Telecommunication Customer Categorization with Novel Data Mining Approach for Effective Communication and Increase Profitability

Ameer Mohamed Aslam Sujah[1], Kapila Tharanga Rathnayaka R.M.[2]
[1]Department of Computing & Information Systems, Sabaragamuwa University of Srilanka,Belihuloya,Srilanka.
[2]Department of Physical Sciences & Technology, Sabaragamuwa University of Srilanka,Belihuloya,Srilanka.
Corresponding Author:Ameer Mohamed Aslam Sujah

Abstract: *Telecommunication industry plays a vital role in the modern fast-moving world. At the same time, the industry is highly competitive because of multiple providers provide different solutions to their consumers. As a result, customers are rapidly moving from one service provider to another. Furthermore, human communications have been moving far from traditional calls and text messages to alternatives. Therefore, mobile operators are under real revenue threats as well as the risk of losing their potential customers. To solve this kind of issues, they need to increase their capabilities on understanding customer behavior patterns and preferences, in order to achieve a high level of customer profitability and revenue. The major aim of this study is to cluster the customers based on profitability and develop a model to predict future customer's profitability level and clustering the customers to provide different promotional packages. The current study is carried under the three phases. Initial phase is comparison of different approaches in K-means algorithm and choose the best one by using Within Cluster Sum of Square (WCSS) and processing time. The second phase is focusing on clustering the customers based on their behaviors by using best algorithm (K-means++) and develop the Artificial Neural Network (ANN) model to predict future customer's profitability level. Finally, choose one of the early clustered customer group and apply the best algorithm (K-means++) to provide different promotional packages. Dataset consists of 10,000 prepaid customer details with 15 different variables to cluster, train and test the model. Comparison of WCSS and process time, K-means++ is the best approach for clustering. Confusion matrix used to evaluate the performance of ANN model and constructed model gives the accuracy of 97.3%. Existing researches use unsupervised or supervised learning algorithms separately. But this study integrates both algorithms and getting high accuracy result. Therefore, this model well fit for telecommunication industries.*

Keywords: *Profitability, Clustering, Neural Network, K-means, Telecom*

Date of Submission: 11-09-2019 Date of acceptance: 26-09-2019

I. Introduction

1.1Background

Telecommunication industry plays a vital role in the modern fast-moving world. It helps to make digital transformation globally. At the same time, the telecommunication industry is highly competitive because of multiple telecommunication service providers provide different solutions to their consumers. As a result, customers are rapidly moving from one service provider to another depends on their requirements. Furthermore, a significant amount of human communications (especially in the younger generation) have been moving far from traditional calls and text messages to alternatives such as Skype, Google Duo, FaceTime, instant messaging and social media [1].

Therefore, mobile operators are under real revenue threats as well as the risk of losing their potential customers. Top level managers are finding proper solution to solve this kind of issues. As a telecommunication service provider, they need to increase their capabilities on understanding customer needs, behavior patterns [2] and preferences, in order to stay competitive, achieve a high level of customer profitability and continue the revenue growth in long run. Because, customers play the key role of any successful business which provides products or services [3]. Customer segmentation based on their different behaviors is one of the best ways to understand the customer requirement and patterns.

The term customer profitability [3] is defined as, the profit that the telecommunication company makes from serving a customer or group of customers for a specified period of time. Make best customer relationship management is the way to maintain our customers for a continuous long run. Most successful businesses continuously do the research and development part for their customers in the fields of customer identification, customer attraction, customer retention, and customer development to achieve a high level of customer relationship [4]. All the competitors are trying to make more customer profit to survive in their business. Therefore, executive level managers are exploring to identify an efficient way to do most appropriate customer clustering methods for mining profitability.

Customers are categories into different groups based on their behavioral patterns such as high profitable customers, average profitable customers and low profitable customers and provide different benefits to the different group of customers based on their profitability level increases the customer satisfaction.

The telecommunication industry produces massive amounts of data every day which need to be mining to discover hidden information for effective prediction, exploration, diagnosis and decision making. Without a proper mechanism to handle this massive amount of data create difficulties to extract knowledge from it. Traditional data analyzing tools have limited number of capabilities and it is not suitable for handling big data in telecommunication industry. Because, lack of getting the most precision results with customization, as well as the performance issues throughout the traditional Weka, and SPSS tools. As a solution, this research study introduces new methodology to predict different clusters for customer profitability using data mining technologies [6] and machine learning algorithms. Also, provide a solution for customer segmentation and introduce promotional packages to the different level of loyality customers [7].

Data Mining (DM) is a powerful technique which help organization to discover the patterns and trends in their customer's preferences and well-known tool for customer relationship management. Data mining methodology has made a vast range of contribution for researchers to extract hidden knowledge and information. The major DM process uses data exploration technology to extract data, create predictive models, and verify the stability and effectiveness of the models. The K-means method segments customers into clusters based on important factors (Example: - their billing, loyalty, and payment behaviors) to make decisions. Most of the researches are apply the DM techniques to make decisions [4].

This study has focused onmining profitability of telecommunication customers and customer segmentation with novel data mining approach. There are different factors which influence the customer profitability level. Based on the above understanding, this study used 15 different customer behaviors for clustering the customers, predict the customer profitability level and analyze promotional packages. K-means, K-means plus plus, and Artificial Neural Network (ANN) are the different machine learning algorithms which are applied to this research. Dataset used in this study contains 10,000 prepaid telecommunication subscribers.

1.2 Significance of Study

There are very few researches done regarding mining profitability of telecommunication customers with customer segmentation. This research study uses two different approaches of clustering algorithm and find out the best one by comparing their performances. Apply the best algorithm to cluster the entire customer community into n number of clusters and analyze those clusters separately. This clustering techniques helps to make better decisions about their customers by the top-level management.

Provide beneficiary packages to the high profitable customers, value added customer care service to the loyality customers, and provide free promotions to the unstable customers to stabilize in the service provider are some benefits that can be acquire by clustering the customers. Analyze different customer behaviors and construct a machine learning model to predict the customer profitability level andpromotional package-basedclustering are the two other scopes of this study. Through this model telecommunication companies can identify their different profitable level of the future customers and their essential requirements to provide promotions and loyality facilities. As a result, company can make positive thoughts among the customers and get more profit for a long run.

1.3 Aim and Objectives

The major aim of the research is to develop customer prediction model based on their profitability level and customer segmentation by using selected customer behaviors.

The specific objectives of this research study are,

i. To compare two different approaches of K-means clustering algorithm and choose the best one by analyzing the processing time and Within Cluster Sum of Squares (WCSS) values.

ii. To develop customer prediction model for different profitability level. ANN is used for implementing the model. This model helps to predict and evaluate the present and future customer profitability level based on their different behaviors.

iii. To improve the accuracy of the customer prediction model.

iv. To develop a clustering methodology for provide different promotional packages based on their specific behaviors.

1.4 Research Questions

Identified research questions can be explained as,

i. What are the factors that directly impact the telecommunication customer's profitability and how to identify the most appropriate factors?
ii. Which machine learning algorithm uses to efficient clustering and how to evaluate the algorithm performance?
iii. How to implement customer prediction model for different profitability level with high level of accuracy?
iv. How to implement the clustering methodology for providing different promotional packages and how to choose the most appropriate factors?

II. Literature Review

Several databases are retrieved to find the relevant studies for the research topic. Finally, twenty-one publications are filtered which are more appropriate to the specified topic. Those specified twenty-one researches are conducted on the topics of mining profitability, customer segmentation, and loyalty customer behaviors. Those terms are considered as the search terms for the mapping study. Briefly discuss the previous researches that are related to a similar topic.

Customer segmentation plays a vital role to understand the customers. Since the enterprises uses different methods to describe customer behavior. Such as, data mining, RFM method, customer value matrix and customer lifetime value method. Customer segmentation are based on personal features of the customer like age, sex, education, etc. RFM analysis can be done using data mining methods specially clustering methods.

In 2014, Alhilman, and their team used CRISP-DM (Cross Industry Standard Procedure for Data Mining) techniques in their research [9]. Business understanding, data understanding, data preparation, modeling, evaluation, and deployment are the crucial steps of this methodology. The major objective of the study was to analyze telecommunication service provider X's customer behavior and predicting customer categories that will be used to improve customer loyalty, increase revenue and profitability. The research outcome helped to determine that who moved from productive (generating revenue) category to unproductive category (not generating revenue) and take recommended actions like up selling, cross selling, customer education, and switching to another package best-suited customer's need [10].

Qirong et.al propose a novel graph edge-clustering algorithm (DGEC) that can predict unique behavioral groups, from rich usage datasets in 2016. A behavioral group is a set of nodes that share similar edge properties reflecting customer behavior but are not necessarily connected to each other and therefore different from the usual notion of graph communities. It helps to improve our understanding of consumer relationship and behavior [2].

In 2012, Jonathan Magnusson and Tor Kvernvik proposed a methodology for identifying influential subscribers in a telecommunication network based on several social network analysis metrics (SNA). They used machine learning algorithms. Such as decision trees, logistic regression, a neural network to implement their model. The solution has been implemented on Hadoop platform to support scalability and to reduce hardware cost [8].

HasithaIndikaArumawadu et.al proposed a methodology to evaluate the customer profitability using K-mean clustering techniques in the telecommunication industry in Sri Lanka in 2015. They use RFM (Recency, Frequency, Monetary) factors to implement their clustering model. Throughout these factors, they predict certain common behaviors of the telecommunication customers [4]. As a result, they got four clusters. Such as High profitable customers, Profitable customers, Medium profitable customers, and Low profitable customers.

ABC and RFM (Recency, Frequency, and Monetary) analyses are used to evaluate the clients according to the purchasing behavior and Kohononen-maps and K-means clustering methods are applied to divide customers into four segmentations in another research in 2010 by Jan Panuš et.al. The Objective of this research is to allow the selected company to effectively use their customer's previous data that should achieve a more effective marketing campaign and an increase of the satisfaction of their customers [11].

In 2017, I. K. Savvas et.al provide a solution to understand the customer behavior of the telecommunication industry. DBSCAN and k-means are the two distributed versions of clustering algorithms were used. Both algorithms are cluster data according to its characteristics. DBSCAN is a density-based spatial clustering algorithm and clustering the data based on the minimum size of participating objects per cluster. K-means clusters the data objects according to the pre-desired number of groups. The two algorithms are different roads to group the data objects. The experiment outcome proved that the algorithm's efficiency to predict customer behaviors from a massive amount of data [5].

In 2012, Salmiah Mohamad Amin et.al proposed a methodology to identify the contributing factors of customer loyalty towards telecommunication service provider in Malaysia. They gather data from 185 telecommunication subscribers among university students using a self-administered questionnaire and multiple regression analysis was performed to analyze the contributing factors of customer loyalty. Their research findings disclosed that there is a positive relationship between switching cost, corporate image, trust, and perceived service quality with customer loyalty. Perceived service quality was found to be the key factor in affecting user's customer loyalty [14].

2.1 Customer Segmentation based on Data Mining

Categorizing methods based on experiences, statistics or partitioning are the traditional procedure of customer segmentation. They can't fulfill the requirements of high-complex analysis which enterprises requires to make decision about their business [16].Nowadays, data mining techniques are using to customer segmentation. It provides best results for extracting meaningful information from huge amount of data. Businesses cannot understand their strength and weakness by using raw data. The results of data mining use not only to increase income but also improve the customer relationship. Rapid expansion of customer data the traditional segmentation methods come valueless. While applying data mining technology, Companies can sort, handle and analyze massive amount of complex customer data [17].The complete structure of data mining system has four main stages: identification of business objectives, data preprocessing, data mining and modeling process, model evaluation and expression as shows in figure 1 [17].

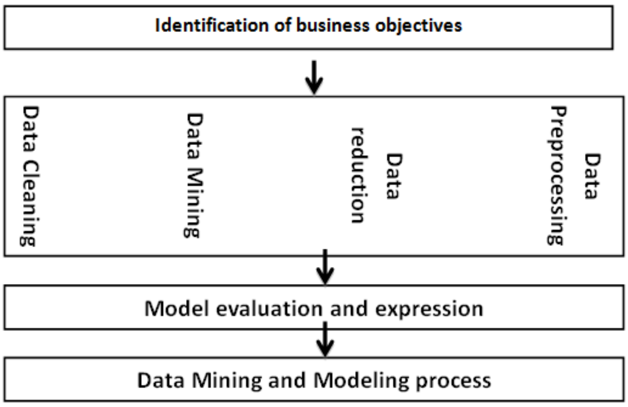

Identification of business objectives
Data Cleaning
Data Mining
Data reduction
Data Preprocessing
Model evaluation and expression
Data Mining and Modeling process

Figure 1 Flow chart of data mining

Gong and Xia proposed a specific method by using data mining technologies to customer segmentation for supermarket industry. Customer purchasing behavior used to segment and formulate the model that used by enterprises to understand the customers and make more scientific future decision [17].From Lai's statement, clustering analysis is the most frequently used customer segmentation technique in data mining. Clustering analysis used to categorize customers based on the differentiating factors. Such as, ages, sex, monthly incomes, occupations, education levels, etc. Also, clustering analysis can make the different levels of importance with different variables in the classifying process; those things can be used to efficient decision making [16].

2.2 K-means clustering method

Because of the simplicity in implementation and fast execution, the K-means algorithm used in customer segmentation, pattern recognition and information retrieval [20]. Clustering results are affected by the choice of initial point, and therefore the solutions obtained are always local optimum, not global optimum [21]. K is used as an input for the predefined number of clusters. An average location of all the members of a special cluster shows means. Each point has been donated to a cluster whose center (or centroid) is nearest. The center is the average of all the points in the cluster. K-Means algorithm calculates its centers iteratively. The steps of the K-means algorithm are given below,

i. Select random K points to be the seeds for the centroids of k clusters.
ii. Assign each example to the centroid closest to the example, forming in this way k exclusive clusters of examples.
iii. Calculate new centroids of the clusters. For that purpose, average all attribute values of the examples belonging to the same cluster (centroid).
iv. Check if the cluster centroids have changed their "coordinates". If yes, start again form the step 2. If not, cluster detection is finished, and all examples have their cluster memberships defined.

K-Means may produce tighter clusters than hierarchical clustering. But there are some disadvantages in using this technique. In each run it does not show the same result, since the final clusters depend on the first random assignment. Another disadvantage of this algorithm is that comparing quality of the clusters produced is so difficult. It is also not useful and appropriate for non-globular clusters.

III. Methodology

3.1 Requirement Analyze

Requirement analysis is the first and foremost step to determine user expectations for any development project. There are several ways for requirements gathering. Questionnaires, interviews, analyzing existing documents, and observations are some of the instances. This research requirements were gathered by analyzing existing documents in selected telecommunication service provider in Sri Lanka and discussion with the managerial people who are working in the industry. The dataset consists of 10,000 prepaid subscribers' details with 15 attributes. Their attribute descriptions are listed in the Table 1.

Table 1 Attribute descriptionand their representation

Description	Representation
Average monthly revenue.	Sri Lankan Rupees
Average number of monthly minutes of usage.	Minutes
Average number of roaming calls	Numeric
Average number of customer care calls	Numeric
Average number of attempted calls	Numeric
Total number of months in the service (Customer Lifetime)	Numeric
Total number of calls taken over the life of the customer	Numeric
Total minutes of outgoing call over the lifetime of the customer	Minutes
Total revenue earned through lifetime of the customer	Sri Lankan Rupees
Geographical location of the customer	Nominal
Device's web access capability	Nominal
Usage of Credit cards	Nominal
Marital status (Single or Married)	Nominal
House ownership details	Nominal
A five-digit number that uses to uniquely identify the customer	Numeric

3.2 Data Preprocessing

Dataset cannot be use directly without doing the preprocessing part. Incomplete data, duplicate data, and noisy data causes inaccurate results throughout the research. Therefore, preprocessing is mandatory for get accurate results. In this study Anaconda Navigator tool was used for data preprocessing. Anaconda Navigator is a free open source distribution for Python and R programming languages. This tool used in machine learning and data science projects to complete the tasks with minimal difficulties. It is a collection of various applications. Such as, Jupyter notebook, Spyder, RStudio, Glueviz and VS Code.

Initially, retrieve the telecommunication customer dataset from Comma-Separated Values (CSV) file to Spyder IDE by using python programming language and store the required data columns in different variables. Fill the missing values, encode the categorical data, outlier removal, and feature scaling are the steps that can be done sequentially. Let's look at one by one.

Fill the Missing Values: There are two types of variable can contain missing values such as numeric data variables and nominal data variables. Missing values of numeric data variables can be filled by using mean values and missing values of nominal data can be filled with mode value of the column.

Encode the Categorical Data: Machine learning algorithms have the capability of handling numeric data to process and analyze. Therefore, the data columns which contains categorical data must be transfer into numerical format before apply machine learning algorithms.

Outlier Removal: Outliers also affect the final outputs which can be generated by different algorithms.Examine the outliers by using the below equation with the help of Inter-Quartile Range (IQR). If the value v overcome the below range that the data, consider as outlier value.

$$(Q1 - 1.5 * IQR) <= v <= (Q3 + 1.5 * IQR)$$

Feature Scaling: Dataset consist of different attributes with different range of values. It is also affecting the final outcomes whenever calculating distance value between those two variables. Therefore, all the attributes are converted into the similar range and it's called feature scaling.

3.3 Data Mining Algorithms

Data mining process uses different algorithms to carry out data analyze, pattern recognition and make prediction to our dataset. There are large number of data mining algorithms are available for different purposes. This research project mainly focusing on three different algorithms namely K-means, K-means plus plus and ANN.

K-means Algorithm

This algorithm comes under unsupervised learning and used by the situation of unlabeled dataset. It's one of the most popular clustering algorithms because of its simplicity. Find groups (clusters) is the goal of this algorithm for a given dataset. The steps of K-means algorithm as follows,

Step 1: Randomly select K cluster centers (centroids). Let's assume that they are C_1, C_2, C_3, …., C_K.

Step 2: Assign each input value to the closest centers by calculating the Euclidean distance (ED) between the data point and each cluster centroids.

$$ED = \sum_{i=0}^{n} \sqrt{(xi - ci)^2}$$

i - number of points in the cluster.

xi - i^{th} point in cluster.

Step 3: Find the new centroids by calculating the average of all the data points that assigned to that cluster.

Step 4: Repeat the step 2 and step 3 until none of the cluster’s assignment change.

K-means Plus Plus Algorithm

This algorithm works same as the K-means algorithm. Random initial centroid selection is the problem in K-means algorithm. K-means plus plus algorithm has the solution to random initial centroid selection. Let’s say D(x) represents the Shortest distance from a data point to the closest center that we have already chosen. The K-means plus plus algorithm defines as the following steps,

Step 1: Take one center c_1, chosen uniformly at random from X.

Step 2: Take a new center c_i, choosing x € X with probability $D(x)^2$ / ∑ $D(x)^2$

Step 3: Repeat step 2 until taken k centers altogether.

Step 4: Proceed as with the standard K-means algorithm.

Artificial Neural Network

The concept of artificial neural network is based on the biological neural networks like the brain. Neuron is the basic structure of neural network. ANN comes under the deep learning technologies. It is focusing on solving the complex pattern recognition problems in different industries.Standardized neural network contain one input layer, one hidden layer and one output layer. Input layer is the first layer of neural network and it provides an interface with the environment. All the computations are done in the hidden layer and the output layer stored the results. Data travels through the successive layers and the outcome is available at the output layer.

Multi-layer neural networks are most popular among different types of neural networks due to more than one hidden layered structure helps to solve complex problems than the single hidden layered neural network. Artificial neural networks are generated through the below steps,

Step 1: Load the dataset

Step 2: Define the artificial neural network

Step 3: Model compilation

Step 4: Train the model

Step 5: Evaluate the model

3.4 Proposed System

This study consists of four phases. The first phase implements the profitable customer clusters based on their behaviors by using two different clustering algorithms and select the best one by comparing the processing time and WCSS value. In the next phase, cluster the customers into n number of clusters by using the best algorithm which is selected from the first phase. Third phase focusing on developing the customer prediction model and in the final phase clustering the specific group of customers into different number of clusters to provide different promotional packages.

In the initial phase, preprocessed telecommunication customer data stored in CSV file format. This dataset consists of 15 attributes including the customer id to uniquely identify the customers. Next step importing the python libraries which require to carry out the processes. pandas, numpy, and matplotlib are some of the libraries used within the development process. Correlation analyses used to find out the most appropriate attributes for clustering from the list of 15 attributes. Separated significant attributes are used in K-means and K-means plus plus algorithm to evaluate the best one by comparing the algorithm processing time and WCSS value for different number of clusters.Use the dataset into K-means algorithms and changing the k values. Processing time and Within Cluster Sum of Squared (WCSS) are stored as an output of different input k value. Again, do the same procedure for k-means plus plus algorithm and get the output values. Compare and analysis the processing time and WCSS value for both algorithms in different conditions. Finally, as a result figure out best algorithm which have lowest processing time and minimum WCSS value for different conditions.

Second phase uses the algorithm which is selected as the best one in the previous phase. Apply that algorithm into our dataset and get the clustering result (WCSS value) by changing the K values. Plot the distortion curve graph and figure out the best K value by using elbow method. Next step, cluster the entire dataset by using the best k value and separate the customers into different levels of profitability. Calculate the final weighted centroid value for each cluster and generate profitability as a target variable for the entire dataset. Final weighted centroid values can be used by top level management to carry out analyzing part for the cluster results with different attributes to decision making.

Third phase uses the target variable which is generated by the previous step for developing profitability prediction model by using ANN. The ANN model consists of input layer, 3 hidden layers and output layer. Import require python libraries and read the CSV file. Then split the dataset into independent variables and target variable. Then, allocate the dataset for training and testing purposes. Generally, 80% of the dataset used to train the model. Trained model is evaluated with the help of test dataset. Continuously change the training and testing dataset until achieve high accuracy to the ANN model. Then the model predicts the telecommunication customer profitability level while we are inputting the customer attributes.

Final phase is clustering the customers into different segments for providing different promotional packages based on their usage behaviors. It is a necessary task that whenever the top-level management take decision to provide different promotional packages to the specific customer category. Separate the customers who are coming under a specific profitability level used for the clustering part. Clustering part is similar to the previous phase and managerial people make efficient decisions related to promotional packages with the help of this clustering technique. The Figure 2 shows the flow chart of the proposed system.

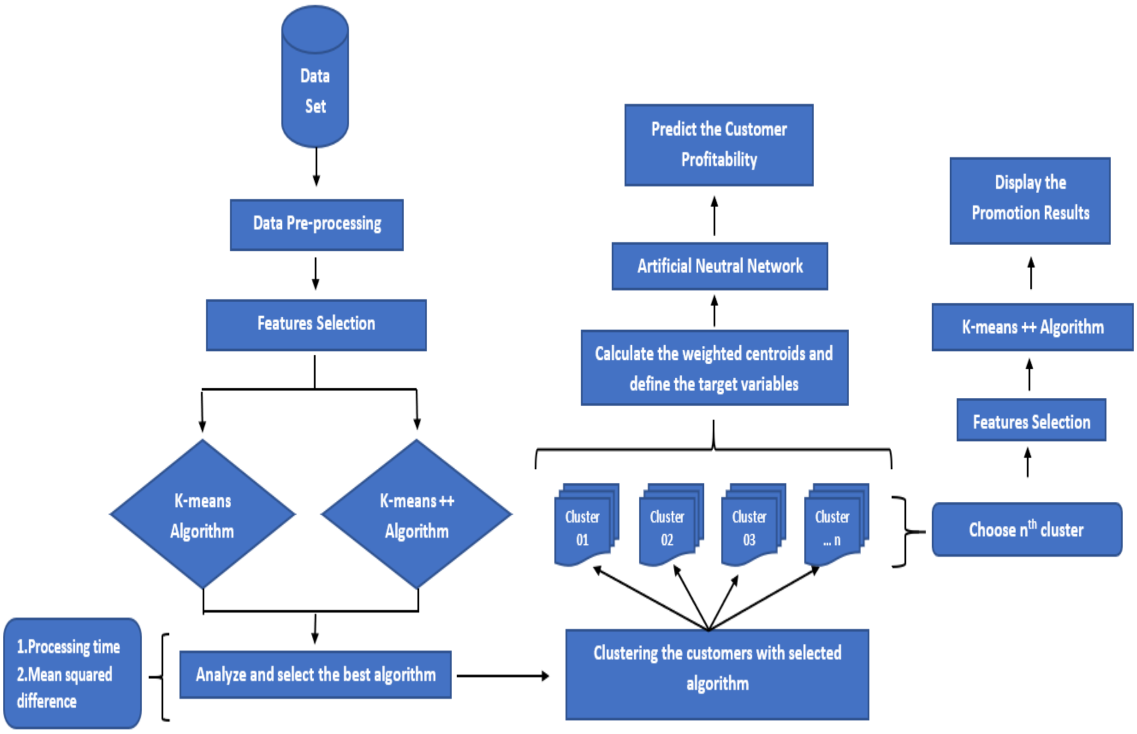
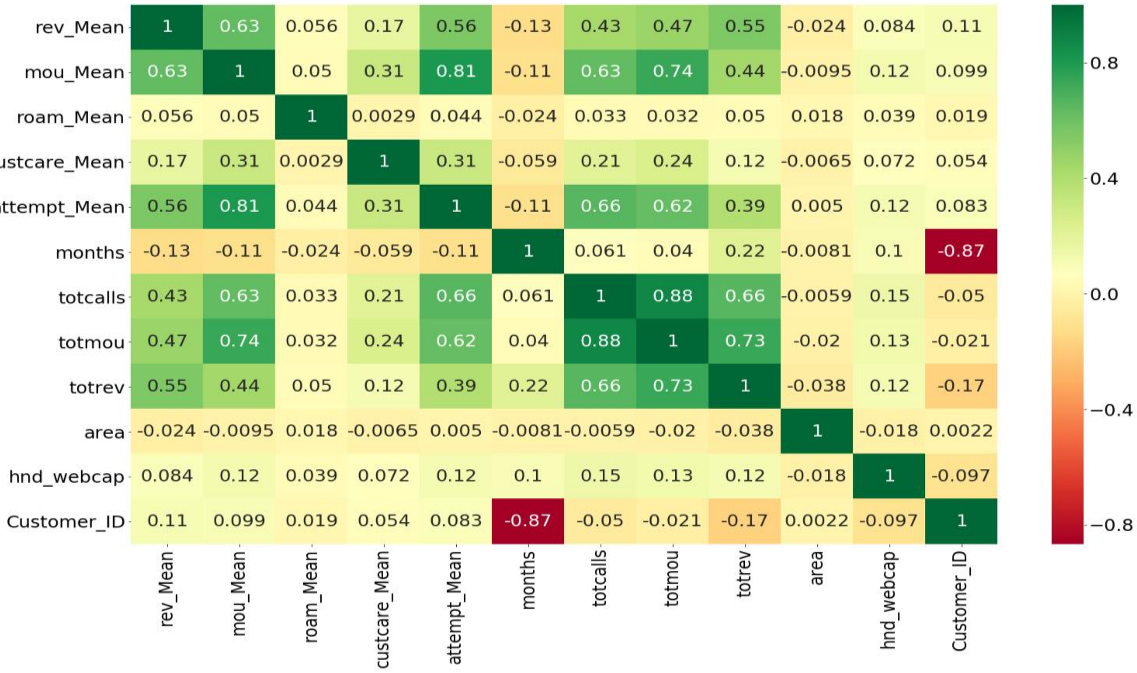

Data Set
Data Pre-processing
Features Selection
K-means Algorithm
K-means ++ Algorithm
1.Processing time
2.Mean squared difference
Analyze and select the best algorithm
Clustering the customers with selected algorithm
Cluster 01
Cluster 02
Cluster 03
Cluster ... n
Calculate the weighted centroids and define the target variables
Artificial Neutral Network
Predict the Customer Profitability
Choose n^th cluster
Features Selection
K-means ++ Algorithm
Display the Promotion Results

Figure 2 Flowchart of the proposed system

IV. Results and Discussion

4.1 Features Selection

Applied the correlation analysis for the 12 attributes which are identified as a significant factor based on the existing documents for the clustering part and analyze the correlation results for choose most influential factors among the 12 attributes. Figure 3 shows the relationship between the 12 attributes by using the heatmap.



	rev_Mean	mou_Mean	roam_Mean	custcare_Mean	attempt_Mean	months	totcalls	totmou	totrev	area	hnd_webcap	Customer_ID
rev_Mean	1	0.63	0.056	0.17	0.56	-0.13	0.43	0.47	0.55	-0.024	0.084	0.11
mou_Mean	0.63	1	0.05	0.31	0.81	-0.11	0.63	0.74	0.44	-0.0095	0.12	0.099
roam_Mean	0.056	0.05	1	0.0029	0.044	-0.024	0.033	0.032	0.05	0.018	0.039	0.019
custcare_Mean	0.17	0.31	0.0029	1	0.31	-0.059	0.21	0.24	0.12	-0.0065	0.072	0.054
attempt_Mean	0.56	0.81	0.044	0.31	1	-0.11	0.66	0.62	0.39	0.005	0.12	0.083
months	-0.13	-0.11	-0.024	-0.059	-0.11	1	0.061	0.04	0.22	-0.0081	0.1	-0.87
totcalls	0.43	0.63	0.033	0.21	0.66	0.061	1	0.88	0.66	-0.0059	0.15	-0.05
totmou	0.47	0.74	0.032	0.24	0.62	0.04	0.88	1	0.73	-0.02	0.13	-0.021
totrev	0.55	0.44	0.05	0.12	0.39	0.22	0.66	0.73	1	-0.038	0.12	-0.17
area	-0.024	-0.0095	0.018	-0.0065	0.005	-0.0081	-0.0059	-0.02	-0.038	1	-0.018	0.0022
hnd_webcap	0.084	0.12	0.039	0.072	0.12	0.1	0.15	0.13	0.12	-0.018	1	-0.097
Customer_ID	0.11	0.099	0.019	0.054	0.083	-0.87	-0.05	-0.021	-0.17	0.0022	-0.097	1

Figure 3 Attributes correlation analyze with heatmap

Customer's lifetime revenue (totrev) is the factor which is direct impact with the profitability of a customer. Therefore, analyzed the correlations between total revenue and rest of the attributes. It helps to find out highly influential factors. Based on the output of correlations this research study finalized the attributes such as totrev, totmou, totcalls, rev_mean, mou_mean, attempt_mean, months, custcare_mean and roam_mean.

4.2 Algorithm Comparison

Selected attributes are used to compare the algorithm performance. Preprocessed 10,000 customer dataset used as an input for algorithm comparison. Measure the processing time and WCSS values of k-means and k-means plus plus algorithm by changing the k value from 10 - 100. K represents the number of clusters. Plotting the graph helps to analyze and get the result that the k-means plus plus algorithm takes less processing time and minimal WCSS value. Figure 4 shows that how the processing time and WCSS value changes with different k values for both algorithms.

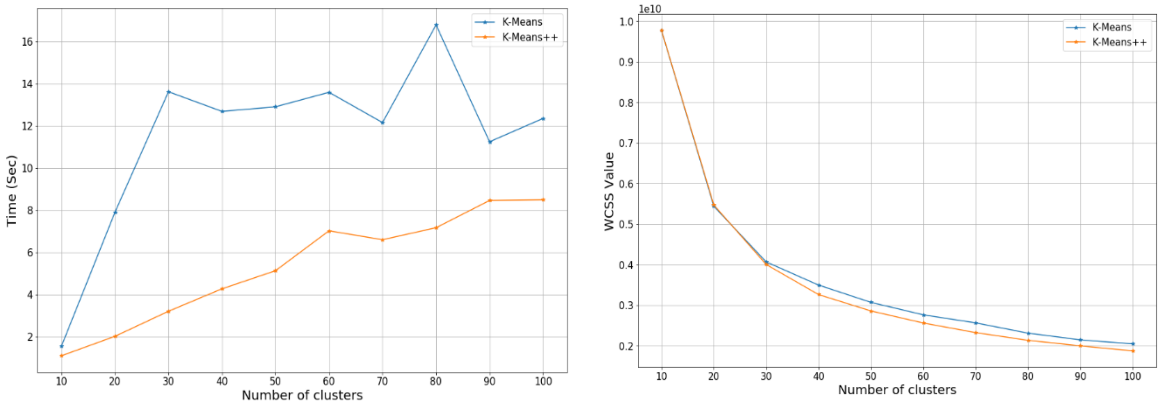


Figure 4 Processing Time and WCSS Value Comparison with Different K Values for Both Algorithms

Best algorithm chooses by comparing the processing time and WCSS values. Algorithm which takes lowest processing time and minimum WCSS value becomes the best algorithm. From the above experiment results k-means plus plus algorithm takes the place for best algorithm.

4.3 Profitable Customer Clustering

K-means plus plus algorithm used for clustering the customers into different profitability level based on their usage behavior. Initially, find out the best K value for the above customer dataset by using distortion curve. By analyzing the distortion curve got the result of (K = 5). Elbow method is applied to identify the best K value. Figure 5 display the distortion curve for the above customer dataset.

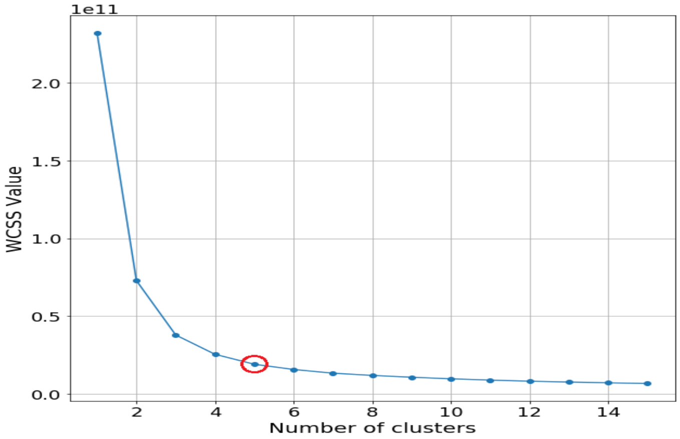


Figure 5 Distortion curve for customer dataset

Apply the K-means plus plus algorithm and use the previously predicted K (=5) value for clustering. Nine different customer behaviors are inputted and clustered into five different segments. Below table 2 shows the centroid values for each cluster with nine attributes.

Table 2 Attribute with clustered centroid values

Attribute Name	Cluster Category				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
REV_MEAN	58.3496	39.6528	52.7503	29.8933	46.1721
MOU_MEAN	596.662	176.737	408.919	60.4388	288.499
ROAM_MEAN	0.0513635	0.0391618	0.0415779	0.0368175	0.0449196
CUSTCARE_MEAN	0.673508	0.255757	0.502459	0.114919	0.415227
ATTEMPT_MEAN	158.716	62.0579	122.123	23.7388	94.0042
MONTHS	31.8369	30.9317	31.4772	31.1717	31.4007
TOTCALLS	5477.9	1825.3	4286.02	656.4	3019.64
TOTMOU	15496.9	4333.95	10997.3	1511.94	7425.97
TOTREV	1830.85	1132.68	1609.03	857.804	1373.26

Table 2 clearly shows the final clustering results with respect to attributes. This step calculates the final weighted cluster centroids. In this research, cannot take equal weight for different attributes. Therefore, considering the correlations, weights are calculated. Weights of the attributes can be shown as w_1 to w_9 and centroid values for a specific cluster with different attributes can be shown as C_{k1} to C_{k9}. In addition, $w_1+w_2+w_3+\ldots+w_9 = 1$. Then, calculated the final weighted cluster centroids (V_k) by using the below equation. K refers to cluster numbers (K^{th} cluster).

$$V_k = w_1(C_{k1}) + w_2(C_{k2}) + \ldots. + w_9(C_{k9})$$

The table 3shows the final weighted cluster centroids with respect to the different clusters.Customer profitability levels are determined by calculating the final weighted cluster centroid values (V_k) of each cluster. By analyzing the above table (Table 3), cluster - 1 is determined as a most profitable customer category and cluster - 4 contains the low profitable customers. So, the results are decided,

Cluster 1: Highest profitable customers.
Cluster 3: Profitable customers.
Cluster 5: Average profitable customers.
Cluster 2: Low profitable customers.
Cluster 4: Lowest profitable customers.

Table 3 Final weighted cluster centroids

Name of the Cluster	V_k
Cluster 1	4130.94
Cluster 2	1357.96
Cluster 3	3070.47
Cluster 4	591.388
Cluster 5	2166.51

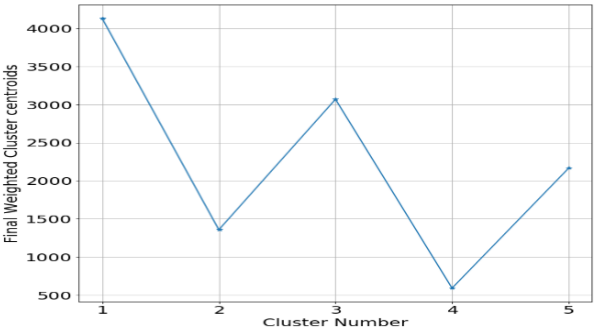


Figure 6 Behavior of final weighted cluster centroids

Figure 6 shows that how the final weighted cluster centroids are changing with different clusters. Top level managers evaluate the below graph with customer quantity of each cluster to make significant decision about their customer and their profitability level. This research experiment gives the output of cluster – 1 contains 876 customers (8.76%), cluster – 2 contains 2591 customers (25.91%), cluster – 3 contains 1559 customers (15.59%), cluster – 4 contains 2889 customers (28.89%) and cluster – 5 contains 2085 customers (20.84%). Based on the about results top level managers understand that who are the most profitable customer and who are the low profitable customers to their organization. It helps to make decisions about to increase their company profit by providing different promotional packages.

4.4 Prediction Model for Customer Profitability Level

This research study implemented as an integration of supervised and unsupervised learning algorithms. Previous phase used unsupervised learning algorithm and this phase used supervised learning algorithm. Supervised learning algorithm required target variables to train the model for future predictions. Those target variables are generated by using the previous phase. The clustered results of K-means plus plus algorithm used as a target variable.

Three hidden layers are used in this artificial neural network model. Continuously change the size of training and testing dataset to get high accurate model. Model received the high accuracy while using 80% of the dataset as a training dataset and got the accuracy of 97.3%. The confusion matrix shows the results of ANN model. This model helps to understand the relationship between the actual results and predicted results. The table 4 shows the confusion matrix of this research experiment for the dataset that the size of 2,000 customer.

Table 4 Confusion matrix of the prediction model

		Predicted				
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Actual	Cluster 1	180	0	0	0	0
	Cluster 2	0	517	0	0	4
	Cluster 3	18	0	271	0	0
	Cluster 4	0	11	0	561	0
	Cluster 5	0	1	20	0	417

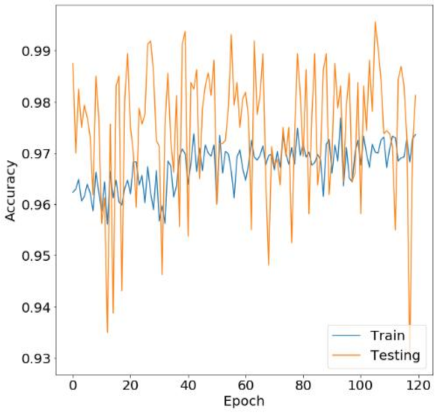


Figure 7 Accuracy on the training and validation dataset over training iterations

This ANN model train and test with different number of iterations to find out the best trained model. The model got high accuracy while choosing 120 iterations (Epoch = 120). The figure 7 shows that how the accuracy level changed in each iteration.This profitability prediction model used by the managerial people in the telecommunication industry to recognize profitability level of their customer who have different usage behaviors.

4.5 Customer Clustering to Recommend Promotional Packages

This is the final phase of the research. This part also contains clustering by using K-means plus plus algorithm. Results of this clusters recommends different promotional packages do their consumers by using different behaviors. This part totally different from the previous clustering technique. Because, specific number of customers are considered as a dataset for this clustering technique. As an example, service provider X plan to improve their profitability by providing different promotional packages to their customers who provide lowest profit to the organization. Initially they need to identify the lowest profitable customers to complete the task (This can be done by the previous clustering technique). After that they separate the lowest profitable customers and use different attributes to cluster them into different segments and finally provide different packages by analyzing the cluster results.

Initially separate the lowest profitable customers (Cluster - 4) into a new dataset and find out the attributes which are chosen as significant attributes for promotions by the managerial people. Based on the existing studies, average minutes of outgoing call duration and customer device category (Smart Phone user or Normal Phone user) chosen as a targeted attribute for the clustering. Distortion curve used to find out the best K vales for the clustering by using K-means plus plus algorithm. Figure 8 shows the distortion curve for the different numbers for K value.According to the distortion curve, the value six chosen for k and cluster the customers into six segments. This result plotted in a graph and analyze to provide promotional packages. X axis represent the average call duration and Y axis represent the device category. The figure 9 shows the clustered results.

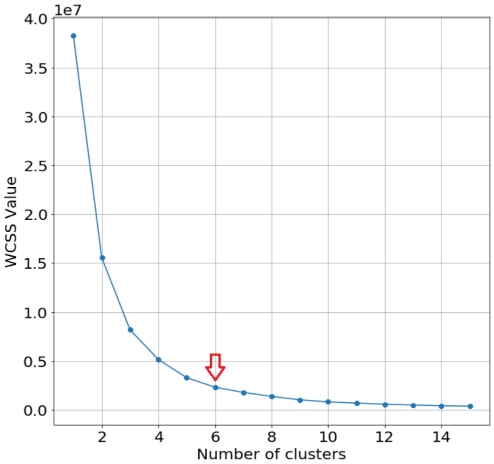


Figure 8 Distortion curve for promotional clusters

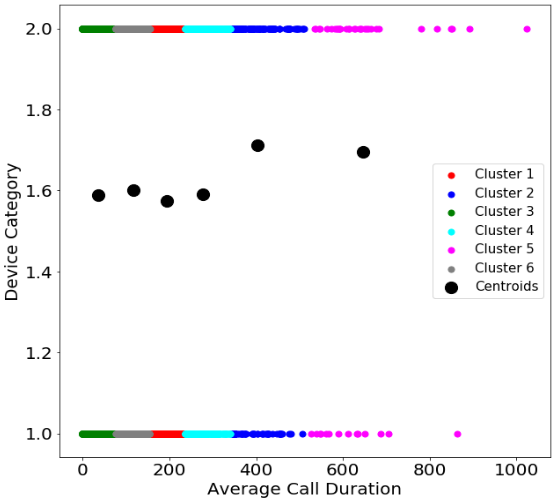


Figure 9 Promotional cluster results

According to the graph, lowest profitable customers are categories into 6 different groups. So, service provider can provide same promotional package with different percentage (Example of 10%, 15%, 20%, 25%, 30% and 35%) among the customers based on their average call durations. Here, cluster - 5 got 35% promotions (Highest Promotion) and cluster - 3 got 10% promotions (Lowest Promotion) because of their usage behavior.Likewise, provide different promotional package based on their device category. Mobile service provider decided to give 10MB data for each completed one minute of outgoing calls. This package gives benefits to the customer category who are coming under the smart device users. So, the customers who are using normal devices are dissatisfied with this promotion. But this clustered result helps to satisfy those customers too. While providing alternative promotional packages to the normal device users by using this clustered result. Clustered results are highly accurate because of that already clustered customers are again clustered into different promotion groups. Likewise, it can be possible to provide different promotions based on their different behaviors.

V. Conclusion

Telecommunication industry faces different challenges to analyze the customer behavior based on their different profitability level. Because the size of customers who are consuming the service rapidly changes. Better idea about their customer category based on the profitability level helps to make efficient decision on critical situations. If an organization's monthly revenue falls, they need recover from this situation quickly. This problem can be solving by segmenting their customers into different profitability level and introducing different promotions to the lowest profitable customers. So, the lowest profitable customers also continue the service after the promotions. It will increase the benefit to the organization. Success of a business measured by satisfying their customers and fulfill their requirements. To achieve this vision as a telecommunication service provider they need a very accurate prediction model. Different telecommunication industries are investing their values to develop this kind of model to their organization.

This research study recommends the profitability prediction model by using unsupervised learning and supervised learning algorithms. Artificial neural network is used to develop the prediction model and the target variable are chosen from the clustering algorithms. Therefore, this model got an accuracy of 97.3%. Manager can provide different promotional packages to their customers by using this prediction model and they can identify their customer’s profitability level individually.

VI. Recommendation

This experiment developed the clustering model by using the different number of attributes which are gathered as a dataset from a specific telecommunication service provider in Sri Lanka. Some of the attributes are not provided by considering their organizational ethics. Therefore, any research considering those attributes and including those attributes gives most precision results and highest accuracy.The prediction model developed by using artificial neural network. Any research considers the different prediction algorithms to develop prediction model gives more benefits. This research experiment used 10,000 customer details. This research study recommends developing a model for handling big data in the future.

References

[1]. P. K. Chang and H. L. Chong, "Customer satisfaction and loyalty on service provided by Malaysian telecommunication companies," in International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, 2011.

[2]. Q. Ho, W. Lin, E. Shaham, S. Krishnaswamy, T. A. Dang, J. Wang, I. C. Zhongyan and A. Shi-Nash, "A Distributed Graph Algorithm for Discovering Unique Behavioral Groups from Large-Scale Telco Data," in 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, Indiana, USA, 2016.

[3]. M. Xu, Y. Qiu and J. Qiu, "Mining for profitable customers," in International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA, USA, 2003.

[4]. H. I. Arumawadu, R. M. K. T. Rathnayaka and S. K. Illangarathne, "Mining Profitability of Telecommunication Customers Using K-Means Clustering," Journal of Data Analysis and Information Processing, pp. 63-71, 2015.

[5]. I. K. Savvas, C. Chaikalis, F. Messina and D. Tselios, "Understanding customers' behaviour of telecommunication companies increasing the efficiency of clustering techniques," in 25th Telecommunication Forum (TELFOR), Belgrade, Serbia, 2017.

[6]. J. Qian and C. Gao, "The application of Data Mining in CRM," in 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Dengleng, China, 2011.

[7]. D. Wang and X. Zhang, "Mobile user stability prediction with Random Forest model," in International Conference on Data Science and Advanced Analytics (DSAA), Shanghai, China, 2014.

[8]. J. Magnusson and T. Kvernvik, "Subscriber classification within telecom networks utilizing big data technologies and machine learning," in 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Beijing, China, 2012.

[9]. J. Alhilman, M. M. Rian, W. Marina and K. Margono, "Predicting and clustering customer to improve customer loyalty and company profit," in 2nd International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 2014.

[10]. D. Pyle, Business Modeling and Data Mining, San Francisco, CA, USA: Morgan Kaufmann, 2003.

[11]. J. Panuš, H. Jonášová, K. Kantorová, M. Doležalová and K. Horáčková, "Customer segmentation utilization for differentiated approach," in 2016 International Conference on Information and Digital Technologies (IDT), Rzeszow, Poland, 2016.

[12]. V. Punyangarm, P. Y. a. P. Charnsethikul and S. Lertworasirikul, "A Credibility Approach for Fuzzy Stochastic Data Envelopment Analysis (FSDEA)," in 7th Asia Pacific Industrial Engineering and Management Systems Conference, Bangkok, Thailand, 2006.

[13]. H. Shin and S. Sohn, "Multi-attribute scoring method for mobile telecommunication subscribers," Expert Systems with Applications, vol. 26, no. 3, pp. 363-368, 2004.

[14]. S. M. Amin, U. N. U. Ahmad and L. S. Hui, "Factors Contributing to Customer Loyalty Towards Telecommunication Service Provider," Procedia - Social and Behavioral Sciences, vol. 40, pp. 282-286, 2012.

[15]. J. A. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression," Journal of Business Research, vol. 60, p. 656–662, 2006.

[16]. Xin-an Lai, "Segmentation study on enterprise customers based on data-mining technology," IEEE First International Workshop on Database Technology and Applications, pp. 247-250, 2009.

[17]. H. Gong and Q. Xia, "Study on application of customer segmentation based on data mining technology," IEEE Conference on ETP International Conference on Future Computer and Communication, pp. 167-170, 2009.

[18]. V. Aggelis and D. Christodoulakis, "Customer Clustering using RFM analysis," 9th WSEAS International Conference on Computers, Special Session Data Mining, Techniques and Application, 2005.

[19]. T. Yuanli and S.H.A.O. Liangshan, "Customer segmentation based on Ant clustering Algorithm," IEEE Second Conference On Computational Intelligence and Neural computing (CINC), vol. 1, pp. 133-136, 2010.

[20]. X. Qin, Y. Huang and G. Deng, "Improved K-Means algorithm and application in customer segmentation," IEEE International Conference on Web Information Systems and Mining, pp. 13-16, 2010.

[21]. T. Haining, XuJuanjuan and Z. Bian, "Research on Index System of Dynamic Customer Segmentation: Based on the Case Study of China Telecom," IEEE International Conference on Information Management and Engineering, ICIME,pp. 441-445, 2009.

Ameer Mohamed Aslam Sujah, "Telecommunication Customer Categorization with Novel Data Mining Approach for Effective Communication and Increase Profitability" IOSR Journal of Computer Engineering (IOSR-JCE) 21.5 (2019): 55-66