Customers Segmentation for Profitability Enhancement Using Data Mining Technique: The case of ethio telecom

PREPARED BY: TAJUDIN MOHAMMED

ADVISER: 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 Telecommunication Engineering

Addis Ababa University
Addis Ababa, Ethiopia
February 21, 2020

Declaration

I declare that the thesis comprises my own work in compliance with internationally accepted practices. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Tajudin Mohammed

Name

Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Customers Segmentation for Profitability Enhancement Using Data Mining Technique: The case of ethio telecom

By: **Tajudin Mohammed**

Signed by :

Adviser	Ephrem Teshale (PhD)	Signature	_____	Date	_____
Evaluator	_____	Signature	_____	Date	_____
Evaluator	_____	Signature	_____	Date	_____

ABSTRACT

Customer segmentation is dividing of customers into groups of individuals that have common characteristics or traits. By segmenting customers based on their usage behavior, telecom companies can better target and classify their customers, provide the services that meet their expectations and increase profitability. On the contrary, companies with improper segmentation or luck of segmentation facing the problem of providing the exact product or service to meet the actual customer needs. Incorrect profit prediction and wastage of resource utilization are the main problems of ethio telecom which results from poor customer segmentation.

To mitigate the segmentation problem this study focuses on segmenting telecom customers based on their usage behavior using unsupervised clustering techniques. K-means algorithm was used to cluster the Call Detail Record (CDR) data. Before clustering CDR data were collected, relevant attributes selected and preprocessing techniques such as data cleaning, data aggregation, data integration, and data formatting were performed. In addition, four datasets were formed by summarizing the data on a monthly base.

The experimentation results in eight different clusters. These clusters were analyzed using quantile score techniques. The clusters were ranked and mapped with customer segmentation type. Among the clusters, the cluster with 236 subscribers was scored the highest in terms of duration, frequency and money. As a result, this cluster was chosen as a platinum customer type. They are highly profitable customers, vital to affect its revenue and need to serve well by the company.

KEYWORDS

CDR,cluster, Customer,K-means, Segmentation, unsupervised

ACKNOWLEDGMENTS

Above all Praise be to Allah, Lord of the Worlds, my gratitude goes to ALLAH who is the governor of the present and the future world. He has been always with me in all bad and good times and will always be anywhere.

The completion of this thesis marks the end of a very important stage in our academic life that without the constructive advice and support of many people would have been impossible to accomplish. Therefore, we would like to take this opportunity to express our appreciation to all of you who had the time and energy to help us throughout this process.

First of all, we would like to thank our case company ethio telecom for providing us the opportunity to undertake this study.

We would also like to express our gratitude to our advisor Dr. Ephraim Teshale for his remarkable dedication, useful insights, helpful guidance and comments during our work.

Last but not least, we would like to take this opportunity to thank our partners, friends and family for their constant support and understanding during this autumn.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ARFF	Attribute Relation File Format
ARPU	Average Revenue Per User
B2B	Business to Business
B2C	Business to Customer
CBS	Convergent Billing System
CDR	Call Detail Record
CLV	Customer life value
CRM	Customer Relation management
CSV	Comma Separated Values
DFM	Duration,Frequency and Monetary
EDR	Event Detail Record
ISD	Information System Division
KDD	knowledge discovery in databases
ML	Machine Learning
RFM	Recency Frequency Monetary
ROI	Return on Investment
ROI	Return on Investment
SOM	Self Organized Map
SSE	Sum of Square Error

1

INTRODUCTION

Customer segmentation can be described as the process of organizing the customer database through classification into different classes. Alternately, it involves assigning each and every customer to a distinctive group, based on their characteristics [1]. The objective of segmentation is to understand the existing customers and use this information to gain new customers, lower the operating costs, boost the service and increase profits. The attributes of segmentation can be based on demographics, geographical or behavioral characteristics, and marketing them as a group. Therefore, in a group, the members express similar requirements, which are not always consistent. Segmentation needs aggregation, ordering, and analyzation of customers' information [2].

The goal of most companies is profitability growth. In order to reach this goal, companies should provide an analysis of how to manage a relationship with their customers and offer appropriate corresponding marketing strategies. Providing transaction-based segmentation on service and customer satisfaction tell that, price is not the only measure to affects customer purchasing decisions, but also customer and company are important to agree on product or service value and good customer services [3]. Therefore, organizations should not seek to develop a product to satisfy their customers, though companies should track customer purchase behavior and present distinct ways about already available products or services for each segment. So customer segmentation established from user behavior is essential for developing successful marketing strategies, which in turn cause creating and maintaining profitability advantage [3].

The fast-growing, tremendous amount of data, collected and stored in a large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large data repositories

become "data tombs" data archives that are seldom visited [4]. However, data mining is used to process of searching and analyzing these data in order to find hidden, but potentially useful information [5] [6]. It involves selecting, exploring and modeling large amounts of data to uncover previously unknown patterns and ultimately understandable information from large databases.

Knowing the exact service to meet the diversified customers' needs is a major issue for the telecom companies. companies with improper segmentation or luck of segmentation facing the problem of providing the exact product or service to meet the actual customer needs. Incorrect profit prediction and wastage of resource utilization are the main problems of ethio telecom which results from poor customer segmentation. Ethio telecom currently follows a product or business-centric approach to increase its revenue which is one size fits all methods. Also it used core segmentation methods such as Enterprise, Residential. . . as shown in Table 1.1.1, which does not correlate well with actual usage behavior, customers experience and there need.
Due to the introduction of data mining methods, there are ways to perform proper segmentation based on the actual behavior of the customers from the available customer CDR data. This research will focus on ethio telecom customer segmentation through the clustering data mining approach using CDR and will provide usage-based customer segmentation which helps to improve profitability practice.

1.1 PROBLEM STATEMENT

In broad terms customer segmentation classifying customers according to their similarity. Thus customers can be segmented in many ways based on segmentation pillar such as demography, psychographic, behavioral and geographical. Telecom Company's such as Vodafone and Tehran telecom follow usage based behavioral segmentation types for their customers. These two companies improve their revenue because of their customer segmentation strategy. In general ways, telecom companies choose segmentation pillars based on their objectives. Ethio telecom, in particular, used core segmentation methods and segments its customers as a

postpaid, prepaid, hybrid, and grouped as enterprise and individuals customers as shown in Table 1.1.1 during service acquisition.

Table 1.1.1: Mobile customer base

ethio telecom Mobile Customer Base				
Category	**Hybrid**	**Postpaid**	**Prepaid**	**Grand Total**
Enterprise	87	418,979	311,852	730,918
Individual	118	103,096	54,887,048	54,990,262
Grand Total	**205**	**522,075**	**55,198,900**	**55,721,180**

The segmentation techniques used by ethio telecom are too general and not a data-driven approach. Such segmentation does not provide information on customer actual usage, meet the diversified need of the customer and develop tailored marketing strategies for each segment, in addition, it provides static results. As a result, depicted in table 1.1.2, it is difficult to achieve planned profitability strategy or revenue forecast. The company achieved 75% to 80% average revenue for the last eight years. Because of this the company unable to achieve to maximize its average revenue.

Table 1.1.2: ethio telcom revenue forcast and actual earned [7]

Ethio Telecom annual revenue planned and actual, from fiscal year 2011 to 2018/19 (in billion Ethiopian Birr)					
Year	Forecast	Actual	Year	Forecast	Actual
2011/12	17	12.777	2015/16	37	28.371
2012/13	22	16.644	2016/17	40	33.313
2013/14	25	17.358	2017/18	44	37.699
2014/15	30	21.5	2018/19	47	38.5

In order to improve this problem, telecom companies should use state of the art customer segmentation using proper data mining techniques from their customer

CDR data. Typically telecommunication customer CDR data provides hidden information about customer behavior.[8].

Generally, customers are the most important asset of an organization to improve its profitability. There is no business vision that remains profitable without satisfied customers and develops a good relationship. That is why many telecom organizations plan and employ a clear strategy for treating customers. Nevertheless, there is a gap in the strategy effectiveness in ethio telecom. So based on the observation, it is important to study ethio telecom customer segmentation to improve its profit using a data-driven approach with suitable data mining techniques.

In this thesis, we try to analyze actual customer usage CDR data using an unsupervised clustering algorithm and associated with there customer segmentation type. Therefore the overall study tries to answer the below two research questions.

Research questions

1. How data-driven approach is useful than general(core) customer segmentation?
2. What are the relevant attributes used to segment behavioral customer segmentation?

1.2 OBJECTIVES

1.2.1 *General Objective*

The general objective of this research is to perform customer segmentation by analyzing customer CDR data for profitability improvement using data mining techniques and algorithm.

1.2.2 *Specific Objectives*

This study has the following specific objectives:

- To identify appropriate unsupervised learning technique and algorithm for customer segmentation.
- To prepare the data for analysis by making preprocess into a suitable format for the selected algorithms.
- To identify segmentation properties and select the relevant attributes that will help to improve profitability.
- To identify appropriate segmentation type by analyzing their historical CDR data using the selected clustering algorithm.
- To recommend future works based on the research findings.

1.3 SCOPE OF THE STUDY

There are several ways in which profitability improvement can be committed to customer segmentation. But this study uses local call usage of mobile subscribers due to a better contribution of company revenue and use peak hours' time and off-peak hours' time of DFM analysis. It is crucial to find the level of customers type towards the company based on the transaction history of customer data [9]. These study emphases ethio telecom customer segmentation by defining different customer types to enhance the profitability of each customer. Segmentation has been prepared by actual mobile customer data collecting from the database of ethio telecom Convergent Billing System (CBS) and Customer Relation management (CRM) system.

1.4 SIGNIFICANCE OF THE STUDY

In addition to the scientific contribution, telecom operators will find the following benefits:

- This Study contributed to customer characteristics and their relationship to customer loyalty, satisfaction, and profitability.
- Investigation is supported by customer data (CDR) and data mining techniques, which gives the opportunity to contribute to the field of CRM analysis by applying general knowledge in a specific environment.
- Ethio telecom can make use of the results as describing profitable customers levels and develop effective strategies.
- It also helps to better understand the structure of profitable relationships and to realize implications to better manage a customer's profitable bond with the company.

1.5 LITERATURE REVIEW

Telecommunication activity consists of managing and collecting a huge amount of data or event information about its customers. Hence, millions of subscribers, in different places can perform hundreds of events in a short time resulting in billions of transactions to be recorded. To manage such data significant research has been conducted to understand the customer. This section reviews some of the literature which is more focused on customer segmentation using different data mining techniques, algorithms and analysis methods that need to be involved.

In [10] discussed about four telecommunication companies that are routinely generated and store huge amounts of high-quality data, have a very large customer base and operate in a rapidly changing and highly competitive environment. However, these companies also face a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data,

and the need to predict very rare events. To know the customer better the author used one of the unsupervised data mining techniques called clustering. And describe clustering mobile customer based on CDR and Self Organized Map (SOM) to easily visualized and understood customer utilization behavior. The algorithm used for clustering is only K-means to identify which cluster is premium and which one is non-premium to show more about profit maximization.

According to [11] customer clustering and segmentation two important used techniques for marketing and customer relationship management. Classification and pattern extraction from customer data is also very important for business support and decision making. Due to advances in computing and information storage areas, large companies are provisioning up a vast volume of data and the traditional mathematical models are difficult to predict the segmentation and customer patterns. To overcome these difficulties in an organization, the new type of technique is required that has intelligence and capability to solve knowledge scarcity and it is called data mining. By conducting demographic clustering and segmentation within the behavioral segments, the author can define tactical marketing campaigns and select the appropriate marketing channel and advertising. Then target those customers who have high-profit, high-value and low-risk using customer clustering (demographic clustering algorithm).

In [12] dynamic information about customer's behavior is required to segment and personalize products and services. And for making an applicable business strategic plan for each group to achieve the highest customer satisfaction with maximum revenue. Since (CDR/Event Detail Record (EDR)) have valid and dynamic information about customer behavior, applying CDR/EDR is a key success factor to identify Customer life value (CLV) and customer segmentation than personal information extracted from a repository such as a billing system. In order to evaluate the huge CDR/EDR data to customer segmentation and CLV, the author applies K-means clustering as a clustering tool. To overcome the high competitiveness, they often need to design a distinguishable marketing strategy based on different customer behavior. The author proposed how effectively it can apply CDR/EDR data to customer segmentation, customer life cycle, loyalty, churn

and cross-selling by considering the past contribution, potential value, and churn probability at the same time.

As per [13] telecom databases have a huge amount of raw data CDRs which generated every day, but they are not suitable for knowledge extraction. So, data transformation is essential to construct an appropriate data set. The data-driven segmentation approach can support marketing strategies to tailor their marketing plans like increasing in Average Revenue Per User (ARPU) and decreasing in marketing expenses. Market segmentation enables organizations to separate a heterogeneous market into homogenous groups of customers with distinct behavior. So mass marketing is not applicable to get a competitive advantage from the customer. Therefore, telecom industries should define different marketing strategies that targeted different segments of customers to improve their business performance. The author describes how customer segmentation based on behavioral and beneficial features like a two-dimensional approach could have a better outcome compared to considering all features in one phase. Also, they developed various marketing strategies based on the customers' behavior to increase revenue.

As per [14] the basic premise of RFM is that customers who have purchased more recently, more frequently and have spent more with your company are your best prospects for future direct marketing campaigns. Like data mining/response modeling, the goal of RFM analysis is to increase marketing ROI. Analysis using RFM dramatically improved profitability by capturing 71% of buyers (3,214/4,522) while mailing only 46% of their customers (22,731/50,000). And the return on marketing expenditures using RFM was more than eight times (69.7/8.5) that of a mass mailing.

1.6 METHODOLOGY

In this study, to segment customers, one of the data mining technique called unsupervised learning approaches is selected to reach the desired knowledge. The researcher used clustering algorithms called K-means. These algorithms are selected

based on previously made researches recommendations for customer segmentation. Main steps under this research methodologies are classified:

- **Review literature or research papers** – literatures and research papers were reviewed for getting information about the problems, solutions and knowing which type of work was done by others on this topic and their methods.
- **Data collection and preprocessing** - In order to get the data from ethio telecom, we have got a formal supporting letter from Addis Ababa Institute of Technology (AAiT) and deliver it to ethio telecom. Accordingly, voice, SMS and data CDR of mobile users are collected and preprocessed. Then four datasets were prepared and converted to the appropriate format.
- **Tools and model experiment** –The model was experimented using the open-source "Weka" Machine Learning (ML) tool. The dataset with the clustering model was tested.
- **Result Analysis and Interpret** - The output results were analyzed with customer segmentation properties using the database script. Then it is interpreted, ranked and map with customer type to see their profitability level.

1.7 THESIS ORGANIZATION

This research paper contains six Chapters. Following this introductory Chapter, customer segmentation, why it is important, its benefit, and types of customer segmentation are discussed in Chapter two. Then Chapter three covered the discussion on data mining and its methods that are used in this study. Chapter four is about data preparation or preprocessing, experimentation and analysis. The fifth Chapter focuses on the result of the experiment and discussion. The last Chapter covered conclusions and future work.

2

CUSTOMER SEGMENTATION

Every business organization's success depends on the consummation of the customers. Whenever a business is new or not, customers always come "first" and then the profit [13][15]. Those companies that are succeeding in fulfill the customers entirely will remain in the top position in a market. Today's business companies have known that customer segmentation to fulfill their customer need is the key component for the success of the business. At the same time, it plays an important role to expand the market value [15]. In general, customers are those people who used goods and services from the market that meet their needs. Therefore, companies should define their needs with the feature of the product that satisfies the customer, maintains the long-term relationship and its profit [16].

2.1 WHAT IS CUSTOMER SEGMENTATION?

Customer segmentation is dividing of customers into groups of individuals that have common characteristics or traits. These characteristics may be age, gender, life stage, or behaviors (such as interests and spending habits). But also customer segmentation is about more than matching customers with suitable product offers and changing the way you communicate with your customers based on what you know about them. Moreover, it's about identifying your most profitable customers and tailoring your products and services to meet their specific needs. Ultimately, customer segmentation is about creating relevant spending experiences that build brand and loyalty for the companies[17][18].

2.2 IMPORTANCE OF CUSTOMER SEGMENTATION

Customers are becoming more sophisticated in how they navigate their spending choices. Targeting the wrong customers can cost the companies by not just in wasted money, also it brought higher operational costs associated with processing product returns and handling customer service. In contrast, targeting the right customers can give off higher conversion rates, higher average order values, and increased profits. It can also lead to brand promotion and word-of-mouth advertising, valuable product insights and greater overall customer satisfaction [15][19].

For identification of the relevant market, [20] the company should serve and divide the market into groups of customers (segments), who are likely to show similar purchase behavior. Segmenting your customers can help you to focus on your marketing efforts, increase profits sharing and overall customer satisfaction.

Segmentation is crucial in any today's marketplace as customers become knowledgeable about how they interact with the company's product or service and who they would like to purchase from. Today world has to move on from the mass marketing to a new age where customers, both Business to Business (B2B) and Business to Customer (B2C) companies consider understanding who they are, their specific needs and behavior [21].

In general, the concept behind customer segmentation is not new, however, now companies have more data available from a variety of sources and more sophisticated tools. This enables to deliver a more detailed and in-depth analysis of customer segmentation [21].

2.3 BENEFITS OF CUSTOMER SEGMENTATION

Define Audiences – Segmentation allows to understand your customers and identify customers who are the most profitable or valuable to offer the greatest future opportunity.

Personalization – The capability to group your customers, adapt communications and promotions based on their natural needs and preferences to provide a personalized service that customers expect.

Improved ROI – Understanding your customers makes you communicate effectively by proper channel. Targeted offers enable cost saving by targeting customers which more likely to improve the rate of response and Return on Investment (ROI).

Customer Satisfaction – Customer targeted product offerings and communication provide a better-personalized relationship that improved customer engagement and satisfaction. Attracting new customers costs 10 times more than retaining existing customers, which is extremely important for company ROI or profit.

Prospecting – Once you have identified your best customers' type, you can target the customer who represents your best opportunity prospects and futures.

Identify Growth Opportunities - Grouping customers based on common needs enables you to identify the most profitable segments and that provide the largest opportunity. Once you understand your customer segments and the value to the business you can identify specific strategies for each customer such as encourage, maintain and protect.

2.4 HOW CAN SEGMENTATION BENEFIT BUSINESSES?

Segmentation allows businesses to make better use of their marketing budgets, and demonstrate better knowledge of your customers' needs and wants. It can also help:

Marketing efficiency – Dividing a large customer base into small to more manageable pieces, making it easier to identify your target audience and launch operations to the most relevant people using the most relevant channel.

Determine new market opportunities –During the process of grouping customers into common behavior, companies may find a new market segment, which could alter its marketing focus and strategy to fit.

Better brand strategy – After identified the key motivators for your customers,

such as design, price, and practical needs, the companies can brand its products appropriately.

Improve distribution strategies – Identifying when product distributions strategies can shape, such as what type of products are sold at particular channels.

Customer retention – Using segmentation, companies can identify groups that require more attention to customers with the highest potential value. It can also help with creating targeted strategies that contained customers' attention and create positive and high-value experiences with its brands.

Behavioral segmentation – It shows, how customers are using your products or service and what type of user they are. It can also include customer transactional behavior, like spend by category, and price point, browsing, and sentiment. Buying behavior – including product usage, brand loyalty, and the benefits they seek from the product or service and identify your most and least profitable customers.

2.5 TYPES OF CUSTOMER SEGMENTATION

Depends on the type of businesses and key drivers that affect the company's products or services sales performance, they can choose to segment their customers by different key segmentation variables, such as behavioral, geographical, demographic and psychographic.

2.5.1 *Behavioral Segmentation*

Behavioral segmentation is the process of dividing the total customer into smaller homogeneous groups based on customer buying behavior. It analyzes characteristics on the basis of buying patterns of customers like usage frequency, brand loyalty, benefits needed during any occasion. It is useful for determining buying behaviors and what changes may affect these behaviors. It can also tell a company how many of its customers use more service repeatedly as opposed to repeating low. If product or service repeat purchases are low, the company may need to focus on improving quality or building brand loyalty [13][22].

If companies have the right data relating to customer segments and purchasing behavior, they can up-sell to the right person at the right time. Successful up-selling and cross-selling should incorporate by automated data-driven demand forecasting and pricing. Moreover, companies can also give staff training for the right people to do an effective job to deliver at the customer end to help in increasing revenue. Furthermore, companies need to put themselves in the shoes of the customers. This means thinking about different customer profiles and their buying behavior as well as product and service preferences. Without an understanding of customer segment characteristics, pricing is little more than guesswork [23].

Based on customer behavior of profit analysis [24] potential profit variables are duration, the number of transaction and the money spent. Due to this customers were segmented in terms of the period using the time spent during transaction (duration), purchase frequency and total purchase expenditure. The parameter was also set to 8, since eight (2x2x2) possible combinations of inputs (DFM). It can also be assigned ↑ or ↓ to the average for D, F, M values of a segment data. If the value is exceeded from the average value, an upward arrow included otherwise downward arrow was included. Such approaches assigned each customer in one of the following eight groups or clusters [25].

cluster	RFM Pattern	Customer Type
C1	R↑F↑M↑	Best
C2	R↑F↓M↑	Valuable
C3	R↑F↑M↓	Shopper
C4	R↑F↓M↓	First Time
C5	R↓F↑M↑	Churn
C6	R↓F↑M↓	Frequent
C7	R↓F↓M↑	Spenders
C8	R↓F↓M↓	Uncertain

Figure 2.5.1: RFM cluster adopted from [25]

The brief explanation mentioned in the above clusters is as follow:
Best Customers: It refers to those who bought recently, with a high buying frequency in a definite period of time, with high monetary value in each transaction.
Valuable Customers: Are those who bought recently, with a low buying frequency in a definite period of time, however with high monetary value in each transaction.

Shopper Customers: It refers to those who bought recently, with a high buying frequency in a definite period of time, however with low monetary value in each transaction.
First Time Customers: It refers to those who bought recently, but with a low buying frequency in a definite period of time and with low monetary value in each transaction.
Churn Customers: It refers to those with a high buying frequency in a definite period of time and high monetary value in each transaction, nevertheless, they haven't bought recently for some specific reasons.
Frequent Customers: It refers to those who haven't bought recently and the monetary value of their transaction is not that significant, however, they buy frequently in a definite period of time.
Spenders Customers: It refers to those who haven't bought recently and they don't buy frequently in a definite period of time, nevertheless, the monetary value of their transactions is very significant.
Uncertain Customers: It refers to those who haven't bought recently, with a low buying frequency in a definite period of time and with low monetary value in each transaction. This segment is the most insignificant customer with the lowest buying characters.

2.5.2 *Demographic Segmentation*

Demographics are personal characteristics used to categorize consumers. Demographic characteristics include age, gender, income level, and marital status. This information helps the company to develop a marketing strategy that appeals to individuals with this demographic profile. The company easily categorize and measures the wants of the consumers on the basis of demographic factors, compared to the variables of other segmentation strategies. It also helps to clarify your vision, has more direction with future advertising plans, and optimize your resources, time, and budget [13].

2.5.3 *Geographic Segmentation*

Geographic segmentation is an element that complements a marketing strategy to target products or services on the basis of where their consumers exist in. Division in terms of countries, states, regions, cities, colleges or areas is done to understand the customers and market a product/service accordingly. Moreover, people in different parts of the world show different characteristics. So marketing strategy created by dividing these different characteristics on the basis of geographic factors such as economics, food habits, clothing habits, languages, traditions, and many other traits [13].

Geographic segmentation is an effective methodology used by organizations with large local or international markets to better understand the location-based attributes that comprise a specific target market. Because customers that live in different geographic regions typically display varying needs, wants, and cultural characteristics, specifically targeted for more efficient and better marketing [13].

2.5.4 *Psychographic Segmentation*

Psychographic segmentation is the psychological aspects that influence customer buying behavior such as lifestyle, social status, opinions, and activities. It is used when you break your customer groups down into units as it affects their beliefs, values, and reasons for being. Psychographic segmentation is important to position the same product differently for different types of people. It prevents you from falling into one size fits all marketing. It also attracts a different group of customers with the same product without making material changes to its [13].

3

DATA MINING

Today a telecommunication network generates a large amount of data daily, which holds information about network, customers, its service, handsets, usage, and quality [8][13]. The amount of data is too big for complete manual analysis and the important information might not be discovered. However, knowledge discovery techniques provide means for disclosed hidden information from the data [8][26]. Data mining is a fundamental part of knowledge discovery in databases (KDD) which expresses a process of changing raw data into valuable information and discover meaningful patterns and rules from large data sets [4][8].
In this chapter, the data mining technique called clustering is introduced in detail, types of clustering methods will be discussed and finally, the cluster compatibility with the problem domain has been discussed.

3.1 DATA MINING

Data mining provides customer insight, which is vital for establishing an effective customer segmentation strategy. It can direct to personalized interactions with customers and therefore increased satisfaction and profitable customer relationships through data analysis. Moreover, it can support individualized and optimized customer management throughout all the phases of the customer life cycle, i.e. from the purchasing and creating of a strong relationship to the prevention of attrition and the winning back of lost customers. So companies struggle to get a greater market share and a greater share of their customers. In simple words, they are responsible for getting, developing, and keeping the customers. Data mining techniques can help in all these tasks [27].

Data mining is learning from data and consists of a set of rules and equations that can be used to identify useful data patterns, understand and predict behaviors. It can be grouped in to two main classes according to there goal [1][27].

1. Supervised Learning

2. Unsupervised Learning

Supervised and unsupervised learning are two quite different techniques of learning. As the names describe, supervised learning involves learning with some supervision from an external source whereas unsupervised learning does not.

Supervised learning often referred to as pre-defined labeled data [4]. To train a supervised model, classified data are needed, and the model then attempts to conclude a function or instruction set that can predict with new examples. Supervised machine learning methods include logistic regression, neural networks, decision trees, gradient bosting machines, random forests of trees, support vector machines and many more. Its limitation is [15] need high domain knowledge experts on the classification of each training sample.

Unsupervised methods are used when there are no prior sets of observations [4] [15]. Unlike supervised learning, unsupervised learning methods cannot be directly applied to a classification problem. It focuses on finding hidden patterns in data. This means that the purpose of these methods is to find patterns in the data that can help to give a structured representation.

Unsupervised learning involves clustering techniques that try to group a set of objects and find whether there is some relationship between the objects. Data mining has other techniques like association technique which mind data to identify the relationship between items in the same transaction and sequential patterns that mine data which frequently appears together [4].

For this study, since there is no label data were provided, we are interested in the unsupervised learning category by considering the customer should not be segment based on actual usage behavior [16]. Also, remove distrustful behavior in our dataset, i.e. data points that are significantly different from the others(called outliers).

3.2 CLUSTERING

At a very young age, people start to sort things based on their matches. This ability to categorize objects together, which helps people to realize what members of the different populations have common things and how they differ with others. Data mine experts have used this approach to perform clustering [8]. Therefore, Clustering is a technique to identify meaningful natural groupings of records and group customers into distinct segments with internal unity. It is also a task of allocating data into a number of subgroups that are similar within the group and dissimilar to other groups [4][8].

The data clustering is a process of partitioning to find a pattern, points, or objects. Objects within a valid cluster are more similar to each other than the objects outside the cluster. As per [28] defines cluster analysis as "a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics".

Clustering is an unsupervised technique which does not have predefined labeled data. Nowadays lots of areas are using many varied kinds of clustering algorithms to separate datasets into groups and have good quality results. Clustering methods differ in the choice of the objective function as well as the distance matrix used and the approach to construct the dissimilarity matrix. Clustering algorithms can be broadly categorized as partition and hierarchical. Other categories have also emerged, based on different data set [28].

To clarify some naming conventions used in this paper, the term segmentation has usually used in marketing while clustering is not restricted to any business domain and much more widely used term. In this thesis, both these terms are used interchangeably.

The clustering process in Figure 3.2.1 are describe how the data are clustered. First, the raw data which is dirty or noisy by nature are collected, then these collected data are preprocessed. Secondly based on the attributes, data are put on matrix format then, clustering algorithm will cluster the input data based on the number of clusters specified by the user.

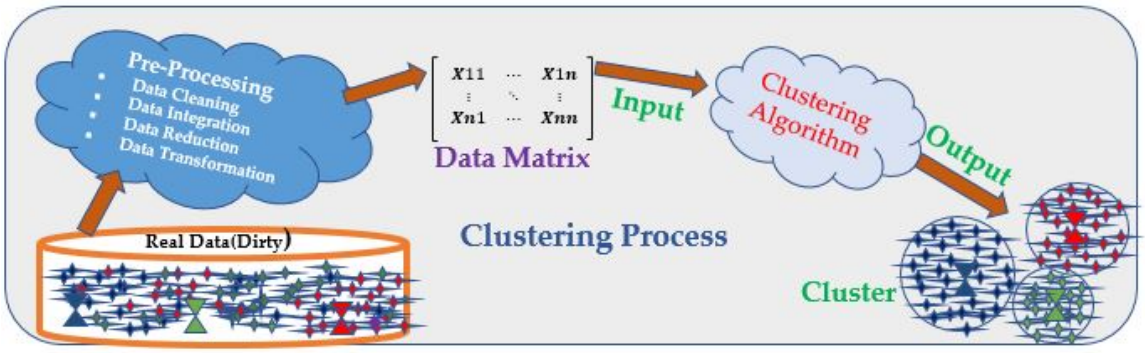


Figure 3.2.1: Clustering Process [8]

3.3 CLUSTERING TECHNIQUES

In data mining field, several ways exist to cluster data. Commonly applied clustering techniques are known as partition and hierarchical clustering [8].

3.3.1 *Partitioning Clustering*

Partitioning clustering methods divide the data object set into clusters where every pair of object clusters is either distinct (hard clustering) or has some members in common (soft clustering). Partitioning clustering begins with a starting cluster partition which is iteratively improved until a locally optimal partition is reached. The starting clusters can be either random or the cluster output from some clustering pre-process, such as hierarchical clustering. In the resulting clusters, the objects in the groups together add up to the full object set [29][30].

Characteristic for partition clustering is divide the data set into several clusters where each data sample can only belong to one cluster. To achieve this partition clustering typically tries to minimize the distance within clusters and maximize it between clusters [8][15].

Partition clustering algorithms [31] partition the n objects into a set of k non-overlapping groups. K is an input parameter that gives in how many clusters you want to partition. Partitioned clustering algorithms use an iterative method to group the data into a K number of clusters by minimizing the objective function.

In partitioning algorithm, if we want to create 2 clusters then we have to identify 2 points, we call them seed point. To find out the nearest seed point to all the points, we need to find the distance between a point and both seed points and assign it to the nearest seed point. In this process, the choice of seed point is an important consideration. Incorrect choice of seed points may give us an incorrect solution.

Partitioning methods produce distinct nonoverlapping clusters. Often, the methods are known as non-hierarchical clustering procedures because of only a single data partition is produced. The partitioning methods can be distinguished by five characteristics [32].

The first characteristic involves the selection of the initial starting partition or seed points. k-means methods use randomly selected data elements as starting partitions, while others allow the user to specify starting seeds.

The second and third characteristics deal with the type of cluster assignment pass made through the data and the statistical criterion used to assign the points to the clusters. K-means algorithms make a single pass, assigning each point in turn to the nearest cluster centroid, while others make multiple passes and update the centroids after each point assignment. Similarly, the statistical criteria range from simple distance measures between point and a cluster centroid to attempt to optimize rather a complex matrix borrowed from multivariate normal distribution theory.

The final two features involve whether a fixed or variable number of clusters will be formed, and the eventual treatment of outliers in the solution. Most methods the user to identify the number of clusters, and outliers are obligatory the dateset to join one of the clusters present in the solution [33].

The advantage of Partition clustering is that it will not have null clusters. One thing to be kept in mind while choosing seed points is to be sufficiently far away from each other so that correct clusters are formed[33].

3.3.2 *Hierarchical Clustering*

The hierarchical clustering method works by grouping a given set of data in a hierarchical way, which is formed as a tree cluster called a dendrogram. This tree of cluster consists of nodes and each node contains children cluster [34]. In other words, the aim of the hierarchical clustering method is to create a series of nested partition which can be represented via tree or hierarchy of cluster. The lowest level of the tree includes points that each one of them exists in its own cluster; while the highest level includes all points in one cluster [35].

The hierarchy can be formed in top-down (divisive) or bottom-up (agglomerative) fashion and necessarily it does not need to be extended to the extremes. In the divisive clustering method, it starts with all the data points contained as one cluster and recursively split each cluster into smaller clusters that contain consistent sub-clusters, till each point belongs to a unique cluster or some other pre-defined termination condition is reached. An agglomerative method of clustering, where starting with a unique cluster consisting of a single data element as its own cluster and merging the most similar pair of clusters iteratively, till a final cluster is achieved which contains all data elements or till some other pre-defined termination condition is reached [15] [28].

3.4 CLUSTERING ALGORITHM

Clustering is one of the machine learning technique that involves the grouping of data points. In a given set of data points, clustering algorithm can use to classify each data point into a specific group. It is a task for which many algorithms have been proposed. Different techniques are in favor for different clustering purposes [36]. So an understanding of both the clustering problem and the clustering technique is required to apply a proper method to a given problem. Every methodology follows a different set of rules for defining the 'homogeneity' among data points. There are many numbers of clustering algorithm known. However, in this specific study, selected clustering algorithms for customer segmentation, which

are used and recommended by many literatures is k-Mean clustering algorithms, and have been discussed in the below section in detail.

K-means Algorithm

K-means is well-known and widely used to automatically partition a data set into K groups [35]. It is unsupervised learning algorithm, which is used when you have unlabeled data (i.e., data without defined categories or groups). It is one of most simple clustering algorithm which is used to solve the problems of clustering by forming clusters iteratively. The aim of this algorithm is to find groups in the data, with the number of groups represented by the variable K. In K-means algorithm it has been needing to define numbers of clusters (i.e. K cluster) at the beginning. And any K points from the dataset are selected to be centroid. Then for each point calculate centroid-data point distance. Based on these distances, the point is associated with the nearest centroids. All the data points are divided into a number of clusters based on the distance of data points from the centroid of the cluster. Centroid is a unique point for each partition and the point from where distance is calculated for each data point.

The distance can be calculated using Manhattan distance, Euclidean distance, Cosine similarity, etc, but the most popular one is Euclidean distance. Once all the data points are placed, all K centroids are calculated again and got the new centroid. The new centroid is mean of all points in the cluster. Then all data points are reassigned to cluster with respect to new centroids by calculating centroid data point distance. This is done iteratively until a certain criterion is satisfied.

K-means algorithm partition the data points into the set of k clusters where each data point is assigned to its closest cluster. This method is defined by the objective function which tries to minimize the sum of all squared distances within a cluster, for all clusters.

The objective function is defined as equation 3.1

$$arg_s min \sum_{x=1}^{k} (\sum_{x_j \varepsilon s_i} ||x_j - \mu_i||^2) \tag{3.1}$$

Where x_j is a data point in the data set, S_i is a cluster (set of data points and μ_i is the cluster mean(the center of cluster of S_i and K is number o cluster).

Algorithm: The k-Means algorithm is explained in the following steps:

1. Define number of clusters and then select same number of data points as centroids.

2. Calculate the distance of a point from all centroid as shown in equation 3.2. Assign the point to cluster with minimum centroid-point distance.

$$Eculudian(b, c) = \sqrt{(x_b - x_c)^2 + (y_b - y_c)^2} \tag{3.2}$$

Where: x_b is the data point in the dataset, x_c is the selected centroid value of x_b, y_b is the observed dataset and y_c is the centroid value of y_b.

3. Repeat step 2 for all points.

4. Calculate the mean of all point in a cluster and assign it as new centroid for that cluster.

5. Repeat from step 2, until desired clusters or certain criteria are satisfied.

Finally, this algorithm minimized intra cluster distance (objective function also known as squared error function), and automatically maximized inter cluster distance (distance between cluster).

Advantages:

1. K-means clustering is simple, very fast, robust and easy to understandable and implement. If the dataset is well prepared from each other data set, then it gives best results.

2. The clusters do not having overlapping character, produces denser clusters and are also non-hierarchical within nature.

Disadvantage:

We can't assure the result might not globally optimal and the value of K required to be determined beforehand. So we can think the value only if we have a good idea about our dataset and if we are working with a new dataset then the elbow method can be used to determine the value of K.

4

CUSTOMER SEGMENTATION MODEL

In this Chapter, the methodologies used for the usage-based customer segmentation are discussed in detail. The system model in Figure 4.0.1 refers to how the research was organized in order to meet the research objectives, from data collection until model experimentation, analyzing and map with customer segmentation type.

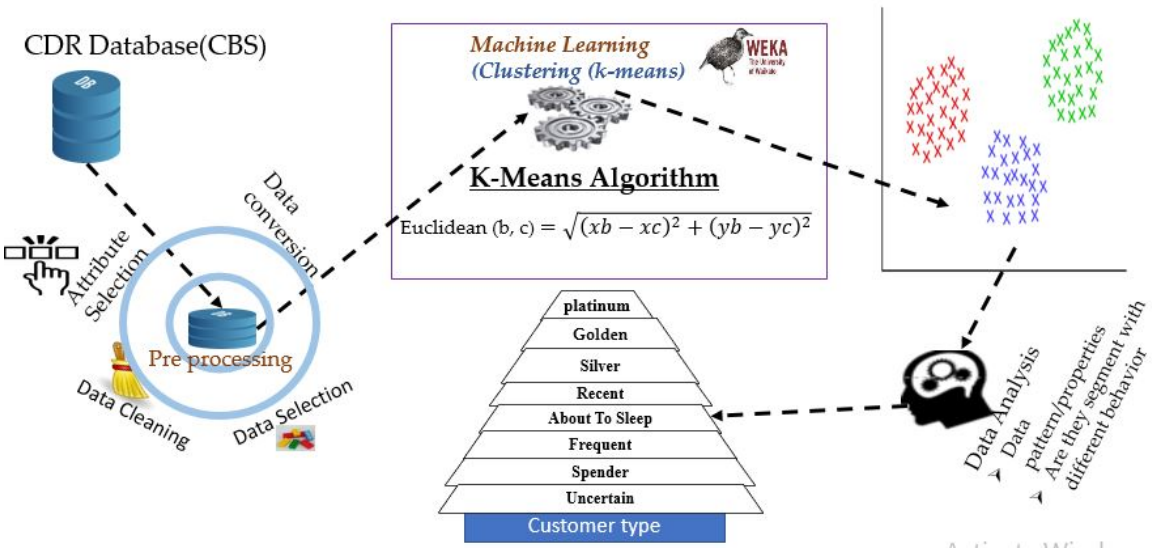


Figure 4.0.1: System Model

Initially, the pre-processing stage of analysis has been done in the Microsoft SQL Server 2017 developer. Secondly, the desired useful attributes of customer data have been produced with the appropriate format. Then using selected clustering algorithms (K-means), clustered into different groups. Finally analyzed the result to mapping with the appropriate customer category. This all forms the process of customer segmentation.

4.1 DATA PREPARETION

Data preparation (data preprocessing) means manipulation of data in the form suitable for further analysis and easy interpretation. It is a process that incorporates several different data preparation tasks and which is difficult to make fully automated. Many of the data preparation activities are repetitive, tiresome, and time-consuming. It has been estimated about 60-80% of the time spent on a data mining process [4][37].

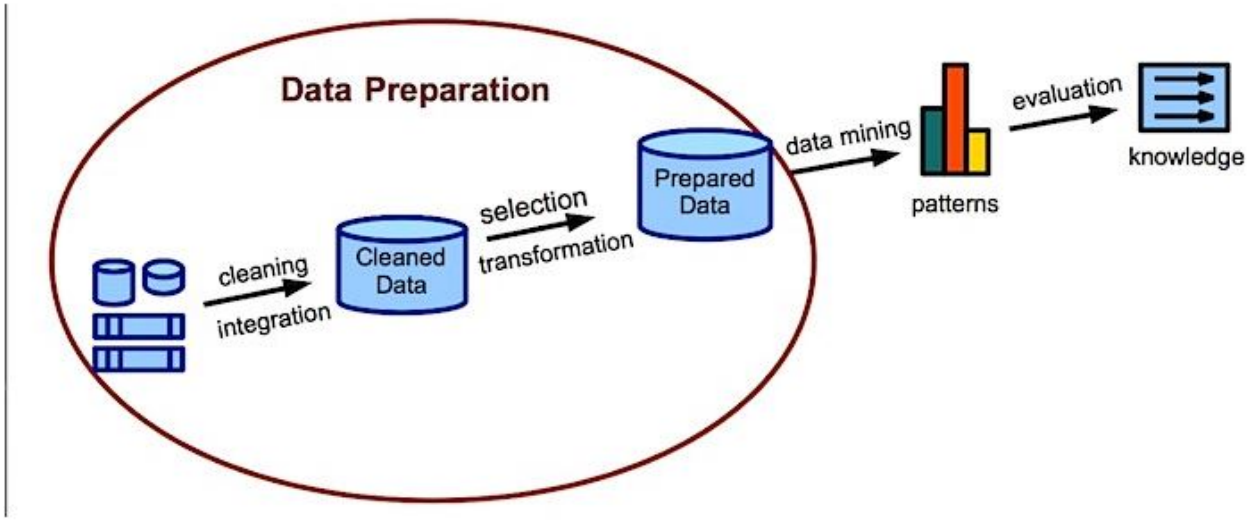


Figure 4.1.1: Data Preparetion adopted from [37]

As shown in figure 4.1.1 data preparation is important to make effective data mining. Poor quality data usually result in inappropriate and untrustworthy data mining results. Proper data preparation increases the quality of data and data mining results. The well-known saying "garbage-in garbage-out" is appropriate to this domain [37].

Extracting a descriptive set of features to construct a clustering model for a particular task is challenging in data mining. Current research has shown data mining algorithms affected by inappropriate and duplicated information [15]. When there is more irrelevant and redundant data, the information getting from this data is noisy and defective and which brings the clustering phase not becomes natural or

homogeneous. Good practices of data preparation help to prepare the raw data for meaningful analysis. Then the data mining algorithms can produce a structural description of the information contained in the data. It also defines, process and makes suitable to a data mining technique. So it is an important initial step in segmentation and analysis using data mining and plays a significant role in the whole process.

The objective of this thesis is making clustering from historical CDR data and mapping with customer segmentation type based on their usage behavior, which helps to see how to improve the profitability of the company. Thus the collected big amounts of CDR data need to be structured to form patterns and scenarios of customer usage behavior. Because real data's are usually vast in size, noisy, and may collect from different sources, therefore, knowing the data in detail is a vital requirement for data preparation [15].

For this specific thesis, CDR's of active sample mobile subscriber's usage, which is available in the ethio telecom billing system, for three month period has been collected. The collected data stored, processed in the database and finally generated the output in the format of the experiment required.

4.2 DATA COLLECTION

The data required for this study has been requested from ethio telecom using a formal cooperation letter from AAiT. After the approval of data access request from the responsible division, department, and section, the required CDR data collection started from ethio telecom Information System Division (ISD) and Billing section on CBS database. But due to the bulkiness of the record and resource constraint on CBS database, storing all CDR for a long period is challenging. So by finding the free server and discussing with different ISD sections for facilitation, the data started to store in a separate machine which makes it easy for the data preparation process and also guarantees for the protection of the business operations.

By considering this constraint, two-month CDR collected in dedicated server and one-month data collect from CBS directly, after identifying the sample subscriber.

Everyday CDR dump files pushed to the server. As shown in Figure 4.2.1, the received text files have been imported to a database through a manual data loading tool. This process continues from Jun 2019 until August 2019.

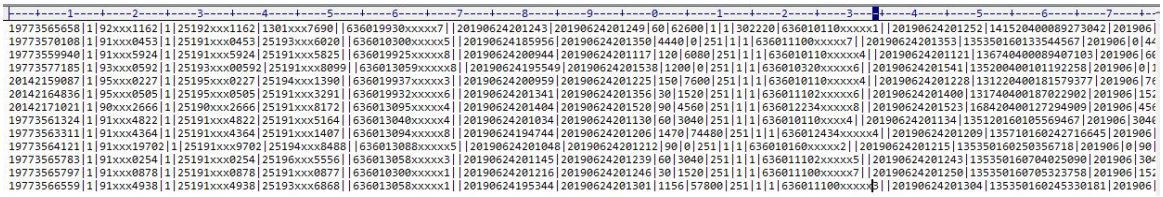


Figure 4.2.1: Raw CDR Dump File Section

The data type need for this research is: a voice CDR data, which is captured the information during Call. SMS data which stands for short message service and is also called text messaging. Internet usage data, which is the amount of information extracted or downloaded and uploaded [38].

Voice CDR as the name indicates, holds each and every record of calls made by the customer with details like calling number, called number, date and time of call, duration, amount charged and other details.

SMS CDR holds the detail information of the sent customer like calling number, called number, date and time, the amount of text content, the money charged, etc.

Data CDR also holds the internet connection and disconnection date and time, connecting number, the amount of download and upload, the money charged, the duration spend and etc.

4.3 DATA UNDERSTANDING

Data understanding is the primary task to uncover the hidden information in it. This step is critical to make it easy for unseen problems during the next phase. It involves accessing the data and exploring it using tables and graphics by data extracting tools. It also found activities such as identifying available features, investigate their value, and valuing their importance for this specific research. The data understanding phase of data mining involves taking a closer look at the data available for mining. In this process carefully studying the data and how it is

constructed were needed together with the domain experts' involvement. Associations of the data with the target problem and specific data mining tools were selected. The data needed in this thesis is collected from ethio telecom, the telecommunication service provider in Ethiopia.

4.3.1 *Call Detail Record*

Every time a call is placed on the telecommunications network of ethio telecom, descriptive information about the call is saved as a call detail record. The number of call detail records that are generated and stored is too large. For instance, in ethio telecom customers generate over 175 million call detail records every day and 3 months of call detail records are available online. This means that the millions of call detail data will need to be stored at any time. Call detail records include the necessary information to describe the significant characteristics of each call. At a minimum, each call detail record will include the originating/outgoing and terminating/called numbers, the date and the time of the call, the time used(duration), and money charged. Call detail records are generated when calls were complete and will be available almost immediately.

As shown in Table 4.3.1, the collected CDR contains a total of 33 fields. Some fields are empty values like Calling IMEI, others to contain duplicate values such as Billing number and Calling number. Most of them are generated for billing purposes like Charge, Call Fee, Account Item ID, Rate ID, Billing Date, Billing Offering ID and Billing Cycle ID. Upload traffic and Download traffic contains internet usage. CDR_ID uniquely identifies each CDR and RE_ID used to differentiate service types like Voice, SMS and Internet Data usage records. CDR_TYPE included for distinguishing call originating, terminating or forwarding call types. For privacy reasons some fields such as called, calling or Billing numbers have been hidden.

In contrast call detail records cannot be used directly for data mining since the aim of data applications is to extract knowledge at the customer level, not at the individual phone call level [18][39]. Thus, the call detail records related to a customer

must be summarized into a single record that describes the customer's calling behavior. The choice of summary variables (features) is also critical in order to obtain a convenient description of the customer. To define the variables, it has been considering the minimum set of variables that describe the complete behavior of a customer like when, what, how often, who, etc. which can help to understand this process [39].

Table 4.3.1: CDR Fields Description

No	Attributes	Description
1	CDR_ID	CDR Sequence Number
2	RE_ID	CDR type ID for voice, SMS and Data
3	BILLING_NBR	Billing Number
4	CDR_TYPE	Call type Id(The types of call 0: local call , 1: toll call within a charging area, 2: toll call between charging areas ,3: international toll call)
5	CALLING_NUMBER	Calling Number(call initiate number)
6	CALLED_NUMBER	called number (call destination number)
7	CALLING_IMEI	International mobile equipment identity
8	CALLING_IMSI	IMSI of the calling party(The value is reported by the network side)
9	THE_THIRD_PARTY_NUMBER	Third Party Number
10	CALL_START_TIME	Call start time(the time when call start)
11	CALL_END_TIME	Call end time(the time when call end)
12	CALL_DURATION	Call duration
13	CALL_FEE	the actual money deducted(charge amount)
14	CALLED_COUNTRY	country area code of called number
15	CALLING_CARRIER	Calling carrier

Continued on next page

Table 4.3.1 – continued from previous page

No	Attributes	Description
16	CALLED_CARRIER	Called carrier
17	CALLING_DISTRICT	Cell ID of the calling party, which is reported in an IDP message
18	CALLED_DISTRICT	Cell ID of the called party, which is reported in an IDP message
19	STATUS_DATE	Billing date
20	CALLING_SUB_ID	Calling subscriber ID
21	BILLING_CYCLE_ID	Billing cycle ID
22	CHARGE_1	Charge amount of you spend
23	CHARGE_2	Charge amount of you get disscount
24	RATE_ID1	Rate ID
25	ACCOUNT_ITEM_ID1	Account item ID
26	UPLOAD_TRAFFIC	Upload trafficc
27	DOWNLOAD_TRAFFIC	Download traffic
28	BILLING_OFFERING_ID	Billing offering ID
29	ERROR_CDT_TYPE	Error CDR Indicator
30	CALLFORWARDINDICATOR	Call Forward Indicator
31	HOTLINEINDICATOR	Hot Line Indicator (voice mail)
32	CALLING_TRUNK_ID	Calling Trunk ID
33	CALLED_TRUNK_ID	Called Trunk ID

4.3.2 *Data Selection*

The data mining process needs a suitable volume of historical data from customer usage for analysis. Usually, the data repository before making integration of the data contains much more data than actually required. From the available data, necessary require data to be selected and stored. Data selection is the process where the data relevant to the analysis is retrieved from the database [35]. The data selection step helps to establish the criteria on which the data will be chosen [35]. This thesis ultimate target is to study how to segment and identify the customer type to improve profitability from their usage behavior through historical data. So based on the number of the subscriber which is 90% of ethio telecom subscriber, its revenue generation contribution which is 69% revenue contribution and consulting literatures [35], mobile subscriber data is selected. And also the following criteria have been set to data selection.

1. From 29.6 million active mobile subscribers, 20,780 distributed mobile subscriber is selected using probability sampling techniques and with the help of DBMS RANDOM VALUE function from oracle database.
2. Currently, based on the number of subscriber coverage and its greatest revenue generation contribution, mobile subscriber data was selected.

So the data selection is identified based on the above two criteria using writing database script and after that process it. The row data found for this sample subscriber is 11,574,477 record and these selected CDRs are further preprocessed.

4.3.3 *Data Sampling*

The success of large data depends fundamentally on the skill of expert knowledge (the data scientist) to create sense and make insights from these riches of data. The process of generating actionable insights, called data exploration, is time-consuming and a difficult task. Data exploration of a big dataset usually re-

quires first generating a small and representative data sample that can be easily managed, viewed and interpreted to generate insight [40].

Sampling is an important step to data reduction techniques because it allows being representative of a large data set by taking a much smaller random data sample. It is often used as managing big data processing problems. Since taking all the collected records in the dataset is difficult tasks to mine the data from big data and challenging for the data mining tools and ML algorithm. It also requires modern and advanced hardware and ML algorithm. Typically, the clustering algorithms show poor performance when allocating with unfair dataset and result in a bias towards the majority group [4] [26].

An insufficient sample size of the segmentation base can have serious negative consequences on segment retrieval. Segment retrieval can be substantially improved by increasing sample size minimum from 10 to 30 times the number of variables. This improvement levels off subsequently but is still noticeable to extend the sample size up to 100 times the number of variables. Improvement in the segment at high sample size levels occurs only if more sample data is used [41].

For this thesis, like in table 4.3.3, from $29,582,489$ active mobile subscriber numbers, $20,780$ mobile subscriber numbers were selected using random sampling with stratified sampling techniques. They also generated about 11.5 million records for three months as shown in table 4.3.2.

Table 4.3.2: Total CDR collected

Number of CDR collected(for 20,780 subscriber)				
service	**May**	**June**	**July**	**Total**
Voice	2,718,887	2,878,304	2,987,717	8,584,908
Data	924,835	425,748	354,144	1,704,727
SMS	554,709	275,291	454,842	1,284,842
Total	**4,198,431**	**3,579,343**	**3,796,703**	**11,574,477**

Table 4.3.3: Sample subscribers

Prefix of subscriber	**Active numbers**	**Random sampling ratio**	**Sample numbers**
90	3,235,885	10.94%	2,273
91	5,494,709	18.57%	3,859
92	3,720,838	12.58%	2,614
93	3,000,577	10.14%	2,107
94	3,338,561	11.29%	2,346
95	1,852,126	6.26%	1,301
96	2,768,593	9.36%	1,945
97	2,382,822	8.05%	1,673
98	2,803,775	9.48%	1,970
99	984,603	3.33%	692
Total	**29,582,489**	**100%**	**20,780**

4.4 DATA PREPROCESSING

The base for customer segmentation is collecting suitable usage data before any process begins. To make accurate follow up steps and reliable, the collected data must be useful and complete.

As per [42] the raw data accumulated in operational systems are prone to various kinds of errors. They are often noisy, inconsistent, incomplete, or outdated data. Therefore, the data should be preprocessed prior to mining to reduce the mining efforts and to enhance stability and interpretability. So one of the important stages of data mining or machine learning is preprocessing. It describes any type of processing used to prepare the raw data for another processing procedure. It transforms the data into a format that easily and effectively processed for the purpose of the experiment. For this study, several database scripts are used for preprocessing and the scripts are found in Appendix A.1.

4.4.1 *Data Cleaning*

The existence of incomplete data has to be investigated carefully while performing a clustering analysis. Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies must be extracted in this stage [42]. Also, errors, bad designed, unnecessary attributes or fields must be removed.
For this study, before the data cleaning phase, it was 11,574,477 customers' transaction records. Later, 1,881,181 records found as missing values like CDRs with errors, zero duration, and zero fee values and duplicate data and then discarded them from the database. Finally cleared 9,693,296 records for 20,780 unique mobile subscribers during three months were ready for the next stage.

4.4.2 *Feature Selection and Data Reduction*

In machine learning, feature selection is an important part of selecting a subset of relevant features (variables) in model construction [43]. It is the process of choosing the data metrics to input as features for ML algorithm by using domain knowledge. Feature selection plays a key role in clustering; using meaningful features that capture the changeability of the data, which is essential for the algorithm to find all of the naturally-homogeneity occurring groups. Also, it reduced the complexity of analysis and increase segmentation effectiveness. The number of attributes was reduced, and the selection process has been done with the help of domain experts and by consulting different literatures [42].
Removing attribute in feature selection stage started by discarding some of the attributes which have redundant values, like fee and charge, billing number and calling number, CDR type and Re id, having zero or one and other constant value, attributes having empty or null value like calling IMEI, called district, calling trunk id, called trunk id, rate id1, etc. In addition, non-relevant attributes are also removed by consulting domain experts and related literatures. Accordingly from the total of 33 attributes, only 10 of them are selected.

The below Table 4.4.1 describes the selected attributes, the data type, and their description.

Table 4.4.1: Selected Attribute Fields, Data Types and Description

No	**Attributes**	**Data type**	**Description**
1	RE_ID	NUMBER(2,0)	CDR type ID for voice, SMS and Data
2	CALLING_NUMBER	VARCHAR2(60 BYTE)	A number initiating or originating the call
3	CALLED_NUMBER	VARCHAR2(60 BYTE)	The number call is terminated or received
4	CALL_START_TIME	DATE	Data and time of call initiation
5	CALL_END_TIME	DATE	Data and time of call terminated
6	CALL_DURATION	NUMBER(10,0)	Call duration in second
7	CALL_FEE	NUMBER(10,0)	The actual money deducted(charge amount)
8	CALLED_COUNTRY	VARCHAR2(10 BYTE)	country area code of called number
9	UPLOAD_TRAFFIC	VARCHAR2(60 BYTE)	Upload trafficc
10	DOWNLOAD_TRAFFIC	VARCHAR2(60 BYTE)	Download traffic

4.4.3 *Data Agrigation*

Data aggregation is another part of preprocessing in which information is gathered and expressed in a summary form for analysis. A common aggregation purpose is to get more information about particular groups based on specific variables like duration, revenue, frequency and date and time. It takes a set of data and merges it into a single value. For example, the number of unique active voice

subscribers within the last 30 days, or the number of times on a given product feature were used in the peak hour or off-peak hour time. When viewed this over time, aggregated metrics show a movement of changes in customer behavior [4][35]. As discussed before, CDR tells information at the level of individual phone calls. However using data mining applications we can extract knowledge at the global, customer or user level by aggregating into a single record [1][37].

Aggregation of variable detail also sometimes works well, especially if a dataset has several aggregation levels. It's often beneficial to have variable aggregations along with some common metrics. For instance, aggregating the second usage to minute, minute to hourly of peak and off-peak hour time, daily to monthly (depending on the needs of the business problem) that may produce better estimates. Common aggregation periods tend to introduce less noise and produce a less residual variance. The aggregation of attributes for this thesis is described in Table 4.4.2. Which reduce noisy and irregularity, and has been followed the following steps:

1. The collected and cleared CDRs are aggregated at the customer level.

2. For voice usage aggregated by peak and off-peak hour time per month.

3. For SMS and data service since there is no fee difference in time, they are aggregated only by month.

4. To make optimum it also take total average usage of the customer.

4.4.4 *Data Integration*

Data mining often requires data integration, which means the combining of data from multiple data stores. Proper integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help to improve the quality and speed of the subsequent data mining process. An attribute such as fee and charge may be redundant if it can be derived from another attribute or set of attributes.

Table 4.4.2: Derived Attributes with Description

Attribute	**Data type**	**Description**
SERVICE NUMBER	VARCHAR	Service number
PEAK_FREQ	NUMBER	Number of call originated at peak time(voice)
PEAK_DUR_INMI	NUMBER	Sum of duration in minute at peak time(voice)
PEAK_FEE	NUMBER	Charge fee at peak time(voice)
OFF_PEAK_FREQ	NUMBER	Number of call originated at off peak time(voice)
OFF_PEAK_DUR_INMI	NUMBER	Sum of duration in minute at off peak time(voice)
OFF_PEAK_FEE	NUMBER	Charge fee in Birr at off peak time(voice)
SMS_FREQ	NUMBER	Number of SMS originated
SMS_FEE	NUMBER	Total SMS Charge fee
DATA_FREQ	NUMBER	Number of data connected
DATA_USAGE_INMB	NUMBER	Actual data usage in Mega bayet
DATA_CHARGE_FEE	NUMBER	Actual data charge fee
AVG_FREQ	NUMBER	Total Average number of call originated
AVG_DUR_INM	NUMBER	Total Average duration in minute
AVG_FEE	NUMBER	Total Average fee charged in Birr

For this thesis, the database contains a separate three tables: Voice, SMS, and Data information table, which consists of CDR type, calling number, called number, call start time, call end time, call duration, call fee, called country, upload traffic, download traffic. In order to make the integration, the three table information must be merged to achieve full transaction of a customer table.

4.4.5 *Data Transformation*

In this step, some string variables need to be converted into numeric variables and some codes should be replaced by text. The other tasks in this step are data aggregation [44]. So total purchase usage of a customer in a period of time must be aggregated for performing consequent processes.
The filed for call start time and call end time presented in table 4.4.3 is used to identify peak and off-peak time of a number of transaction (frequency), duration in minute used, and the amount of money spent in birr separately. The field duration values were on seconds in original data and converted to minutes, and the filed call fee values were also on cents and converted to birr. The filed uploads and download value was in a byte and converted to megabyte. This all helps to minimize the load for the algorithms for proper segmentation.

Table 4.4.3: Peak and Offpeak time range

Time period	**Range for trafiic**	**Fee per minute**
Peak time	From 7:01 am to 9:59 pm Monday to Saturday	0.5
Off peak time	From10:00 pm to 7:00 am Monday to Saturday + Sunday + public holidays	0.35

Off-peak hours' time in ethio telecom context, the network traffic assumed to be idle and created to encourage the subscriber to use at that time with less price.

Peak hours' time means the network traffic or conjunction is high time period.

4.4.6 *Data Formating*

In a classic machine learning task, data is denoted as a table of instances. Each instance of a transaction is described by a fixed number of measurements or features. Formatting is a process of re-engineering the input dataset into a format that are suitable by the particular ML algorithm. Commonly attributes are nominal or numeric data type [45]. In this study, all the records used are numeric format, and consistency of format is checked during aggregating the raw data to derived attributes carefully in the entire document.
Before dealing with the experimentation the dataset has to be formatted in a way that suitable for the tool to be used for modeling. In this study WEKA 3.8.2 is used which requires file formats like Comma Separated Values (CSV), Attribute Relation File Format (ARFF) and ARFF format preferred to use. Since the preprocessing is done using the database and it can provide the data in such format.

4.5 EXPERIMENTATION

For this study, it had been performed four sets of experiments. Three experiments were clustered usage per month and the fourth experiment was the total of the three-month usage per each customer. Finally, the cluster data analyzed by service type and consolidate with their customer type or customer segmentation type in the next chapter.
The dataset used in this experiments are classified as MAY dataset, JUNE dataset, JULY dataset, and MAJUJL dataset. Further explanation is given below:

- All the four data sets have 20,780 number of aggregated subscriber instance.
- 15 aggregated attributes.
- The dataset is set for the month of May, Jun, July 2019 and there total.
- It is included Voice, SMS, and Data usage of sample subscribers.

4.5.1 *Define the Number of Cluster*

There is no common rule to find the solution for an optimum number of clusters for any given dataset. However, input parameters are the way to deliver information from a stored process. For K-means the number of K is the input parameter for clustering [8]. As discussed in section 3.4, the disadvantage is that K-means always make to the K amount of clusters even if there would not exist this number of clusters. However, the discussion made in subsection 4.5.1.1, one method to validate the quality of the cluster is defining the number of clusters using the elbow method. Then clustering could be made more reliable by applying this method, which helps to decide the number of clusters. Thus, before starting the clustering process for analysis K-means always need to adjust the number of clusters.

4.5.1.1 *Elbow Method*

The main idea behind clustering methods, like k-means clustering, is to describe clusters which are minimized the total intracluster variation. It measures the compression of the clustering and it has to be minimized as small as possible. Although k-means functioned fine on this dataset method, and it is important to recall that a downside of k-means is difficulties to specify the number of clusters, K before what the optimal K is. Choosing the number of clusters may not always be obvious in real-world scenarios, specifically, if we are working with a big dataset that cannot be visualized easily [39].

The elbow method is one of the useful graphical tool methods for the evaluation of the optimum number of clusters k for a given dataset. Spontaneously, It has been saying that, if k increases, the within-cluster Sum of Square Error (SSE) ("alteration") will decrease. This is because the data will be nearer to the centroids they are given. The elbow method identifies the value of k where the alteration begins to decrease most quickly, which is become perfect if we plot the alteration for different values of k. The Elbow method describes the total SSE as a function of the number of clusters. Once we choose a number of clusters in this way; adding another cluster doesn't improve much better the total SSE [39].

Steps to define the optimal number of cluster

1. Compute selected clustering algorithm by varying the range of values k (1-12) on the same dataset.

2. For each k, compute SSE .

3. Plot the curve of SSE together with the number of clusters k.

4. As shown in Figure 4.5.1, the location of the knee (bend) in the plot is generally considered as an indicator of the appropriate or an optimum number of clusters. For this thesis dataset, the optimal number of clusters is preferred to be eight(8).

No of cluster	SSE
2	76.0933
3	70.8821
4	65.4043
5	57.3939
6	47.7475
7	35.8155
8	25.0749
9	22.9268
10	21.1708
11	18.0749
12	15.7475

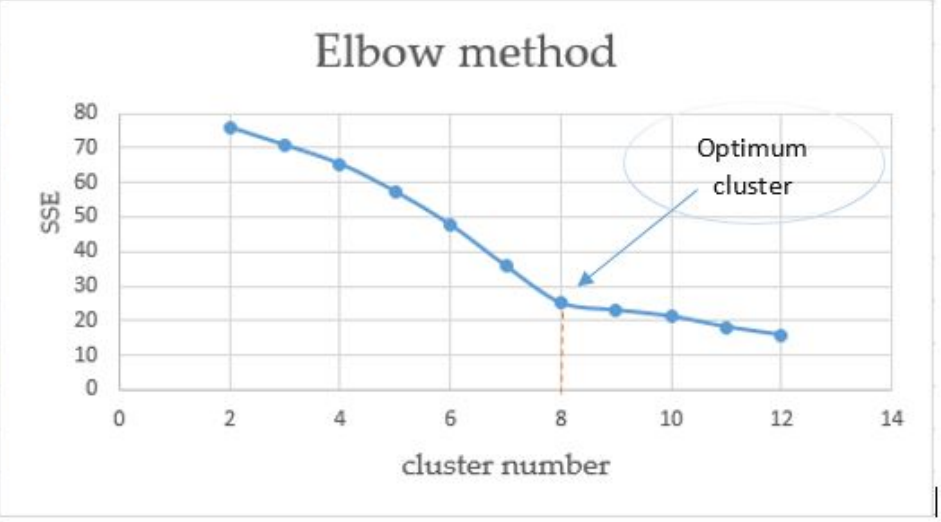


Figure 4.5.1: Optimum number of cluster [39]

$$SSE = \sum_{x=1}^{k} \sum_{o \varepsilon c_i} d(o, cen_i)^2 \tag{4.1}$$

After defining the number of clusters, the unlabeled dataset record has been clustered into eight. Then examined customer segmentation properties as discussed in Section 2.5 and during a conversation with domain expert on the real scenario of ethio telecom customers call behaviors.

4.5.2 *Experiment Using K-means Algorithm*

Four datasets were used during clustering and, these datasets are presented in section 4.5. As mentioned in section 4.5.1.1 cluster size was set as eight. Then the test was run on the laptop having 8GB RAM and 2.6 GHz processor. After running the four datasets the clustered data are as follows:
Using MAY dataset, from the total of 20780 instances 5866 (28%) number of instance grouped under cluster 0, 1257 (6%) number of instance grouped under cluster 1, 351 (2%) number of instance grouped under cluster 2, 3055 (15%) number of instance grouped under cluster 3, 267 (1%) number of instance grouped under cluster 4, 484 (46%) number of instance grouped under cluster 5, 421 (2%) number of instance grouped under cluster 6 and the rest 107 (1%) instance under cluster 7. The algorithm took 44 number of iteration and 1.22 second.
Using JUNE dataset, from the total of 20780 instances, 5828 (28%) instance grouped under cluster 0, 2901 (14%) instance grouped under cluster 1, 331 (6%) instance grouped under cluster 2, 493 (2%) instance grouped under cluster 3, 620 (3%) instance grouped under cluster 4, 104 (1%) instance grouped under cluster 5, 328 (26%) number of instance grouped under cluster 6 and the rest 4137 (20%) instance under cluster 7. The algorithm took 47 number of iteration and 1.6 second until all data object are clustered.
Using JULY dataset, from the total of 20780 instances 5828 (28%) number of instance grouped under cluster 0, 2901 (14%) number of instance grouped under cluster 1, 331 (6%) number of instance grouped under cluster 2, 493 (2%) number of instance grouped under cluster 3, 620 (3%) number of instance grouped under cluster 4, 104 (1%) number of instance grouped under cluster 5 ,328 (26%) number of instance grouped under cluster 6 and the rest 4137 (20%) instance in cluster 7. The algorithm took 47 number of iteration and 1.6 second until all data object are clustered.
Using MAJUJL dataset, from the total of 20780 instances 4141 (20%) number of instance grouped under cluster 0, 5513 (27%) number of instance grouped under cluster 1, 358 (2%) number of instance grouped under cluster 2, 1060 (5%) number of instance grouped under cluster 3, 1804 (9%) number of instance grouped under

cluster 4 ,236 (1%) number of instance grouped under cluster 5, 5279 (25%) number of instance grouped under cluster 6 and the rest 2389 (11%) instance under cluster 7. The algorithm took 121 number of iteration and 3.21 second until all data object are clustered.

The graphical results of the four dataset are shown in Figure 4.5.2. Additionally, the detail output information is summarized in tables A.1.1, A.1.2, A.1.3 and A.1.4.

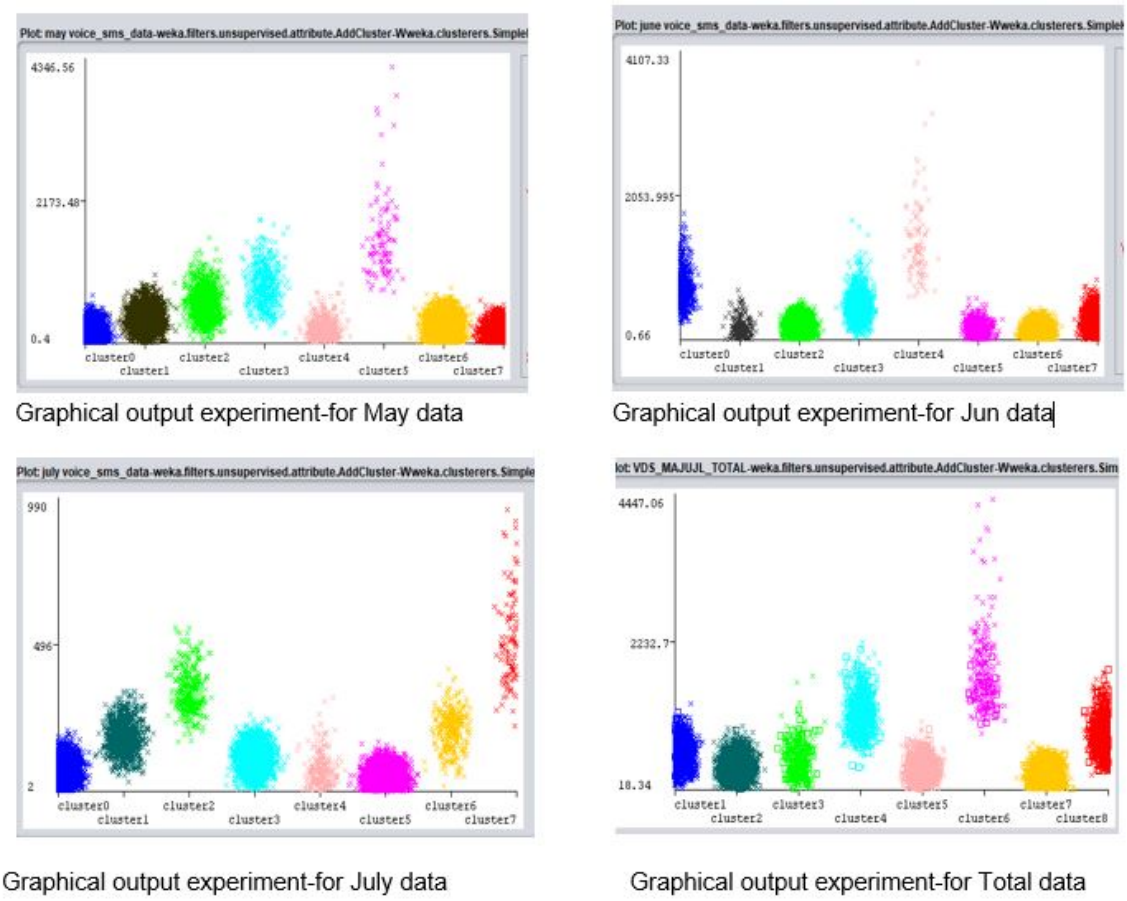


Figure 4.5.2: Graphical result of the four experiments

Following the clustering experiment, further analysis was performed based on clustering behavior. To do this the values of the total instance found in each cluster data were again grouped. Get the average value of duration, frequency, and monitory of each cluster, then make them ordering and gives score using quantile methods as a parameter to differentiate customer type using database script. Finally based on the defined customer type parameter the clustered instances were mapped to each type of customer.

5

RESULT AND DISCUSSION

This Chapter will present and discuss the results found during the experiment made in Chapter 4. A model with k-means was tested using real CDR. These data analyses further classified into four. One to three are analyzed based on service type usage of sampled subscriber and the fourth one is prepared based on the total of the three service type usage. Accordingly, the results obtained from the k-means algorithm with four datasets are explained in the following subsection. Finally, in Section 5.2 covered the discussions on results.

5.1 RESULT

The result obtained from the clustering model with k-means is presented in Figure 5.1.1. The dataset was clustered into eight groups. The clustered data records further analyzed and their results will be discussed in each subsection.

Analysis of clustered result

The analysis of customer DFM initiates by grouping customers based on their buying behavior, in terms of how much time spent during they bought, how often did they bought, and how much monetary value they contributed to the company. In DFM Analysis, three parameters were analyzed, each denoted by the letters D, F, and M. To satisfy the need of knowing true customer value, analysis of just one parameter will give an inaccurate report of the customer base, so the customer segmentation value can not be reliable. That's why at least three parameters of customer's purchase behavior are analyzed and computed for each customer.

Now, it is time to do a clustering analysis using the three variables. To do this, we have four options. First, compare customers' DFM values with voice usage average

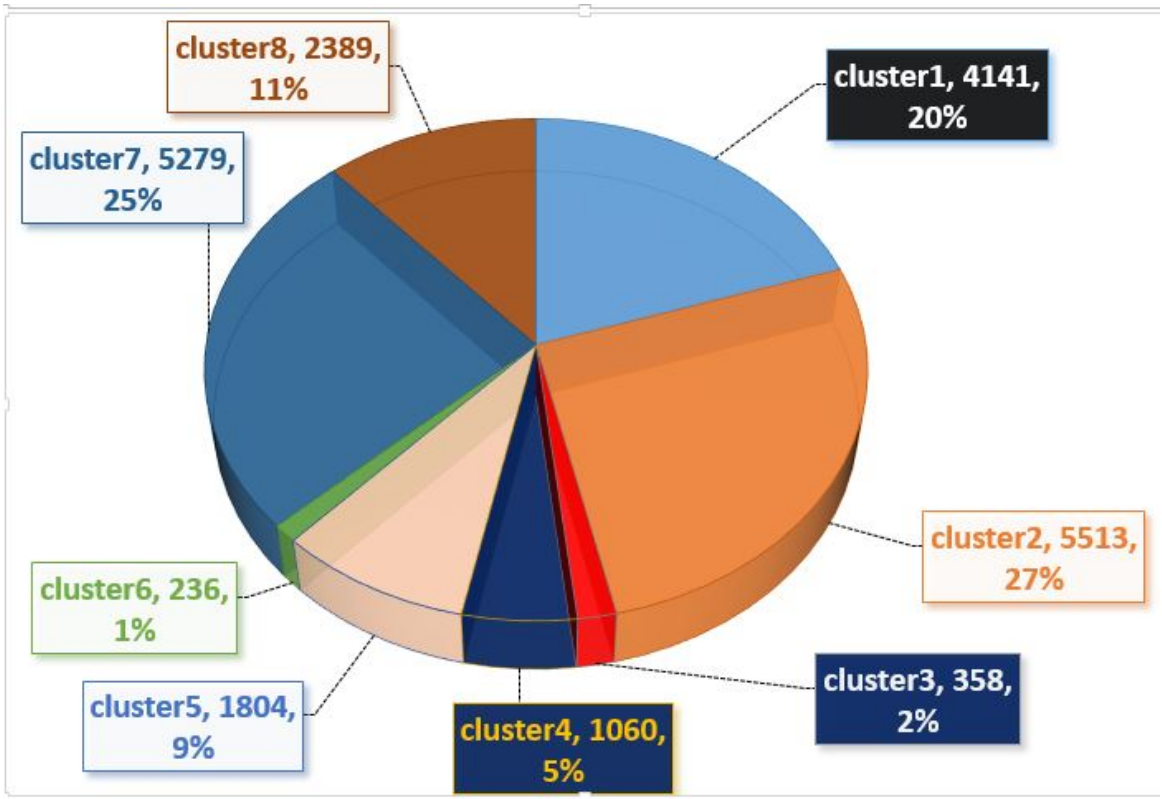


Figure 5.1.1: Clustered instance using k-means algorithm

DFM values. Second, compare customers' DFM values with data average FMBM values. The third one compare Frequency and Monetry (FM) values with SMS average FM values and the last one is comparing a clustering analysis by the total of DFM value shown on Figures 5.1.2, 5.1.3, 5.1.4, and 5.1.5 respectively.

Using this method we can generate eight customer segments. And customers are divided along the following key dimensions: duration or the amount of time spent, frequency of transaction and earned monetary amount. The total average values of the customers DFM parameters are concatenated and measured as the base for this segmentation.

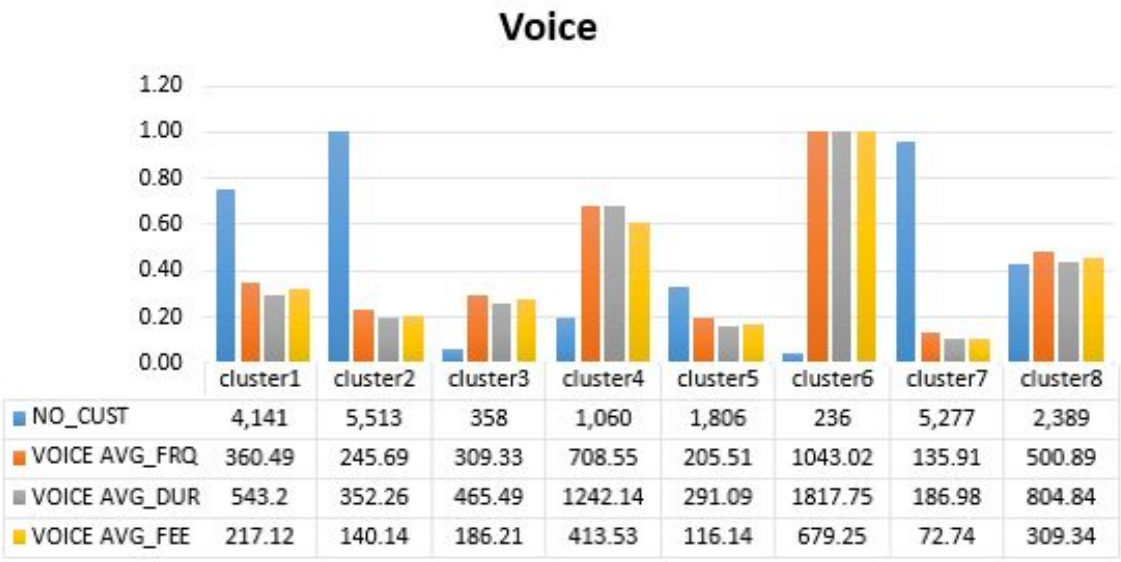
5.1.1 *Result Obtained Using Different Service Type*

The segmentation result obtained from different types of service with k-means are presented in the Figures 5.1.2, 5.1.3, 5.1.4,and 5.1.5. Each data were taken from

the clustered dataset result and further analyzed and discussed in the following subsection.

5.1.1.1 *Experiment result in Voice usage*

The result using voice CDR dataset shown in figure 5.1.2, points that cluster 6 is better to voice users than all other clusters, cluster 7 which contain 5,277 subscriber is the least voice users of all clusters and the rest six clusters were between them.

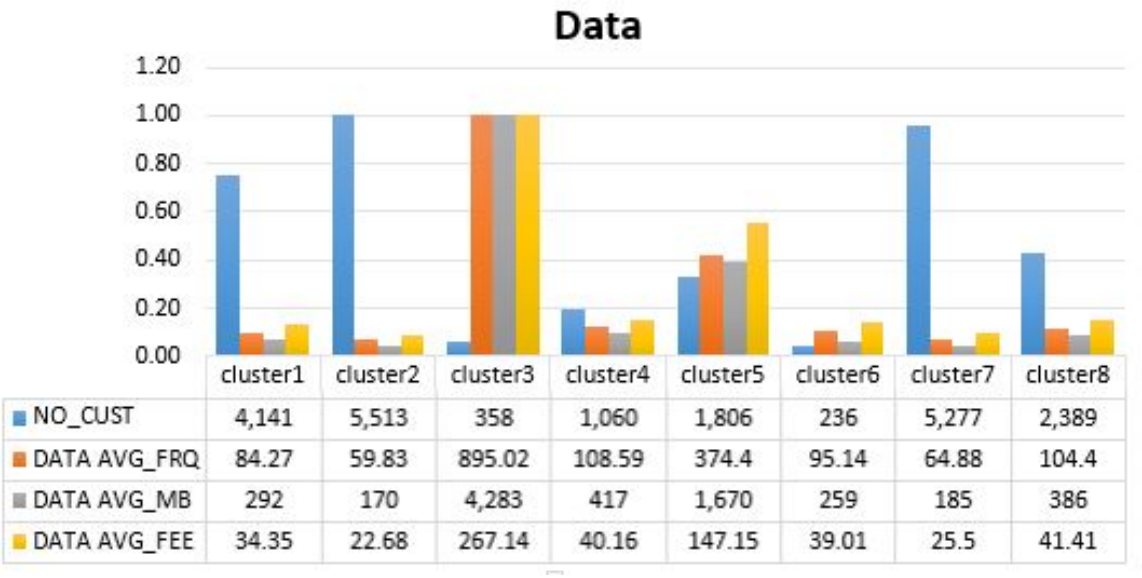


	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8
NO_CUST	4,141	5,513	358	1,060	1,806	236	5,277	2,389
VOICE AVG_FRQ	360.49	245.69	309.33	708.55	205.51	1043.02	135.91	500.89
VOICE AVG_DUR	543.2	352.26	465.49	1242.14	291.09	1817.75	186.98	804.84
VOICE AVG_FEE	217.12	140.14	186.21	413.53	116.14	679.25	72.74	309.34

Figure 5.1.2: Clustring with voice

5.1.1.2 *Experiment result in Data usage*

In a similar manner, the result using Data CDR dataset shown in figure 5.1.3 directs, cluster 3 is better data users than the rest of other clusters and cluster 2 which contain 5,513 subscriber is the least data users of all clusters in this data set.



	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8
NO_CUST	4,141	5,513	358	1,060	1,806	236	5,277	2,389
DATA AVG_FRQ	84.27	59.83	895.02	108.59	374.4	95.14	64.88	104.4
DATA AVG_MB	292	170	4,283	417	1,670	259	185	386
DATA AVG_FEE	34.35	22.68	267.14	40.16	147.15	39.01	25.5	41.41

Figure 5.1.3: Clustring with data usage

5.1.1.3 *Experiment result in SMS usage*

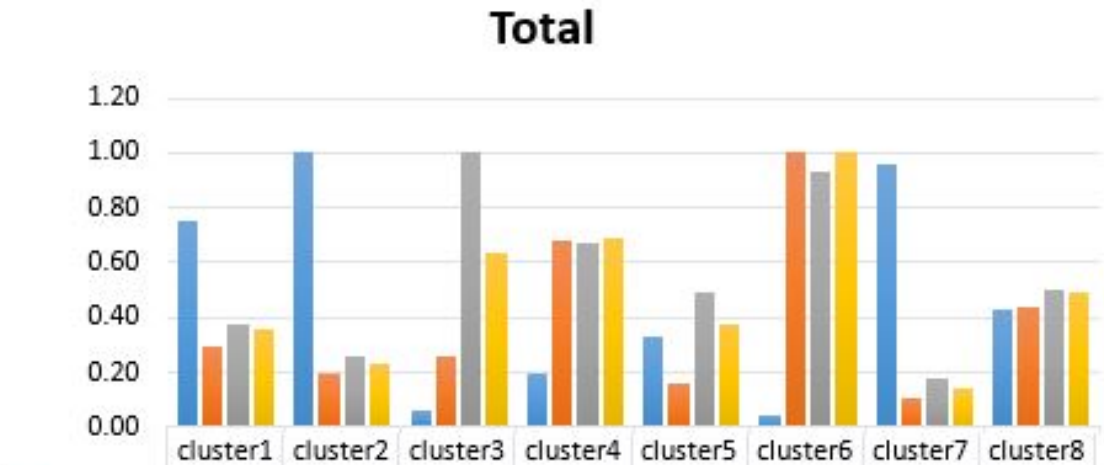
As shown in figure 5.1.4 and similar to the above way the result using SMS CDR dataset shows that cluster 3 is better SMS users than the rest of other clusters. The other all cluster which contains a different number of subscribers are approximately similar in a number of transaction and amount of money serve.

	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8
NO_CUST	4,141	5,513	358	1,060	1,806	236	5,277	2,389
SMS AVG_FRQ	57.66	54.96	131.15	80.25	95.14	61	53.97	58.6
SMS AVG_FEE	16.09	15.4	26.53	20.27	22.07	15.96	15.58	16.48

Figure 5.1.4: Clustring with SMS

5.1.1.4 *Experiment result in total usage*

After combing the total average values of whole service, the result using Total CDR dataset shown in figure 5.1.5, cluster 6 which contains 236 subscriber is better company service users than the rest of other clusters and cluster 7 which contain 5,277 subscriber is the least service users of all clusters during this three specific month.



	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8
NO_CUST	4,141	5,513	358	1,060	1,806	236	5,277	2,389
AVG_DUR	543	352.26	465	1,242	291	1817.75	186.98	805
AVG_FRQ	463.98	323.84	1248.07	843.89	611.62	1158.49	218.78	624.82
AVG_FEE	256.83	167.95	460.86	497.45	270.65	723.58	103.43	356.24

Figure 5.1.5: Clustring with Total usage

5.1.2 *Quantile Method*

After analyzing each cluster usage behavior the next step is to give rank or score for each duration, frequency and monetary value of each cluster using quantile methods. Quintiles are like percentiles, but instead of dividing the data into 100 parts, you divide it into 5 equal parts. It works well with any industry since ranges are picked from data itself and to create cut-off points from a given dataset. It distributes customers evenly and it is the recommended method to calculate DFM score. It also has been used for different business analytics and marketing insight solutions [14]. The flow of the quantile method shown in figure 5.1.6.

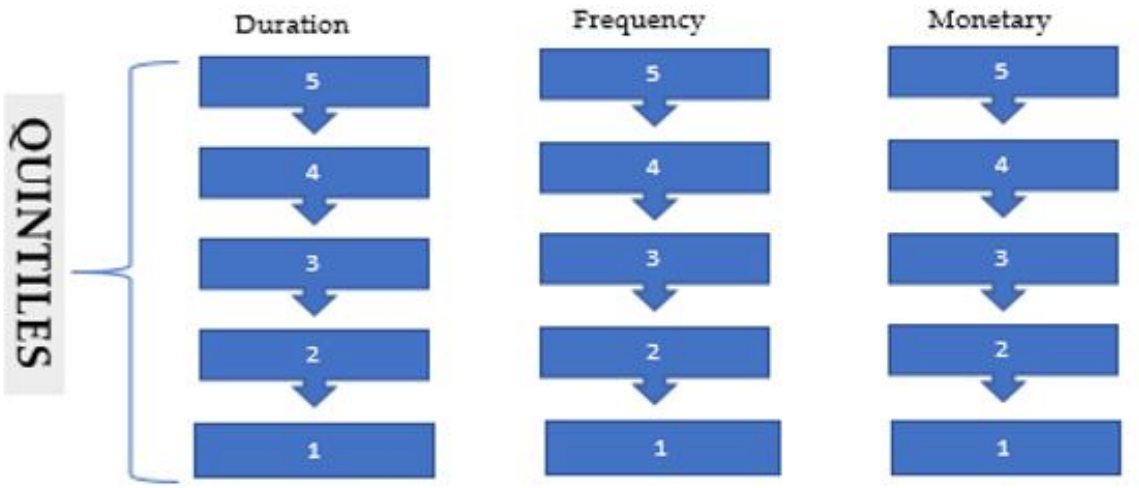


Figure 5.1.6: Quantile Methods adopted from [14]

If the data value rank ends up with 5 bands for each of the D, F, and M-values. The larger the score for each value, the better it is and the ways to do this called Concatenation. It is done by linking the three scores together. For example cluster 3 has a duration score of 2, a frequency score of 5 and a monetary score of 3. Its DFM score result is 253. By using this methodology, give a score for each cluster and the maximum value is 555 and the minimum begins by 111. Lastly, each customer cluster is placed into its corresponding segment type based on their scores [14].

5.1.3 *Applying DFM Score*

This is an important step for DFM analysis. DFM values and scores are different. Value is the actual values of D/F/M for each cluster, whereas the score is giving the number from $1-5$ based on the value using quantile methods as discussed in subsection 5.1.2. For instance, the score for cluster 2 who has a duration value of 352.26 (Figure 5.1.5), was the 2nd quantiles of score range, which is from $271-420$ minute range usage. So its duration score is 2. By repeating this calculation for every cluster, we can get the result on table 5.1.1.

Table 5.1.1: Quantile score of the three service

GROUP	NO_CUST	VOICE			DATA			SMS		Parameter for Voice	Parameter for DATA	Parameter for SMS
		FRQ_RANK	DUR_RANK	FEE_RANK	FRQ_RANK	DUR_RANK	FEE_RANK	FRQ_RANK	FEE_RANK			
cluster1	4,141	2	2	2	1	1	1	2	3	D↑F↓M↓	D↓F↑M↓	**F↓M↓**
cluster2	5,513	2	1	1	1	1	1	2	3	D↓F↑M↓	D↓F↓M↓	**F↓M↓**
cluster3	358	2	2	2	**3**	**3**	**5**	**4**	**4**	D↓F↑M↑	**D↑F↑M↑**	**F↑M↑**
cluster4	1,060	4	4	3	1	2	1	3	3	D↑F↓M↑	D↑F↓M↑	**F↓M↑**
cluster5	1,806	1	1	1	2	2	4	3	3	D↓F↓M↑	D↑F↑M↓	**F↑M↓**
cluster6	236	5	5	5	1	1	1	2	3	**D↑F↑M↑**	D↓F↑M↑	**F↓M↓**
cluster7	5277	1	1	1	1	1	1	2	3	D↓F↓M↓	D↓F↓M↑	**F↓M↓**
cluster8	2,389	3	3	3	1	1	1	2	3	D↑F↑M↓	D↑F↓M↓	**F↓M↓**

After having this score for each cluster, all it needs to do is a concatenation of each score value because valuing customers based on a single parameter is insufficient to judge. Moreover refereeing customer value on just one aspect will give you inaccurate information on your customer segmentation. That's why DFM analysis associations with three different customer attribute to rank customers.
Finally the concatenation result map with recommended segment type (customer type) as discussed in chapter 2 section 2.5 and the result is shown in Table 5.1.2 and 5.1.3.

Table 5.1.2: Concatination pattern and customer type

GROUP	NO_CUST	Duration		Frequency		Monetary		Score Concatination	Pattern	Customer Type
		AVG_DUR	SCORE	AVG_FREQ	SCORE	AVG_FEE	SCORE			
cluster6	236	1817.8	5	1158.5	5	723.6	5	555	D↑F↑M↑	best(platinium)
cluster4	1,060	1,242	4	843.89	4	497.5	4	444	D↑F↑M↓	Loyal or Valuable(Golden)
cluster8	2,389	805	3	624.82	3	356.2	3	333	D↑F↓M↑	Potential Loyalist(silver)
cluster3	358	465	2	1248.1	5	460.9	3	253	D↑F↓M↓	Recent Customers
cluster1	4,141	543	2	463.98	2	256.8	2	222	D↓F↑M↑	About To Sleep
cluster5	1,806	291	1	611.62	3	270.7	2	132	D↓F↑M↓	Frequent(Customers Needing Attention)
cluster2	5,513	352.26	1	323.84	2	168	2	122	D↓F↓M↑	spender(At Risk)
cluster7	5,277	186.98	1	218.78	1	103.4	1	111	D↓F↓M↓	uncertain(Hibernating)

Table 5.1.3: Customer type with its usage activities

GROUP	**NO_CUST**	**Pattern**	**Customer Type**	**Activities**
cluster6	236	**D↑F↑M↑**	best(platinium)	Bought more duration, buy often and spend the most!
cluster4	1,060	D↑F↑M↓	Loyal or Valuable(Golden)	Spend good money and duration with often. Responsive to promotions
cluster8	2,389	D↑F↓M↑	Potential Loyalist(silver)	Recent customers, but spent a good amount and bought more than once.
cluster3	358	D↑F↓M↓	Recent Customers	Bought or spend more duration but not often
cluster1	4,141	D↓F↑M↑	About To Sleep	Below average duration, frequency and monetary values. Will lose them if not reactivated
cluster5	1,806	D↓F↑M↓	Frequent	Above average duration, frequency and monetary values. May not have bought very recently though
cluster2	5,513	D↓F↓M↑	spender(At Risk)	Spent big money and purchased often. But long time ago. Need to bring them back
cluster7	5,277	D↓F↓M↓	uncertain(Hibernating)	low time spenders and low number of orders, low money spent

5.2 DISCUSSION

The overall objective of this study is analyzing customer call usage behaver to map with appropriate customer segmentation type using CDRs to enhance profit. So that in order to meet this objective, well known and recommended by literatures clustering algorithm has been selected and which is called the K-means algorithm. In addition to the algorithm, the study shown how the data-driven approach is useful and find relevant CDR attributes that have to play a major role in the process of segmentation. Therefore with this in mind, we will discuss the result obtained during the experiment in line with our objective and research question whether we meet or not.

To begin from the algorithm, clustering algorithms used throughout this study was k-means for classifying the types of customer which help to improve the profitability of telecom companies. K-means clustering is a good technique to segment subscribers into different groups. The main disadvantage of K-mean clustering is difficulty in finding the start / initial centroids. However by using the elbow method it has been found the optimal number of initial centroids. Then according to the above section 4.5.1.1 which means after computing SSE to identify a specific range that features a minimal decrease in average diameter, the optimal number of the cluster has been found as eight for this dataset.

This analysis is a better method to describe telecom customers' usage data. Based on the experiment can determine DFM values for each cluster. To improve effectiveness we used average DFM values for the final result of each cluster. Then according to the final values and using quintile methods, divided customers into 8 customer types. Such as Best (Platinum), Loyal or Valuable (Golden), Potential Loyalist (Silver), Recent Customers, About to Sleep, Frequent (Customers Needing Attention), spender (At Risk), and uncertain.

Making segmentation, the company can make better strategic decisions for each group. To see customer behaviors moving through the cluster, at least perform segmentation once in three months. Because high profitable customers may decline their level from high to medium or to a low level. And also some customers might increase their level from another customer type to loyal. So, analyzing customers'

usage behavior is another characteristic of mining based on historical data. It will help to stop the reduction of the customer's usage.
Usually, companies decide the profitability of the customers from there revenue. However, only taking revenue is not a good way for decision making. They have to see more other criteria. So, we used DFM analysis methods to address the above problem. DFM is the better way to describe different behaviors of customers. In telecom also only considering the revenue of the customer is not a good way. They have to be considered the duration and frequency of the customers too. Because some customers have brought higher revenue but are not calling frequently. Some customers have lesser revenue but they are calling so frequently. If they call frequently they always keep in touch with the company. So we have to concern about them.

summarize by Answering the research question

Answering the first research question, Research question one "How data-driven approach is useful than general(core) customer segmentation?" When we use customer segmentation using a data-driven approach. we can group customers according to there actual usage behavior and map easily with a customer type Such as Best (platinum), Loyal or Valuable (Golden), Potential Loyalist (silver), Recent Customers, About to Sleep, Frequent (Customers Needing Attention), spender (At Risk), and uncertain. which directs our target according to the user level to enhance the profit and must be performed at least in three months. However, when we use general(core)segnetation such as enterprise, residential , SOHO, etc does not provide information on the actual usage, provides static results.

Answering the second research question, research question two "What are the relevant attributes used to segment behavioral customer segmentation?", we found useful to represent data information as aggregated. Aggregated data resulted from transforming raw data into new attributes/variables that measure a certain goal concept, that is not yet explicit in the raw data. In the process of aggregating data how much time spent, the number of the transaction during peak and offpeak time and the amount of money spent during peak and off-peak time, and the

service type used were the main concerns. From the real CDR data, we found the following main attribute to define the customer usage property.

Table 5.2.1: Selected Attribute

Attribute	**Description**
RE_ID	CDR type ID for voice, SMS and Data
CALLING_NUMBER	Calling Number(call initiate number)
CALLED_NUMBER	called number (call destination number)
CALL_START_TIME	Call start time(the time when call start)
CALL_END_TIME	Call end time(the time when call end)
CALL_DURATION	Call duration
CALL_FEE	The actual money deducted(charge amount)
CALLED_COUNTRY	Called country(country area code of called number)
UPLOAD_TRAFFIC	Upload trafficc
DOWNLOAD_TRAFFIC	Download traffic

6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

Effective customer segmentation is critical for any company to handle their customer in better ways and providing services based on their use and value, accordingly, improve its revenue. Customer segmentation is a technique for grouping customers based on homogeneity with respect to any measurement, whether it is the customer wants, purchases, channel preferences, interest in certain product offers, or the profitability of a customer, etc. However, the dynamic growth of telecom service and their customer become challenges to manage customers using core segmentation methods. In addition, for telecom service providers like ethio telecom has performed segmenting customers using core segmentation methods such as enterprise and residential customers. such kinds of segmentation could not provide information on the actual usage behavior and that leads to not getting revenues according to their revenue forecasts or plan. Therefore, this study intends to segment customer using data-driven approaches and this approach's lead to develop successful customer segmentation that helps to increase revenue by satisfying and attracting customers need.
In this study, a machine learning-based approach for customer segmentation has been selected and an effort has been made to analyze customer segmentation using unsupervised clustering algorithms called k-means. A CDR data were collected from ethio telecom and tested. In order to determine the number of clusters to be optimal, one of the intra-cluster evaluation methods called elbow method has been used for optimal cluster evaluation. Accordingly, 8 clusters have been constructed.

In the data preparation phase, a total of 11,574,477 CDR data within three-month for 20,780 sample mobile customers were collected. This data was further pre-

processed and reduced to 9,693,296. Out of 33 CDR data attributes 10 attributes were selected with the help of domain experts and consulting related literature. Following this, the data were further reduced to 20,780 instances by aggregating and integrating based on calls made on monthly bases. Four datasets have been constructed, the three datasets were constructed by services types such as Voice, data and SMS and the fourth dataset constructed by integrating the three-service dataset. Then, the data were transformed into an appropriate ARRF file format and the clustering experiment has been done using the K-means algorithm. The clustered data were recorded for further analysis.

Eight clusters, cluster 1 to cluster 8 were formed as seen in Table 6.1.1. These clusters were analyzed in line with customer segmentation property. The eight formed clusters were ranked and labeled with segmentation types such as Best (platinum), Loyal or Valuable (Golden), Potential Loyalist (silver), Recent Customers, About to Sleep, Frequent (Customers Needing Attention), spender (At Risk), and uncertain. For instance, Cluster 6 score high DFM values and ranked as first and labeled as best (Platinum) type. Cluster 7 also score low DFM values and ranked as last and labeled as uncertain. The detail is shown in Table 5.1.2.

Table 6.1.1: The clusterd instance

Cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8
4,141	5,513	358	1,060	1,806	236	5,277	2,389

In general, by understanding the characteristics of customers and their profitability, telecom companies can make varies decisions to improve the quality of service and increase their revenue. Thus, companies need to have different customer handling strategies based on their category. High profit or platinum customers have a vital effect on the company's income because such customers use the service better than others. So, companies must provide better offers such as discounts, promotions, and gifts on their special days like birthdays, holiday anniversaries, and etc to satisfy them.

Telecom customers call to call center to get any information they need. However, to get this service sometimes they have waiting a long time over the telephone

to contact customer officers. This long queue creates customer dissatisfaction on the services. As a strategy, for platinum and loyal customer, the company should avoid long queues by assigning a special shortcode link to access the service in every short time. In addition, it needs to assign more skilled and experienced customer employees to handle them. Companies cannot ignore other customer types too. Because they contribute to the companies' income. If the company motivates them to use the service, they will join to platinum customer group in the future. Therefore, the company's needs to give attention by arranging different motivation techniques using promotions and make discounts to satisfy them. Also, the company can send some messages (text) to reminding promotions and discounts.

6.2 FUTURE WORK

This thesis work focuses on using local call usage data of mobile subscribers by assuming local call usage of the mobile subscriber are vital on revenue generation contribution and a high percentage in the number of customers from other service subscribers. As future work includes another service product such as roaming, International call, broadband, and fixed-line. In addition, by using unseen CDR attributes during this research, construct different customer segmentation to compare a user's present behavior to the same user's past behavior.

BIBLIOGRAPHY

[1] T Konstantinos and C Antonios, *Data mining tech-niques in crm: Inside customer segmentation*, 2009.

[2] C. Bounsaythip and E. Rinta-Runsala, "Overview of data mining for customer behavior modeling," *VTT Information Technology Research Report, Version*, vol. 1, pp. 1–53, 2001.

[3] D Qiasi, B Minaei-Bidgoli, and G Amooee, "Developing a model for measuring customer's loyalty and value with rfm technique and clustering algorithms," *The Journal of Mathematics and Computer Science*, vol. 4, no. 2, pp. 172–181, 2012.

[4] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[5] J. Behnamian and R. Asgari, "Bi-objective customer segmentation using data mining technique (a case study in sima-choob)," *Iranian Business Management*, vol. 7, no. 4, pp. 841–864, 2015.

[6] J. Granat, "Data mining and complex telecommunications problems modeling," *Journal of Telecommunications and Information Technology*, pp. 115–120, 2003.

[7] E. telecom, *Ethio telecom's annual revenue, ethiopia, from fiscal year 2011 to 2018*, 2011 to 2018. [Online]. Available: `https://www.statista.com/statistics/749670/ethiopia-ethio-telecom-revenue/`.

[8] E. Mattila, *Behavioral segmentation of telecommunication customers*. Dataveteskap och kommunikation, Computer Science and Communication . . ., 2008.

[9] J.-T. Wei, S.-Y. Lin, and H.-H. Wu, "A review of the application of rfm model," *African Journal of Business Management*, vol. 4, no. 19, pp. 4199–4206, 2010.

[10] A. Arora and D. R. Vohra, "Segmentation of mobile customers for improving profitability using data mining techniques," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 4, pp. 5241–5244, 2014.

[11] D. Rajagopal *et al.*, "Customer data clustering using data mining technique," *arXiv preprint arXiv:1112.2663*, 2011.

[12] S. Aheleroff, "Applying call and event detail records to customer segmentation and clv," *International Journal of Information*, vol. 3, no. 8, 2013.

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

[14] J. R. Stafford, "Rfm: A precursor to data mining," *Stafford ABSG*, 2009.

[15] G. S. Linoff and M. J. Berry, *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011.

[16] D. Gaspar, I. Coric, and M. Mabic, "Data mining in customer profitability analysis," 2015.

[17] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[18] K. Tulankar and R. Wajgi, "Clustering telecom customers using emergent self organizing maps for business profitability 1," 2012.

[19] A. Kumar, *Day marketing research-7th*, 2000.

[20] C. D. Cuschieri, "Customer profitability analysis in a local five star hotel: A case study," B.S. thesis, University of Malta, 2010.

[21] G. Maji, L. Dutta, and S. Sen, "Targeted marketing and market share analysis on pos payment data using dw and olap," in *Emerging Technologies in Data Mining and Information Security*, Springer, 2019, pp. 189–199.

[22] G. Carmichael, Y.-W. Chen, and C. Luo, "Data-driven segmentation of consumers' purchase behaviour in the retail industry," in *2018 4th International Conference on Information Management (ICIM)*, IEEE, 2018, pp. 215–219.

[23] A. Aziz, *Customer segmentation basedon behavioural data in e-marketplace*, 2017.

[24] D. Birant, "Data mining using rfm analysis," in *Knowledge-oriented applications in data mining*, IntechOpen, 2011.

[25] M. Mohammadian and I. Makhani, "Rfm-based customer segmentation as an elaborative analytical tool for enriching the creation of sales and trade marketing strategies," *International academic journal of accounting and financial management*, vol. 3, no. 6, pp. 21–35, 2016.

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

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

[28] A. Nagpal, A. Jatain, and D. Gaur, "Review based on data clustering algorithms," in *2013 IEEE Conference on Information & Communication Technologies*, IEEE, 2013, pp. 298–303.

[29] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[30] A. Kassambara, *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA, 2017, vol. 1.

[31] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2013.

[32] D. Sisodia, L. Singh, S. Sisodia, and K. Saxena, "Clustering techniques: A brief survey of different clustering algorithms," *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, vol. 1, no. 3, pp. 82–87, 2012.

[33] S. K. Kingrani, M. Levene, and D. Zhang, "Estimating the number of clusters using diversity," *Artificial Intelligence Research*, vol. 7, no. 1, pp. 15–22, 2018.

[34] I. A. Venkatkumar and S. J. K. Shardaben, "Comparative study of data mining clustering algorithms," in *2016 International Conference on Data Science and Engineering (ICDSE)*, IEEE, 2016, pp. 1–7.

[35] M. J. Zaki, W. Meira Jr, and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[36] S. Kaushik, "An introduction to clustering and different methods of clustering," *Analytics Vidhya*, vol. 3, 2016.

[37] S. Chakrabarti, E. Cox, E. Frank, R. H. Güting, J. Han, X. Jiang, M. Kamber, S. S. Lightstone, T. P. Nadeau, R. E. Neapolitan, *et al.*, *Data mining: know it all*. Morgan Kaufmann, 2008.

[38] F. M. Bianchi, A. Rizzi, A. Sadeghian, and C. Moiso, "Identifying user habits through data mining on call data records," *Engineering Applications of Artificial Intelligence*, vol. 54, pp. 49–61, 2016.

[39] A Bascacov, C Cernazanu, and M. Marcu, "Using data mining for mobile communication clustering and characterization," in *2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, IEEE, 2013, pp. 41–46.

[40] J. A. R. Rojas, M. B. Kery, S. Rosenthal, and A. Dey, "Sampling techniques to improve big data exploration," in *2017 IEEE 7th symposium on large data analysis and visualization (LDAV)*, IEEE, 2017, pp. 26–35.

[41] S. Dolnicar, B. Grün, and F. Leisch, "Increasing sample size compensates for data problems in segmentation studies," *Journal of Business Research*, vol. 69, no. 2, pp. 992–999, 2016.

[42] M. Kamel, "Data preparation for data mining," in *Encyclopedia of Data Warehousing and Mining, Second Edition*, IGI Global, 2009, pp. 538–543.

[43] D. D. Gutierrez, *Machine learning and data science: an introduction to statistical learning methods with R*. Technics Publications, 2015.

[44] R Soudagar, *Customer segmentation and strategy definition in segments*, 2012.

[45] M. A. Hall and L. A. Smith, "Feature subset selection: A correlation based filter approach," 1999.

A

APPENDIX

A.1 SCRIPTS FOR SAMPLING AND DATA COLLECTION

Script for Random Sampling

DBMS_RANDOM : Generating Random Data (Numbers, Strings and Dates)

*SELECT MSISDN FROM (SELECT * FROM service_no where msisdn like'2519%' ORDER BY DBMS_RANDOM.VALUE) WHERE rownum = 20780;*

SCRIPT FOR DATA COLLECTION

a)script for collecting CDR of sampling subscriber

*create table taj_voice as select * from TIS.voice_source_table where substr(msisdn,4,9) in(select service_no from taj_20sh);*

*create table taj_SMS as select * from TIS.sms_source_table where substr(msisdn,4,9) in(select service_no from taj_20sh);*

*create table taj_data2 as select * from TIS.data_source_table where substr(msisdn,4,9) in(select service_no from taj_20sh);*

b)script for voice peak and off-peak houre usage

peak houre voice usage

select msisdn,count(MSISDN)Freqency,round(sum(call_duration)/60,2)Actual_DUR_INM, round(sum(call_fee)/10000,2) amt_INBIRR, to_char(call_start_time,'yyyy-mm-dd')dateD, to_char(call_start_time,'hh24')dateh from taj_voice where to_char(start_time, 'mm/dd/yyyy')not in('07/07/2019', '07/14/2019', '07/21/2019', '07/28/2019') and to_char(start_time, 'hh24') > '07' and to_char(start_time, 'hh24') < '22' and acct_res_id1='1' group by

to_char(call_start_time,'hh24'),to_char(call_start_time,'yyyy-mm-dd'),msisdn;) x, (select a.price_id, a.price_name, b.acct_item_type_name, c.acct_res_name from price@LINK_SCU_CC1 a, acct_item_type@link_scu_cc1 b, acct_res@link_scu_cc1 c, rate_plan@link_scu_cc1 d, rate_plan_type@link_scu_cc1 e where a.rate_plan_id = d.rate_plan_id and a.acct_item_type_id = b.acct_item_type_id and b.acct_res_id = c.acct_res_id and d.rate_plan_type = e.rate_plan_type and b.acct_item_type_name = '3G Price') y where x.priceid = y.price_id;

off-peak houre voice usage

select msisdn,count(MSISDN)Freqency,round(sum(call_duration)/60,2)Actual_DUR_INM, round(sum(call_fee)/10000,2) amt_INBIRR, to_char(call_start_time,'yyyy-mm-dd')dateD, to_char(call_start_time,'hh24')dateh from taj_voice where to_char(start_time, 'mm/dd/yyyy') in ('07/07/2019', '07/14/2019', '07/21/2019', '07/28/2019') or to_char(start_time, 'hh24') < '07' or to_char(start_time, 'hh24') > '22' and to_char(start_time, 'mm/dd/yyyy') not in ('07/07/2019', '07/14/2019', '07/21/2019', '07/28/2019') and acct_res_id1='1' group by price_id1) x, (select a.price_id, a.price_name, b.acct_item_type_name, c.acct_res_name from price@LINK_SCU_CC1 a, acct_item_type@link_scu_cc1 b, acct_res@link_scu_cc1 c, rate_plan@link_scu_cc1 d, rate_plan_type@link_scu_cc1 e where a.rate_plan_id = d.rate_plan_id and a.acct_item_type_id = b.acct_item_type_id and b.acct_res_id = c.acct_res_id and d.rate_plan_type = e.rate_plan_type and b.acct_item_type_name = '3G Price') y where x.priceid = y.price_id;

script for data usage

select msisdn,count(a.msisdn)Frequancy,round(sum(a.data_usage)/60,2) usage_INM, round(sum(a.upload_traffic + a.download_traffic)/1024/1024,4) Data_usag_INMB, round(sum(call_fee)/10000,2) charge_FEE from taj_data a where call_fee <> 0 and a.upload_traffic is not null– and a.download_traffic <> 0 –and call_fee = 0 group by msisdn;

script for sms usage

create table SMS_total as select a.msisdn,count(a.MSISDN)SMS_Freqency, round(sum(a.call_fee)/10000,2)SMS_FEE from taj_sms a group by a.msisdn;

c)script for agragation

peak off-peak agregation

create table voice_peakoffpeak_total as select msisdn, peak_freqency,peak_dur_inm,peak_fee, offpeak_freqency,offpeak_dur_inm,offpeak_fee, sum(peak_freqency + offpeak_freqency)Total_frequency, sum(peak_dur_inm + offpeak_dur_inm)Total_DUR_INM , sum(peak_fee + offpeak_fee) total_fee from voice_peak_offpeak group by msisdn, peak_freqency, peak_dur_inm,peak_fee,offpeak_freqency,offpeak_dur_inm,offpeak_fee

voice data and sms agregation

CREATE TABLE VDS_MAJUJL_TOTAL AS SELECT a.msisdn,SUM(A.PEAK_FREQENCY + B.PEAK_FREQENCY + C.PEAK_FREQENCY)PEAK_FREQ_TOTAL, SUM(a.peak_dur_inm + B.peak_dur_inm + C.peak_dur_inm)PEAK_DUR_TOTAL, SUM(A.PEAK_FEE + B.PEAK_FEE + C.PEAK_FEE)PEAK_FEE_TOTAL, SUM(A.OFFPEAK_FREQENCY + B.OFFPEAK_FREQENCY + C.OFFPEAK_FREQENCY)OFFPEAK_FREQ_TOTAL, SUM(a.OFFpeak_dur_inm + B.OFFpeak_dur_inm + C.OFFpeak_dur_inm)OFFPEAK_DUR_TOTAL, SUM(A.OFFPEAK_FEE + B.OFFPEAK_FEE + C.OFFPEAK_FEE)OFFPEAK_FEE_TOTAL, SUM(A.TOTAL_FREQUENCY + B.TOTAL_FREQUENCY + C.TOTAL_FREQUENCY) VOI_FREQ_MAY_JUN_JUL, SUM(a.TOTAL_dur_inm + B.TOTAL_dur_inm + C.TOTAL_dur_inm)VOI_DUR_MAY_JUN_JUL, SUM(A.TOTAL_FEE + B.TOTAL_FEE + C.TOTAL_FEE)VOI_FEE_MAY_JUN_JUL, SUM(A.DATA_FREQ + B.DATA_FREQ + C.DATA_FREQ)DATA_FREQ_TOTAL, ROUND(SUM(a.DATA_USAG_INMB + B.DATA_USAG_INMB + C.DATA_USAG_INMB),2)DATA_USAG_INMB_TOTAL, SUM(A.DATA_CHARGE_FEE + B.DATA_CHARGE_FEE + C.DATA_CHARGE_FEE) DATA_FEE_TOTAL, SUM(A.sms_freq + B.sms_freq + C.sms_freq)SMS_FREQ_TOTAL, ROUND(SUM(A.sms_FEE + B.sms_FEE + C.sms_FEE),2)SMS_FEE_TOTAL FROM july_cluster A,JUN_CLUSTER_FF B,MAY_CLUSTER_FF C WHERE A.MSISDN=B.MSISDN AND B.MSISDN=C.MSISDN GROUP BY A.MSISDN

Table A.1.1: May cluster

Group	No_Cust	VOICE			DATA			SMS	
		Avg Freq	Avg. Dur.	Avg Fee	Avg Freq	Avg Mb	Avg Fee	Avg Freq	Avg Fee
cluster1	5,867	49.78	67.4	26.48	42.52	143	13.67	24.39	8.54
cluster2	2,901	224.37	346.11	138.4	31.74	92	9.93	25.6	8.97
cluster3	1,331	342.57	596.93	227.63	41.76	115	11.84	27.73	9.13
cluster4	493	543.54	879.93	294.52	39.3	105	11.68	25.89	9.32
cluster5	620	100.99	148.7	57.88	482.75	2,052	152.81	64.04	10.66
cluster6	104	838.75	1602.77	666.93	34.29	61	11.25	30.18	11.08
cluster7	5,328	129.58	190.32	77.77	27.5	82	8.65	26	8.9
cluster8	4,136	51.21	70.54	27.43	14.25	18	3.17	25.7	8.97

Table A.1.2: June cluster

Group	No_Cust	VOICE			DATA			SMS	
		Avg_Freq	Avg_Dur.	Avg_Fee	Avg_Freq	Avg_Mb	Avg_Fee	Avg_Freq	Avg_Fee
cluster1	475	475.11	744.79	271	54.07	226	21.25	12.12	3.2
cluster2	220	95.22	130.27	51.49	777.01	2,264	193.33	111.96	10.67
cluster3	5,453	93.75	130.46	52.35	25.25	91	10.06	10.58	2.87
cluster4	1,363	282.23	459.66	182.21	48.7	246	20	12.08	3.11
cluster5	106	827.37	1392.44	460.58	62.15	399	26.34	20.17	3.55
cluster6	1,215	60.09	83.14	33.04	308.1	1,911	154.46	30.6	6.24
cluster7	8,954	34.11	46.73	17.65	25.3	90	10.35	10.79	2.85
cluster8	2,994	172.73	253.71	103.26	39.28	156	16.53	11.62	3.04

Table A.1.3: July cluster

Group	No_Cust	VOICE			DATA			SMS	
		Avg_Freq	Avg_Dur	Avg_Fee	Avg_Freq	Avg_Mb	Avg_Fee	Avg_Freq	Avg_Fee
cluster1	5,842	60.45	83.9	32.21	16.2	32	5.88	24.33	4.97
cluster2	1,301	182.6	266.6	102.42	19.04	54	9.04	18.73	4.05
cluster3	356	336.23	429.47	153.99	20.47	83	13.55	18.3	3.74
cluster4	3,086	108.16	166.01	64.56	18.37	43	7.77	21.45	4.5
cluster5	244	67.19	104.77	51.72	110.66	739	108.76	27.91	5.78
cluster6	9,515	22.53	29.82	11.01	13.53	27	5.01	21.06	4.44
cluster7	337	205.63	564.54	211.68	36.11	172	29.46	19.5	4.03
cluster8	99	518.51	925.98	376.06	29.27	118	20.59	18.48	3.77

Table A.1.4: Total cluster

Group	No_Cust	VOICE			DATA			SMS	
		Avg Freq	Avg. Dur.	Avg Fee	Avg Freq	Avg Mb	Avg Fee	Avg Freq	Avg Fee
cluster1	4,141	360.49	543.2	217.12	84.27	292	34.35	57.66	16.09
cluster2	5,513	245.69	352.26	140.14	59.83	170	22.68	54.96	15.4
cluster3	358	309.33	465.49	186.21	895.02	4,283	267.14	131.15	22.53
cluster4	1,060	708.55	1242.14	413.53	108.59	417	40.16	80.25	20.27
cluster5	1,806	205.51	291.09	116.14	374.4	1,670	147.15	95.14	22.07
cluster6	236	1043.02	1817.75	679.25	95.14	259	39.01	61	15.96
cluster7	5,277	135.91	186.98	72.74	64.88	185	25.5	53.97	15.58
cluster8	2,389	500.89	804.84	309.34	104.4	386	41.41	58.6	16.48