Customer Segmentation for Value and Retention Using Data Mining: In the Case of Ethio-Telecom Mobile Service

BY: GIRUM AFEWORK

ADVISOR: 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

September 21, 2021

Declaration.

I, Girum Afework, declare that this thesis document and the work presented in it are my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also adhere to all academic honesty and integrity

Girum Afework

Name Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

This is to certify that the thesis prepared by Girum Afework, entitled *Customer Segmentation for Value and Retention Using Data Mining: In the Case of Ethio-Telecom Mobile Service* and submitted in partial fulfillment of the requirements for the degree of Master of Science Telecommunication Engineering complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by:

Advisor Ephrem Teshale (PhD) Signature ______________ Date ______________.

Evaluator ____________________ Signature ______________ Date ______________.

Evaluator ____________________ Signature ______________ Date ______________.

____

Dean, School of Electrical and Computer Engineering

ABSTRACT

Market competition is becoming intense among telecom service providers around the world. To gain a competitive advantage in the industry, service providers must address and meet their customers' needs and demands. To survive in today's competitive market, companies must analyze and interpret their customers' usage behavior, as well as plan related market strategies to retain customers, be profitable, and build long-term relationships.

One of the most effective ways to engage with each customer is through segmentation. Customer segmentation can help organizations identify more effective marketing strategies for each segment by leveraging the power of data mining clustering technology.

In this paper, the unsupervised clustering technique with k-means algorithm is applied and customer detail record (CDR) and customer information data are used to segment Ethio-telecom customers for the purpose of retain existing customers and increase customer value by treating each customer segment according to their usage behavior.

The collected data is cleaned and preprocessed in an Oracle database and then the aggregated data is used to calculate the optimal cluster number using the k-means and elbow methods. Based on the selected attributes the dataset is segmented into five groups by Weka knowledge discovery tool.

Each cluster segment is scored and mapped with the type of customer segmentation based on three-month average usage data, frequency, longevity, and service interruption time. Clusters 2 and 4 account for 5% of the total customer size but cover 72% of the company revenue, whereas Clusters 1 and 5 account for 76% of the total customer size but contribute significantly less than the others. Finally, based on the analysis result, a marketing strategy for each segment is proposed.

Keywords: Data Mining, Clustering, CDR, elbow method, k-means

Acknowledgement

I would like to thank my advisor, Dr. Ephrem Teshale, for your constant encouragement and support. My committee members, Dr. Surafel Lemma and Dr. Sosina Mengistu. Your suggestions and feedback have been invaluable to me.

My heartfelt gratitude goes to my wife Seble Mesfin and our parents for their unwavering support and encouragement over the years, as well as Ethio Telecom for providing the opportunity and covering all expenses. Thank you very much.

Finally, I'd like to express my gratitude to the Lord Jesus. Thank you, Almighty God, for your unending love and mercy.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CDR	Call Detail Record
SMS	Short Message Service
CLV	Customer lifetime value
SOHO	small office/home office
SME	Small and Medium Enterprise
CRM	Customer Relationship Management
RFM	Recency, frequency, monetary value
ROI	Return of investment
SOM	Self-organizing map
ML	Machine Learning
EM	Expected Maximization
CLARA	Clustering for Large Application
CLARANS	Clustering for Large Applications based on Randomized Search
CSV	Comma Separated Values
WSS	Within-Cluster Sum of Square
KDD	Knowledge Discovery in Databases
BIRCH	Balance Iterative Reducing Clustering using Hierarchies
PAM	Partitioning Around Medoids
FCM	Fuzzy C-Means

1 INTRODUCTION

Customer segmentation is the process of grouping customers with similar characteristics, behaviors, and values. Customer retention is one of the primary goals of customer segmentation, as it aids in increasing loyalty and avoiding customer churn. Every company must understand their customers' behavior and implement various marketing strategies to retain and satisfy the demand of each segmented group in order to be successful [1].

Customers are the foundation of every business organization's survival and retaining customers is one of the most critical challenges to deal with. It is critical to identify newly emerging trends and customer behavior in order to adapt and change the company's business processes.
The main objective of every business organization is to gain a better business advantage in the competitive market and only providing quality service cannot guarantee its success. Customer satisfaction is one of the metrics that customers to be loyal for every business organization, so it needs to be investigated correctly by the service providers [2].

Telecommunications sectors have evolved rapidly over time. Because of mobile technology, the industry has emerged as the world's fastest growing market and generate huge volume of Customer related data so organizing, analyzing, and interpreting this data by different Scientific approach and discover a unique pattern of the data, give clear view and insights to the service providers about the behavioral patterns of existing clients, which may result in the identification of new prospective customers [3]. In today's competitive business environment, companies that retain their valuable customers spend less to generate sales than companies that acquire new customers. Therefore, customer retention should be a key goal of any business strategy [4].

When there is strong competition in a market, it becomes increasingly difficult to create the same marketing strategy for everyone. So, service providers should tailor their products and services to the needs of their customers.

Nowadays, telecom service providers are facing a variety of challenges, including rapidly changing customer behavior patterns, financial and technological challenges. Identifying emerging trends and translating them into actionable strategies is critical for long-term smooth customer relationship and sustaining profitability.

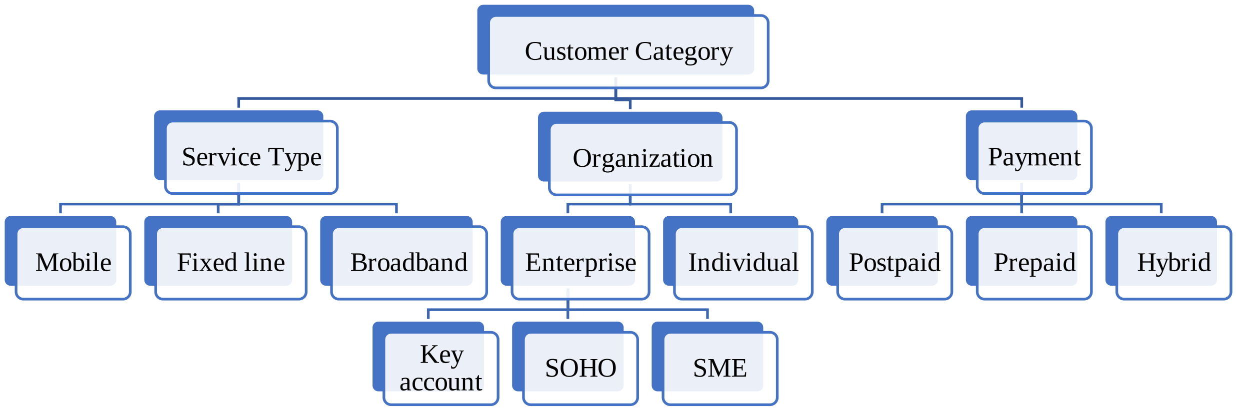
In this paper CDR, customer payment settlement information that affect the service continuity of a customer, and the customer longevity as a company's subscriber data are used as an input for clustering customer with K-means unsupervised learning techniques in Weka data mining tool.

1.1 STATEMENT OF THE PROBLEM

Ethio-telecom is an incumbent service provider that offers major mobile, fixed, and broadband telecom services throughout the country without competition; however, under the new law, additional service providers will enter the market, and fierce market competition is expected soon [5]. Currently the company has around 54.9M+ mobile service subscribers, including data and internet users, 374K+ fixed broad band subscribers, and 912K+ fixed service subscribers and 56.2+ million total customers as of July 2021 [6].

Based on the number of employees in a company, Ethio-telecom segment its enterprise customers into three categories: If a firm has more than 50 employees, it is classified as a key account; if it has 5 to 49 employees, it is classified as a SOHO; and if it has less than 5 employees, it is classified as a SME [7]. Other than that, As shown in Table 1-1 The company segment its customers based on generic names and market survey that is subjective in nature and limited to marketing department expertise. Segmenting its customers based on real customer characteristics, usage and behavior help the company better understanding of customers to support the identification of new marketing opportunities, design of customized product offering strategies to existing customers according to each segment identified wants and needs, to develop special marketing strategy to address the segmented customers other than mass marketing campaign and service package development [5].

Table 1-1 Ethio-Telecom customer category



There are some papers available on customer segmentation, and their work focuses on how to increase the profitability of the company in the era when the company was the sole service provider, but market competition is on the horizon, and the real issue could be how to retain existing customers, sustain in the business, and gain a competitive advantage over other service providers.

Customer retention refers to a company's ability to keep its customers for an extended period of time in order to continue using its services rather than switching to competitors. To retain customers and compete with other service providers, it is wise to know and identify potential and loyal customers, discover, and enhance customer value through personalized products and services, which helps to increase company profit and retain valuable customers while decreasing customer churn [8] [9].

A large volume of customer information and usage data is stored in the company's large depository of customer relationship management (CRM) and billing solution databases. Based on this data, the company can use various algorithms and techniques to segment its customers by interpreting the customers' hidden patterns and behavior.

We used data mining clustering techniques to learn customer data for this study, which will help us identify different types of customer groups and prepare different marketing strategies based on each segment's hidden characteristics, unique pattern, and behavior, with the goal of increasing customer value and retaining existing customers.

1.2 OBJECTIVES

1.2.1 General Objective

The main objective of this thesis is to analyze mobile customers usage and information data and to perform customer segmentation for customer value and retention using data mining clustering techniques.

1.2.2 Specific Objectives

The general objective is subdivided into the specific objectives listed below.

- To explore and select the right data mining tool and technique for customer segmentation.
- To cluster the customers based on the customer data and align with customer model.
- To analyze the cluster group behavior and propose a marketing approach for each segment.

1.3 SCOPE OF THE STUDY

The scope of this thesis is limited to a sample of 30000 Ethio-telecom mobile service users and the collection of three-month CDR and customer information from billing and customer relationship management sections with the goal of identifying different customer types and proposing different marketing solutions to retain each customer group.

1.4 SIGNIFICANCE OF THE STUDY

The following benefits may be gained by reading this thesis paper.

- The research findings can be used as a reference point for service providers to optimize their marketing decisions.

- It helps practitioners to use this work as baseline and improve the segmentation process by adding additional dimensions.
- It helps operators in investigating their existing data to find valuable information using data mining techniques.

1.5 LITERATURE REVIEW

The telecommunications industry provides a wide range of services to a large number of customers in the world, but the growing number of service providers creates fierce market competition in the industry, and this circumstance has influenced customers to switch to other competitors for a variety of reasons. Researchers have been working for decades to develop models and techniques for customer segmentation based on data mining and machine learning approaches. This review focuses on studies that are related to this research and are primarily concerned with customer segmentation.

Reference [1] conducted a comparison study on the eight most commonly used algorithms for predicting customer churn. According to the study, the selected models are evaluated on a public dataset of customers, and ensemble – based learning methods are suggested as a reference point. Customers' diverse behaviors force telecom industries to differentiate their strategies in order to meet their needs, and data-driven approaches can assist them in gaining perspectives from the industry's massive amount of data [3]. They use a behavior and beneficial segmentation method with usage-based features to segment customers' data using the K-Means Clustering method, and they believe it is better to use a two-dimensional approach rather than a merged one, and they recommend some marketing strategies for each segment.

In data mining technology, there are numerous clustering methods, including fuzzy clustering, system clustering, dynamic clustering, and K-means clustering. The K-means identification, on the other hand, is the most frequently applied method. Reference [10] used K-means clustering method to segment mobile customers based on their call detail records and analyzes their behaviors. The researcher proposes the model for China mobile marketing and implement the model then compares marketing effectiveness before and after using the clustering and proves that new promoting strategy has visible effects.

According to [11], Data mining can assist companies to identify meaningful trends, habits, and patterns in their customer, product, or service, resulting in improved relationships with customers and a reduction in operational risks. The study proposes a new procedure for extracting meaningful rules by combining quantitative values of RFM attributes and the K-means algorithm into rough set theory, arguing that it can effectively improve the drawbacks of a single data mining technique.

There are numerous challenges in the telecom industry, including a competitive market, high churn rates, and massive data collection. These are some of the reasons why data mining techniques are used in telecommunications companies to manage customer relationships [12]. Customer segmentation is defined as the process of identifying groups with common characteristics, with the main goal of understanding the customer structure and gaining customer insight to enable the design and development of appropriate marketing campaigns and pricing strategies.

Reference [2] explains that Customers are an integral part of an organization. If they didn't exist, there would be no business. That is why it is critical for any company to cultivate long-term customers. The key to success in any industry is a well-defined customer service strategy that builds close relationships with customers and uses those relationships to make it easier for customers to do business with you rather than your competitors. Clustering can identify previously unknown customer groups and define groups based on the characteristics shared by the members. Once such groupings are discovered, they can be profiled, and each group labelled with a name.

As per [10], The global economy and telecom market are increasingly intertwined, making it more difficult for domestic mobile communications providers to compete. The rapid growth of mobile users, the diversification of telecom companies' services, and the increasing competition in the industry have all raised the quality standards that customers demand, so customer acquisition and retention techniques are critical to telecom service providers' survival.

Most studies use CDR data to segment customers in order to understand customer behavior, but this study includes additional attributes such as customer payment settlement and service interruption history, as well as customer lifetime to understand their relationship with the service providers.

1.6 THESIS ORGANIZATION

This paper is organized in six chapters. In the introduction section, the purpose of the paper is stated. The second chapter explores into customer segmentation, including its definition, purpose, significance, and various types. Machine learning and its technology as it relates to customer segmentation are discussed in Chapter three. Chapter four contains an experiment analysis that demonstrates the entire process of the experiment used in this project. The experiment's findings are detailed in Chapter Five. Finally, in chapter six, we will look at the conclusion and discussion.

2 CUSTOMER SEGMENTATION

Customer segmentation, according to [2] [13], is the process of categorizing customers into distinct, meaningful, and homogeneous segments based on various attributes and common characteristics using customer data. These may include, among other things, customer usage, demographic, geographic, psychographic, and technological data. Furthermore, it is a business strategy that helps businesses to identify and gain a competitive advantage from their customers.

2.1 PURPOSE AND IMPORTANCE OF CUSTOMER SEGMENTATION.

The main purpose of customer segmentation is to identify various customer groups and create a channel to reach out to each of them in the most cost-effective and preferred way possible in order to gain a competitive advantage, enhance customer service, and improve market performance [14]. There are also several reasons for customer segmentation but here are most common in telecom and other industries:

- To Retain customer and value maximization
- To improve our understanding of current and potential customers
- To address business challenges and identify new opportunities
- To tailor products, services, and customer experiences
- To improve customer loyalty
- To improve product development process
- To optimize campaign performance
- To identify the most valuable customers.
- To improve Business Focus

Without a thorough grasp of how a company's best current customers are categorized, a company frequently lacks the marketing mix required to effectively allocate and spend its limited people and capital resources. The goal of segmentation is to decide how to relate to customers in each segment to maximize the value of each customer and to understand and to retain to the business. retaining customers should be a key aspect of any business organization.

2.2 ADVANTAGES OF SEGMENTATION

Segmentation has several advantages. At its most fundamental level, segmentation forces firms to evaluate where they are today and where they want to be in the future. As a result, companies are being compelled to analyze what they excel at in compared to competitors, as well as make an effort to understand what customers want. Segmentation offers for important fresh perspectives and insights, as well as the ability to review and reconsider their positions [15].

Organizations can utilize segmentation to make better use of their marketing resources, get a competitive advantage over competitors, and, most importantly, establish a better understanding of their customers' requirements and needs [16].

Below are some advantages of segmentation.

- **Market Focus strategy**

It is easier for a company to focus on which segments they want to target with specific services and campaigns once specific market segments have been identified. The market competitiveness of a company grows when a company focuses on certain segments, it ensures that the proper segment is targeted with the right product, resulting in the best ROI.

- **Customer retention**

Customers who require special attention, those who often churn, and those with the highest potential value can all be identified through customer segmentation. It can also assist in the development of targeted initiatives that attract customers' focus and result in good, high-value interactions.

- **Competitive advantage**

Because of the high level of competition in the telecom industry, a mix of market segmentation approaches might help to assure sources of income. Companies can acquire insights into customers' behavioral tendencies in different geographical locations by combining geographical segmentation and behavioral segmentation, for example.

- **Customer Satisfaction**

Companies can serve a variety of customer needs by offering different bundles and incentives when they have defined segments. Depending on the needs, wants, and characteristics of each group, different types and promotional and campaign activities will be used. Customer-specific product offerings and communication provide a more personalized connection, resulting in increased customer engagement and satisfaction.

- **Improve distribution strategy**

Knowing where and when customers shop can help inform product distribution plans, such as the kind of products supplied at specific locations.

As a company, it is critical to know in advance how any particular marketing action is going to influence the customer since the goal is to maximize the value of each customer to the company.

2.3 BASIC TYPES OF CUSTOMER SEGMENTATION.

Because market conditions are constantly changing, firms evaluate customer segmentation solutions' efficacy by validating them using segmentation criteria. In today's telecommunications sector, there are four different kinds of segmentation strategy.

2.3.1 Demographic segmentation

Demographic segmentation is one of the most popular types of segmentation. It refers to segmenting customers based on observed variations in persons. Age, sex, marital status, family size, occupation, education level, income, race, nationality, and religion are some of these characteristics. This sort of segmentation has the advantage of being relatively straightforward to get and affordable. Demographics are frequently coupled with other segmentation techniques to develop the target segment that are most likely to acquire or use a company's products or services.

However, as a result of market globalization and advancements in the fields of information technology and e-Commerce, demographic variables play a lesser significance in customer segmentation nowadays [17] [3] .

2.3.2 Behavioral Segmentation

It is the most common segmentation type and mainly focus on the purchasing habits of the customers like customer loyalty, engagement level, and service usage. In behavioral segmentation, which is the most effective, customers' attitudes, perspectives, interests, and other connected traits are considered [18]. You must be aware of your customers' actions in order to perform behavioral segmentation. These activities could be related to how a client interacts with your brand or to activities that occur outside it. Focusing on purchasing habits and a few other behavioral indicators can help segment customers more effectively. Product ownership and usage statistics, which are normally maintained and accessible in the organization's databases, are among the required data. Customers are classified based on their observed behavior and usage habits.

Companies also employ behavioral segmentation to identify future consumer leads and prospects in the market who are more likely to buy and use their product and services.

Companies divide buyers into groups based on their knowledge of, attitude toward, use of, or response to a product in behavioral segmentation. The classification of customers into heavy, medium, and light users is a frequent use of behavioral segmentation. Heavy users are a company's essential customers, and they should always be taken seriously. Marketers frequently employ tactics to boost product consumption among light users as well as non-users. Creating market divisions based on distinct customer advantages is one of the most effective uses of behavioral segmentation [19].

2.3.3 Geographic segmentation

Geographic customer segmentation refers to the separation of customers based on their location. In this scenario, customer segments might be defined as a country, region, or city, as well as specific properties in certain towns. Both large and small organizations employ this kind of customer segmentation.

Companies segment their customers by geographic area for a variety of reasons, including evaluating differing demands in different regions and tailoring communications to regional variances in language, interests, conventions, and other characteristics. It is also extremely beneficial in the field of digital marketing.

Geographic segmentation is utilized in the case of small enterprises to target certain customers in a given area. It assists small businesses in organizing their budgets and preventing wasteful spending as well as unfavorable consumer perceptions.

Geographic segmentation has the advantage of allowing each customer to be readily assigned to a geographic unit. As a result, targeting communication messages and selecting communication channels to reach the targeted geographic groups is simple. The main disadvantage is that just because people live in the same country or region does not mean they share similar features that are important to businesses [15].

2.3.4 Psychographic segmentation

Psychographic segmentation deals with characteristics that are more mental and emotional customer tastes like personality traits, interests, beliefs, values, attitudes and lifestyles, Habits, Hobbies, Opinions, Personality, Social status. Psychographics is the study of how to better understand customers by combining psychology and demographic data. Buyers are classified into distinct categories based on personality qualities, lifestyle, or values in psychographic segmentation. Within the same demographic category, people might have divergent psychographic characteristics [20]. Companies need a little bit deeper analysis to segment its customers based on psychographic pattens. Interviewing existing clients and Observing customer data may help to understand its customers.

3 DATA MINING

The technique of mining large volumes of data for discovering patterns and information is known as data mining. It, also known as database knowledge discovery, has grown as an interdisciplinary field in computer science. Databases, data warehouses, the internet, and other data repositories are examples of data sources. To examine enormous digital collections, known as data sets, the field integrates statistics and artificial intelligence technologies with database management system. Business, engineering, scientific research, and government security all employ data mining [21] .

3.1 DATA MINING IN TELECOM

Because of the vast number of phone conversations, the telecom industry creates a lot of data. As a result, the adoption of data mining techniques in the business benefits both the operator and the client by improving services and giving crucial customer information [22].

In the Telecom industry, data mining techniques have been successful in detecting hidden characteristics, patterns, and unexpected trends in very large datasets created in its network and customer databases. Data mining is the process of analyzing and systematically extracting crucial and previously unknown relationships and associations between data in order to gain a better knowledge of the problem. In order to forecast emerging patterns, data mining applies well-established techniques for analyzing current and historical data [22]. In general, it involves evaluating data already stored in databases to solve problems.

Telecom companies can use data mining techniques to improve their competitive advantage in the market and increase consumer value. Association rules, classification, clustering, rule induction, genetic algorithm, decision tree, and neural network are some of the key techniques applied in Data Mining [21] [8].

3.2 DATA MINING AND KOWLEDGE DISCOVERY IN DATABASE

The notion of evaluating huge volumes of data and extracting relevant information leads to the knowledge discovery process, which extracts meaningful patterns, rules, and models from raw data and makes discovered patterns understandable. KDD is a analytical technique for extracting meaningful and practical 'knowledge' from database data. Some academics consider data mining to be a synonym for knowledge discovery from data, while others consider it to be a critical step in the knowledge discovery process [21].

The essential steps involved in the Data discovery process are outlined here, according to [21] and others.

- **Data Cleaning**

It is the process of searching for missing data and removing noisy, irregular, redundant, and low-quality data from a data set to improve the data's reliability and effectiveness.

- **Data Integration**

Multiple data sources can be merged and organized into meaningful sets based on their availability, significance, and quality. Cleaning and integrating data is a preprocessing phase that results are being stored in a database.

- **Data Selection**

It is the process of identifying and extracting useful data from a database that is relevant to the analysis activity.

- **Data Transformation**

In this step data are prepares to be enter into the data mining algorithms. Hence, the data are transformed and consolidated into forms appropriate for mining by performing summary analysis as well as aggregatory activity.

- **Data Mining**

It is an essential process where algorithms are applied to extract meaningful data patterns from the transformed data.

- **Pattern Evaluation**

It is the process of identifying interesting patterns that represent knowledge based on patterns that are easily understood, valid, possibly valuable, new, and if it validates a hypothesis that the user was looking to confirm.

- **Knowledge Representation**

This is the final step in the knowledge discovery process, and it requires the knowledge extracted from the pattern evaluation step to be applied to a specific domain using visualization and knowledge representation techniques and formats such as trees, tables, graphs, reports, and matrices.

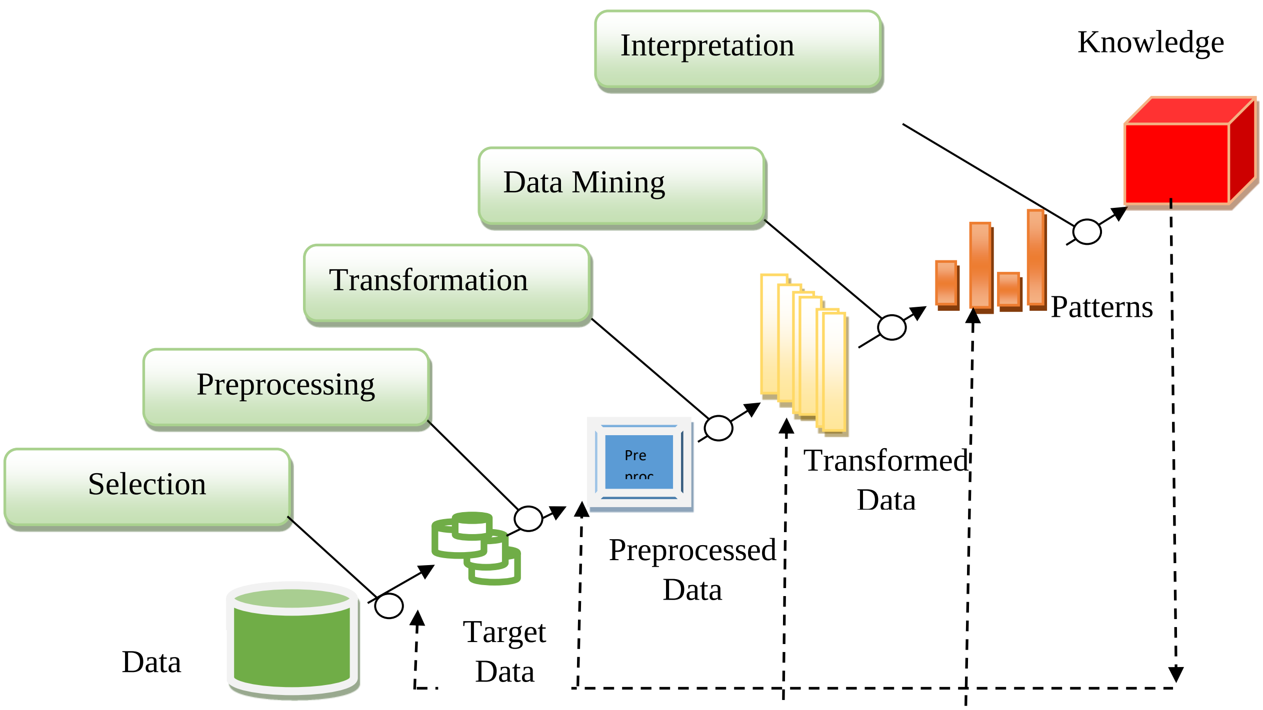


Figure 3-1 Basic steps of the KDD Process [23].

3.3 LEARNING APPROACHES

In machine learning, there are two primary learning approaches: supervised and unsupervised learning.

3.3.1 Supervised learning

Supervised learning is a technique that uses labeled datasets to train algorithms that can reliably classify data or predict events. Using labeled inputs and outputs, the model may assess its correctness and learn over time. This method can be utilized in situations where we know both the input and output values [24]. The goal of supervised learning is to enable machine learning algorithms to predict the output class for each incoming data sample whose classification is unknown.

In data mining, supervised learning mainly used to solve two types of problems: classification and regression.

- **Classification**

It is a supervised learning problem that involves in predicting a class label. Different algorithms are used to classify test data accurately. It identifies certain entities in the dataset and tries to derive conclusions about how they should be labeled or described.

- **Regression**

To explore the relationship between dependent and independent variables, regression employs an algorithm to predict a real value for each item in numerical label. The magnitude of the difference between the real-valued label predicted and the true or accurate one is used to calculate error in regression, rather than the equality or inequality of these two values.

3.3.2 Unsupervised learning

Unsupervised learning is a sort of machine learning in which models are trained on unlabeled data and then allowed to act on it without being supervised. The Algorithm receives just unlabeled training data and find the hidden patterns and insights from the given data then generates predictions for all points that are hidden. The algorithms can find hidden patterns in data without relying on outside intervention. It can be difficult to quantitatively evaluate the performance of an algorithm because no labeled data is usually available.

Unsupervised Learning can be classified in **Clustering** and **Associations** problems.

- **Clustering**

Clustering is a data mining technique that groups unlabeled data into homogeneous groups based on similarities or differences. Clustering is frequently used to analyze very large data sets.

- **Association**

An association rule is a type of unsupervised learning method for determining associations between variables in a large database. It identifies the group of elements in the data set that occur together. [25].

For this thesis work a clustering unsupervised learning approach is considered. We prefer to use the Weka knowledge discovery tool because it is simple, has a user-friendly interface, and provides a variety of tools for data preparation, classification, regression, clustering, association rules mining, and visualization [26].

3.4 CLUSTER ANALYSIS

Clustering is a type of machine learning unsupervised learning strategy. It is the technique of grouping a set of data objects into clusters with similar attributes between objects within a cluster while being highly dissimilar from those in other clusters. The conclusions were drawn from data sets that do not have a labelled output variable in the unsupervised learning approach. The data is partitioned using this technique. As a result, clustering can lead to the discovery of previously unknown patterns in data [21]. It's a type of exploratory data analysis that allows us to look at multivariate data sets.

3.4.1 Basic Clustering Methods

Clustering is a technique for identifying groups of similar objects in a data set, which can be used to characterize the distribution of the data. currently there are four types of clustering algorithms widely used: partitioning- based, hierarchical-based, density-based, and grid-based methods [27] [28].

3.4.1.1 Partitioning Methods

Partition clustering, also known as the centroid-based method, is a type of clustering that divides the data into non-hierarchical groups. This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. It optimizes cluster quality by iteratively performing on any objective function. Each Cluster has minimum one data point. Also no data point can be present in more than one cluster [21] [29]. There are many algorithms that come under partitioning method some of the popular ones are K- K-means, K-modes, K-medoids, PAM, CLARA, FCM, and CLARANS.

3.4.1.2 Hierarchical methods

An algorithm that creates a hierarchy of clusters is referred to as a hierarchical clustering approach. This algorithm starts with each data point being assigned to its own cluster. Then the two clusters that are closest to each other are combined into one. When there is just one cluster left, the process ends. Based on how the hierarchical division is created, a hierarchical technique can be categorized as either agglomerative or divisive. Hierarchical clustering algorithms are: BIRCH, CURE, ROCK, Chameleon and Echidna [29].

3.4.1.3 Density-based methods

Density-based clustering algorithms consider clusters as dense regions of objects in the data space and clusters are separated by regions of low density. These algorithms assign each object a density value based on the number of items that it shares a radius with. A dense item is defined as one that has a density greater than a certain threshold and is first generated as a cluster. If two clusters have a dense neighbor in common, they are merged. Density based clustering algorithms are: DBSCAN, OPTICS, DENCLUE, DBCLASD [27].

3.4.1.4 Model-based methods

The model-based clustering method seeks to maximize the correspondence between the given data and a predefined mathematical model. The assumption is that the data was generated by a model, and the attempt is made to recover the original model from the data. The recovered model establishes clusters and assigns objects to it. This clustering technique leads to a way for automatically determining the number of clusters using standard statistics procedure, resulting in a robust clustering method.

The two major model-based clustering approaches are statistical and neural network approaches. And algorithms include Self-organizing map (SOM), EM (Expected Maximization), COBWEB, and CLASSIT.

3.4.1.5 Grid-based methods

Grid-based clustering techniques start by creating a consistent grid mesh over the problem space domain. For clustering, the original data space is transformed into a multi-resolution grid structure with a defined size. Clustering algorithms include: STING, CLIQUE, Wave-cluster and Opti Grid [29].

Figure 3-2 shows the total clustering techniques mentioned above and its algorithms.

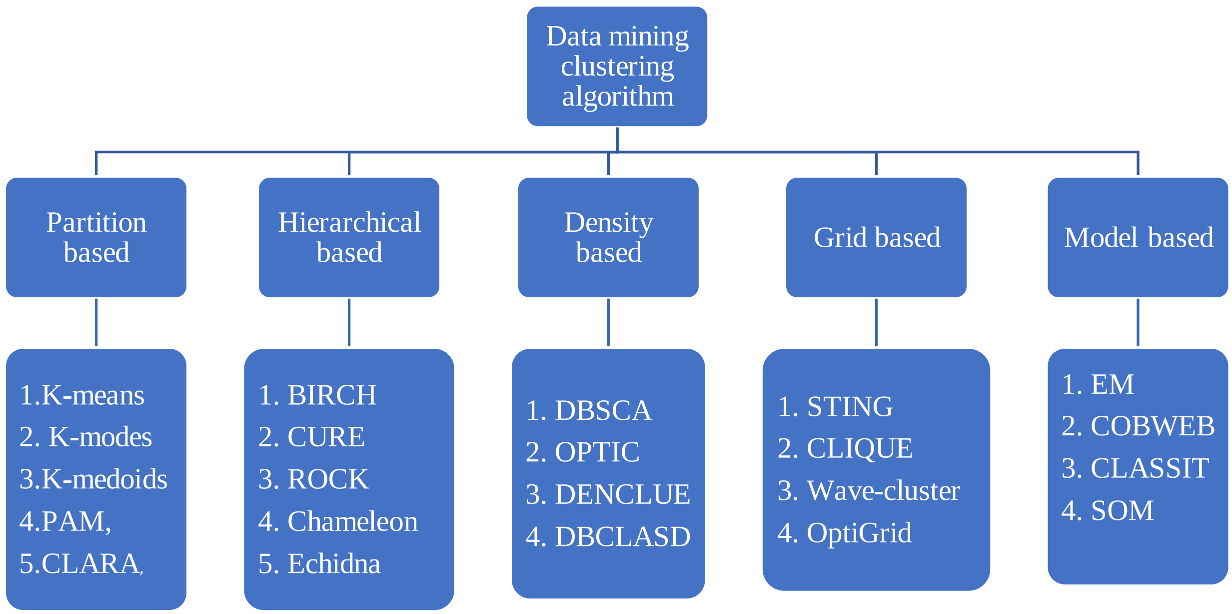


Figure 3-2 Clustering Taxonomy [29]

3.5 K-MEANS ALGORITHM

In the telecom industry, a variety of data mining algorithms can be used to process the massive amounts of data generated in their network databases (k-means, Random Forest, logistic regression, etc.). Nonetheless, the k-means algorithm is the most efficient and widely used for data processing [22]. It is used to solve clustering problems by forming clusters iteratively. We must define the number of Clusters at the start of the K-means algorithm. Then, at random, specify the points as cluster centers, and for each point, calculate the centroid-data point distance. All instances are assigned to the cluster center closest to them based on this distance. The centroid is the point from which the distance between each data point is calculated. The distance can be calculated using several methods, such as Manhattan distance calculation, Euclidean distance calculation, or cosine similarity calculation, and so on. After all the data points have been placed, the K centroids are calculated again. The mean of all points in a cluster is equal to the centroid of that cluster. Then, using the centroid-data point distance, all data points are reassigned to clusters with respect to new centroids. Iteration continues until the same points are assigned to each cluster in repetitive cycles, at which point the cluster centers have stabilized and will remain the same [28] [29].

The k-means algorithm works in the following way [28]:

1. Determine the number of clusters K.

2. Choose the same number of data points as the centroids.

3. Determine the distance of a point from each centroid. Assign the point to the cluster with the shortest centroid-point distance.

4. Repeat step 3 for all points, then compute the mean of all points in a cluster and assign it as the cluster's new centroid.

5. Continue iterating until the centroids do not change.

The objective function for k-means algorithm is sum of square distance. i.e.:

$$E = \sum_{i=1}^{i=k} \sum_{j=1}^{ni} \left\| x_{ij} - m_i \right\|^2$$

x_{ij} is the j^{th} data point of i^{th} cluster.

m_i is the center of i^{th} cluster.

n_i is the number of data points of i^{th} cluster.

4 EXPERIMENT ANALYSIS

4.1 PREPARATION AND COLLECTION OF DATA.

The First step for this experimentation was prepare the working environment and select the type of data that are being needed for this thesis. The following equipment's and software's are deployed and used.

1. Windows server 2012 r2 version.
2. Oracle 12c database.
3. 2TB Mounted Storge.
4. Oracle SQL Developer Version 20.4.1.
5. Weka version 3.9.3.

The working environment is set up at Arada Ethio-telecom data center. First, Windows Server 2012 R2 version on the 2 TB disk and 8 GB RAM server installed, then configured the Oracle 12c database which is compatible with the Windows Server 2012 R2 platform and installed Oracle SQL Developer to interface with the database and install Weka Knowledge discovery tool version 3.9.3 for the analysis and finally prepared the network environment to be used remotely.

Figure 4-1 shows the system process flow we follow for this thesis

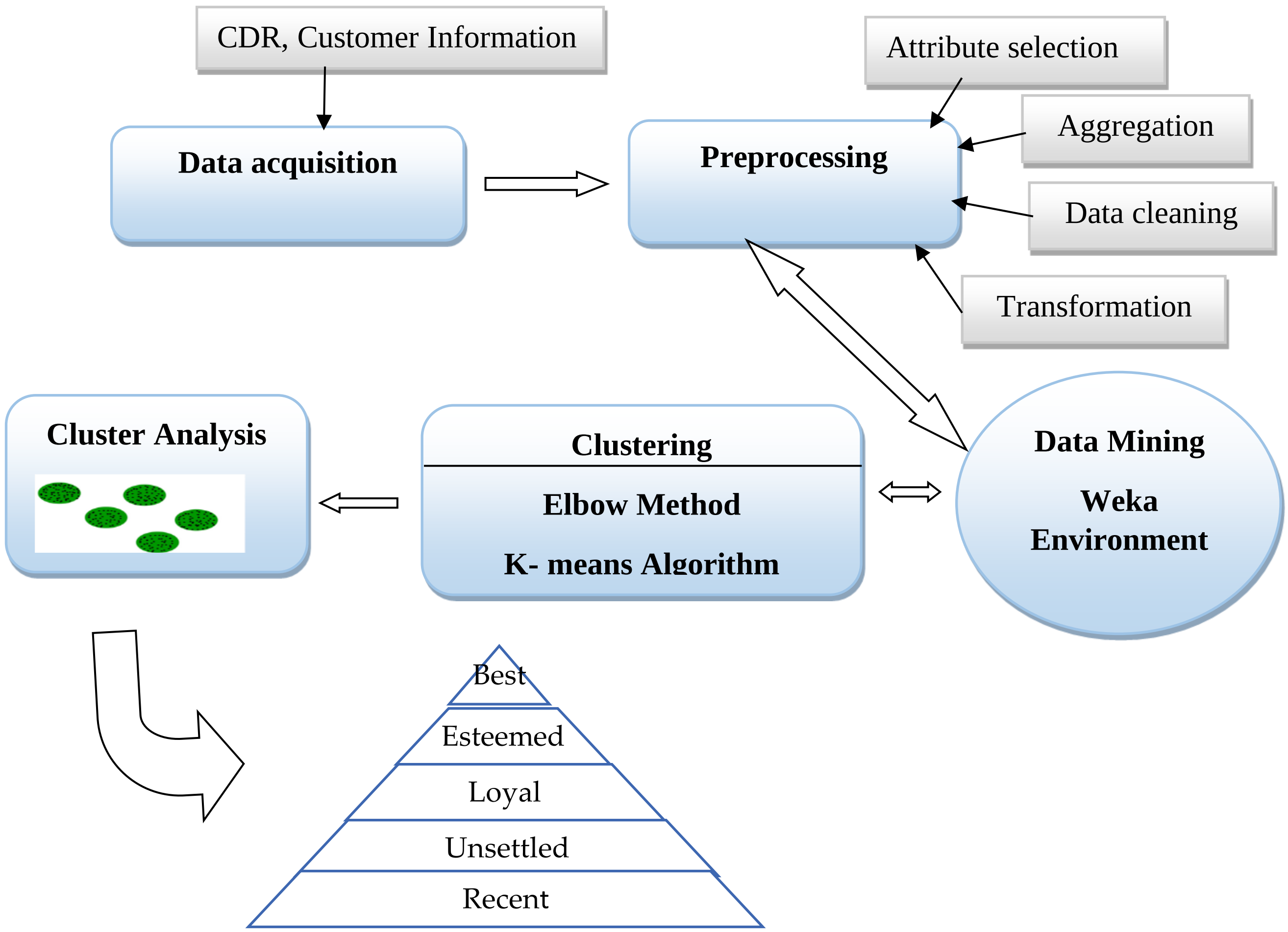


Figure 4-1 System Flow

4.1.1 Data collection

Data collection is a systematic process of gathering relevant data for the specific purpose. As per the internal security policy of Ethio-Telecom, we signed the non-disclosure agreement with the company and made a formal data request for the respective section of the company. After getting the approval of security section, we started the process of collecting data. Due to the request approval procedure, the size of the customer number and the amount of data, the collection process required a lot of time and energy.

As of June 2021, there are 54.3 million mobile customers are registered in Ethio-telecom internal systems [30]. But as per the observation on the system, only 33 million customers are in active status and billable during this thesis research period.

Due to the limitation of resource and the huge size of usage data, sample customers are selected then the three months of March, April, and May of 2021 Customer Detail Record of SMS, Voice, and Internet Data usage of Ethio-telecom mobile customers and customer status data are collected from Ethio-Telecom Billing report databases.

4.1.2 Data Sampling

Sampling is one of the most important steps in large data analysis, since it reduces the amount of data needed to analyze. It's used to get a representative sample from a much larger data set. It helps a researcher to discover trends and patterns in small sets of data Intelligently and methodically instead of interpreting a huge volume of data that is hard to manage.

In this thesis Stratified random data sampling technique is used to select the right number and portion of different segment of mobile customer of Ethio-telecom. 30,000 mobile customers with 12 million records are sampled for this research thesis.

Table 4.1 Sample customer size

Number Group	TOTAL SIZE	RATIO	SAMPLE SIZE
90******	3,009,155	9.05	2,716
91******	5,681,992	17.09	5,128
92******	4,649,889	13.99	4,197
93******	4,372,466	13.15	3,946
94******	3,800,080	11.43	3,430
95******	1,372,087	4.13	1,238
96******	3,701,094	11.13	3,340
97******	1,963,510	5.91	1,772
98******	2,369,792	7.13	2,139
99******	2,319,059	6.98	2,093
Total	33,239,124	100	30,000

4.1.3 Customer Information data

The Customer relationship management system contains customer related information in its repository databases. The main information includes customer name, address, age, gender, Customer creation time, customer status, payment information and others. Customer engagement is one of the metrices to measure the loyalty of a customer. This metrics works by assigning every customer a score based on their individual activity and usage of the services then categorize customer based on their score. The longevity and a smooth relationship of a customer with its service provider is another indication of customer loyalty.

For this thesis, a customer longevity and customer loyalty status information of 30,000 mobile customers is collected from the CRM system databases.

4.1.4 Customer detail Record

Every telecom service provider may capture and record transactions between entities in telecommunication networks, such as calls with their caller number, time of call and terminating numbers and messages with their originating and terminating numbers and also internet data transfer with their download and upload size as well as other technical data [31].

In Ethio-Telecom, all mobile customers, both prepaid and postpaid, events are recorded in mobile switching center as a network log then sent to billing system for further processing. Each customer detailed record contains more than 70 fields when first created in network side but only 33 fields fetch for the billing purpose. but some fields are empty and created for future use.

Table 4-1 CDR Attributes

Attribute ID	Attribute Name	Description
1	CDR_ID	Event Instance Id
2	RE_ID	CDR Type
3	BILLING_NBR	Billing Number
4	CDR_TYPE	Call type id
	CALLING_NUMBER	Call Originating Number
5	CALLED_NUMBER	Call receiving Number
6	CALLING_IMEI	Caller imei number
7	CALLING_IMSI	Caller imsi number
8	CALL_START_TIME	Call Start time
9	CALL_END_TIME	Call end time
10	CALL_DURATION	Duration of a call
11	CALL_FEE	Service fee
12	CALLED_COUNTRY	country code of called number
13	CALLING_CARRIER	Calling operator
14	CALLED_CARRIER	Called operator
15	CALLING_DISTRICT	Caller local cell
16	CALLED_DISTRICT	Peer Party Cell
17	STATUS_DATE	Billing date
18	CALLING_SUB_ID	Calling subscriber ID
19	BILLING_CYCLE_ID	Billing cycle ID
20	CHARGE_1	Usage amount
21	CHARGE_2	Usage with discount
22	RATE_ID1	Rate ID
23	ACCOUNT_ITEM_ID1	Account item ID
24	UPLOAD_TRAFFIC	Upload traffic

25	DOWNLOAD_TRAFFIC	Download traffic
26	BILLING_OFFERING_ID	Bill offering ID
27	ERROR_CDR_TYPE	Error CDR Indicator
28	CALLFORWARDINDICATOR	Call Forward Indicator
29	HOTLINEINDICATOR	Hot Line Indicator
30	CALLING_TRUNK_ID	Calling Trunk ID
31	CALLED_TRUNK_ID	Called Trunk ID

For this thesis, three months of March, April, and May 2021 customer detail records are collected from Ethio-Telecom billing system. A total of 12.7 million records of voice, SMS, and internet data are collected with CSV file format for the selected 30,000 customers that randomly sampled from the total of 33.3 million total Active customers and insert into the database for performing data preprocessing.

Table 4-2 Collected CDR

S.No	**Service Type**	**Data Record**
1	VOICE	9,777,218
2	DATA	2,358,926
3	SMS	610,443
	TOTAL	12,746,587

4.2 DATA PREPROCESSING

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by machine learning tools and algorithms and involves preparing the data for further processing.

Raw data should first be cleaned and organized so it's easier to find patterns and discover insights. The process of preparing data is crucial to the success of any data science project. [21]

The process has several key tasks including Feature selection, data cleaning, integration, reduction, and transformation.

4.2.1 Feature selection

Feature selection is a process of determining which features or variables are relevant for a particular model. it is an important technique in clustering and segmentation. If we use the right set of features, we can find all the natural groups with the same pattern or characteristics that exist in the data. It also makes the analysis simple and more effective.[32]

For this thesis, after discussing with marketing experts in the company and thesis advisors as well as by taking some reference from various literatures of related ideas, the following attributes are selected from the customer detail record and customer information records that are relevant for customer value and retention.

Table 4-3 Selected Attributes

No.	**Attributes**	**Description**
1	RE_ID	Service type
2	CALLING_NUMBER	Call initiator
3	DATA_USAGE (MB)	Data usage
4	VOICE_USAGE(Min)	Voice usage
5	SMS_USAGE	Sent SMS
6	TOTAL_FEE	Total fee
7	SERVICE_INTERRUPT_TIME	Suspend repetition time
8	LONGITIVITY	How long a user being a telecom Customer.

4.2.2 Data cleaning

When there is a high volume of data from multiple sources involved, there may be errors and anomalies in the data. the main idea of this step is cleaning data by removing outliers, replacing missing values, smoothing noisy data, and correcting inconsistent data. The cleaning process helps to eliminate redundancies and avoid inconsistencies and ensure that data is accurate and relevant. [31]

The collected customer detail record and customer information data had a lot of issues. Some fields were incorrectly recorded, there was usage information but do not have service fee and vice versa, the usage information and service fee are not matched, and there were duplicate and missing values. These issues are carefully handled by replacing the missing values by calculating the usage and insert the right data into the missing columns and discard records that has no information.

4.2.3 Data Aggregation and integration

Data aggregation and integration is the process of combining and summarizing data created in different applications or on different systems into a single file. Customer detail records and customer status information are gathered for this thesis from the billing system and the customer relationship management system, respectively.

Three months of voice, SMS, and internet data, as well as customer information in various formats and data types, are collected in a csv file. Convert the data into the Excel format to transform it. The initial data preparation was done on the Excel application, and then it was inserted into the database.

In the database, eleven tables are created for the three-month CDR record of voice, SMS, internet data, and customer status information.

After cleaning the data, each data set file is aggregated per month before being integrated at the customer level and formatted in numerical form to prepare for the algorithm.

4.3 SEGMENTATION ANALYSIS

Clustering is a technique for categorizing data based on its attributes. Data that has similar characteristics are grouped together. To assess clustering quality, there are two methods available. The first is an external evaluation, in which the truth labels in the data sets are known ahead of time, and the second is an internal evaluation, in which the evaluation is done with the data set itself without the true labels [33]. The clustering of the customer dataset is subjected to internal evaluation in this thesis. The Elbow method and total within-cluster sum of squares are used to validate the number of clusters for internal performance measures.

Based on customer usage and customer status information, the customer segmentation clustering for value and retention experiment was carried out. For clustering analysis, Weka knowledge discovery tool and k-means algorithm are used in this thesis.

The three-month data are aggregated in total of voice, SMS, and internet data with the customer status and longevity. Finally, all data sets are integrated as a whole and each customer's usage per month was clustered in the analysis.

4.3.1 Elbow Method

A fundamental step in any unsupervised clustering is determining the optimal number of clusters into which the data can be clustered. The Elbow Method is a popular method for determining the optimal value of k.

In the Elbow method, the number of clusters (K) varying from 1 to the defined number. For each value of K, the WSS would be calculated, and it is the sum of squared distance between each point and the centroid in a cluster. When the WSS plot with the K value, the plot looks like an Elbow. As the number of clusters increases, the WSS value will start to decrease. When we analyze the graph, we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters [22]. The total WSS measures the clustering's compactness and should be as small as possible.

In this thesis the following steps are used to calculate the optimal number of clusters.

- Compute k-means clustering algorithm for different values of k (1 to 10 clusters).
- For each k, calculate the total within-cluster sum of square (WSS).

$$\mathrm{WSS} = \sum_{i=1}^{N_c} \sum_{x \in Ci} d(x, xCi)\ 2 \qquad (1)$$

- Plot the curve of WSS according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. For this experiment the optimal number of k is 5.

Clusters	WSS
1	1.43E+12
2	0.55E+12
3	0.30E+12
4	0.20E+12
5	0.19E+12
6	0.07E+12
7	0.04E+12
8	0.04E+12
9	0.03E+12
10	0.02E+12
11	0.02E+12

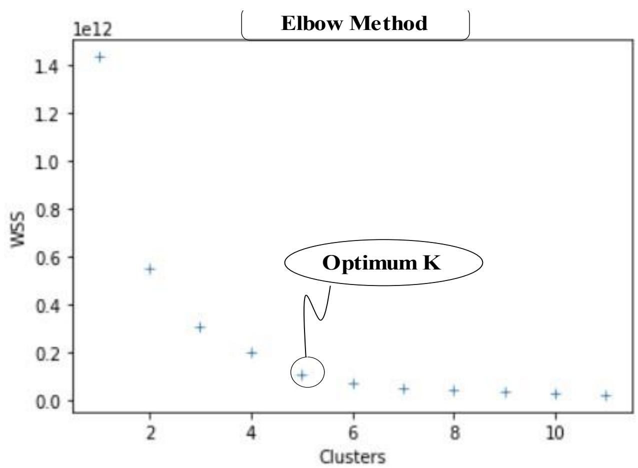


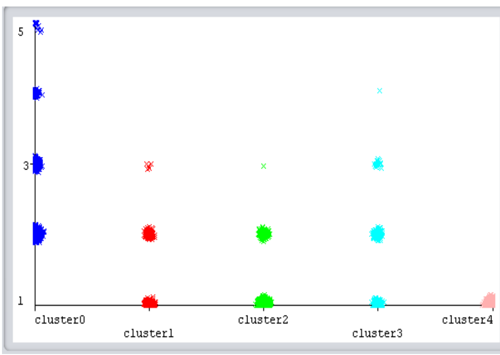
Figure 4-2 Optimal Cluster

4.3.2 Cluster using K-means Algorithm

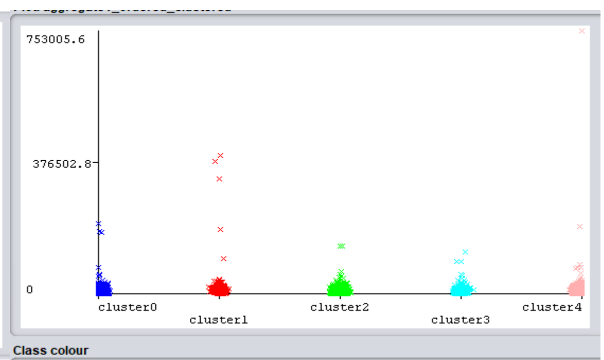
The three-month dataset of voice, SMS, and internet data is aggregated at the customer level and integrated with customer longevity and service interruption status information. After determining the optimal cluster number, the dataset is uploaded to the Weka knowledge discovery tool, and the test is run in the prepared environment using the k-means algorithm and Euclidean distance.

The 3000 customers are clustered into five groups and the clustering process takes 0.37 seconds with 40 iterations. The first group contains 9109 customers, which is 30% of the total customer base, 474 customers are assigned to the second group, which is 3%, 5628 customers are assigned to the third group, which is 18%, 804 customers are assigned to the fourth group, which is 3%, and finally 13712 customers are assigned to group five, which has the largest customer size compared to the other groups.

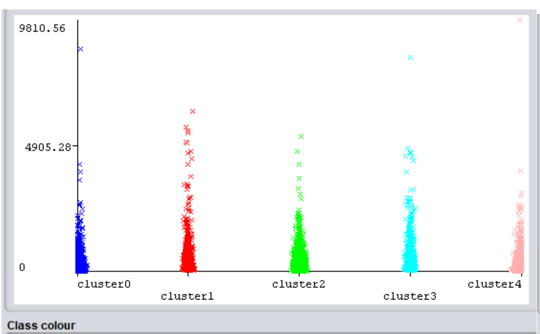
The graphical representation of the clustering according to each aggregated attribute is presented below.



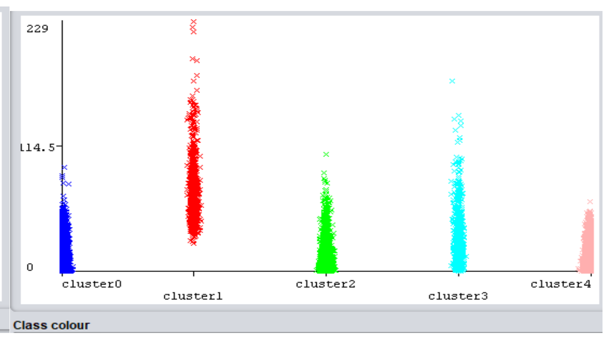
Interrupt time



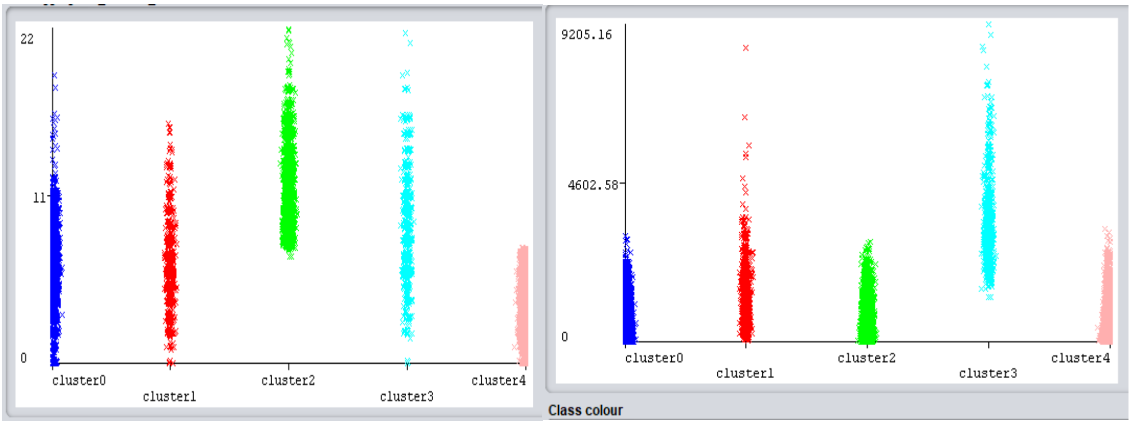
Internet(data) usage



Total fee



SMS usage



Service year Voice usage

Figure 4-3 Graphical representation of attributes

5 RESULT AND DISCUSSION

This chapter presents and discusses the findings of the study as it relates to the purpose of the study. Section 5.1 describes the clustering result of aggregated CDR and customer status data by the k-means algorithm, and Section 5.2 covers the discussion points of the experimental analysis.

5.1 RESULT.

The analysis was carried out based on the attributes selected from the CDR and customer status information that is explained in detail in chapter 4. Based on the optimal cluster number, the dataset was clustered with the k-means algorithm in to five groups. Then each group is analyzed based on the average usage of three-month internet data, SMS, and voice that incorporate with the service interruption time of each customer and longevity of the customer in that group.

As shown in Figure 5-1, the total 30000 sample customers selected for the experiment are clustered into five groups using k-means algorithm. The overall time required to cluster the dataset was 40 seconds, with 44 iterations. The cluster results show that 46 percent (13713) of customers are grouped into cluster5 and 30 percent (9109) of customers are grouped into cluster1, but when compared to other groups, a small number of customers are grouped into cluster2 (747) and cluster4 (824).

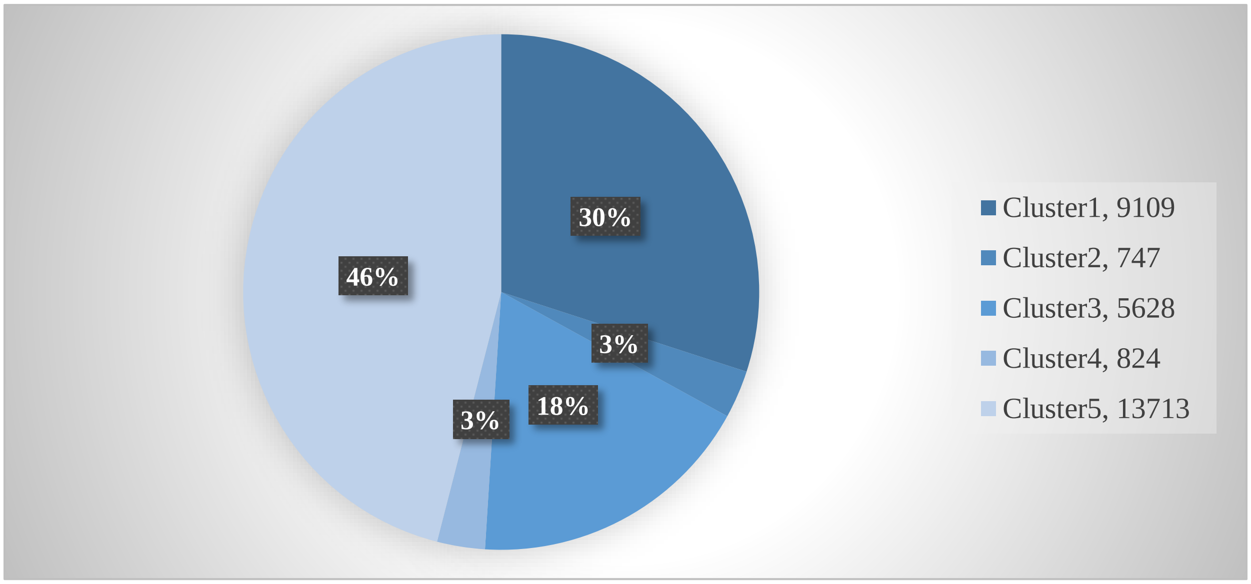


Figure 5-1 Customer cluster using K-means.

5.1.1 Usage based cluster

One of the clustering attributes used in this experiment for Customer segmentation process for retention and value is based on the customer usage event, which is gathered as a CDR from the billing part of telecom company.

Figure 5-2 depicts how the dataset is clustered based on the average SMS, data, and voice usage of each cluster, as well as the average usage fee derived from the average usage.

Based on data usage segmentation, cluster2(4405MB) customers are the most data consumers, followed by cluster4(2662MB) customers. Cluster5 customers use the least amount of data per month, and the rest fall somewhere in the middle.

Based on voice attribute segmentation, the results show that Cluster4(3354 Min.) customers are the most vocal service users, followed by Cluster2 (1289 Min.). The other two clusters use voice service in a way that less than half of cluster2 customers do. Cluster5(275 Min.) customers use the least amount of voice services.

According to the results of SMS attribute segmentation, Cluster2(74) customers interact highly with other service users by using SMS service. Customers of Cluster4(29) also contribute significantly to the company's revenue. However, the other cluster group makes only a relatively insignificant use of the SMS service.

The other segmentation attribute used based on usage is the average fee derived from the average usage of the three services of SMS, voice, and internet data. both cluster2(564) and cluster4(505) customers contribute a significant amount of revenue to the company each month. The other groups also share a fair amount of revenue.

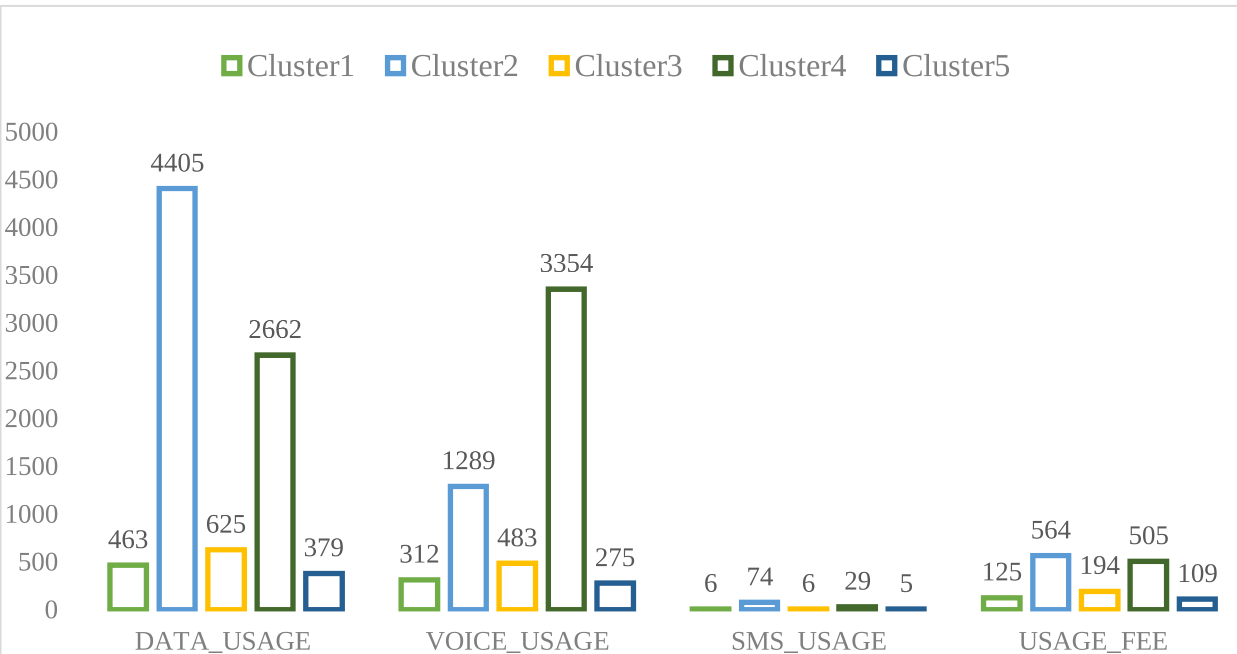


Figure 5-2 Usage based cluster

5.1.2 Status based cluster

Other attributes that are used in the analysis were customer longevity and service interruption time for each customer in the event that they do not fulfill their payment obligation during their customer lifetime.

Figure 5.3 depicts the cluster based on the average longevity and customer service interruption time of a customer as a registered user over the course of their lifetime. Cluster4 (11 years) is the company's longest-standing customers, followed by Cluster2 (9 years), which has a moderately longer time relationship. Clusters 1 and 3 both have an average age of 6 years But Cluster5 customers, on the other hand, are relatively new customers.

When it comes to service interruption time, Cluster5 and Cluster2 have the shortest average service interruption time when compared to the other groups, indicating that these customers pay their bills on time. Customers in Clusters3 and Cluster4 had their service interrupted 1.4 and 1.3 times, respectively, due to late payments. But Customers in Cluster1 had a poor track record of paying their bills and maintaining good relationships with the company.

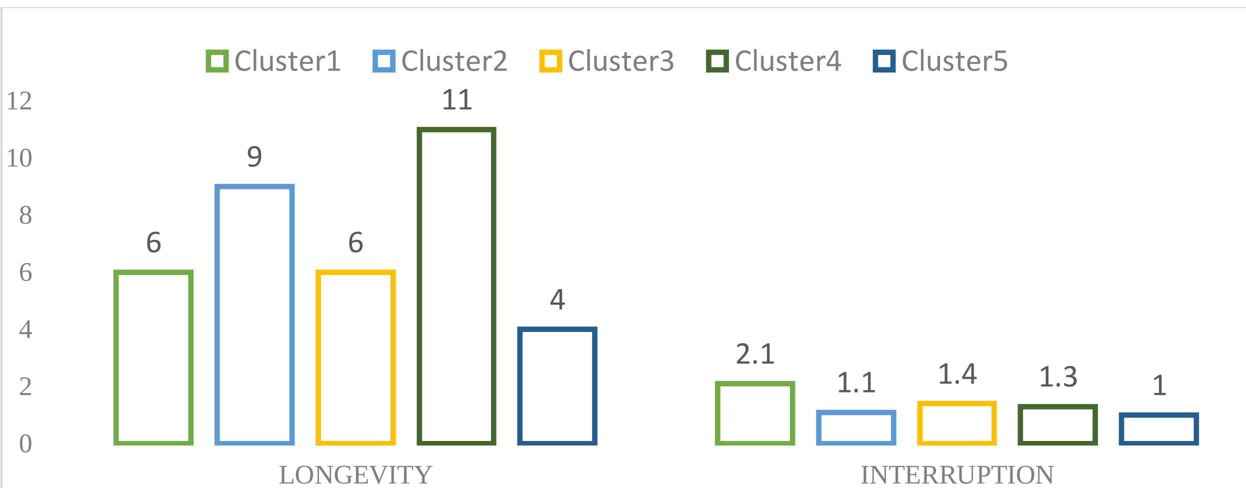


Figure 5-3 Status based Cluster

5.1.3 Total Result

Figure 5.4 depicts the outcome of combining the total average values of the three services and the status of the customers. Cluster2 has the highest data and SMS usage, as well as moderately high voice usage, while Cluster 4 has the highest voice usage, as well as relatively higher data and SMS usage. These groups have been customers for a relatively long time (6 and 8 years, respectively). Even though these two groups represent only 6 percent of total customers, they generate 72 percent of total revenue for the company.

Cluster1 is the second most populous group (30 percent). However, these customers had a very poor track record with the company in terms of service continuity. Due to service interruption, their service usage is also low.

Customers grouped into cluster3, 18 percent of a total customer, have a very long relationship with the company, with an average of ten years, and they create a fair share of payment of 12% of total company revenue. When compared to other groups, their usage statistics show that they use all services in the same manner. However, as evidenced by their records, these customers have a history of poor payment settlement.

Cluster5 accounts for 45 percent of customers, but they contributed the least amount of revenue to the company and their service was not interrupted due to payment because they are comparatively new customers.

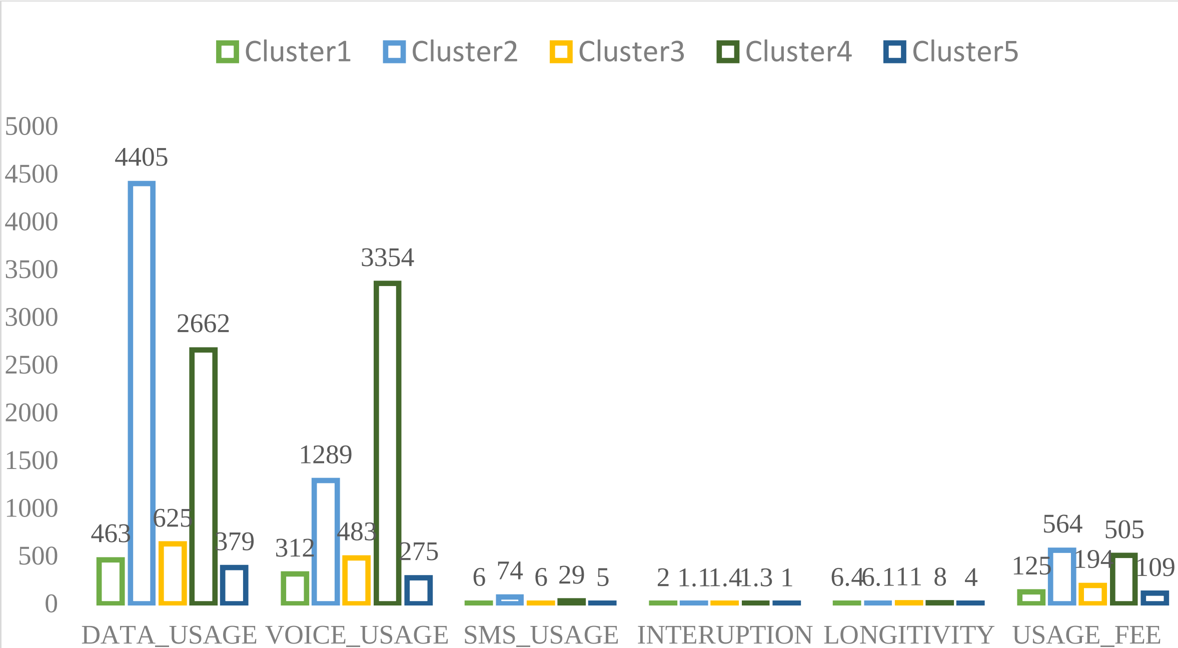


Figure 5-4 Total Clustering

5.1.4 Ranking value model

Scoring is a method of manipulating data in preparation for analysis. One of the ranking tools for evaluating proposed segmented customers is RFM [34]. The customer RFM analysis begins by categorizing customers based on their usage behavior, such as how recently they used the service, how frequently they used it, and how much monetary value they contributed to the company. It is a strategy for retaining existing customers rather than attracting new ones [14]. However, for this thesis project, we modified the dimensions by removing the recency because we use the same three-month event period and adding two new dimensions: customer longevity and service interruption time.

so, the variables duration, monetary, interruption, and longevity from the data set are used to understand the customer's behavior based on usage and status. Weights are assigned based on the value of each group's variable, from lowest to highest in the cluster set. Then, for each group in the database, generate a weighted score ranging from 1 to 5. Finally, As shown in Table 5-1, and Table 5-2 summarize the entire data set by concatenating each score and ranking each group based on the outcome and align with the model.

Table 5-1 Score of the total Attribute

CLUSTER	CLUSTER SIZE	DURATION	MONETARY	INTER-RUPTION	LONGIVITY	SCORE
Cluster1	9109	2	2	1	3	2213
Cluster2	747	5	5	4	3	5543
Cluster3	5628	3	3	2	5	3325
Cluster4	804	4	4	3	4	4434
Cluster5	13712	1	1	5	1	1151

Table 5-2 Cluster Type link with the score

CLUSTER	CLUSTER SIZE	SCORE	CLUSTER TYPE
Cluster2	747	5543	Best (Platinum)
Cluster4	804	4434	Esteemed (Gold)
Cluster3	5628	3325	Loyal (Silver)
Cluster1	9109	2213	Unsettled (Bronze)
Cluster5	13712	1151	New (Normal)

5.1.5 Marketing Approaches

After Clustering customers into groups and understanding the behavior of each segment, possible marketing strategies are suggested based on the analysis of each customer segment. Table 5-3 shows some marketing strategies for each cluster to increase customer value, total revenue, and retention rate.

Table 5-3 Marketing Suggestion

CLUSTER	**CLUSTER TYPE**	**SUGGESTION**
Cluster2	Best (Platinum)	Special customer window. Service continuity guarantee. Special data package.
Cluster4	Esteemed (Gold)	special customer window. Service continuity guarantee. Special Voice package.
Cluster3	Loyal (Silver)	Award discount, special package. Loyalty program.
Cluster1	Unsettled (Bronze)	Analytical marketing. special discount offers. Address customer problem.
Cluster5	New (Normal)	Advertising and promotion. Further investigation needed.

Customer segmentation can help organizations identify more effective marketing strategies by utilizing the power of machine learning clustering technology. Customers in Clusters 2 and 4 are high-value customers who require careful marketing approaches and should be treated differently because the company's survival may depend on them. The company should formulate a different payment settlement method to deal with them and never stop their service due to late payments or other reasons and should set up a fully equipped Customer Service window for them that is available 24 hours a day, seven days a week, and offer them long-term data packages as well as a special total package.

Cluster3 customers are long-standing customers from the incumbent industry's era. Across the whole spectrum of services, this sector exhibited consistent service utilization. After using a product for a while, customers usually develop brand loyalty. Adjust your messaging based on how loyal they are to your brand. These customers also require special attention, need to provide a specific custom service such as a loyalty bonus plan or the development of a special retention strategy that includes a special data, voice, and SMS package to keep them motivated to use the services for an extended period during the competitive epoch.

Customers in cluster1 seems to be potential migrators, and special consideration should be given to this group to retain and take care of them in respect their behavior. Its better review these customers data in real time to better understand their needs. More promotional plans, including special discount offers and short-term service packages are required to keep them in the customer relationship with the company, but their behavior indicates that they will migrate among providers, so further investigation is required to determine the true cause of their situation.

The final cluster group (cluster 5) are new customers, and their usage behavior and customer relationship should be improved. A high volume of advertising and promotion can help to motivate new customers, and offering customized packages, surprise gifts, and incentives can also help to attract them.

In order to determine the root cause of low usage behavior, the company should monitor and analyze these customer group metrics.

5.2 DISCUSSION

Customer segmentation is essential for understanding different customer behaviors and determining which customer groups are important to the company. After categorizing customers, the organization should devise a unique marketing strategy to address each cluster group to maximize customer value and retention rate.

If telecom companies want to increase their market share, they should consider developing marketing strategies that are specifically tailored to different customer segments.

The main purpose of this thesis is to segment telecom customers with the data mining technique based on mobile event usage and customer status information. K- means algorithm is used for clustering purpose. K-means is a simple and widely used unsupervised machine learning algorithm. Before clustering, elbow method is used to get the optimal cluster number which is five.

The dataset was clustered into five groups using the algorithm, and the clustering analysis was made based on the result and score the value to map with the customer type. Best (Platinum), Esteemed (Gold), Loyal (Silver), Uncertain (Bronze), and New (Normal). For retaining the existing customers (finding new customers is more expensive than keeping current customers) and increase customer value, the company should address the findings and create a specific communication channel for each cluster group. It is not advisable to develop a marketing strategy before segmenting your market. And, as demonstrated above, there is a specific approach that provides companies with more opportunities to differentiate themselves and appeal to specific groups based on common behavior. Finally, the marketing suggestion is placed after discussing with marketing experts and knowledge grasp from different related literatures.

6 CONCLUSION AND RECOMMENDATION

6.1 CONCLUSION

As technology advances, the telecommunications industry can offer more products and services. This means that competitive pressure will increase, and if you want to compete, you must understand your customers' usage behavior and true need.

Customer segmentation is critical for comprehending various customer behaviors and determining which customer groups are most important to the company and which customer groups require special attention. Following customer categorization, the organization should develop a distinct marketing strategy to address each cluster group to maximize customer value and retention rate.

As a result of Ethiopia's telecom reform, emerging telecom service providers will face new and intense market competition. It can use their call data records (CDRs) and customer related information to find out more about their customers' activities and improve the quality of their services. If service providers want to increase their market share, they should consider developing marketing strategies that are specifically tailored to different customer segments.

In this paper, data mining techniques were used to cluster customers based on mobile event usage and customer status. The dataset was gathered from various company departments, preprocessed, and then integrated into the database. The elbow method, an intra-cluster evaluation method, was used for optimal cluster evaluation, and the aggregated data was clustered in the Weka knowledge discovery tool using the unsupervised k-means clustering algorithm.

As a result, Cluster2 and Cluster4 customers are the company's most frequent users of telecom services and generate the most revenue. Cluster1 customers' behavior indicates that they are involved in a serious payment dispute with the company. It appears possible to switch to the best customer category by creating the appropriate channel for cluster3 customers. Finally, based on our extensive analysis, a marketing strategy is proposed for each cluster group in order to retain every customer and gain a competitive advantage in the industry.

6.2 RECOMMENDATIONS

Telecommunications industries generate a variety of large amounts of data. Data mining technology can assist various sectors of the industry, such as marketing, security, and network, in identifying hidden behavior and patterns in generated data.

This thesis work focuses primarily on using CDR data to gain insights for segmenting customers based on their usage behavior and recommending a business strategy for properly handling and retaining existing customers. In the future, we can use data from multiple providers to determine true churn behavior among service providers, and it helps in overcoming the limitation of only using data from one service provider. Also, there are numerous data mining applications available to be used in telecom security, network, and marketing to gain a competitive advantage by utilizing these massive amounts of telecom data.

BIBLIOGRAPHY

[1] S. F. Sabbeh, “Machine-Learning Techniques for Customer Retention : A Comparative Study,” vol. 9, no. 2, pp. 273–281, 2018.

[2] H. Ziafat and M. Shakeri, “Using Data Mining Techniques in Customer Segmentation,” vol. 4, no. 9, pp. 70–79, 2014.

[3] A. Namvar, M. Ghazanfari, and M. Naderpour, “A customer segmentation framework for targeted marketing in telecommunication,” *Proc. 2017 12th Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2017*, vol. 2018-Janua, pp. 1–6, 2017, doi: 10.1109/ISKE.2017.8258803.

[4] W. Verbeke, D. Martens, C. Mues, and B. Baesens, “Building comprehensible customer churn prediction models with advanced rule induction techniques,” *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2354–2364, 2011, doi: 10.1016/j.eswa.2010.08.023.

[5] E. C. Authority, “Ethiopian Telecommunications Sector Stakeholder Consultation No . 001-2019 Public Notice,” no. 001, pp. 1–16, 2019.

[6] E.Telecom, “Ethio telecom 2013 EFY (2020/21) Annual Business Performance Summary Report,” *https://www.ethiotelecom.et/ethio-telecom-2013-efy-2020-21-business-performance-summary-report/*, 2021. .

[7] A. Damtew, “Assessment of Marketing Strategy Practices in Ethio Telecom : From the Company Perspectives,” no. october, 2013.

[8] S. Y. Hung, D. C. Yen, and H. Y. Wang, “Applying data mining to telecom churn management,” *Expert Syst. Appl.*, vol. 31, no. 3, pp. 515–524, 2006, doi: 10.1016/j.eswa.2005.09.080.

[9] B. Luo and S. Lin, “Research on the Anonymous Customer Segmentation Model of Telecom,” *Proc. 2020 IEEE 5th Inf. Technol. Mechatronics Eng. Conf. ITOEC 2020*, no. Itoec, pp. 1026–1031, 2020, doi: 10.1109/ITOEC49072.2020.9141572.

[10] Q. Lin and Y. Wan, “Mobile Customer Clustering Based On Call Detail Records For Marketing Campaigns,” *Int. Conf. Manag. Serv. Sci.*, pp. 1–4, 2009, doi: doi: 10.1109/ICMSS.2009.5302716.

[11] C. H. Cheng and Y. S. Chen, “Classifying the segmentation of customer value via RFM model and RS theory,” *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 4176–4184, 2009, doi: 10.1016/j.eswa.2008.04.003.

[12] D. Ćamilović, “Data Mining and Crm in Telecommunications,” *Serbian J. Manag.*, vol. 3, no. 1, pp. 61–72, 2008.

[13] F. Abdi and S. Abolmakarem, “Customer Behavior Mining Framework (CBMF) using clustering and classification techniques,” *J. Ind. Eng. Int.*, vol. 15, no. s1, pp. 1–18, 2019, doi: 10.1007/s40092-018-0285-3.

[14] Gordon S. Linoff and Michael J. A. Berry, *Data Mining Techniques. Third Edition For Marketing, Sales, and Customer Relationship Management*, Third Edit., vol. 53, no. 9. Wiley, 2013.

[15] S. Dolnicar, B. Grün, and F. Leisch, *Management for Professionals Market Segmentation Analysis Understanding It, Doing It, and Making It Useful*. 2011.

[16] E. Mattilla and S. Arnborg, “Behavioral Segmentation of Telecommunication Customers Behavioral Segmentation of Telecommunication Customers,” vol. fgg, 2008.

[17] Z. Wang, L. Tu, Z. Guo, L. T. Yang, and B. Huang, “Analysis of user behaviors by mining large network data sets,” *Futur. Gener. Comput. Syst.*, vol. 37, pp. 429–437, 2014, doi: 10.1016/j.future.2014.02.015.

[18] A. S. M. S. Hossain, “Customer segmentation using centroid based and density based clustering algorithms,” *3rd Int. Conf. Electr. Inf. Commun. Technol. EICT 2017*, vol. 2018-Janua, no. December, pp. 1–6, 2018, doi: 10.1109/EICT.2017.8275249.

[19] O. . C. . FERRELL and M. D. HARTLINE, *Marketing Strategy*, SIXTH EDIT. Mason, OH : South-Western Cengage Learning, 2011.

[20] K. Philip and K. . Kevin Lane, *Marketing Management. (14th ed.) Prentice Hall.* 2012.

[21] H. Jiawei, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed. Morgan Kaufman, 2011.

[22] A. Bascacov, C. Cernazanu, and M. Marcu, “Using data mining for mobile communication clustering and characterization,” *SACI 2013 - 8th IEEE Int. Symp. Appl. Comput. Intell. Informatics, Proc.*, pp. 41–46, 2013, doi: 10.1109/SACI.2013.6609004.

[23] F. Usama, P.-S. Gregory, and S. Padhraic, “Knowledge Discovery in From Data Mining to Databases,” *AI*, vol. 17, no. 3, 1996, doi: 10.1609/aimag.v17i3.1230.

[24] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, vol. 0. 2012.

[25] M. Kaur, “A COMPARATIVE STUDY OF TECHNIQUES TO PREDICT CUSTOMER CHURN IN TELECOMMUNICATION INDUSTRY,” 2017.

[26] D. Baumgartner and G. Serpen, “Large Experiment and Evaluation Tool for WEKA Classifiers,” no. April 2017, 2009.

[27] W. Liao, Y. Liu, and A. Choudhary, “A Grid-based Clustering Algorithm using Adaptive Mesh Refinement,” *7th Work. Min. Sci. Eng. Datasets SIAM Int. Conf. Data Min.*, pp. 61–69, 2004, [Online]. Available: http://users.eecs.northwestern.edu/~choudhar/Publications/LiaLiu04A.pdf.

[28] K. Sanse and M. Sharma, "Clustering methods for Big data analysis," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 4, no. 3, pp. 642–648, 2015.

[29] Iyer Aurobind Venkatkumar and Sanatkumar Jayantibhai Kondhol Shardaben, "Comparative study of Data Mining Clustering algorithms," 2016, [Online]. Available: http://bdbib.javerianacali.edu.co:2208/document/7823946/.

[30] Ethio-telecom, "https://www.ethiotelecom.et/," *OUR STATISTICS*, 2021. .

[31] D. Steinberg, "SESSION A TOUR OF ADVANCED DATA MINING Chair (s)."

[32] H. (National U. of S. Liu, H. (Osaka U. Motoda, R. Setiono, and Z. Zhao, "Feature Selection : An Ever Evolving Frontier in Data Mining," *J. Mach. Learn. Res. Work. Conf. Proc. 10 Fourth Work. Featur. Sel. Data Min.*, pp. 4–13, 2010.

[33] A. Vysala and D. J. Gomes, "Evaluating and Validating Cluster Results," pp. 37–47, 2020, doi: 10.5121/csit.2020.100904.

[34] W. Xiong, L. Chen, Z. ZhanG, and Z. Qiu, "RFM value and grey relation based customer segmentation model in the logistics market segmentation," *Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008*, vol. 5, no. 1, pp. 1298–1301, 2008, doi: 10.1109/CSSE.2008.79.

APPENDIX

Customer Segmentation for Value and Retention Using Data Mining: In the Case of Ethio-Telecom Mobile Service

Ephrem Teshale (PhD) ephremteshale@gmail.com, Girum Afework girum2020@gmail.com
School of Electrical and Computer Engineering,
Addis Ababa Institute of Technology, Addis Ababa University

***Abstract*—** Market competition is becoming intense among telecom service providers around the world. To gain a competitive advantage in the industry, service providers must address and meet their customers' needs and demands. To survive in today's competitive market, companies must analyze and interpret their customers' usage behavior, as well as plan related market strategies to retain customers, be profitable, and build long-term relationships.

One of the most effective ways to engage with each customer is through segmentation. Customer segmentation can help organizations identify more effective marketing strategies for each segment by leveraging the power of data mining clustering technology.

In this paper, the unsupervised clustering technique with k-means algorithm is applied and customer detail record (CDR) and customer information data are used to segment Ethio-telecom customers for the purpose of retain existing customers and increase customer value by treating each customer segment according to their usage behavior.

The collected data is cleaned and preprocessed in an Oracle database and then the aggregated data is used to calculate the optimal cluster number using the k-means and elbow methods. Based on the selected attributes the dataset is segmented into five groups by Weka knowledge discovery tool.

Each cluster segment is scored and mapped with the type of customer segmentation based on three-month average usage data, frequency, longevity, and service interruption time. Clusters 2 and 4 account for 5% of the total customer size but cover 72% of the company revenue, whereas Clusters 1 and 5 account for 76% of the total customer size but contribute significantly less than the others. Finally, based on the analysis result, a marketing strategy for each segment is proposed.

Keywords: Data Mining, Clustering, CDR, elbow method, k-means

I. INTRODUCTION

Customer segmentation is the process of grouping customers with similar characteristics, behaviors, and values. Customer retention is one of the primary goals of customer segmentation, as it aids in increasing loyalty and avoiding customer churn. Every company must understand their customers' behavior and implement various marketing strategies to retain and satisfy the demand of each segmented group in order to be successful [1].

Customers are the foundation of every business organization's survival and retaining customers is one of the most critical challenges to deal with. It is critical to identify newly emerging trends and customer behavior in order to adapt and change the company's business processes.

The main objective of every business organization is to gain a better business advantage in the competitive market and only providing quality service cannot guarantee its success. Customer satisfaction is one of the metrics that customers to be loyal for every business organization, so it needs to be investigated correctly by the service providers [2].

Nowadays, telecom service providers are facing a variety of challenges, including rapidly changing customer behavior patterns, financial and technological challenges. Identifying emerging trends and translating them into actionable strategies is critical for long-term smooth customer relationship and sustaining profitability.

CDR, customer payment settlement information that affect the service continuity of a customer, and the customer longevity as a company's subscriber data are used as an input for clustering customer with K-means unsupervised learning techniques in Weka data mining tool.

II. RELATED WORK

The telecommunications industry provides a wide range of services to a large number of customers in the world, but the growing number of service providers creates fierce market competition in the industry, and this circumstance has influenced customers to switch to other competitors for a variety of reasons. Researchers have been working for decades to develop models and techniques for customer segmentation based on data mining and machine learning approaches.

Reference [1] conducted a comparison study on the eight most commonly used algorithms 1) Regression analysis: logistic regression. 2)

Decision tree–CART. 3) Bayes algorithm: Naïve Bayesian. 4) Support Vector Machine 5) Instance – based learning: k-nearest Neighbor. 6) Ensemble learning: Ada Boost, Stochastic Gradient Boost and Random Forest. 7) Artificial neural network: Multi-layer Perceptron. 8) Linear Discriminant analysis for predicting customer churn for the purpose of customer retention. According to the study, the selected models are evaluated on a public dataset of telecom customers to evaluate and benchmark the performance of the models and ensemble – based learning methods are achieved the highest performance and suggested as a reference point.

Reference [10] used K-means clustering method to segment mobile customers based on their call detail records and analyzes their behaviors. The researcher proposes a model and marketing strategies for China mobile marketing, then implements the model, comparing marketing effectiveness before and after implementing the clustering and suggested strategies, and demonstrating that the new promoting strategy has visible effects.

Reference [28] give an overview of the various clustering methods along with their classification, general working and also provides a theoretical comparison among them.

III. SEGMENTATION ANALYSIS

After discussing with marketing experts in the company and advisors as well as by taking some reference from various literatures of related ideas, the following attributes are selected from the customer detail record and customer information records that are relevant for customer value and retention.

Table 3-1 Selected Attributes

No.	Attributes	Description
1	RE_ID	Service type
2	CALLING_NUMBER	Call initiator
3	DATA_USAGE (MB)	Data usage
4	VOICE_USAGE(Min)	Voice usage
5	SMS_USAGE	Sent SMS
6	TOTAL_FEE	Total fee
7	SERVICE_INTERRUPT_TIME	Suspend repetition time
8	LONGITIVITY	How long a user being a telecom Customer.

The collected customer detail record and customer information data had a lot of issues. Some fields were incorrectly recorded, there was usage information but do not have service fee and vice versa, the usage information and service fee are not matched, and there were duplicate and missing values. These issues are carefully handled by replacing the missing values by calculating the usage and insert the right data into the missing columns and discard records that has no information.

Then the three-month CDR dataset of voice, SMS, and internet data is aggregated at the customer level and integrated with customer longevity and service interruption status information. After determining the optimal cluster number of 5 with Elbow method, the dataset is uploaded to the Weka knowledge discovery tool, and the test is run in the prepared environment using the k-means algorithm and Euclidean distance.

The 30000 sample customers are clustered into five groups and the clustering process takes 0.37 seconds with 40 iterations. The first group contains 9109 customers, which is 30% of the total customer base, 474 customers are assigned to the second group, which is 3%, 5628 customers are assigned to the third group, which is 18%, 804 customers are assigned to the fourth group, which is 3%, and finally 13712 customers are assigned to group five, which has the largest customer size compared to the other groups.

Figure 3-1 Customer cluster using K-means

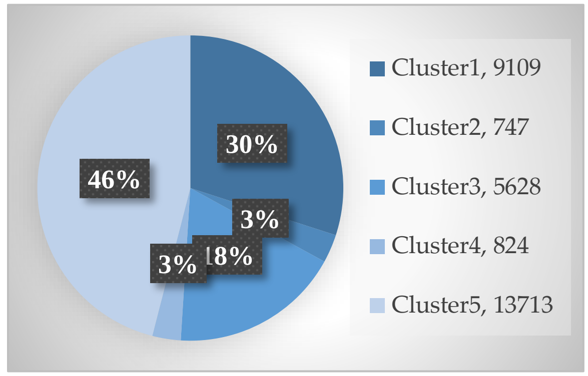


Table 3.1 depicts the outcome of combining the total average values of the three services and the status of the customers. Cluster2 has the highest data and SMS usage, as well as moderately high voice usage, while Cluster 4 has the highest voice usage, as well as relatively higher data and SMS usage. These groups have been customers for a relatively long time (6 and 8 years, respectively). Even though these two groups represent only 6 percent of total customers, they generate 72 percent of total revenue for the company.

Cluster1 is the second most populous group (30 percent). However, these customers had a very poor track record with the company in terms of service continuity. Due to service interruption, their service usage is also low.

Customers grouped into cluster3, 18 percent of a total customer, have a very long relationship with the company, with an average of ten years, and they create a fair share of payment of 12% of total company revenue. When compared to other groups, their usage statistics show that they use all services in the same manner. However, as evidenced by their records, these customers have a history of poor payment settlement.

Cluster5 accounts for 45 percent of customers, but they contributed the least amount of revenue to the company and their service was not interrupted due to payment because they are comparatively new customers.

Table 3.2 Total Clustering

CLUSTER	CLUSTER SIZE	DATA USAGE	VOICE USAGE	SMS USAGE	INTTER RUPT TIME	LONGITIV ITY(years)	TO
Cluster1	9109	462.6923	311.6074	5.9155	2.1255	6.4028	125.
Cluster2	747	4404.756	1288.8608	73.66	1.0645	6.1272	563.
Cluster3	5628	624.6268	482.8564	6.1205	1.4515	10.7026	193.
Cluster4	804	2662.3962	3353.8683	28.699	1.2771	8.551	504.
Cluster5	13712	379.3874	274.9916	4.5385	1	3.8303	108.

After Clustering customers into groups and understanding the behavior of each segment, possible marketing strategies are suggested based on the analysis of each customer segment.

Customer segmentation can help organizations identify more effective marketing strategies by utilizing the power of machine learning clustering technology. Table 5-3 shows some marketing strategies for each cluster to increase customer value, total revenue, and retention rate.

Table 3-3 Marketing Suggestion

CLUSTER	CLUSTER TYPE	SUGGESTION
Cluster2	Best (Platinum)	Special customer window. Service continuity guarantee. Special data package.
Cluster4	Esteemed (Gold)	special customer window. Service continuity guarantee. Special Voice package.
Cluster3	Loyal (Silver)	Award discount, special package. Loyalty program.
Cluster1	Unsettled (Bronze)	Analytical marketing. special discount offers. Address customer problem.
Cluster5	New (Normal)	Advertising and promotion. Further investigation needed.

Customers in Clusters 2 and 4 are high-value customers who require careful marketing approaches and should be treated differently because the company's survival may depend on them. The company should formulate a different payment settlement method to deal with them and never stop their service due to late payments or other reasons and should set up a fully equipped Customer Service window for them that is available 24 hours a day, seven days a week, and offer them long-term data packages as well as a special total package.

Cluster3 customers are long-standing customers from the incumbent industry's era. Across the whole spectrum of services, this sector exhibited consistent service utilization. After using a product for a while, customers usually develop brand loyalty. Adjust your messaging based on how loyal they are to your brand. These customers also require special attention, need to provide a specific custom service such as a loyalty bonus plan or the development of a special retention strategy that includes a special data, voice, and SMS package to keep them motivated to use the services for an extended period during the competitive epoch.

Customers in cluster1 seems to be potential migrators, and special consideration should be given to this group to retain and take care of them in respect their behavior. Its better review these customers data in real time to better understand their needs. More promotional plans, including special discount offers and short-term service packages are required to keep them in the customer relationship with the company, but their behavior indicates that they will migrate among providers, so further investigation is required to determine the true cause of their situation.

The final cluster group (cluster 5) are new customers, and their usage behavior and customer relationship should be improved. A high volume of advertising and promotion can help to motivate new customers, and offering customized packages, surprise gifts, and incentives can also help to attract them.

In order to determine the root cause of low usage behavior, the company should monitor and analyze these customer group metrics.

IV. CONCLUSION

As technology advances, the telecom industry can offer more products and services. This means that competitive pressure will increase, and if you want to compete, you must understand your customers' usage behavior and true need.

Customer segmentation is critical for comprehending various customer behaviors and determining which customer groups are most important to the company and which customer groups require special attention.

Following customer categorization, the organization should develop a distinct marketing strategy to address each cluster group to maximize customer value and retention rate.