same digital radio frequency without causing interfering for each other. Cell phones have an in-built antenna which is used to send packets of digital information back and forth with cell-phone towers via radio waves. Mobile phones connect to a cell tower in the area, and instead of connecting to another phone it connects to the Internet and can fetch or retrieve data.

Table 2.1: Comparison among 2G, 3G and 4G (Adopted from [23])

Technology Features	**2G**	**3G**	**4G**
Deployment	1990 - 2004	2004-2010	Now
Data Bandwidth	64Kbps	2Mbps	1Gbps
Technology	Digital cellular	CDMA 2000, UMTS, EDGE	WiMax, LTE, Wi-Fi
Service	Digital voice, SMS, Higher Capacity Packetized data	Integrated high quality audio, video and data	Dynamic information access, Wearable devices
Multiplexing	TDMA, CDMA	CDMA	OFDM
Switching	Circuit, packet	Packet	All Packet
Core Network	PSTN	Packet network	Internet

2.3.2 *Telecom Service in Ethiopia*

Telecommunication service was introduced in Ethiopia in 1894. As of the writing of this paper, ethio telecom formerly known as Ethiopian Telecommunications Corporation (ETC) is the only service provider in the country which is wholly owned by the government.

Mobile service has begun in 1999 with a capacity of 36, 0000 subscribers in Addis Ababa [24]. Currently, there are around 42 million mobile customers [25]. Mobile Internet is one of the services offered by ethio telecom. Also known as mobile data service, mobile Internet enables content delivered to mobile devices such as smartphones and tablets over a wireless cellular connection. 2G (GPRS), 3G (HSDPA) and 4G (LTE) technologies are used for providing mobile Internet service in Ethiopia. Table 2.2 shows the coverage of technologies deployed in Ethiopia.

Table 2.2: Mobile Communications Technologies Deployed by ethio telecom

Technology	**Coverage**
GSM	85.5% countrywide
GPRS	85.5% countrywide
EDGE	66 % countrywide
3G	66% countrywide
4G LTE	1.1% countrywide

Ethio telecom gives two options for using mobile data service. Pay as you go mobile Internet allows customers to pay for what they used in real time with a flat tariff rate of 0.20 cents per MB. Mobile packages based on customers' needs are also available. These are daily, night, weekly, weekend and monthly packages.

Mobile Broadband(MBB) customers of ethio telecom are broadly grouped as residential and enterprise. Residential customers are private users and enterprise customers are those subscribed as a business. From the total of 21,739,665 active data users, 99% are residential customers. Table 2.3 describes the distribution of Mobile Broadband (MBB) customers based on the type of service subscription. 1X, GPRS or M2M users are called Narrowband (NB) users and were not considered in the study. There are about 297,570 (NB) users.

Table 2.3: Mobile Broadband (MBB) customers of ethio telecom

Type of Subscription	**Residential**	**Enterprise**
3G Data Only	71,879	29,177
3G On Mobile	21,238,909	91,355
LTE Data Only	10,786	11,351
LTE On Mobile	253,560	29,730
EVDO	339	2,579
Total	21,575,473	164,192

2.4 TYPES OF DATA IN TELECOM

Telecom data is data coming from telecommunication networks. Each device that uses networks in telecom system is producing telco data that operator can collect. For example, mobile phone with mobile subscription leaves digital traces almost all the time. This happens because functioning mobile phones are all the time connected with operator's infrastructure [26].

Telecom operators collect huge amount of data daily. The interactions users have with the network can be a useful source of information to understand both the customers and the network. Telco data is a data source that has multiple advantages like, it is fast real-time data, is easily collectable and has wide user range. It can reveal specific habits of users and help in speeding up network operation activities.

Data collected from mobile phones can be categorized in many ways. Based on the type of the entity collecting the data, it can be classified as either data collected by the network operator (DCNO) or data collected by the sensors on the phone (DCSP) [21]. Call detail records (CDRs) are type of(DCNO). Network operators collect this data for monitoring the network traffic and for billing purposes.

Similarly, authors in [27] define data stored by the mobile service providers as groups of mobile datasets generated as a result of interactions between a mobile user and the mobile network systems. A user interacts with a mobile operator to sign-up for services, to make a call, to send text, to pay a bill, to visit a website and so on. These interactions forms location and movement data, financial and economic data, identity and demographic data, social/browser data, usage, sentiment and trends, and diagnostic/ambient conditions.

One of the concerns of this thesis is to show the possibility to utilize datasets already available in telecom providers data warehouses for proper market segmentation of mobile Internet customers. Therefore, at least at this stage of the research work, data will be exclusively based on the CDR and customer demography data collected from the operator's network. Next, the two types of datasets are briefly discussed.

2.4.1 *Call Detail Record*

According to Peterson [28], CDR is defined as a telephone record-keeping system, usually used for accounting and administrative purposes, that tracks and records details about incoming and outgoing calls such as the call duration, caller and/or callee, time of day, etc.

CDR data contains mobile network events from , voice calls, data connections, value-added services and all other types of services. Telecom companies generate huge amount of CDR every day because they record all usage events by their customers. Although the main purpose of collecting CDR is to generate financial transactions like bills of subscribers, the data is full of other precious information that can be used for marketing, fraud detection, location-based analysis, user behavior identification and so on.

CDRs include lots of information that can be used to analyze location, financial, demographic, identity and other user behaviors. CDRs are relatively simple to aggregate, analyze, and interpret because they follow standard formats. Therefore, it can be said that the CDR dataset will remain the single most important source of information for research efforts [27].

In ethio telecom, billing CDRs are found from the Convergent Billing System (CBS). CBS is Huawei's billing data management system. It processes and stores charge data of all services offered by the company.

2.4.2 *Demographic Information*

Demographic information is part of customer data. Customer data consists of basic, demographic, psychographic, and other relevant information about customers. The basic information is customer's general data like name, address and phone number. Demographic information includes age, gender, marital status and education. Data related to customer's values, activities, and interests is the psychographic information [29].

CRM databases are network components telecom companies use to store customer data. In ethio telecom, although there is a CRM system, its application is limited to keeping basic information and demographic information. Neither of which are readily obtainable for research purpose. First, these data are not properly filled. Second, it is not possible to know whether the information stored about a customer implies the actual user of the service or not. The psychographic information is nonexistent as far as ethio telecom's experience is concerned.

3

CLUSTER ANALYSIS

Chapter three gives an overview of cluster analysis. Unsupervised learning techniques are discussed first. Then, basic types of clustering methods, clustering algorithms used by telecom service providers for customer segmentation and how to evaluate clustering algorithms are presented.

3.1 UNSUPERVISED LEARNING ALGORITHMS

Clustering algorithms are part of unsupervised machine learning techniques. Machine learning investigates how computers can learn (or improve their performance) based on data [30]. In unsupervised learning, you do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information. It deals primarily with unlabeled data. In unsupervised learning you are not taught but you learn from the data. Whereas, in supervised learning, algorithms are trained using labelled data.

In machine learning, clustering and classification are the two commonly used tasks for data categorization and analysis. The former one is unsupervised learning algorithm while the latter is supervised learning algorithm. Clustering labels previously unlabeled objects or data with unknown classes. Therefore, clustering is sometimes called unsupervised classification. But in classification, data objects are grouped based on already assigned labels or known classes. The following table 3.1 better explains the difference between classification and clustering.

Table 3.1: Clustering vs Classification

Basis for Comparison	**Clustering**	**Classification**
Type of learning	Unsupervised	Supervised
Data needs	Unlabeled	Labeled
Training set	Not required	Required
Aim	Find similarities with a dataset	Assign new input into a class
Algorithm example	K-means, Expectation Maximization	Regression, Decision Trees
Use case	Marketing, Recommendation System	Spam filtering, sentiment analysis

3.2 OVERVIEW OF CLUSTERING

Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters [30]. Clustering is a technique used for identifying similar groups of data objects or clusters. It has been used in many fields. The broad areas where cluster analysis is used include market segmentation, information retrieval, pattern recognition, fraud detection, web search analysis, bio informatics and image processing.

Cluster analysis is sometimes called segmentation because its goal is to segment data based on similarity and dissimilarity of the points in the data. For example, in market segmentation, customers which share common features are organized into the same cluster.

Clustering is also used in building better customer relation management (CRM) . Customer information is better organized into manageable groups or segments with the help of clustering algorithms. Here the main goal of clustering algorithms is to differentiate homogeneous and heterogeneous groups in the customers database. Once segments are created, companies can easily access, process and manage the data. Understanding the difference between groups help organizations develop different approaches for achieving their business goals such as profiling their user base for efficient target marketing.

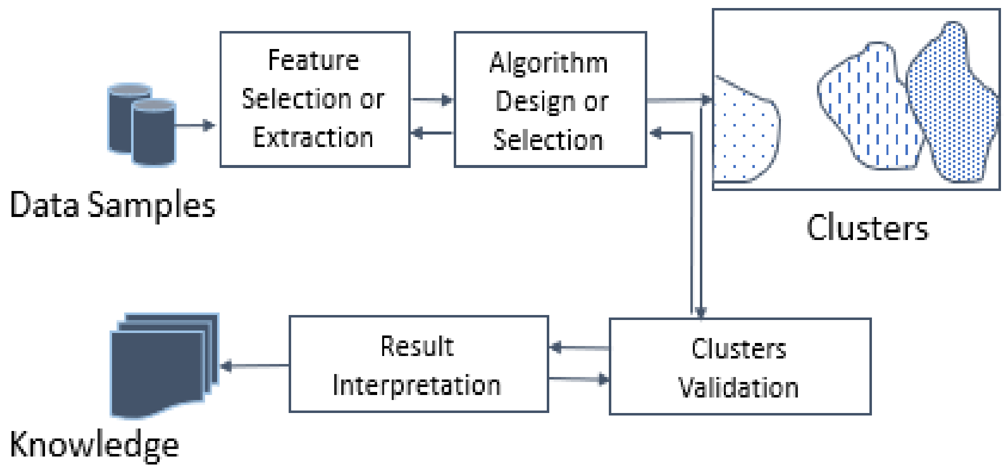


Figure 3.1: The Four Steps in Clustering (Adopted from [31])

Typical cluster analysis consists of four steps as depicted in figure 3.1. The first step is feature selection or extraction is where features are chosen from a set of candidates or extracted from original ones through transformations. The second is clustering algorithm design or selection step which is excuted to select or design appropriate clustering strategy for the specific problem domain. Cluster validation is the third step and it is used to provide a degree of confidence for the clustering results through objective assessment of algorithms. Finally, the goal of clustering is to get meaningful insights from the original data through interpretation of results [31].

3.2.1 *Basic Types of Clustering*

There are four major types of clustering methods, namely, partitioning, hierarchical, density-based and grid-based. Each method is described below.

1. **Partitioning methods** divide dataset into a set of disjoint clusters with the aim of discovering a structure already present in the data. A partitioning method constructs k clusters from a set of n objects where

   $$k \leqslant n$$

   In other words, partitioning algorithms organize the objects of a set into several exclusive groups or clusters. Each group contains at least one object and each object belongs to exactly one group (exclusive cluster separation). Partitioning methods try to minimize dissimilarity within samples belonging to the same cluster while maximizing the dissimilarity between different clusters [30]. Some of the well-known clustering algorithms based on partitioning are K-means, K-medoids, and SOM. K-means is chosen for this thesis because the algorithms is used for practical customers segmentation related applications like the one this thesis tries to answer.

2. **Hierarchical methods** integrate smaller clusters into larger clusters or separate larger clusters. The two types of hierarchical clustering agglomerative (bottom-up) and divisive (top-down) methods. The agglomerative method begins with a separate group being created by each element. It merges objects or groups similar to each other successively until all groups are combined into one (the topmost level of the hierarchy), or a termination condition holds. The divisive approach, also called the top-down approach, starts with all the objects in the same cluster. A cluster is divided into smaller clusters in each successive iteration until each object is finally in one cluster or holds a ter-

mination condition. Hierarchical clustering methods can be distance-based or density- and continuity based [30].

Two-Step clustering can be grouped under hierarchical clustering. As the name implies, the clustering process has two steps. First, the entire dataset is scanned and a large number of small primary clusters are found. But unlike classical hierarchical clustering techniques, it forms subsclusters by resizing the initial input records. In other words, in the first step or the preclustering step assigns records to a limited set of initial subclusters by making a single pass through the data. The second step, further groups the initial subclusters through hierarchical clustering into the final segments [32].

3. **Density-based methods** continue growing a given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold. For example, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise or outliers and discover clusters of arbitrary shape.These are unlike partitioning methods which cluster objects based on distance and are only suitable for finding spherical-shaped clusters [30].

4. **Grid-based methods** quantize the object space into a finite number of cells forming a grid structure. All the clustering operations are performed on the grid structure (i.e., on the quantized space). It has fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. Using grids is often an efficient approach to many spatial data mining problems, including clustering [30].

The following table shows the differences more clearly.

Table 3.2: Overview of Clustering Methods [30]

Method	**General Characteristics**
Partitioning methods	- Find mutually exclusive clusters of spherical shape - Distance-based - May use mean or medoid(etc.) to represent cluster center - Effective for small-to medium-size data sets
Hierarchical methods	- Clustering is a hierarchical decomposition (i.e. multiple levels) - Cannot correct erroneous merges or splits - May incorporate other techniques like microclustering or consider object 'linkages'
Density-based methods	- Can find arbitrarily shaped clusters - Clusters are dense regions of objects in space that are separated by low-density regions - Cluster density: Each point must have a minimum number of points within its "neighborhood" - May filter out outliers
Grid-based methods	- Use a multiresolution grid data structure - Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

3.2.2 *Clustering Algorithms for Segmentation*

In this thesis, K-means and Two-Step clustering algorithms are applied on mobile Internet subscribers CDR dataset. Other researches on customer segmentation were done using these algorithms as already discussed previously. Next, the rationale for choosing these algorithms will be addressed.

K-means clustering algorithm is adopted for the following advantages it offers [5]:

- The algorithm gives a good solution for the clustering problem of data objects with numeric attributes and often terminates at a local optimum.
- The algorithm is relatively scalable and efficient in processing large data sets.
- The algorithm is not sensitive to the input order of data.
- Although, the algorithm lacks the ability to process noisy data, the telecom data are relatively complete and the data pre-processing can make up for it.
- The algorithm is fast in modeling and its results are easier to understand.

Two-Step clustering is chosen because some of the features extracted from the datasets for the thesis are of mixed data (numerical and categorical). Two-Step cluster analysis is said to segment data based on any form of data measurement unlike K-means that require numerical values for effective segmentation. Other advantages include its ability to work with large datasets and in automatically determining the number of clusters [6].

The detail descriptions of the two clustering algorithms are provided below.

3.2.2.1 *K-means*

The K-means clustering technique is said to be the simplest and mostly used clustering algorithm. It is a prototypical unsupervised learning algorithm (one which doesn't assume the availability of labeled training data).

The goal of K-means is to partition dataset into k clusters such that each point in a cluster is similar with points from its own cluster than with points from some other cluster. The algorithm tries to find groups in the data, with the number of groups represented by the variable K. K-means works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The working principle of K-means algorithm can be summarized by the following flowchart adopted from [30]:

Input:

- k: the number of clusters
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;

(2) **repeat**

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;

(5) **until** no change

The K-means algorithm is used to segment a set of observations into a predefined number of k clusters. The algorithm as described starts with a random set of k center-points (c_i). All observations of x are assigned to their nearest center point during each update process. In the standard algorithm, only one assignment to one center is possible. A random center would be selected when several centers are within the same distance of observation.

Afterwards, the center-points are repositioned by calculating the mean of the assigned observations to the respective center-points (see 3.1).

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x \tag{3.1}$$

where x is an observation; C_i the ith cluster; c_i is the centroid of cluster C_i; m_i is the number of objects in the ith cluster . The update process will continue until all observations remain at the assigned center-points and therefore the center-points will no longer be changed. The K-means algorithm tries to optimize the objective function or sum of squared error (SSE) which measures the quality of clustering 3.2. As there is only a finite number of possible assignments for the amount of centroids and observations available and each iteration has to result in better solution, the algorithm always ends in a local minimum.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \tag{3.2}$$

3.2.2.2 *Two-Step*

The Two-Step clustering method is a scalable cluster analysis algorithm designed to handle very large data sets. It can also handle both continuous and categorical variables or attributes. It requires only one data pass. It has two steps 1) preclustering: the cases (or records) are pre-clustered into many small sub-clusters; the precluster stage groups the respondents into several small clusters. 2) hierarchical clustering: cluster the sub-clusters resulting from pre-cluster step into the desired number of clusters. It can also automatically select the number of clusters. The cluster stage uses the small clusters as input and groups them into larger clusters [6].

Two-Step cluster analysis is based on hierarchical clustering. The algorithm identifies groups of cases that exhibit similar response patterns. However, the analyst may specify a percentage of noise (cases not belonging to any cluster). The distance of the respondent to the nearest nonnoise cluster and to the noise cluster decides the membership of the segment. Respondents who are nearest to the noise cluster are considered outliers. Based on well-defined statistics, the procedure can automatically select the optimal number of clusters given the input variables [13]. There are two types of distance measures for Two-Step algorithm: Euclidean distance and a log-likelihood distance. The log-likelihood distance which is a probability based distance is the one that can handle both continouus and categorical types of data. The log-likelihood between two clusters i and j is defined as 3.3:

$$d\,(i,j) = \xi_i + \xi_j - \xi_{(i\ ,j)} \tag{3.3}$$

where

$$\xi_v \ = \ -\, N_v \ (\sum_{k\,=\,1}^{K^A} \frac{1}{2} \log\ (\widehat{\sigma}_k^2 \ + \ \widehat{\sigma}_{vk}^2 \ ) \ + \ \sum_{k\,=\,1}^{K^B} \hat{E}_{\ vk} \ ) \tag{3.4}$$

and

$$\hat{E}_{\ vk} \ = \ \ - \sum_{l\,=\,1}^{L_k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v} \tag{3.5}$$

where ξ_v can be interpreted as a kind of dispersion (variance) within cluster v; K^A is the total number of the continuous variables in the analysis; K^B is the total number of the categorical variables in the analysis ; N_v is number of data records in cluster v; $\hat{E}_{\ vk}$ is the estimated mean of the kth continuous variable in cluster v; N_{vkl} is the number of data records in cluster v whose kth categorical variable takes the lth category.

3.3 EVALUATION OF CLUSTERING ALGORITHMS

Clustering aims to arrange objects into classes that are similar in some way to their members. The researcher needs to decide on the similarity criterion used during the process as well as on the performance criterion of the clustering to determine what constitutes a successful clustering.

After a clustering of the data has been determined, it is important to evaluate its quality. This problem is referred to as cluster validation. In real data sets, cluster validation is often difficult because the problem is described in an unattended manner. Therefore, there may be no specific requirements for the validity of a clustering [33].

Internal validation criteria are used when no external criteria are available to evaluate the quality of a clustering. Unsupervised measures of cluster validity are often further divided into two classes: measures of cluster cohesion (compactness, tightness), which determine how closely related the objects in a cluster are, and measures of cluster separation (isolation), which determine how distinct or separated a cluster is from other clusters [34].

Using silhouette coefficients is one way of internal validation. Silhouette coefficients evaluate the clustering result based on the average distance between a data point and other data points in the same cluster and average distance among different clusters. A graphical representation of the clustering is provided by displaying the silhouettes introduced by Kaufmann and Rousseeuw [35]. Silhouette measure of cohesion and separation is an overall goodness-of-fit measure. Silhouettes are constructed in the following way. Given the number of clusters in the problem, the value s(i) or the silhouette coefficient of the ith object is defined for each object i and these numbers are presented in a plot. The value s(i) is defined by:

$$s(i) = \frac{b(i) - a(i)}{max(\ a(i) - b(i))} \tag{3.6}$$

where a(i) is the average dissimilarity of i to all other objects in A, where A is the cluster in which the object i is first assigned, with d(i,C) being the average dissimilarity of i to all objects of any other cluster C different from A. Object i is assigned to the cluster B that satisfies b(i) = d(i,B). s(i) is always between -1 and 1. When s (i) is close to l, it implies the “within” dissimilarity a(i) is much smaller than the smallest “between” dissimilarity b (i). Therefore, it can be said that i is well clustered.

4

METHODOLOGY

This chapter describes the approach followed to conduct the thesis. All the methods aim to meet the goals of the thesis: understanding mobile Internet users' behaviors by selecting CDR features, evaluating clustering algorithms and obtaining various insights to segment users for marketing. These goals are first achieved by applying data collection, data cleaning, sample selection and aggregation tasks as presented in section 4.1. Then, through features selection and cluster analysis processes as discussed in sections 4.2 and 4.3 respectively. Finally, by giving possible marketing suggestions for different customer segments. Figure 4.1 shows the system model.

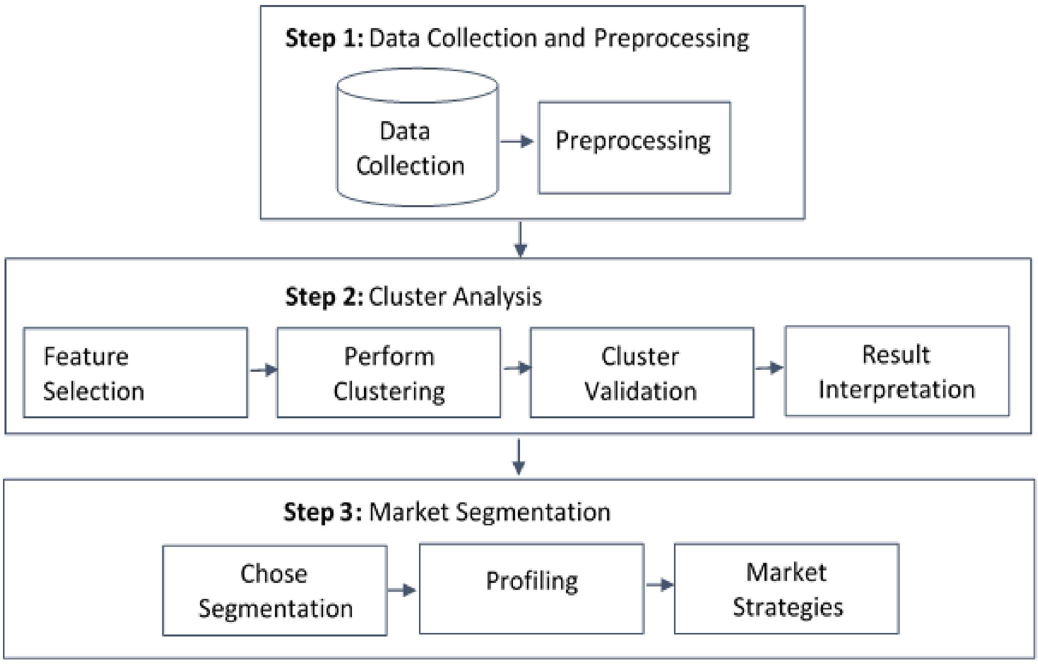


Figure 4.1: System Model

As indicated in the system model, there are three important steps involved in this study. The first step or preprocessing is part of data mining techniques. Data mining helps to extract meaningful information from huge database systems such as CDR and CRM. For this project, datasets were extracted from both systems. Preprocessing was used to refine raw data so that proper data is collected and sampled. Tsiptsis and Chorianopoulos [32] recommend the use of preprocessing to simplify and enhance segmentation process before applying cluster analysis to input data. In the second step, cluster analysis tasks such as feature selection, clustering with selected algorithms, cluster validation and results interpretation were performed. The third (and final) procedure is market segmentation which includes chosing segmentation type, profiling customers and giving market suggestions.

4.1 DATASET COLLECTION AND PREPROCESSING

Data collection is the process of gathering relevant information that helps to address the research question(s). The two major types of data collected for research work are primary and secondary data. Primary data is a type of data the researcher collects for the first time and secondary data is one which has been collected by someone else. The type of data used in this research is secondary data.

Telecom service providers record transactions made between different entities in telecommunication networks. CDR is a type of data they keep for billing purpose. It contains valuable information to mine as described in section 2.4.1. The specific type of record used for this thesis is cellular mobile Internet users CDR which is produced when mobile customers connect with the carrier network to access the Internet with their smartphones and other portable gadgets over cellular networks.

Getting actual usage data of subscribers is not easy because of different reasons like government policies, companies' internal rules and security concerns. Therefore, it is important to describe the steps followed to obtain the permission to use CDR data for the research work. First, a formal letter is obtained from the university that requested ethio telecom for collaborative work . Then, all the necessary arrangements such as meeting and working with the respective sections and personnel were done. Finally, the setup for the data collection is completed which is described in depth as follows.

4.1.1 *Data Collection and Sample Selection*

The main tasks performed to collect and analyze the data are mentioned below:

1. Windows Server 2012 R2 V.6.3.9600 with 8GB RAM and quad-core processor was deployed.

2. 2.5 TB external mounted storage was allocated.

3. Oracle database was installed and configured on the server.

4. Network access configuration that enabled full access of the server from the company's intranet was done.

5. CDR data was uploaded to the database source table from the CBS system.

6. Oracle Developer version 19.1.0.094 installed on the researcher's laptop and was used to access and analyze the data. The laptop used for this experimentation is HP-ProBook 640 G1, Intel (R), Core I7, CPU 3.00 GHz, 8.00GB RAM, and 64-bit operating system.

Figure 4.2 shows how the data collection setup was done. Raw CDR data from the CBS system was pushed to the Temp Server or temporary server. The server has three types of tables (voice, sms and data). Then, CDR dataset of cellular Internet users was extracted.

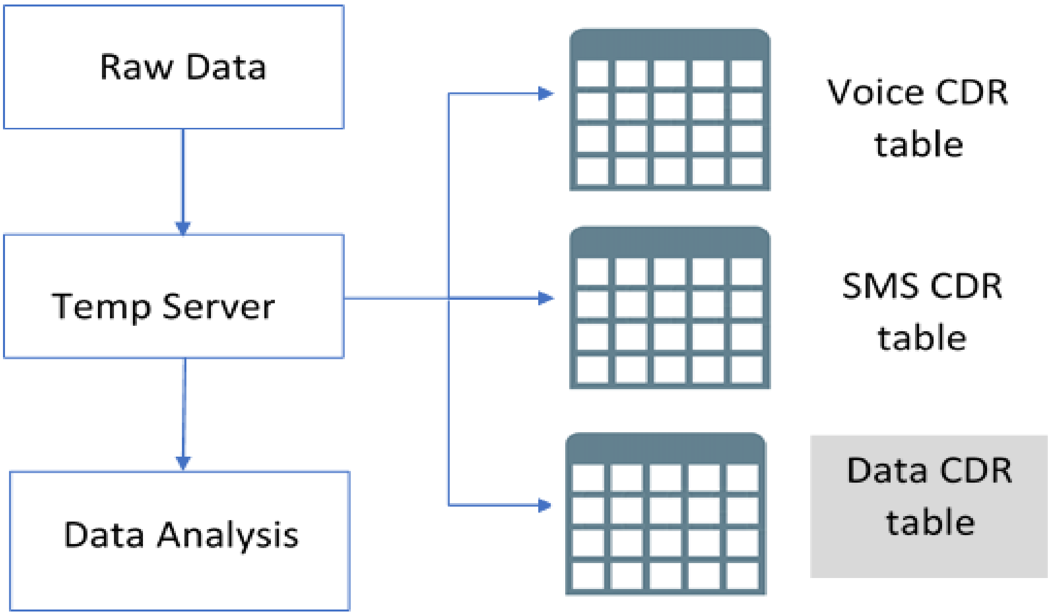


Figure 4.2: Data Collection Steps

Data sample selection: There are millions of customers who use mobile Internet service each day. Therefore, it is difficult to analyze all information stored in the database. In order to solve computational and other resource issues coming from trying to consider records of all customers, sample data was taken using simple random sampling. So 100,000 unique customers drawn from all mobile service prefixes used by the telecom company and those who used cellular Internet service during two months period i.e. from June 25-August 25, 2019 were selected. The CDR file used for this thesis is collected in CSV format through the process described above.

To further improve the sample selection process, the following two criteria were applied: 1. Usage Based Filtering: among the total sample customers selected or from users who connected to cellular data service in the time period of the analysis, those with a minimum monthly usage of 15MB were included. This is done according to the recommendation from marketing experts. 2. Charge Based Filtering: ignore users who are exempted from usage charges. These include employees of the service provider and others who are not disclosed. The two criteria resulted decrease in the number of unique customers considered for the study to be 65,0000. And the total number of usage records(rows) analyzed after final refinement was 8,694,83.

The data selection is based on the main objective of the thesis: understanding mobile Internet users' behavior for market segmentation from CDR data. Having a good knowledge of the dataset is basic requirement to take relevant features for analysis. Billing system specialists were consulted to figure out how some columns and values should be interpreted. Understanding and preparing data is the most time- consuming phase of data mining project.

4.1.2 *Data Preprocessing*

Data preprocessing is perhaps the most crucial task in data mining process [33]. Before doing any analysis work it is important to deal with missing, erroneous or inconsistent data. During data preprocessing stage, the raw data will transform into a format that is easy for the experimentation that follows. For this research, dozens of sql scripts were used during the preprocessing stage. Some major preprocessing tasks were explained below.

Data reduction: helps to obtain a reduced representation of the dataset that is much smaller in volume but producing the same or similar analytical results. In this experiment, the data reduction activities include dropping records with null values or with very high/low values or with charge fee less than 3 Birr (minimum threshold recommended). Clustering algorithms are sensitive to extreme values (outliers), so it is important to observe if the input data is valid before executing the modeling.

Data transformation: is the process of performing normalization and aggregation tasks. Occasion (usage time) of data service connection was the main field selected for transformation purpose. The twenty-four-hour period was divided into four: night (12:00 AM-6:00 AM), morning (7:00 AM-12:00 PM), afternoon (1:00 PM-5:00 PM) and evening (6:00 PM-11:00 PM). Then, the number of times a customer connected to data service during these periods were counted and named as Visit_Count. This aggregation task was done because one of the ways to study behavioral analysis of service users is to divide them based on the periods they use the service. Traditional segmentation techniques relied on characteristics like gender, location or income. Behavioral segmentation takes their decisions like the time they use the service into account.

4.2 FEATURE SELECTION AND EXTRACTION

Feature selection is all about selecting the relevant features or variables or attributes for the alogrithm(s). The key goal of feature selection is to remove the noisy attributes that do not cluster well. Feature selection is generally more difficult for unsupervised problems, such as clustering, where external validation criteria, such as labels, are not available [33]. Involvement of domain experts is useful in real-life clustering tasks to select appropriate features from the raw data. That was the reason the guidance of billing and marketing experts was consulted before selection and extraction of attributes were done.

Table 4.1 shows the seven original CDR features selected from the total thirty three columns extracted from the database. The full descriptions of all CDR columns can be found in appendix A. The dataset used in this research was anonymized to protect customers privacy and to follow data security policy of the company. The feature that uniquely identifies customer MSISDN was dropped after it was replaced by a randomly generated key.

Table 4.1: Original CDR Features

No.	Columns	Data Type	Descriptions
1	MSISDN	Number	Mobile subscriber's primary identifier
2	DATA_DAY	Date	Data Usage Date
3	START_TIME	Time	Time data usage started
4	END_TIME	Time	Time data usage finished
5	DNLD_VOL	Number	Data download volume
6	UPLD_VOL	Number	Data upload volume
7	CHR_FEE	Number	Amount of payment per MB in birr

For the sake of behavioral segmentation, sixteen new features were derived from the original CDR attributes. The features were formed as follows. Each of the four time periods (morning, afternoon, evening and night) were examined for the four major variables formed by aggregation (Visit_Count, Total_Download, Total_Upload and Data_Fee) that describe the characteristics of cellular data users. For example, the following sample table 4.2 clarifies how morning period usage was observed for every customer. Visit_Count implies the total number of times a user accessed data, total_download is the sum of all download volumes, total_upload is the sum of all uploads and data_fee is the total amount of payment per MB in birr. The same logic applied to afternoon, evening and night time periods.

Table 4.2: Derived Features Sample

Usage Period	Features	Description
Morning	VISIT_COUNT	Total count of data connection
	TOTAL_DOWNLOAD	Sum of data download volume
	TOTAL_UPLOAD	Sum of data upload volume
	DATA_FEE	Total usage fee

4.3 segmentation using cluster analysis

The analysis was done with two different algorithms namely the K-means and Two-Step clustering. The algorithms were discussed in chapter three. After proper preprocessing was completed, the users were grouped based on the monthly total data usage volume i.e. the sum of downloads and uploads. Then the algorithms were executed on each group.

The total mobile Internet users a.k.a mobile data service subscribers of the sampled numbers were therefore divided into three as High, Medium and Low Users according to Namvar, et al. [3] and based on suggestions from ethio telecom marketing experts. Also, analysis of total usage volumes from the actual collected data was considered. High Users are those whose total data usage is greater than 2GB, Medium Users are customers whose usage ranges between 500MB and 2GB, and Low Users have total usage less than 500MB and greater than 15MB. The following diagram describes the three groups.

The rest part of the thesis is based on discussion of results of the Low Users group for two main reasons. Firstly, according to past analysis the company performed most of the users fall into this group. Therefore, studying Low Users group properly implies the overall picture of the behavior of potential data users of ethio telecom.

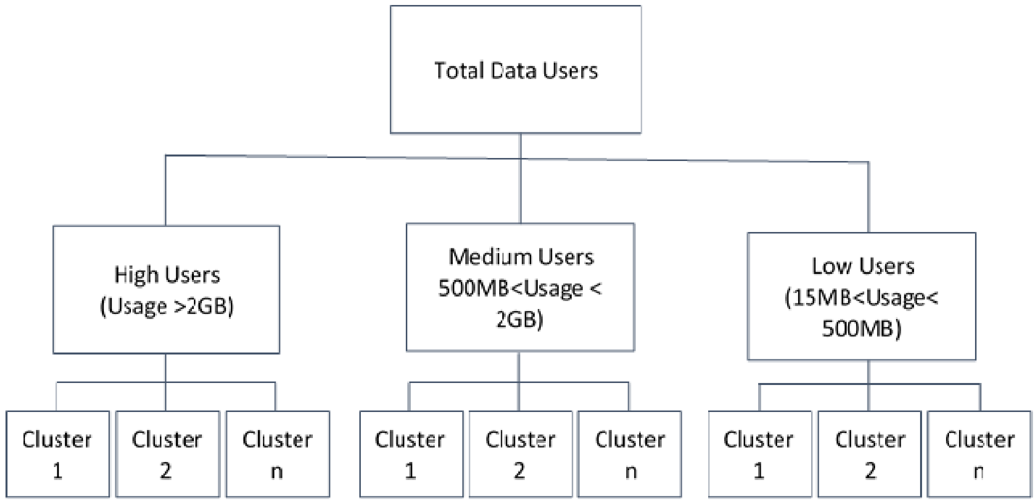


Figure 4.3: Customers Grouping Based on Usage

Secondly, there are marketing possiblities to address low value users through incentives and promotions more than medium and high value users.

4.3.1 *Behavioral Segmentation*

Behavioral segmentation is a type of customer segmentation based on patterns shown by consumers while using a service. One way to understand the behavior customers reveal while they use mobile Internet is to observe the time they prefer to access data. In behavioral segmentation, segments are defined by using appropriate clustering models on usage/behavioral data. The data is found from the organization's warehouse [32].

As shown on table 4.2 above, four of the derived variables came from the aggregated usage records of customers who accessed cellular data during two months period. The variable name M_VISIT_COUNT, for example, implies the total number of times the user connected to the service. 'M' stands for morning period. The same naming applied to the rest three variables.

4.3.2 *Demographic Segmentation*

The CRM database contains information about customers. However, the quality of the data the telecom company or ethio telecom collects regarding demographic information is questionable by database administrators, marketing experts and the person conducting this project due to reasons stated in section 2.4. But it was still possible to obtain information for a limited number of customers. This is done to show the possiblity that the proposed methodology can work.

Important demographic variables like gender, age, and educational background are not fully registered in the database. That means most cells are empty or stored meaningless information like ages greater than 100. Others like marital status and occupation are not available at all. Here, the researcher would like to reassure that the sample numbers are anonymized or doesn't contain personal details like name, address and phone number.

To sum up, for the sample numbers selected, one fifth of the customers demographic data or information of only 20% of the customers was acquired. Numerically speaking, out of the 100,000 numbers requested, only 20,294 had full records of the demographic variables. And only 1,271 of these belong to the low data users' group. The features collected from the demographic data are three: age, gender and educational backgound. After the results of the cluster analysis are obtained, the demographic features of the groups were appended forming 19 features, and the clustering algorithms applied to observe if there are significant differences in the clustering result between the algorithms.

5

RESULT AND DISCUSSION

The results of the thesis and the discussions of the results are presented in this chapter. Section 5.1 explains results of the CDR analysis and in section 5.2 the demographic analysis outputs are shown. Evaluation of clustering algorithms is discussed in section 5.3. Lastly, possible marketing approaches are indicated for the identified segments in section 5.4.

5.1 RESULTS OF CDR ANALYSIS

The cluster analysis task performed in this thesis is based on the extracted features (a.k.a. variables). Thus, the first research objective (selecting appropriate features for analyzing behaviors of mobile Internet customers) is already answered in the previous chapter. This chapter starts by analyzing the results obtained by value based segmentation as shown in figure 5.1.

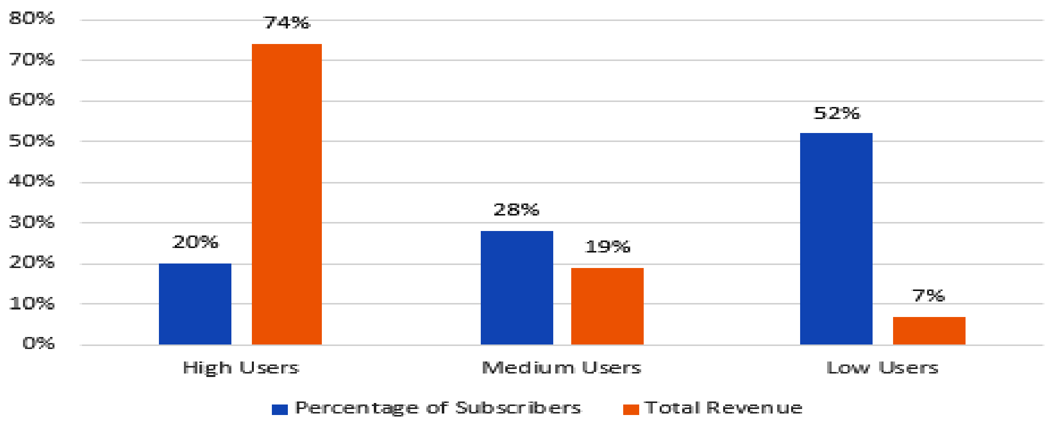


Figure 5.1: Value Based Segmentation

As described in section 4.3, grouping of customers was done based on the total data consumption volume of mobile Internet users for a month. These user groups are high, medium and low users. Unlike [3], where high users use more than 3GB, in our analysis high users are those who use more than 2GB. Medium users fall between 500MB and 2GB. Low Users are users who consume data packages as low as 15MB and as high as 500MB. The grouping is performed to observe the overall characteristics of the users.

The value segmentation implies that High Users which comprises 20% of the customers have generated 74% of the revenue. The Medium Users or the next 28% of the customers brought 19% of the revenue. Whereas, Low Users 52% of the population has value as low as 7%.

Analysis Result with K-means

The preprocessed CDR was clustered using K-means and the result is presented in figure 5.2. As described in section 4.3 M,A,E,N stand for Morning, Afternoon, Evening and Night time periods respectively. For example, M_Visit_Count represents the number of times the customers initiated mobile Internet connection in the morning.

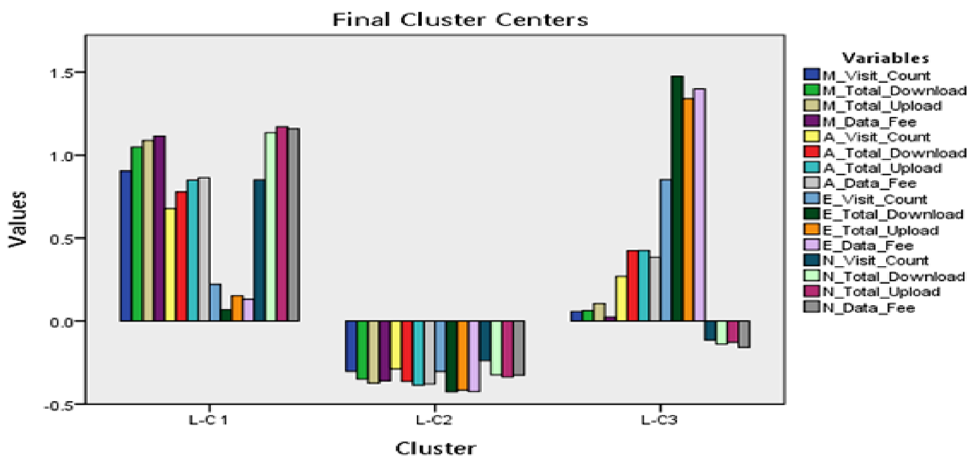


Figure 5.2: K-means Clustering Result

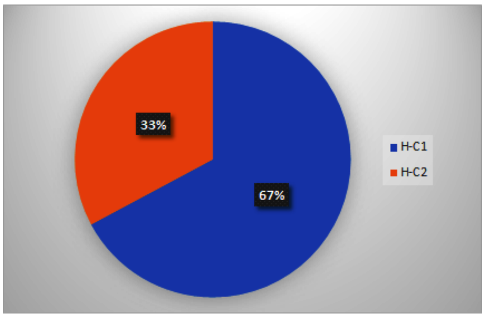
The following major observations can easily be noticed from the clustering result of Lower Users presented in the above bar chart or fig. 5.2:

1. Users in cluster two have the lowest values for all the 16 features.

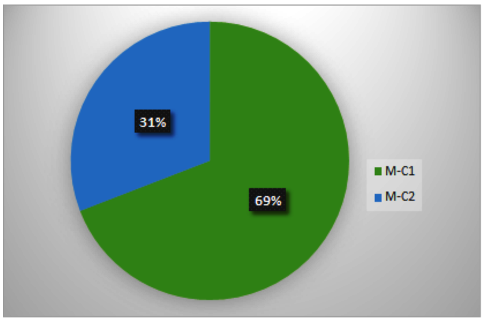
2. Evening downloads are higher for users in cluster three.

3. Nighttime usage is relatively better for L-C1 while users in L-C3 have high evening time usage. The figure showed the possiblity of characterizing user groups or doing behavioral segmentation of customers.

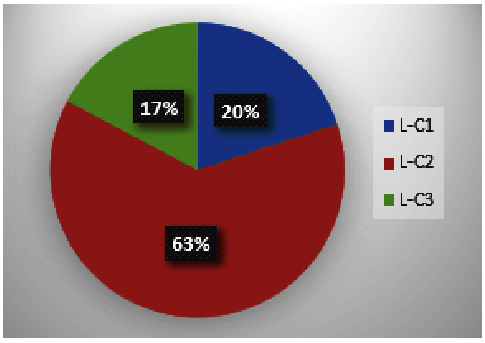
The charts in fig. 5.3 are the results of clustering the three different users group in percent. As the chart implies, for example, clustering low users group resulted three clusters. 48% of the users fall in L_C1 or lower users group cluster one, 25% are under the second cluster and the rest 27% belong to cluster three.



(a) High Users



(b) Medium Users



(c) Low Users

Figure 5.3: Percentage of Clusters

Customers in L_C1 have paid higher fees as compared to the customers in the rest two groups during mornings and evenings. Whereas, customers in L_C2 showed almost equal amount of expenses during morning and afternoon periods. Insights like these can be extracted easily. The following figure 5.4 shows the result of value segmentation of lower value customers with respect to usage period.

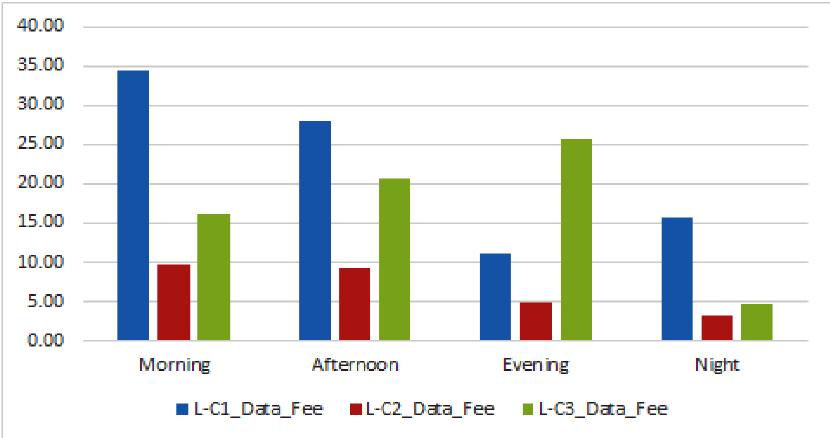


Figure 5.4: Low Users' Data Fee vs Usage Period

Clustering results of the rest two user groups or medium and high users groups can also be observed in the figures 5.5 and 5.6 respectively.

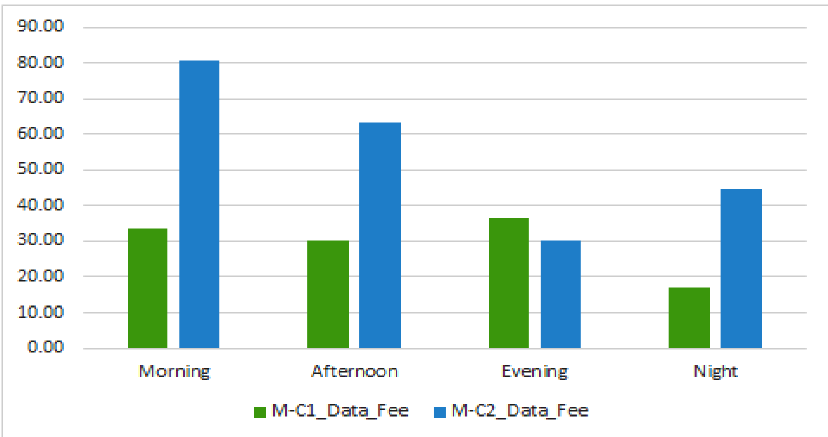


Figure 5.5: Medium Users' Data Fee vs Usage Period

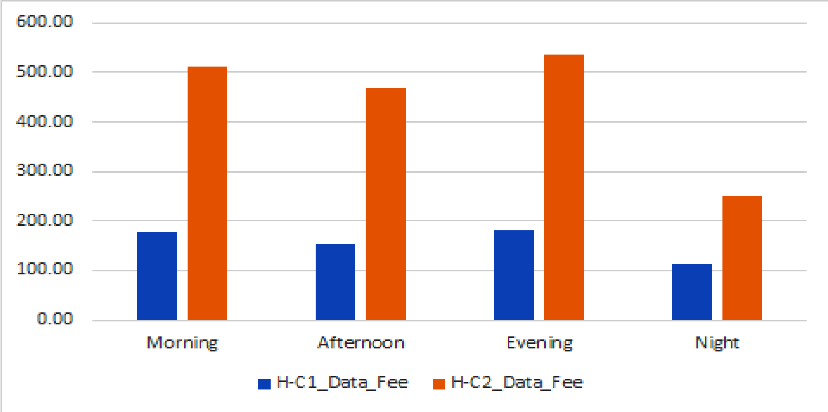


Figure 5.6: High Users' Data Fee vs Usage Period

5.2 RESULTS OF DEMOGRAPHIC ANALYSIS

Demographic analysis is performed to segment customers based on the distribution of age group, gender and educational backgrounds. This information is from the CRM system. Since some of the variables are categorical by nature, and Kmeans algorithm better works for numeric ones, only Two-Step clustering was choosen. Figure 5.7 shows demographic information of cluster 1 of Lower Users(L-C1).

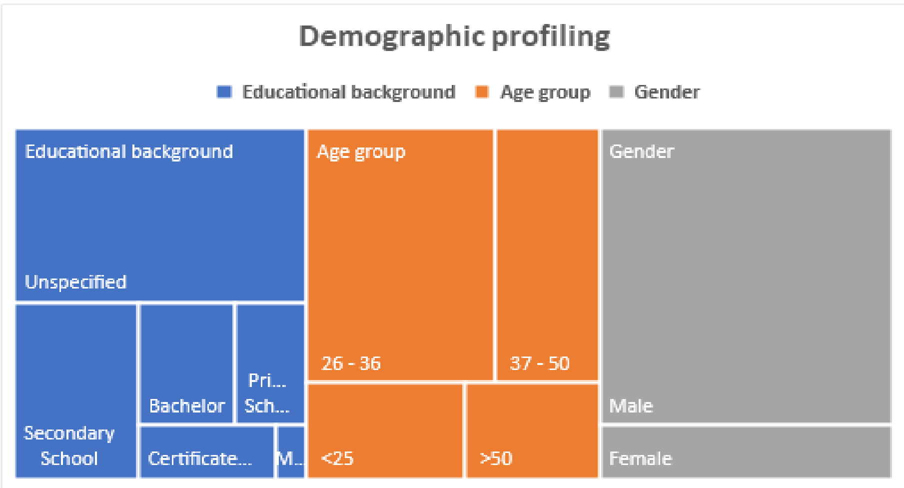


Figure 5.7: Demographic information

It can be seen from the diagram that the educational background of approximately 50% of the users in this category is not specified. But it is improbable that illiterates could use Internet service anywhere in the world most especially in Ethiopia given that using Internet in our country somewhat requires English language skills which most people acquire at least after they joined high school. Other demographic data stored also lacks quality. Figure 5.8 shows educational background of (L-C1) users in percent.

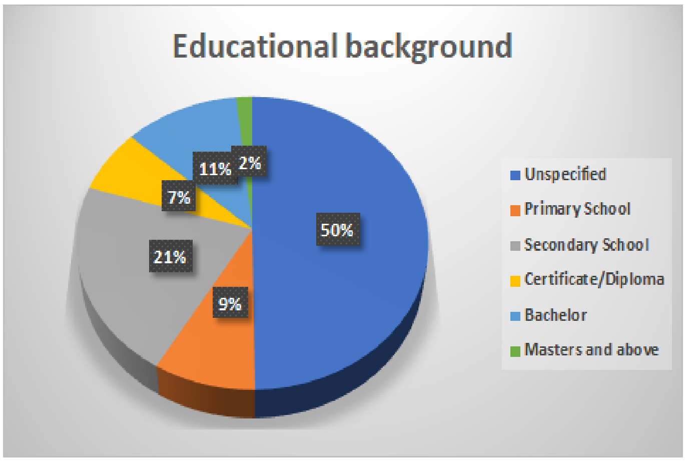


Figure 5.8: Educational background of L-C1

5.3 ASSESSMENT FOR K-MEANS AND TWO-STEP RESULTS

One of the criteria for validating clustering results is to observe how clusters are internally related or separated. The concept of unsupervised measurement of clustering using silhouette coefficients was addressed in section 3.3. The difficulty of evaluating clustering algorithms, given that there are no data labels, was also discussed. But it is still possible to measure the goodness of clustering results with measurement tools like silhouette. Thus, K-means and Two-Step clustering algorithms were evaluated for the CDR data analysis and K-means has a silhouette score was below 0.5 and Two-Step had around 0.6 (figure 5.9) that is in the range of good clustering for the 16 derived features used. For the 19-features the silhouette score for K-means is 0.3 whereas for Two-Step clustering is 0.5 as table 5.1 shows.

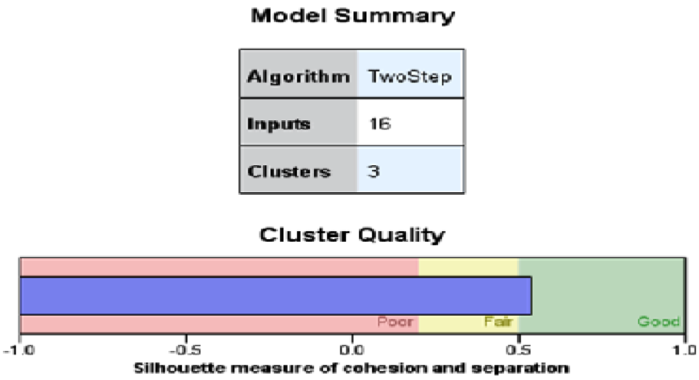


Figure 5.9: Cluster Quality Measure

Silhoutte scores obtained for both algorithms used to cluster the datasets in this thesis can be interpreted according to the theory discussed in section 3. Silhoutte coefficients below 0.5 are grouped under fair and poor cluster quality group. Those in range between 0.5 and 1.0 are considered of good cluster quality.

Table 5.1: Silhoutte Scores for K-means and Two-Step

Number of Features	Algorithm	Silhoutte Score	Cluster Quality
16	K-means	0.5	Good
	Two-Step	0.6	Good
19	K-means	0.3	Fair
	Two-Step	0.5	Good

5.4 SEGMENT IDENTIFICATION FOR MARKETING

This thesis was conducted on CDR and CRM datasets to understand how clustering algorithms can be used to segment customers for marketing purpose. The experiments were done with actual datasets. K-means and Two-Step clustering algorithms were used to cluster users of different groups and a single group was selected for further analysis. In our case, both algorithms cluster Higher User group into two which were named(H-C1 & H-C2), Medium User group into two (M-C1 & M-C2) and the Low User group which were analyzed above for behavioral segmentation into three(L-C1, L-C2 and L-C3).

Based on the analysis results of all clusters, the following possible marketing suggestions are given as it can be referred on table 5.2.

Table 5.2: Possible Marketing Suggestions

Marketing Approaches	**Segment ID**
Discounted Offers	L-C1, L-C2, L-C3
Short term data packages	L-C2, L-C3
Campaign Management	L-C1, L-C2, MC1
Unique Marketing Approach	H-C1
Customer Loyalty Management	H-C1, H-C2
Limited Package with Full Speed Data	M-C1, M-C2, H-C1
Unlimited Package with Limited Speed	H-C1, H-C2, M-C2

Discounted offers and short term data packages were suggested to the Low User groups only because customers in these groups are sensetive to prices. The frequency of their service usage, in the case of this study, the number of times they connect to mobile data service, shows most of them have an interest to freely enjoy the benefits of mobile data had their economic problems do not hold them back. Short term data packages are also to the best of the two lowest groups from all the seven segments distinguished. Students and lower income customers can be addressed with this market approach if proper demographic information was not at stake in ethio telecom. Likewise, it is possible to discuss the marketing suggestions given for customers in the higher value stages.

Market segmention is necessary before developing a marketing plan. And approaches like the one shown above gives better advantages for telecom companies in that every segment group will have a chance to be considered based on the common behavior of the group members.

Due to the increasing use, and even reliance, of mobile Internet services, the target groups of customers is diverse. Therefore, telecom companies are required to create marketing strategies that fits into the desires of many subscribers.

6

CONCLUSION AND RECOMMENDATION

6.1 CONCLUSION

Segmenting customers helps telecom companies during products development, services design, maintenance and operations activities. By applying appropriate segmentation techniques, operators can also meet expectations of specific groups of subscribers.

Clustering algorithms are used to identify groups having similar behavior and to observe the relationship between different groups. The aim of cluster analysis in marketing is to segment customers so that effective personalized marketing strategies are implemented.

Actual usage records of customers or transaction data are stored on CDR database of a telecom network. And demographic information of the customers can be found from CRM server. In this study, by examining the raw usage data from the billing system, new features were generated and transformed into a format applicable to data mining methods.

This thesis shows behavioral segmentation of customers based on newly derived variables is capable of grouping mobile Internet users for marketing purpose. The finding of clusters such as those named lower, medium and higher groups can prove the importance of behavioral clustering.

Hence, two clustering methods or the K-means and Two- Step clustering algorithms were investigated to analyze the behaviors of mobile Internet users based on their real consumption trends. By looking at silhouette score, cluster size distributions and final results of clustering algorithms, the different obtained clusters were evaluated and compared.

To the best of the author's investigation, there is no existing work in the literature that addresses the issue of applying the two algorithms on CDR data set for marketing purpose. To fill this research gap and to satisfy an important practical need, this paper used two known clustering algorithms and suggested possible marketing approaches for different clusters obtained.

6.2 RECOMMENDATION

By using efficient cluster analysis techniques, telecom companies may gain a clear understanding of the interests of their customers for marketing purposes. In addition to that, opportunities to develop improved CRM systems, to advance knowledge of network management and operation tasks, and to find specific habits of group of customers are possible.

Mobile traffic data collected from telecom service providers core networks or from the mobile phones of individual users can also be used with CDRs to get information regarding applications access behaviors and mobility pattern of customers. The analysis of usage patterns of subscribers with different handsets and application choices would be interesting one. Future study can therefore strive to use bigger databases with more fields from different sources to obtain better results from the algorithms.

Eventually, it will be recognized that the work can be significantly improved by including certain parameters for understanding user behavior, rather than just using the amount of monthly consumption and time preferences used in this study.

A

DESCRIPTION OF RAW CDR

A.1 EXPLANATION OF CDR FEATURES

CDR Column Name	Description
CDR_ID	CDR service type (Voice, Data, SMS) identifier
RE_ID	Distinguishes between records of services
BILLING_NBR	Identification of billing
CDR_TYPE	The service type the CDR is generated for
CALLING_NUMBER	Identity of call initiator
CALLED_NUMBER	Identity of call termination
CALLING_IMEI	Equipment identity of calling party
CALLING_IMSI	Subscriber (calling) identity
THIRD_PARTY_NUMBER	Third party equipment (like MSC, SMSC) numbers
CALL_START_TIME	Call start time stamp
CALL_END_TIME	Call end time stamp
CALL_DURATION	Length of duration on call
CALL_FEE	Fee charged for service usage

CALLED_COUNTRY	Country code where call is terminated
CALLING_CARRIER	Carrier identification where call is started
CALLED_CARRIER	Carrier identification where call is terminated
CALLING_DISTRICT	The visited area of calling party
CALLED_DISTRICT	The visited area of called party
STATUS_DATE	CDR status update date
CALLING_SUB_ID	ID of a sub CDR generated after CDR splitting.
BILLING_CYCLE_ID	The time when the charging event occurs
CHARGE_1	Total call fee
CHARGE_2	Reserved call fee
RATE_ID1	Rate service usage id
ACCOUNT_ITEM_ID1	Default account id of the charged party
UPLOAD_TRAFFIC	Volume of data traffic used during uploading
DOWNLOAD_TRAFFIC	Volume of data traffic used during downloading
BILLING_OFFERING_ID	Primary offering id
ERROR_CDR_TYPE	Error type
CALLFORWARDINDICATOR	Call forwarding flag
HOTLINEINDICATOR	Special number to indicate the number is common, barred, free, special-tariff,or voice mail
CALLING_TRUNK_ID	Network id of calling party
CALLED_TRUNK_ID	Network id of calling party

BIBLIOGRAPHY

[1] GSM Association. https://www.gsma.com/r/mobileeconomy/. [Online; accessed 28-June-2019]. 2019.

[2] ITU. *Measuring the Information Society Report Volume*. https://www.itu.int/en/ITU-D/Statistics/Pages/publications/misr2018.aspx, Last accessed on 2019-10-30. 2018.

[3] Anahita Namvar, Mehdi Ghazanfari, and Mohsen Naderpour. "A customer segmentation framework for targeted marketing in telecommunication." In: *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE. 2017, pp. 1–6.

[4] Salar Masood et al. "Customer segmentation and analysis of a mobile telecommunication company of Pakistan using two phase clustering algorithm." In: *Eighth International Conference on Digital Information Management (ICDIM 2013)*. IEEE. 2013, pp. 137–142.

[5] Luo Ye et al. "Telecom customer segmentation with K-means clustering." In: *2012 7th International Conference on Computer Science & Education (ICCSE)*. IEEE. 2012, pp. 648–651.

[6] Aaron Tkaczynski. "Segmentation using two-step cluster analysis." In: *Segmentation in social marketing*. Springer, 2017, pp. 109–125.

[7] Weiyi Gao, Gangyi Qian, and Hong Xu. "SOM clustering analysis for telecommunication customer segmentation." In: *2009 International Conference on Management and Service Science*. IEEE. 2009, pp. 1–4.

[8] Deepali Arora and Kin Fun Li. "Identifying Prime Customers Based on MobileUsage Patterns." In: *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. Springer. 2016, pp. 853–861.

[9] Hasitha Indika Arumawadu, RM Rathnayaka, and SK Illangarathne. "Mining Profitability of Telecommunication Custormers using K-Means Clustering." In: (2015).

[10] Ethio Telecom. *Market Segmentation Report and Strategy*. 2016.

[11] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. "A survey of results on mobile phone datasets analysis." In: *EPJ data science* 4.1 (2015), p. 10.

[12] Michel Wedel and Wagner A Kamakura. *Market segmentation: Conceptual and methodological foundations*. Vol. 8. Springer Science & Business Media, 2012.

[13] E Mooi and M Sarstedt. "A concise guide to market research: The process, data, and methods using IBM." In: (2011).

[14] Gary M Armstrong et al. *Marketing: an introduction*. Pearson UK, 2018.

[15] salesforce. *M*. https://www.salesforce.com/products/service-cloud/best-practices/famous-customer-service-quotes/, Last accessed on 2019-9-15. 2019.

[16] Lihua Wu et al. "Statistic analysis of data access behavior in the mobile Internet." In: *2013 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE. 2013, pp. 89–93.

[17] Shuai Jiang et al. "Big data enabled user behavior characteristics in mobile internet." In: *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE. 2017, pp. 1–5.

[18] Chieh-Hsin Liao et al. "Using big data for profiling heavy users in top video apps." In: *2015 IEEE International Congress on Big Data*. IEEE. 2015, pp. 381–385.

[19] Jie Yang et al. "Characterizing user behavior in mobile internet." In: *IEEE transactions on emerging topics in computing* 3.1 (2014), pp. 95–106.

[20] Duan Hu et al. "We Know What You Are–A User Classification Based on Mobile Data." In: *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*. IEEE. 2013, pp. 1282–1289.

[21] Hulya Yalcin. "Extracting user profiles from mobile data." In: *2017 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*. IEEE. 2017, pp. 1–5.

[22] Annabel Z Dodd. *The essential guide to telecommunications*. Prentice Hall Professional, 2002.

[23] Haard Mehta et al. "0G to 5G mobile technology: a survey." In: *J. of Basic and Applied Engineering Research* 1.6 (2014), pp. 56–60.

[24] Worku Bogale. "A Background Paper on Telecom & Telecom Statistics in Ethiopia." In: *ethio telecom, Ethiopia* (2005).

[25] Ethio Telecom. https://www.ethiotelecom.et//. [Online; accessed 19-July-2019]. 2019.

[26] Jani Haapaniemi. "Telecommunication data monetization." In: (2018).

[27] Syed Raza Raquel Frederick Jake Kendall Ed Naef Philipp Muelbert and Nirant Gupta. "Using Mobile Data for Development." unpublished. 2014.

[28] Julie K Petersen. *The telecommunications illustrated dictionary*. CRC press, 2002.

[29] Viivi Jäppinen et al. "Customer Information Utilization as a Part of Telecom Operator's Customer Relationship Management." In: (2015).

[30] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[31] Rui Xu and Donald C Wunsch. “Survey of clustering algorithms.” In: (2005).

[32] Konstantinos K Tsiptsis and Antonios Chorianopoulos. *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons, 2011.

[33] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

[34] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[35] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.