UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

CUSTOMER SEGMENTATION ON MOBILE MONEY USERS IN KENYA

By

Asha Panyako. Makana

P52/11874/2018

Supervisor

Dr. Evans A. K. Miriti

RESEARCH PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT FOR THE REQUIREMENTS OF THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE, SCHOOL OF COMPUTING AND INFORMATICS, UNIVERSITY OF NAIROBI.

July 2020

DECLARATION

Researcher's Declaration

This project report is my original work and has not been presented in any other institution for the purpose of academic award. All sources, references, literature used or excerpted during elaboration of this work are properly cited and listed in reference to the respective source.

SIGNATURE________________________ DATE ________________________

Asha Makana Panyako

Registration Number: P52/11974/2018

Supervisor's Approval

This project report has been submitted in partial fulfillment for the requirements of the award of the Degree of Master of Science in Computational Intelligence in the University of Nairobi with my approval as the University Supervisor.

SIGNATURE________________________ DATE ________________________

Dr. Evans A. K. Miriti

School of Computing and Informatics

University of Nairobi

ABSTRACT

Customer segmentation enables organizations to partition a market into subsets that have common needs, interests and priorities. This helps businesses to come up with design and strategies that fulfills the customer needs. Usage of mobile money services has been widely adopted in Kenya with over 22 million customers using the service. Mobile money operators are mainly telecommunication companies leveraging on their infrastructure and customer base. Many businesses and individuals in Kenya use mobile money for services such as peer to peer transfers, payment services and financial services such as mobile banking. Previous segmentation of mobile money customers has mostly been done basing on the types of transaction or demographic factors as in banking such as age, gender, assets and location however this does not give a 360 degrees overview of the customer enabling an opportunity for improvement. Mobile network operators own big volumes of data which can enable improvements on segmentation models e.g subscriber's network activity i.e. calls, sms, data usage, demographic factors such as age and behavioral factors such as types and frequency of loans and payments. Leveraging on this data to identify different customer groups and their needs is important for service providers to respond to changing customer demands, cope with fast technological advancements and innovate around local market conditions.

Using network and mobile money data, this study compares various clustering algorithms aiming at identifying the algorithm that creates the most solid customer profiles. Hierarchical clustering, KMeans and affinity propagation algorithms were used to segment customers and compared using internal validation measures. Our dataset comprised of various demographic and behavioral features obtained from a telecommunications company data warehouse. Co-relation between the features was tested enabling us to focus on age, network revenue, amounts transacted on mobile money, frequency of loan uptake, customer and organization transfers, goods and service payments and deposits and withdrawals as our features for modelling. The dataset was then fit into our algorithms. Agglomerative clustering generated seven clusters with a normalized mutual score of 0.5526 and adjusted rand score of 0.5436 and silhouette coefficient of 0.4523. KMeans generated 11 clusters with an NMI score of 0.5168 and adjusted rand score of 0.3315 and silhouette coefficient of 0.2369. Affinity propagation generated the largest number of clusters of 504, had a memory utilization of 91% and took the longest time to execute. This established AP as unsuitable for our dataset. Agglomerative clustering had the best performance in terms of the compactness and connectedness of clusters however clusters obtained from KMeans were more granular as compared to agglomerative clustering segments.

.

ACKNOWLEDGEMENTS

I am grateful to God for enabling me to conduct my studies, research and compile this report. His grace has always been overwhelming.

I am extremely grateful to my supervisor Dr. Evans A. K. Miriti and panelists Dr. Wanjiku Nganga and Dr S.N. Mburu for their continuous guidance, frequent feedback and valuable time and advice throughout my project.

To my late father who has been the greatest enabler of my education for guiding me through the processes of life. I am also grateful to my mother, sisters and brothers for their support and encouragement all through my studies.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AP	Affinity Propagation
ATM	Automatic Teller Machine
B2B	Business to Business
B2C	Business to Customer
C2B	Customer to Business
CBA	Commercial Bank of Africa
CBK	Central Bank of Kenya
DI	Dunn Index
EAC	East African Community
ETL	Extract, transform and load
GSM	Global System for Mobile
KCB	Kenya Commercial Bank
KYC	Know Your Customer
MMT	Mobile Money Transfer
MNO	Mobile Network Operator
M-PESA	Mobile Money
MSME	Micro Small Medium Enterprises
NMI	Normalized Mutual Information
OD	Overdraft
P2P	Peer to Peer
PAM	Partitioning around Medoids
RFM	Recency, Frequency and Monetary
SIM	Subscriber Identity Module
SMS	Short Message Service
SQL	Structured Query Language
T-Kash	Telkom Money
WRFM	Weighted Recency Frequency and Monetary

1. CHAPTER ONE: INTRODUCTION

1.1 Background

Mobile Money services were first introduced in Kenya in 2007. This technology functions by storing money on the Subscriber Identity Module (SIM) which is used to recognize each subscriber. The SIM is placed in a mobile phone and corresponds to an account numbers in banking. Mobile Network Operator (MNO) use digits corresponding to the amount issued by the subscriber to keep the value on the subscribers SIM card. The amount value in the SIM card is then used for conducting transactions such as payments, transfers and withdrawals. However, the cash value is normally held somewhere as for the East African Community (EAC) the real amount is kept in a bank. Mobile phone devices are then used to conduct transactions which are limited by the value in the SIM card (United Nations, 2012).

Kenya and Philippines were the first countries ever in the world to offer mobile banking services. Safaricom, a telecommunication company in Kenya first launched their mobile banking solution in 2007 and named it M-PESA. This services which is one of the major economic drivers is relied upon by most of the market in Kenya and has been highly accepted by the customers. Philippines launched an electronic wallet known as SMART money through which users do most of their banking transactions (Maitai, 2016).

Traditionally, telecommunication companies were used to provide communication across their network using services such as SMS(short message services), voice calls and data for browsing the internet. However they recently introduced the use of mobile phones for money transfer services thus enabling them to be identified as Mobile Money operators (Munyange, 2012). There are several telecommunication companies in Kenya which also serve as mobile money operators. Telkom which started their mobile money transfer (MMT)system called T-Kash in March 2018, Airtel money by Airtel Kenya and M-PESA launched in 2007 by Safaricom.

The 2019 Global digital report states that there are 5.112 billion distinct mobile phone users in the world. Sub-Saharan Africa mobile market as compared to the rest of the world experienced the highest growth. Mobile phone users in the region grew to 456 million users in 2018 with a growth of 4.6% expected by 2025 (gsma, 2019). Most of the developing countries have highly adopted the use of mobile money services such as payments and transfers. This is because in most rural areas bank branches and other conventional banking channels, like ATMs (Automatic Teller Machines), internet banking and fixed-line phones are unavailable. In developed economies there is a high competition for investor and consumer commitment on mobile banking since it is considered as an additional channel among other vast options.

Since the birth of the M-PESA over a decade ago, Kenya has led the African continent in financial innovation with the mobile money service being used by two in five Kenyans. Given the country's love for M-PESA services, entrepreneurs are trying to leverage on the service to solve bigger economic problems. Safaricom's M-PESA is the most dominant player in Kenya on offering mobile money services. According to the 2019 financial report released by the Safaricom, the telco boasts of 31.8 million customers out of which 22.6 million are registered M-pesa users (Safaricom plc, 2019). Report by CBK at the end of financial year 2017/2018 indicated an increase in percentage of 2.68 and 4.84 in volume and value of transactions respectively. Amounts worth KSh 3.7 billion were transacted in financial year 2017/2018 as compared to 3.5 billion transacted in 2016/2017 (Central Bank of Kenya, 2018). Mobile money operators offer various services to their customers which can be grouped into three as follows;

(a) M-transfers: Commonly known as Peer to Peer (P2P) transfers that can be conducted locally or international. It involves transfer of money between users without trading of goods or services.

(b) M-payments: Entails trading of goods and services after money is exchanged via mobile between users.

(c) M-financial services: Bank accounts are connected to the users mobile money wallet to enable transactions such as savings and credits normally accessed at a bank branches. Other banking services such as insurance and micro-finance can also be accessed via M-banking services. (Safaricom plc, 2019).

Transactions can combine two or more service categories such as m-transfers and m-financial transactions e.g, a customer can use their bank account to move money to their mobile money wallet or another bank account without trading goods or services. Additionally, M-PESA transfers can be local or international. Western Union is an example of an institution that facilitates international money transfers through partnerships with mobile money service providers such as M-PESA.

Safaricom requires that customers are registered on M-PESA by M-PESA agents at zero costs. Once registered, the users can be able to deposit/withdraw money, transfer or trade money for goods and services through the Lipa na M-PESA service (Safaricom plc, 2019).

As the company continues to innovate around mobile banking, an additional set of service was recently added in January 2019 where customers are allowed to overdraft(OD) their mpesa wallets also known as ‘Fuliza’. This is an addition to existing loaning services running on the platform such as Mshwari and M-KCB loans through bank partnerships.

Mobile money platforms are networked systems, this refers to systems that often follow Pareto principle where a small group of users is responsible for most of the activity happening in the system. Comprehending key mobile money users is a necessity for mobile money providers, promoters of financial inclusion and policy makers. From the services offered by MNO’s their target customers can either be MSME’s using the service for paying employee wages, paying suppliers or receiving payments from customers or individual customers whose wallets are mostly for sending or receiving money over the network, withdrawals from mobile money agents, bills and services, payment of goods and for loans and savings (Mattsson, 2018). In order to serve their customers better, it is necessary for MNO’s to identify various customer segments within their customer base. Customer segmentation enables organizations to address distinct customer needs and preferences which is also a key marketing strategy for the marketing departments.

Given the rapidly evolving and unstable marketing environment, companies are facing severe competition. To be successful and leaders in the market place, they must provide quality services that address changes in their customers’ demands, wishes, characteristics and behaviors. Thus, instead of viewing the entire the customer base as homogenous and engaging all customers in similar campaigns or marketing incentives, companies need to approach customers differently, depending on their needs, characteristics and behaviors (Bose, 2010).

Given the lack of business-specific digital financial services that are offered, MSMEs and individual customers often use the very same services for business and personal needs. That means that many person-to-person (P2P) transactions may in fact be business transactions. Despite using common products, these segments are likely to transact differently, with different service needs. Data-driven analytics can identify these different usage patterns to inform which users belong to which segment (Buri, 2019).

The ability to identify different groups of mobile money users, their activity on the network, the kind of product they extensively use, and the nature of their influence is key to driving business and innovation in the mobile money sector.

1.2 Problem statement

Segmentation refers to the process of identifying groups of customers, understanding and evaluating their requirements, and defining their profiles. Operators can only select a target market after identifying the unique profiles of their consumers.

Customer segmentation allows companies to partition a market into subsets of consumers that have, or are perceived to have, common interests, needs and priorities then analyze, design and implement strategies with a specific target toward them.

The goal is to eliminate generalized segmentation and individualize everything from the products you offer, to how you offer them and to when you offer them based on actual rather than perceived needs and preferences using customer data.

Majority of bank marketing use segmentation by demographic factors though the correlation with the evolving customer needs due to technology changes is weak. Adding customer behavior and attitudes to the traditional demographic factors could enhance the bank's capacity to address the antagonism between individual service and cost-saving standardization (Kamande, 2018).

Segmentation of mobile money customers has been done majorly on the type of the transaction conducted leading to segments such as acceptors, bulk senders, cash in/cash out, airtime traders and service providers/agents (Buri, 2019). Just as other segmentation models done in relation to the product or category of products, this falls short of getting 360 degrees view of the customer. Whereas most segments are created in regard to demographic factors such as gender, location, age, occupation and education other factors influenced by the advancement in technology are coming up which commands a change in the customer behavior. Psychographic factors such as lifestyle, interests, attitude and social network were not previously considered in traditional segmentation models are a major influencer in the customer behavior and must therefore be put into consideration during segmentation.

MNOs collect large amounts of data from their network and mobile money systems. Leveraging on this data to identify different customer groups and their needs is essential if service providers want to meet rapidly evolving customer demands, adjust to local market conditions and grow with shifting technological advancements. This way MNO's can be able to come up with innovative pricing models, relevant product, good customer services and targeted marketing.

1.3 Objectives of the study

This research's main objective is to analyze and evaluate mobile money data to come up with homogeneous customer groups based on a combination of behavioral factors, demographic factors and psychographic factors. The specific objectives are as follows:

- To determine variables that affect customer segmentation in mobile money operators.
- To identify mobile money customer segments through analysis of distinct customer characteristics among each group.
- To compare clustering algorithms performance statistics to establish the most accurate and efficient clustering algorithm.

1.4 Justification

This study focusses on mobile money customer segmentation. Since inception in 2007 in Kenya mobile money customer base in the telecommunication marketplace has experienced significant growth with M-PESA having a total of 22.6 million registered customers. In traditional banking methods, only demographic factors are considered during customer segmentation. However, this cannot be enough to segment dynamic mobile money users provided the various type of services and channels offered by mobile money. Behavioral, lifestyle, attitude and social network factors need to be considered during segmentation.

The mobile telecommunication industry is very competitive. The mobile network operators are required to design an innovative marketing strategy leveraging on different behavior of their customers to raise their marketing results and revenue. Currently telecommunication companies have in possession call detail records describing customer utilization behavior, mobile money transfer, payments and banking records describing financial transactions of a customer. Clustering analysis based on this information can give more customer insights for both revenue and customer growth.

Customer segmentation enables organizations to: Utilize marketing resources efficiently, target relevant products to distinct consumers and to stay updated on emerging market trends. Using segmentation, we can address to the needs, demands and interests of different customer groups, we can establish whether there is a product or service that fit in segments which act as high opportunity areas, we can evaluate whether product enhancements or creating new products might please the targeted groups. Segmentation often improves average profitability and retention of customers on a cluster by cluster basis. Segmentation also helps to execute personalized marketing plans for each intended segment.

2. CHAPTER TWO: LITERATURE REVIEW

2.1 Customer Segmentation

Customer segmentation is described as the practice of grouping customers into distinct classes. Market segmentation plays a fundamental role in the companies' strategic market planning because goods and services need to be evaluated by taking into consideration the customers' needs and wishes and the fact that different customers have different needs before being produced and retailed.

During segmentation customers are grouped into homogeneous classes based on shared or common attributes (Manero, 2018). The aim of clustering is to understand the customer better, identify the most beneficial consumer segments, enable proper utilization of resources, conception of individualized campaigns and incentives, better positioning of products and services for the consumers and against the competition and creating of value offers. At operational level it steers the companies to lay more emphasis on enhanced customer understanding and to create more efficient relationships with them. Segmentation entails collection, organization, evaluation and analysis of customer data. With proper customer grouping, it is easier to identify the loyalty or reliability of customers to increase the organization's revenue. Through customer profiling companies can be able to split their customers into groups with common characteristics or interests. In most cases customer segments are created based on the below factors;

- Demographic information, such as gender, age, educational level and income.
- Geographic segmentation, which entails geographic position and population density.
- Behavioral segmentation comprehends every factor that explain customer's behaviors from usage rates to expected benefits.
- Psychographic characteristics, e.g. lifestyle, values and interests (Kotler, 2009).

According to Rajagopal et al(2014) there are two main segmentation approaches: the most commonly used segmentation is the segmenting of customers by understanding the needs of the customer which is known as needs-based segmentation. Another type of segmentation is, characteristics-based segmentation, which entails segmenting customers according to their characteristics, behaviors or attitudes.

Majority of the organizations use segmentation based on demographics. In high competition markets such as the telecommunication market, this method is not enough. Companies also need to put into consideration customers' demands, service or payment inclination, consumer patterns and behavior, perception of the product, growth potential and customer migration and probability of leaving the network (Mihai, 2012).

Majority of the banking institutions banks segment customers by traditional methods using demographic and geographic factors such as location, age, gender, occupation and financial variables, like asset levels, credit rating and liabilities. This method of segmentation is far from enough to develop deeper insights needed to understand customers better. In most situations, this information is basic and has little to no correlation with the real needs of customers, who "are much more than the cumulation of their banking deposits, credits and loans, and do not align well to basic or larger demographic characteristics" (Ernst & Young, 2018).

2.2 Traditional methods of customer segmentation.

Previously, clustering models were built around basic and unnuanced, high level demographic data. The general focus was on age, income, education and location to enable an understanding of the customer, however with the advances in technology these methods fall short as more sophisticated data sources such as social media, chatbots and e-commerce are taking over. Without understanding customer preferences and behavior, banks tend to offer wrong products, use inappropriate delivery channels and ultimately fail to meet their customer needs (Afande, 2015).

Studies have shown that four variables are used to segment consumer markets: demographic factors, geographic variables, psychographic elements and behavior variables. Geographic elements refer to the influence of terrain, climate, size of the city, population density and urban/rural regions on customer product demands. Meanwhile, demographic features such as gender, age, race, education, profession, income and marital status are the most commonly used. Behavior variable can be observed by individual attitudes towards a product, purchase time, user profiles, customer benefits, usage rate, loyalty, adoption stage, and so on. Understanding one of the psychographic variables, such as lifestyle variables, we can predict customer's psychology on purchase (Andronikidis, 2008). Studies have also referred to the effect significantly of lifestyle variables on customer's decisions on purchase. Since lifestyle variables are considered more efficient for outlining consumer characteristics and influence customer psychology than other elements, marketers can use lifestyle variables to better predict consumer's psychology on purchase.

2.3 Clustering techniques for customer segmentation

Clustering refers to a Machine Learning technique that entails grouping of data points into similar groups. Given a set of data points, a clustering algorithm can be used to classify each data point into a unique group. In theory, data points found in the same group are supposed to have similar properties and/or features, whereas data points in other groups should have highly unique properties and/or features. Clustering is a common method of unsupervised learning and is a statistical data analysis technique used in many fields (Anon, 2019).

Clustering is a highly regarded subject in data mining. During clustering, the data set is partitioned into some segments and the data points in each segment, that is, the cluster have similarities with each other than to those in other groups. These data points are grouped together by detecting relationships according to the variables found in raw data.

The goal entails finding the suitable number of clusters which are significant and insightful for analysis and evaluation purposes. This process is a continuous and iterative task where huge amounts of raw data are evaluated for similarities, relations and patterns. The uncategorized data is scanned for knowledge that is relevant and then data points are assigned.

There exist different clustering algorithms that differ from every other in regard to the approach they follow to do the clustering of the data points according to their characteristics (Shreya Tripathi, 2018).

- Hierarchical Clustering has two methods, agglomerative (bottom up) and divisive (top down) approach, one of these two present methods can be used for implementation.

- Grid Based Clustering algorithms use the approach of partitioning the data points into grid structures with several cells. This method uses subspace and hierarchical clustering approaches. STING and CLIQUE are some of the grid-based clustering algorithms.
- Partitioned Based Clustering, at the beginning all the data points are taken as a single cluster for this approach. These data points are then grouped into clusters by iteratively aligning these objects between the clusters. K-Means, K-Medoids and K-Modes are examples of the partitioning algorithms.
- Density Based Clustering in this approach, the clusters are identified as regions of higher density than the other parts of the dataset. Objects differentiators are core, noise and border points.

2.3.1 Clustering algorithms

K-Means Clustering

K-means is one of the most commonly-used centroid-based algorithm for clustering (Shreya Tripathi, 2018). Centroid-based clustering algorithms are efficient but highly sensitive to initial conditions and outliers.

With K-means the user is required to first select several classes/groups to use and initialize their respective center points randomly. In order for one to figure out the initial classes to use, it’s required to take a quick look at the data set and try to identify any unique groupings. Each data point is clustered by calculating the distance between that point and each group centroid, and then classifying the point to be in the cluster whose center is nearest to it. Based on these classified points, the group center is recomputed by taking the average of all the vectors in the group. These steps are iterated for a defined number of iterations or until when the group centers remain the same between iterations.

K-means has an advantage of a fast compute time, effectiveness and has the ability of dealing with large amounts of data. However, it presents a challenge where the user has to initially set the number of clusters before computation and it also starts with a random guess of cluster centers and may therefore yield different clustering results on each run of the algorithm. This makes the results to lack consistency and may not be repeatable. Other cluster methods are more consistent.

K Means ++

K-means sensitivity to parameter initialization of the centroids or the mean points poses a disadvantage to the algorithm. So, if a centroid is set to be a “far-off” point, it could end up with zero points associated with it and additionally on the other hand more than one clusters might end up associated with a single centroid. Similarly, in some cases greater than one centroid might be grouped into the same cluster resulting into bad clusters.

To solve the above-mentioned limitation, we use K-means++. This method ensures a smarter initialization of the cluster centers and enhances the quality of the clustering process. K-means++ is similar to the standard K-means algorithm apart from parameter initialization. That is K-means++ is the standard K-means algorithm with an added advantage of smarter initialization of the centroids.

The steps involved are as follows:

- Randomly select the initial centroid from the data set.
- For every data point compute the nearest distance from the previously chosen centroid.
- Thirdly, from the data points select the next centroid such that the possibility of choosing a point as the centroid is proportional directly to its distance from the nearest, previously selected centroid. (i.e. the point with the maximum distance from the closest centroid is most likely to be chosen next as a centroid)
- Repeat through steps 2 and 3 until k centroids have been evaluated.

Affinity Propagation Algorithm

The Affinity Propagation algorithm was published in 2007 by Brendan Frey and Delbert Dueck in Science. In comparison to other traditional clustering algorithms, there is no need to specify the number of clusters prior in Affinity Propagation. In other words, in AP, every data point shares messages to every other point informing its targets of every other target's relative attractiveness to the message sender. Each target then sends answers to all senders with a message informing each sender of its availability to relate with the sender, given the level of attractiveness of the messages that it has gotten from all other senders. Senders respond to the targets with messages informing each target of the target's revised level of attractiveness to the sender, given the availability information it has received from all targets. The message-exchanging process proceeds until a consensus is reached. When a relationship is built between the sender and with one of its targets, that target becomes the point's exemplar. Every point sharing the same exemplar is allocated to the same cluster (Frey, 2007).

The following matrices are computed during affinity propagation:

Similarity matrix

Similarity matrix(S) shows us information on how similar instances are to each other. The euclidean distance can be used to compute the similarity between two instances. The bigger the distance between any two instances, the least the similarity between them.

$$s(i,k) = -\|x_i - x_k\|^2$$

Responsibility matrix

Responsibility r(i , k) quantifies how suited element k is as an exemplar for element i , taking into account the closest contender k' as an exemplar for i.

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t\ k' \neq k}\{a(i,k') + s(i,k')\}$$

The intuition behind formula: $r(i,k)$ can be thought of as relative similarity between i and k. It quantifies how similar is i to k, compared to some k', taking into account the availability of k'. The responsibility of k towards i will decrease as the availability of some other k' to i increases.

Availability matrix

Availability $a(i,k')$ quantifies how appropriate is it for i to select k as its own exemplar, considering the support from the other elements that k should an exemplar.

Criterion matrix

Criterion matrix is calculated after the updating is terminated. Criterion matrix C is the sum of R and A. An element i will be assigned to an exemplar k which is not only highly responsible but also highly available to i.

Hierarchical Clustering

Hierarchical clustering works by creating a tree of clusters from bottom up or top down. This type of clustering is well suited for clustering hierarchical data, e.g. taxonomies. There exists two approaches of hierarchical clustering, Agglomerative clustering and Divisive clustering.

Divisive method

In *divisive* also known as *top-down clustering* approach we assign all the elements of a data set to a single cluster and then split the single cluster to two most dissimilar clusters. We proceed iteratively on each cluster until the point where there is one cluster for each element. Evidence that divisive algorithms create highly accurate hierarchies as compared to agglomerative algorithms in some cases exist however it is conceptually very complex.

Agglomerative method

In agglomerative also known as bottom-up clustering approach we assign each element to its own group. We then compute the similarity such as the distance between each of the clusters and merge the two highly similar clusters. Finally, steps 2 and 3 are repeated until only one cluster is left.

This clustering approach has an advantage since it allows clusters to grow 'following the underlying manifold' rather than clusters being presumed to be globular. It also allows evaluation of the dendrogram of clusters for users to get more information about how the clusters are broken down.

2.4 Related work

(Mihai, 2012), carried out market segmentation based on customers expenditure on credit recharging, voice calls, on messaging and on internet activity. K-mean cluster analysis method was used for subscriber profiling. Their clustering revealed seven customer groups with distinct features and behaviors. The obtained results were used by a telecommunication company to inform marketing strategies that resulted into a better understanding of its customers' needs and eventually yielding to better relationships with the subscribers and improved customer satisfaction. The obtained results also enabled the evaluation and profiling of utilization patterns for services. Furthermore, this research demonstrated that K-means algorithm is efficient for a large customer base. This study only focused on the monetary aspect of the customer by only considering the following features; the cumulation of the amounts used in 6 months, the revenue of the messages sent for a duration of 6 months, the browsing activity value in the 6 months and the revenue from calls made during the 6 months. However, this does not give 360 degrees view of the customer, factors such as the social network, age, location and frequency should have been considered.

A study on consumers expenditure in Kenya done by collecting and analyzing data collected through daily mobile conversations with consumers in Kenya (Kamande, 2018). This study was used to compare different clustering algorithms to determine the method that best segments consumers, and afterwards produce profiles that provide a foundation for marketing and brand strategy using existing demographic data i.e. gender, age, location and main income source. K-Means, partitioning around medoids (PAM) and hierarchical clustering methods were compared using cluster validation techniques. Hierarchical clustering produced four clusters which in comparison with other methods had the best performance. In their study they used additional profile descriptors (customer's spending habits) to establish a deeper understanding of the customer clusters created on expenditure data in Kenya. One of the challenges faced during this study was the issue of missing data which lead to omission of incomplete records. Previous studies also show the inefficiency of hierarchical clustering when dealing with missing data. The algorithm does not break large clusters well, does not handle large volumes of data well and it sometimes poses a challenge of pointing out the correct number of customer groups by the dendrogram.

A study conducted by (Ustundag, 2016), indicated that the weights of RFM (recency, frequency and monetary) features affect rule association results positively during clustering. To create more solid customer segments, they recommended a mixture of WRFM and demographic features. When tested the proposed model had the best outcomes and profiles therefore validating the idea of combining demographic data and weighted recency, frequency and monetary during clustering. The results analysis despite using limited elements of demographic data as variables, indicated the undeniable relevance of demographic factors combined with RFM data. By use of more solid and accurate rules, marketing officers can create more targeted and result oriented campaigns based on valid customer segments.

(Seyed, 2017), introduced a new clustering algorithm named autocluster. Autocluster has combined advantages of partitioning methods (easily can be understood and applied), density-based methods (creating micro-clusters) and agglomerative hierarchical algorithms (developing clusters from micro-clusters). Autocluster has the capability of clustering data records in the true clusters without knowing the number of clusters. Additionally, Autocluster is a robust algorithm which is not sensitive to selecting first data record (in every analysis of a special dataset, it would select data records in a special order, without randomness). Autocluster can be applied properly for customer segmentation problems.

Establishing the similarity and dissimilarity of customer attributes is an essential element of cluster analysis and validation which most of the time requires a lot of knowledge on the context and creativity over what statistical tools can produce. (Singh, 2017), applied K-means clustering algorithm, and the hierarchical clustering approach. They chose the two methods because K-means expects the user to initialize the number of clusters to create, while hierarchical clustering approach does not. Hierarchical clustering was executed first so that they could establish the possible number of clusters which then informed the choice of k in KMeans clustering algorithm. Dendograms in hierachical clustering approach are used for visualization which shows the resulting number of classes. This can however be misleading. From the moment two data points are grouped into the same cluster they retain the same cluster throughout the clustering tree. Hierarchical clustering has a tendency of rigidity which sometimes lead to segmentations which are not optimal in many forms. Therefore, it is advisable to construct the dendogram using various methods, to not only focus on distance metrics but also looking at how the data in the clusters has been aggregated.

2.5 Gaps in literature review

KMeans clustering algorithm is a widely used clustering method because of its efficiency when it comes to computation time, simplicity and the ability to handle large volumes of data. However, it faces a limitation when it comes to parameter initialization i.e. the expected number of clusters and the initial centroid prior to the algorithm execution. K-means ++ was developed as an improvement to K-means algorithm. Whereas K-means initial step is randomly identifying cluster centers then proceeding to search for better centroids, K-means++ starts with randomly selecting one centroid followed by searching for other centers in reference to the first one. KMeans random initialization take less time than KMeans++ but gives poor result. Because of random initialization KMeans faces a risk of getting stuck in a local optimum when the first set of centroids are not well distributed over the data set. KMeans++ centers are distributed over the data it is more likely to have less cost (within cluster sum of square) than random initialization.

Affinity propagation algorithm in contrast to KMeans doesn't need prior initialization of cluster numbers. The algorithm relies on passing messages between data points to determine the similarity, responsibility, availability and the criterion matrix for clustering. Research done by (Serdah, 2016), shows that AP algorithm performed better than K means in terms of accuracy.

Most banking institutions focus on demographic factors for customer segmentation. However, study on factors affecting customer segmentation (Andronikidis, 2008) indicates that considering age, gender, location, occupation and asset value does not show a correlation with actual customer needs. Psychographic factors such as lifestyle variables i.e. purchase, and payment habits describe the customer characteristics more. Other behavioral factors such as frequency and activity on the network should also be included during segmentation to provide 360 degrees view of the customer.

Customer segmentation done on mobile banking users is highly focused on the type of transaction being conducted by the customer, this has resulted in more generalized clusters such as acceptors, service providers, airtime traders, cash in/cash out agents and bulk sender agents (Buri, 2019). This clusters are more generalized with a main focus on the transaction type. A combination of demographic factors, psychographic and behavioral clusters need to be done to establish more exhaustive clusters that focus on characteristics portrayed by the customer.

2.6 Description of proposed solution

Customer segmentation refers to the process of partitioning a customer base into unique groups of individuals who share similar characteristics in ways that are relevant for business development. The goal is to identify customers with similar characteristics. The proposed solution will involve a combination of mobile money transactional data (P2P transactions, C2B transactions, B2C transactions and B2B transactions) via various channels with KYC details (age, gender, location) and GSM data (Voice, sms, data) to come up with the various clusters using three algorithms, K-means, affinity propagation and agglomerative clustering. Clusters from the three algorithms will be analyzed and compared to identify the most accurate algorithm.

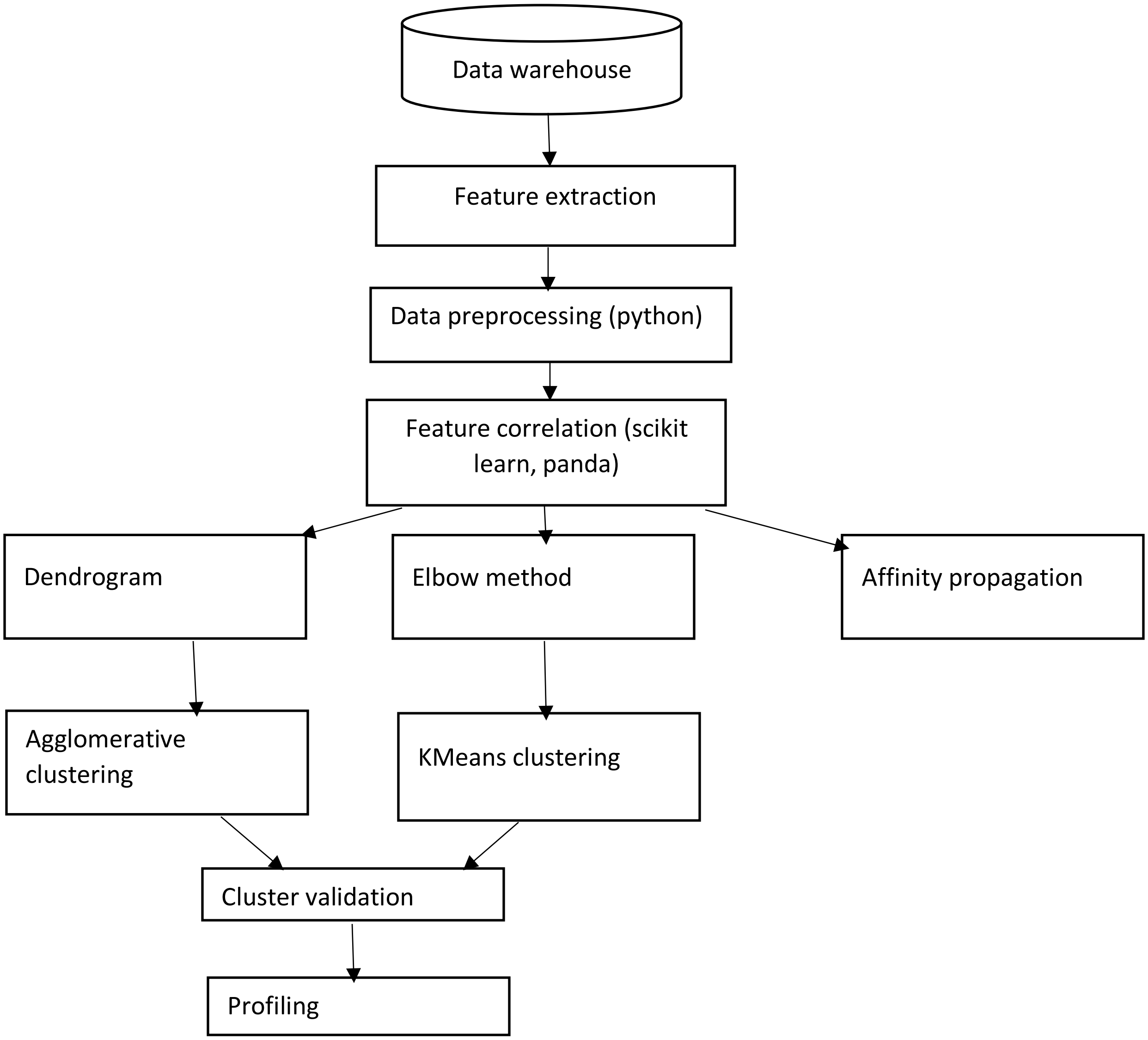

Figure 1:Proposed clustering process

Data Warehouse

The term data warehousing implies the act of collecting and managing data coming from varied sources to analyze and create meaningful business insights. Most organizations get their data from multiple systems, thus using preferred ETL (extract, transform and load) tools data is collected and stored in the data warehouse. Business intelligence (BI) systems rely on the data warehouse for data analysis activities and reporting for various organization units. In our case, the data warehouse consists of customer data from the customer registration system, GSM data from the network and the core billing system and mobile money transactional data from the Mpesa system. The customer data consists of KYC details such as customer name, age, registration location, gender and customer registration number. GSM data consists of call records, data usage, sms and mms details. Mobile money transactional data consists of P2P transactions, C2B transactions, B2C and B2B via various channels.

Data extraction

The data will be extracted from the data warehouse using PL/SQL stored procedures and stored in form of csv files. The extracted features consisting of demographic and psychographic factors include age, location, gender, name, number of calls, sms, mms, recharge, loan and mpesa transactions for a duration of three months, the amount of money consumed on GSM and Mpesa transactions for a duration of three months, count and value of different mpesa transaction types. Current Mpesa transaction types will also be broken into more granular types to be able to identify various types of payments done by the customer e.g. supermarket payments, ecommerce, bank transactions, bill payments etc.

Data preprocessing

Data preprocessing refers to the process of cleaning and preparing data features for modelling. Some of the processes conducted during data preprocessing include;

- Taking care of missing data in the dataset, in our case all instances of missing data will be replaced with nan.
- Encoding categorical data. Transaction types, gender and location will be encoded.
- Feature scaling. This is a fundamental step in data preprocessing. It normalizes features to a range of values between 0 and1. Additionally, feature scaling assists in speeding up the computation time in an algorithm. In our study, amount and frequency figures will be standardized to a range between 0 and 1.
- Removal of duplicate records from the dataset.

Clustering

During the clustering process we aim to maximize on the advantages of K-means and use K-means' elbow method approach to determine expected cluster numbers prior to algorithm execution. Affinity propagation and agglomerative clustering algorithms as discussed in the above sections will also be used to create heterogenous customer clusters. The three algorithms will then be analyzed and evaluated to determine the most efficient and accurate algorithm.

Cluster analysis and validation

Cluster analysis refers to the process of evaluating how good the resulting clusters are.
Some of the factors evaluated during cluster analysis are;

- Cluster number: refers to how many clusters will be obtained from each algorithm.
- Cluster size: number of points in each segment, how large or big a cluster is.
- Average and median distance: value of average or median distances within a cluster.
- Average between: Implies the mean distance between clusters should be as big as possible.
- Average within: Implies the mean distance within clusters should be as small as possible.

For our cluster analysis we will use three internal cluster validation measures namely normalized mutual information, adjusted rand score and silhouette coefficient.

3. CHAPTER THREE: RESEARCH DESIGN AND METHODOLOGY

3.1 Introduction

Chapter three outlines the research methodology applied in this study. It outlines the quantitative approach of this paper. The researcher outlines the data collection methods and information required to fulfill the goals of this research, the design of the research, the research data and population, data collection measures and tools used to collect the data, data preprocessing methods used to prepare and clean the data for evaluation and the proposed data analysis methods.

3.2 Quantitative approach

Research design refers to a framework of techniques and approaches identified by a researcher to combine different attributes of research in a logical manner to efficiently handle the research problem. This step helps the researcher to establishes the plan for gathering, analysis and evaluation of the collected data. It is necessary to have a guideline on how the research questions and the research objectives would be responded to and how the tracking of stages of the research will be done. It evaluates the research purpose, methods and approaches and the time limit. Thus, the research design answers questions on what data needs to be collected and how data collection and analysis should be done.

Research designs for the quantitative approach are either descriptive where subjects are measured only once or experimental where subjects measured before treatment and after treatment.

In this study, we used the experimentation methodology to determine mobile money customer segments. Experimental research is a type of study conducted using a scientific approach, during the research a set of features are kept constant whereas the other set of features are measured as the experiment subjects (Oates, 2006). In situations where the data is not enough to support your decisions, experimentation can be used to unearth the facts. This type of research can generate a lot of information that can enable better decisions. Experimental research reveals a cause and the impact of a phenomenon which implies that effects of a phenomenon are established from an experiment in reference to the cause.

We used the experimental design to:

Choose features, to determine variables that were most relevant to a model given the dataset.

Compare different models, through experimentation we compared the performance of our clustering algorithms using algorithm validation measures.

3.2 Research data

According to the 2019 financial report released by the Safaricom PLC, the telco boasts of 31.8 million subscribers out of which 22.6 million are registered M-pesa customers (Safaricom plc, 2019). With this huge subscriber base, the company collects terabytes of data per day consisting of the subscriber's GSM and mobile money data.

3.2.1 Data generation

This research relies on subscriber data collected from the company's data warehouse. The data consists of call, sms, mms, recharge and registration records and mobile money transaction data from the M-PESA platform. This data describes the subscriber's daily activity on the network including the time of activity, value, duration, location, frequency and payments. This data was collected from the company's various data sources i.e. the billing system, customer registration platform, application platform and the mobile money system. The data was then extracted, transformed and loaded into the data warehouse using informatica and streamsets. Out of the 31.8 million subscribers, we randomly selected a subset of approximately 12,000 mobile money subscribers for our study. This is because of the limited processing and storage capacity and time limits.

3.3 Experimentation environment

Anaconda Distribution open source machine learning environment provides an efficient platform to create data science python or R scripts on all types of operating systems. More than 15 million developers worldwide currently use the anaconda environment. The open source environment enables data scientists to:

- Easily download python and R data science packages
- Download, import and manage libraries, dependencies, and environments using Conda.
- Gives data scientists the scalability and efficient performance of NumPy and pandas etc.
- Create and train machine learning algorithms and deep learning models with scikit-learn, theano and tensorflow.
- Matplotlib provides the ability of data visualization and graph plotting.

We used the anaconda environment on Windows operating system using python programming language for data modelling and algorithm testing and analysis leveraging on the inbuilt libraries as described below.

3.4 Performance evaluation

Cluster validation techniques can be grouped into 4 classes as below;

Relative clustering validation is used for deciding the ideal number of clusters. This method varies parameter values for the same algorithm such the number of until the optimal value is obtained.

External clustering validation entails comparing cluster analysis results of a given algorithm to external results. This technique is mostly used for choosing the best algorithm for a given dataset.

Internal clustering validation methods are used to estimate the optimal number of segments and to determine the best clustering algorithm without any external cluster labels. Internal validation measures evaluate the correctness of clusters by analyzing the internal information obtained during the clustering process.

A special version of internal validation which does a clustering stability validation is used to assess the consistency of a result obtained during clustering by validating against clusters obtained after features from the data set are varied, one at a time.

We used both internal clustering and external clustering validation techniques to assess our clusters. The aim of internal cluster validation is to establish that the data points found in the same group as much as possible similarities and the data points in dissimilar groups are very distinct i.e. for objects in the same cluster, average distance should be as minimum as can be and for distinct clusters, the average distance should be the greatest possible. In most cases, clusters are evaluated on the connectedness, solidity and the separateness of each cluster partitions.

For our cluster analysis we used the below internal cluster validation measures.

Normalized Mutual Information (NMI)

NMI is a measure of the mutual connection between the two attributes. Normalized Mutual Information scales cluster validation results between 0 which indicates no mutual connection and 1 for a perfect score of correlation.

Adjusted Rand Score (ADR) calculates the similarity measure connecting two clusters. Adjusted rand score evaluates all pairs of samples then counts pairs located in the same clusters or different clusters in the true clusters and the predicted clusters.

The Silhouette Coefficient is a metric for judging the similarity of instances in similar clusters and the difference with instances from other clusters. Silhouette coefficient ranges between negative 1 for bad clustering and positive 1 for very compact clustering. Close to zero scores stipulate overlapping clusters. Dense and properly separated clusters generate a higher score which speaks to the standard idea of a cluster.

4. CHAPTER FOUR: RESULTS AND DISCUSSION

4.1 Introduction

Chapter Four illustrates the outcomes from this research in relation to the main objective of analyzing mobile money data to come up with homogeneous customer groups as discussed in chapter one. Three clustering algorithms were tested iteratively evaluating different sets of variables to determine the best cluster outcomes. Performance of the algorithms is a measure of how well they can split the dataset into dissimilar groups (Kamande, 2018), therefore we compared the ability of different cluster algorithms to split mobile money customers basing on age, behavior, mobile money frequency of usage and value and their usage on GSM activities.

4.2 Exploratory data analysis

The dataset collected from the data warehouse comprised of 12,350 distinct mobile money customers for the months of November 2019 to January 2020. The features in the dataset that have been used in this research are

- Age
- GSM gross revenue
- Mpesa gross amounts
- Count of organization credit transactions
- Count of customer credit transactions
- Frequency of cash in transactions
- Frequency of cash out transactions
- Count of organization to organization transfer
- Count of organization transfers
- Count of B2C transactions
- Count of customer to organization transfers
- Frequency of pay bill transactions
- Frequency of merchant payment transactions
- Count of send money transactions
- Count of cash in transactions
- Count of cash out transactions
- Count of financial institution payments

The scatter plots below show how different features are distributed across the entire dataset.

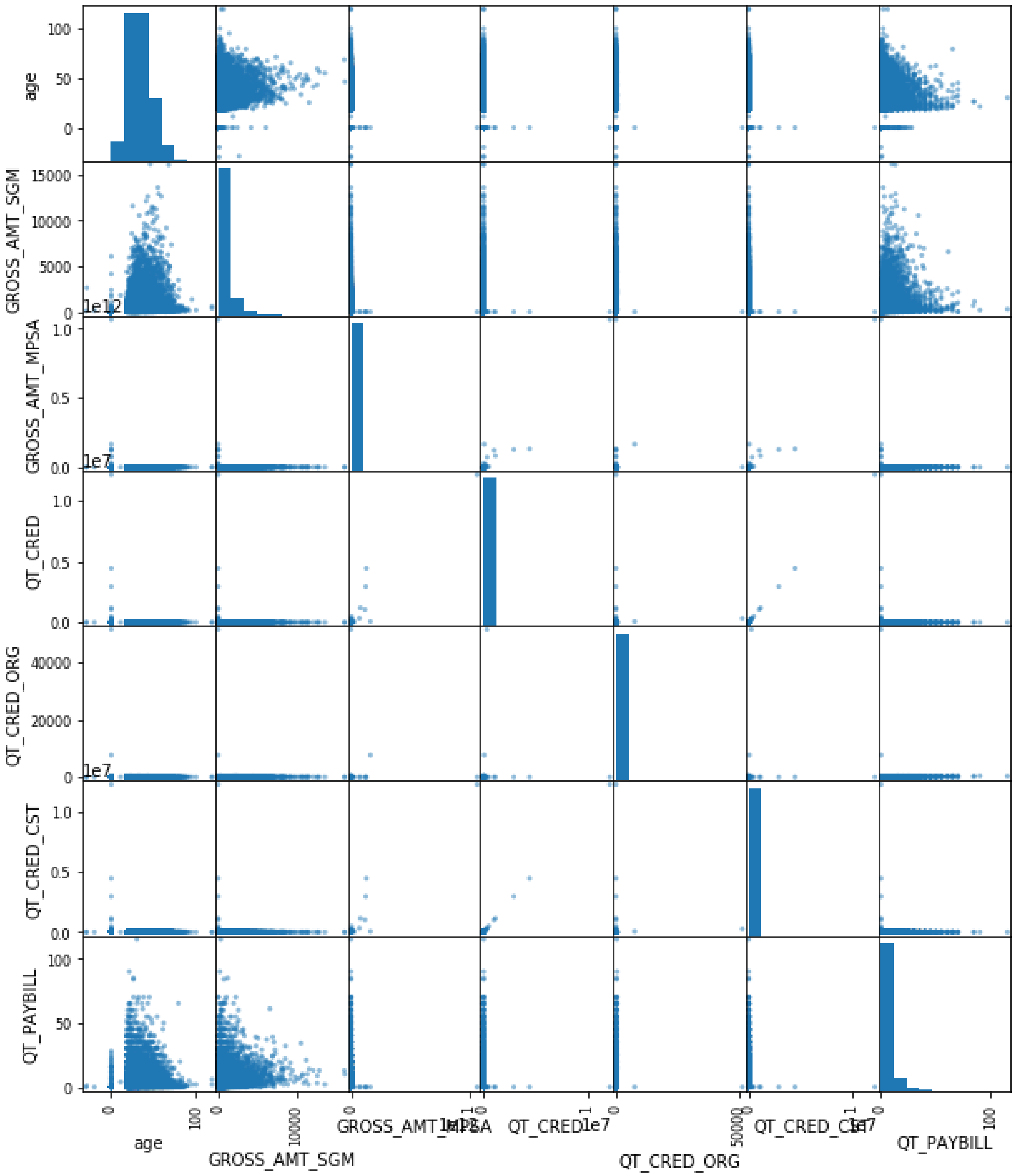

Figure 2: Feature distribution

4.3 Clustering algorithms

Hierarchical clustering

As discussed in the previous chapters hierarchical clustering creates a tree of clusters, in this study we used agglomerative clustering which creates clusters in a bottom to up manner. Each object is at the beginning regarded as a single element cluster just as a leaf in the case of a tree. During every step of the hierarchical clustering algorithm, two of the clusters that are considered to be highly similar are merged into a new larger cluster just as tree nodes. This procedure is repeated until all instances are part of just one big cluster as a tree root. This method results into a tree like graph which is then drawn as a dendrogram as in the diagram below.

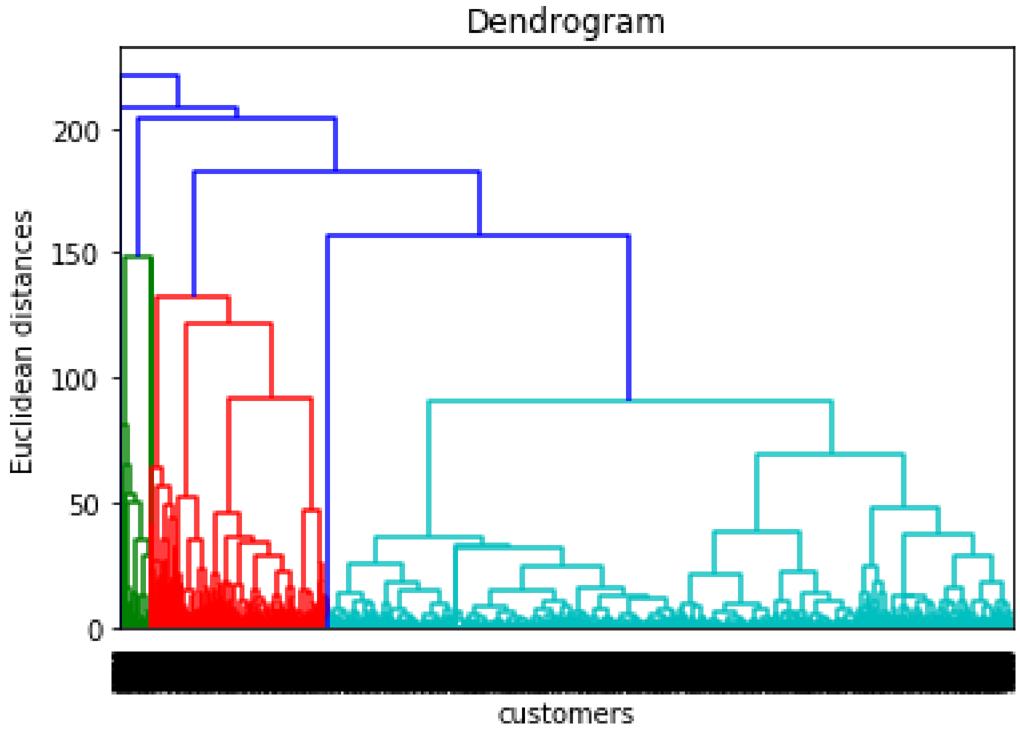

Figure 3:Dendrogram

Each leaf corresponds to one instance from the dendrogram shown above. When you climb up the tree, instances that share similarities are merged into branches, the branches are then fused as you go up. The variation between two instances can be observed from the height of fusion given on the vertical axis. The longer the height of fusion, the more dissimilar the instances are. From the above diagram we observed the highest height of fusion in the region between 100 and 150, giving us 7 clusters.

KMeans clustering

KMeans as described in previous chapters is an example of unsupervised algorithms that partitions data into k number of clusters. With KMeans clustering the user is required to choose prior to the algorithm execution what value of k to use, KMeans is considered as a naïve algorithm.

For this study, to select the ideal number of clusters in relation to the dataset, we used the elbow method. This technique runs the k-means algorithm on the dataset for different possible values for k. For each value chosen as k it calculates an average score given all clusters. The distortion score value is then calculated which is the cumulation of square distances starting from every point to its allocated center.

It is visually possible to select the most appropriate value of k when the overall scores for each model iteration are plotted. The line graph should resemble an arm. The point of inflection as seen in the graph below, also known as the elbow indicates the optimal value of k.

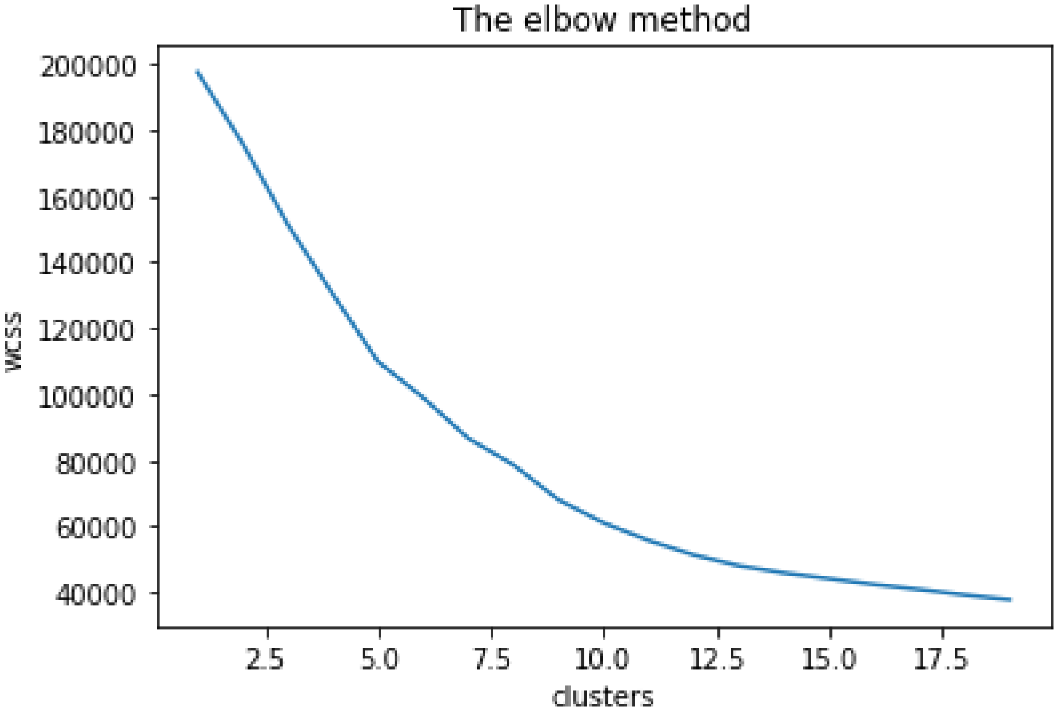

Figure 4:The elbow method

From the above graph, the point of inflection is at k=11 showing that the best value of k from the dataset is 11.

4.5 Algorithm evaluation and comparison

With the increased need by businesses for data driven insights, most companies are adopting data mining techniques to understand their customers better. There exist numerous clustering algorithms for customer segmentation each with its pros and cons depending on complexity, performance and ability to adapt to different datasets. For our mobile money customer segmentation research, we evaluated three algorithms; KMeans clustering, agglomerative clustering and affinity propagation using internal and external validation techniques as shown below.

Overview of clustering algorithms

Algorithm	**Parameters**	**No. of clusters**	**Memory utilization**	**Time taken(mins)**
Agglomerative	Number of clusters, distance, linkage type	7	50%	30
KMeans	Number of clusters	11	49%	30
Affinity propagation	Preference	504	92%	110

Table 1:summary of clustering algorithms performance

From the above results, affinity propagation produced a large number of clusters, had a high memory utilization and took the longest time to segment our dataset. We therefore considered it unsuitable for our dataset. Previous studies have shown affinity propagation as a suitable method for biology and on computer vision problems for example clustering images of human faces and recognizing regulated transcripts (Brendan, 2020).

Comparison by internal and external validation

Algorithm	**Validation method**	**Score**	**Clusters**
Agglomerative	NMI	0.5526	7
	ARS	0.5436	7
	Silhouette	0.4523	7
KMeans	NMI	0.5168	11
	ARS	0.3315	11
	Silhouette	0.2369	11

Table 2: validation measures

Cluster analysis

Method	**Cluster**	**Counts**
Agglomerative	0	2
	1	2440
	2	437
	3	1
	4	9468
	5	1
	6	1
KMeans	0	6166
	1	1451
	2	1
	3	1
	4	3466
	5	1
	6	435
	7	507
	8	1
	9	1
	10	320

Table 3:Customer segments summary

From the above results there is a similarity of clusters obtained from the two algorithms with four clusters having almost the same number of customers. However, clusters obtained from KMeans are more granular as compared to agglomerative with has three compact classes whereas KMeans further splits them into six clusters.

4.6 Profiling

Segment 1

Mobile money users in this category average of 27 years of age, conduct mpesa transactions averaging 27 million Kenyan shillings, their organization and customer credits average 24 and a have few pay bill and merchant store transactions below 5. Average GSM activity is 500 Kenyan shillings. Most likely to be subscribers running businesses.

Segment 2

Users in this category average 37 years of age, have the highest average value on GSM activity of about 2,901 Kenyan shillings. Their organization and customer credits average 52 and pay bill and merchant store transactions above 10. Their frequency of using financial service institutions is approximately 1.1

Segment 3

Only has one customer which is likely to be a financial institution with large volumes of mpesa transactions and over 12 million subscribers crediting the organization.

Segment 4

This is a business to customer organization, crediting approximately 250,000 customers.

Segment 5

Mobile money users in this category average of 49 years of age, conduct mpesa transactions averaging 143,000 Kenyan shillings, their organization and customer credits average 10 and a have few pay bill and merchant store transactions below 3. Average GSM activity is 640 Kenyan shillings. Most likely to be senior citizens conducting very few organization-based transactions.

Segment 6

A business conducting large volumes of mpesa transactions. Mostly organization to organization-based transactions.

Segment 7

This segment consists of small medium enterprises where majority of the transactions are customer credits with mpesa transactions over 500 million Kenyan shillings.

Segment 8

Users in this category average 28 years of age, have the highest frequency on using financial institutions of above 5 transactions. Their organization and customer credits average 70 and pay bill and merchant store transactions above 14. They mostly conduct cash out transactions.

Segment 9

Has the highest number of merchant to customer transactions.

Segment 10

Business to customer type of a business with over 2million customers.

Segment 11

Users in this segment are closely related to segment 8. Averaging 30 years of age, financial institution transactions approximately 4.91. They however have the highest frequency on credits to organizations of about 180 and pay bill and merchant store transactions above 50.

4.7 Discussion

From the above results, we observe that NMI which computes the connectedness between two instances produced us above average results for both agglomerative and KMeans clustering with agglomerative being higher. ADR which calculates the commonness between 2 clusters with a range of -1 and 1 where negative values indicate a bad score and 1 as a perfect score, both algorithms gave us a positive score with agglomerative giving a high value of 0.54. Silhouette Coefficient measures the similarity between instances in their own cluster and how their difference with instances from different clusters. Negative values indicate a bad score while 1 is the perfect score. Agglomerative performed better than KMeans with a value of 0.45. In comparison with a study conducted by (Samuel W. Kamande, 2018) on customer segmentation for consumers in Kenya using their expenditure information, he compared three clustering algorithms Partitioning around Medoids, K-Means clustering and Hierarchical clustering. He established that hierarchical clustering provided the best results with a score of 0.84 despite challenges of missing data.

Previous segmentation of mobile money customers done by (Sinja Buri, 2019) established five customer segments i.e. acceptors, bulk senders, cash in/cash out, service providers/agents and airtime traders. In comparison to the eleven clusters developed from our study, this research was able to create segments not just on the basis of the transaction type but putting in consideration the customer behavior in terms of frequency of use of services, age and their network activity. From our study we are able to identify different types of bulk senders, cash in and cash outs and different service providers.

4.8 Achievements

The main objective of our study as highlighted in chapter one was to analyze mobile money data and come up with homogeneous customer groups based on a combination of behavioral factors, demographic factors and psychographic factors. This has been well fulfilled with both clustering algorithms giving us almost similar clusters with a slight difference in the number of clusters where by KMeans splits three clusters from agglomerative clustering further into six clusters. This study has also helped us establish the benefit of adding other features such as age, network value and frequency of different types of transactions as factors to consider during mobile money customer segmentation.

4.9 Conclusion

Given the wide use of mobile money services with over 22 million customers in the country using the platform, customer segmentation is a necessity for the business to be able to cater for customer needs, innovate and develop target specific services. Segmentation on demographic factors such as age and assets alone cannot fulfill the current growth of financial services. By combining a set of demographic, behavioral and psychographic factors this research was able to profile customers into distinct customer segments that provide a clearer picture on the different characteristics of customers using the service as opposed to the general classification based on the transaction types.

Out of the three algorithms evaluated, agglomerative clustering provided more compact and connected clusters as compared to KMeans. However given how closely related some clusters were, KMeans was able to further split larger segments into granular sets highlighting characteristics that were not visible by agglomerative clustering hence the tendency of rigidity in agglomerative clustering where once instances are grouped into a similar segment they remain in that segment throughout the construction of the entire segment tree.

Affinity propagation produced a large number of clusters, had a high memory utilization and took the longest time to segment our dataset and was therefore considered it unsuitable for our dataset. This can be attributed to its inability of working with a large dataset and proficiency on several computer vision and biology problems (Brendan, 2020).

Clusters obtained from this study indicate three dominant segments which have the largest population and very distinct characters in terms of their mobile money activity on the network.

Further work

I would recommend further work on how the location of a transaction affects the customer segments. Given a lot of missing data on geographic location, this research did not put the location aspect into consideration.

References

Afande, D. F. O., 2015. Market Segmentation by Commercial Banks in Kenya. *Journal of Marketing and Consumer Research .*

Ahmed M Serdah, W. M. A., 2016. Clustering Large-Scale Data Based On Modified Affinity Propagation Algorithm. pp. 22-33.

Andronikidis, A., 2008. Psychographic segmentation in the financial services context: a theoretical framework. *The Marketing Review.*

Anika Singh, P. B. V., 2017. Customer Segmentation through K-Means and Hierarchical Clustering Techniques. *International Journal for Scientific Research & Development,* Volume 5.

Anon., 2019. *https://developers.google.com/machine-learning/clustering/clustering-algorithms.* [Online]
Available at: https://developers.google.com/machine-learning/clustering/clustering-algorithms

Bacila Mihai, A. R. I. M., 2012. Prepaid telecom customers segmentation using the k-mean algorithm.

Bose, I. a. C. X., 2010. Exploring Business Opportunities from Mobile Services Data of Customers: An inter-cluster Analysis Approach. *Electronic Commerce and Applications,* pp. 197-208.

Brendan, N., 2020. *www.learndatasci.com.* [Online]
Available at: https://www.learndatasci.com/tutorials/k-means-clustering-algorithms-python-intro/

Carolina Mattsson, G. S., 2018. *Understanding Key Mobile Money Users,* s.l.: s.n.

Central Bank of Kenya, 2018. *Annual Report and Financial Sttatements,* s.l.: s.n.

Ernst & Young, 2018. Mobile money — the next wave of growth.

Frey, D. D. a. B. J., 2007. Non-metric affinity propagation for unsupervised image categorization.

gsma, 2019. *https://www.gsma.com/r/mobileeconomy/sub-saharan-africa/.* [Online]
Available at: https://www.gsma.com/r/mobileeconomy/sub-saharan-africa/

James Maitai, D. J. O., 2016. Factors Influencing the Adoption of Mobile Money Transfer Strategy in Telecommunication Industry in Kenya: A Case of Safaricom–Kenya Ltd.. *www.iosrjournals.org ,* p. 84.

Khakbaz Seyed, P. M. H. N., 2017. An efficient hybrid clustering algorithm for segmentation:Autocluster. *International Journal of Data Science.*

Khamis Mwero Manero, R. R. a. C. O., 2018. Customer Behaviour Segmentation among Mobile Service Providers in Kenya using K-Means Algorithm. *International Journal of Computer Science Issues,* p. 68.

Kotler, P. a. A. G., 2009. Principals of Marketing.

mercy, a., 2012. THE INFLUENCE OF THE PROVISION OF MOBILE MONEY TRANSFER. *IEEE,* pp. 45-56.

Munyange, M. M., 2012. THE INFLUENCE OF THE PROVISION OF MOBILE MONEY TRANSFER SERVICE ON THE SOCIO-ECONOMIC STATUS OF THE SERVICE PROVIDERS:CASE OF NAIROBI CENTRAL BUSINESS DISTRICT, KENYA. p. 16.

Oates, B. J., 2006. *Researching Information Systems and Computing.* s.l.:SAGE Publications Inc.

Rajagopal, J. K. a. J. A. F. a. R., 2014. Household Energy Consumption Segmentation Using Hourly Data. *IEEE Transactions on Smart Grid,* pp. 420-430.

safaricom plc, 2019. [Online]
Available at:
https://www.safaricom.co.ke/annualreport_2019/assets/Safaricom_FY19_Financial_Statements.pdf

Samuel W. Kamande, E. A. K. M. E. A., 2018. Consumer Segmentation and Profiling using Demographic Data and Spending Habits Obtained through Daily Mobile Conversations. *International Journal of Computer Applications,* pp. 33-42.

Shreya Tripathi, A. B. E., 2018. Approaches to Clustering in Customer Segmentation. *International Journal of Engineering & Technology,* pp. 802-807.

Sinja Buri, M. v. d. W. S. H., 2019. Predictive Modelling and Segmentation for Market Sizing and Product Design.

United Nations, 2012. *Mobile Money for Business Development in East African Community,* s.l.: s.n.

Ustundag, P. A. S. a. A., 2016. Performance evaluation of different customer segmentation approaches based on RFM anddemographics analysis. pp. 1129 - 1157.