MINING PROFITABILITY OF TELECOMMUNICATION CUSTOMERS AND CUSTOMER SEGMENTATION WITH NOVEL DATA MINING APPROACH

A.M.A. Sujah[1] & R.M.K.T. Rathnayaka[2]

Correspondence: ameersujah@gmail.com

ABSTRACT

Telecommunication industry plays a vital role in the fast-moving modern 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 behaviour 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 three phases. The first phase is the comparison of different K-means algorithm and chooses 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 behaviours by using K-means++ algorithm 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 K-means++ algorithm to provide different promotional packages. Dataset consists of 12,000 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 one 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

1. INTRODUCTION

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,

[1] Department of Computing and Information Systems, Sabaragamuwa University of Sri Lanka.
[2] Department of Physical Sciences and Technology, Sabaragamuwa University of Sri Lanka.

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. Therefore, provide different benefits to the different group of customers 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, 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 [5] and machine learning algorithms. Also, provide a solution for customer segmentation and introduce promotional packages to the different level of loyality customers [6].

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 on “Mining 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 and the dataset consist of 12,000 prepaid customer details.

2. METHODOLOGY

2.1 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. Initially, retrieve the telecommunication customer dataset from Comma Separated Values (CSV) file to Spyder IDE 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 in the preprocessing phase.

2.2 Build the Training and Testing Datasets

Machine learning algorithms are learning from data which is given to the system as a training dataset. From the training dataset it will find relationships between the attributes and testing dataset that helps to measure the actual performance of the algorithms. Initially, preprocessed dataset divided into two sets that namely training and testing dataset. Training dataset consists 80% of the entire dataset and rest of the dataset used for testing the prediction model.

2.3 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 previous 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. Next step importing the python libraries which require to carry out the processes such as pandas, numpy, and matplotlib. 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 valued. 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 a best one. Apply that algorithm into our dataset and get the clustering result by changing the K values. Plot the Receiver Operating Characteristic (ROC) 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 the 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 required 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. 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 this 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 1 shows the flowchart of the proposed system.

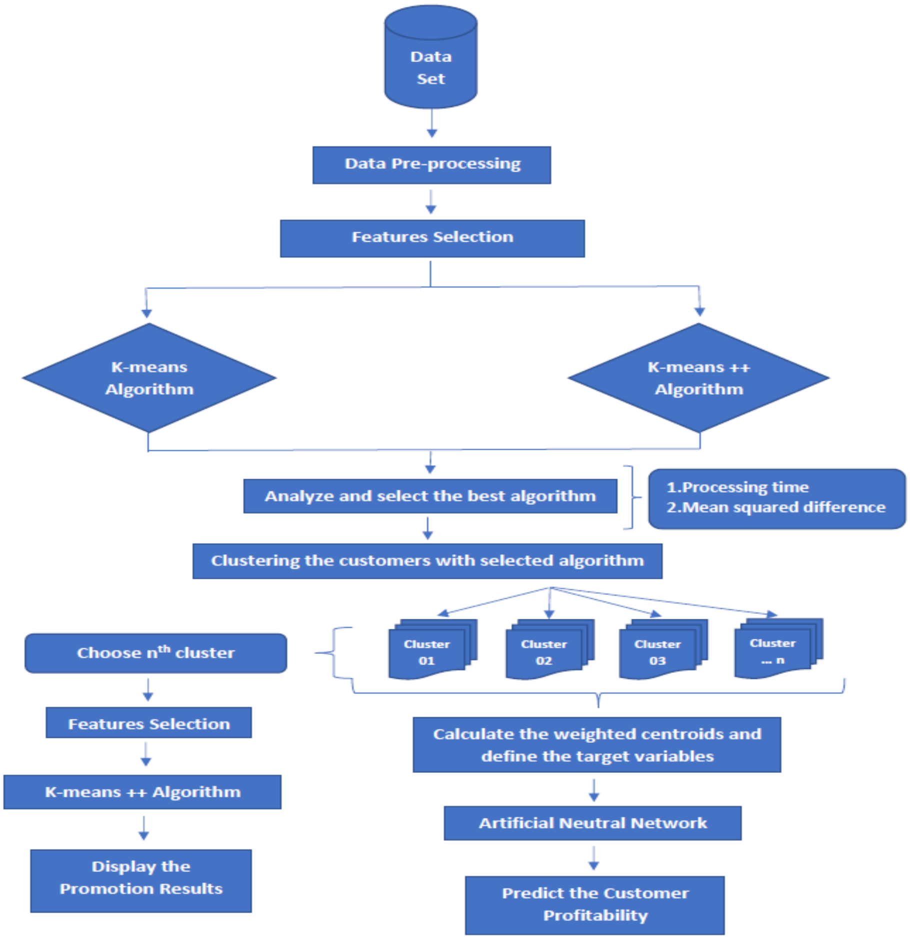
Data Set
Data Pre-processing
Features Selection
K-means Algorithm
K-means ++ Algorithm
Analyze and select the best algorithm
1.Processing time
2.Mean squared difference
Clustering the customers with selected algorithm
Choose nth cluster
Cluster 01
Cluster 02
Cluster 03
Cluster ... n
Features Selection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

Figure 1 Flowchart of the Proposed System

3. RESULTS AND DISCUSSION

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 2 shows that how the processing time and WCSS value changes with different k values for both algorithms.

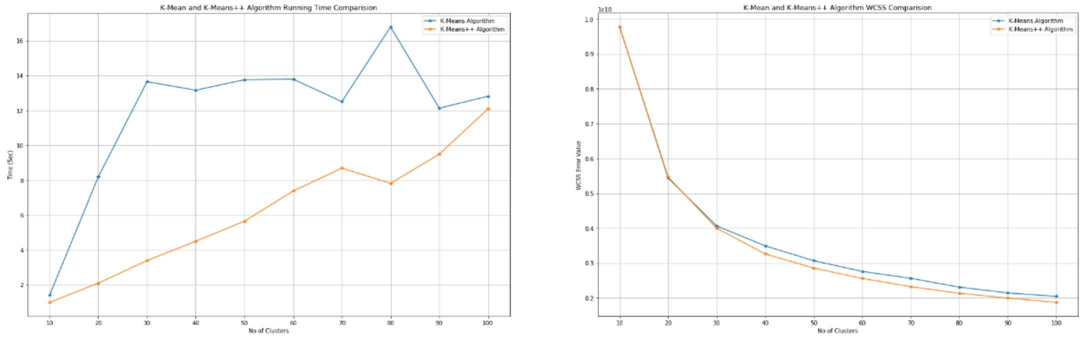

Figure 2 Processing Time and WCSS Value Comparison with Different K Values for Both Algorithms

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.

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 ROC curve. By analyzing the curve got the result of k = 5. Elbow method is applied to identify the best k value. Figure 3 display the ROC curve for the above customer dataset.

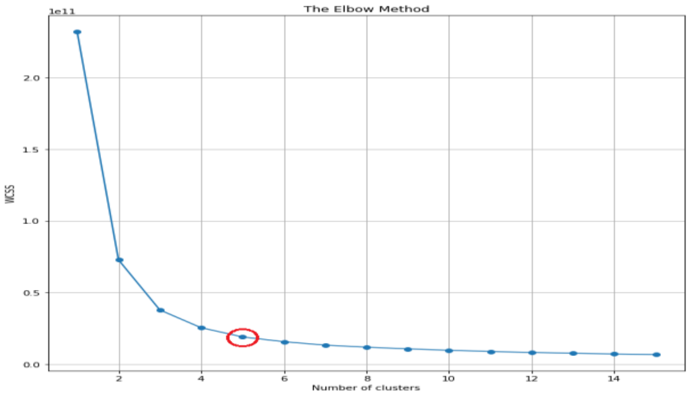

Figure 3 ROC Curve for Customer Dataset

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 1 shows the centroid values for each cluster with nine attributes.

Table 1 Attributes 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

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 (kth cluster).

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

Customer profitability levels are determined by calculating the final weighted cluster centroid values (v_k) of each cluster. By analyzing the table 2, cluster - 1 is determined as a most profitable customer category and cluster - 4 contains the lowest 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 2 Final Weighted Cluster Centroids

Cluster's Name	V_k
Cluster 1	4130.94
Cluster 2	1357.96
Cluster 3	3070.47
Cluster 4	591.388
Cluster 5	2166.51

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.

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 ANN 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 3 shows the confusion matrix of this research experiment for the dataset that the size of 2,000 customer.

Table 3 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

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). 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.

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.

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. ROC curve used to find out the best k vales for the clustering by using k-means plus plus algorithm. According to ROC curve, the value five chosen for k and cluster the customers into five 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 4 shows the clustered results.

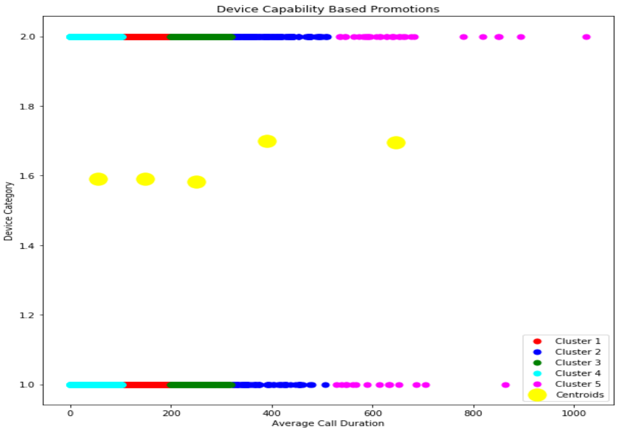

Figure 4 Promotional Cluster Results

According to the graph, lowest profitable customers are categories into 5 different groups. So, service provider can provide same promotional package with different percentage (example of 10%, 15%, 20%, 25% and 30%) among the customers based on their average call durations. Here, cluster - 5 got 30% promotions (highest promotion) and cluster - 4 got 10% promotions (lowest promotion) because of their usage behavior and 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.

4. 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 recommends the best clustering algorithm by comparing the K-mean and K-means plus plus. Processing time and WCSS values are used to evaluate the algorithms. The customers are clustering into five groups by using nine different customer behaviors. This clustering results are high accurate because of comparing large number of attributes. Research study achieved the high accuracy of 97.3% for profitability prediction model. This accuracy achieved by using the clustered result as a target variable. Finally, this research experiment recommends different promotions to their customers according to their behaviors. Manager can provide different promotional packages to their customers by using this prediction model and they can identify their customer’s profitability level individually.

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 algorithms to develop prediction model gives more benefits. This research experiment used 10,000 customer details. This research study recommends develop a model for handling big data in the future.

5. REFERENCES

[1] Chang, P., & Chong, H. (2011). Customer satisfaction and loyalty on service provided by Malaysian telecommunication companies. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*. Bandung, Indonesia: IEEE. Retrieved from https://ieeexplore.ieee.org/document/6021730/authors#authors

[2] Qirong, H., Wenqing, L., Eran, S., Shonali, K., & Jingxuan, W. (2016). 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* (pp. 1353-1362). Indianapolis, Indiana, USA. Retrieved from https://dl.acm.org/citation.cfm?id=2983354

[3] Min, X., Yuhui, Q., & Jing, Q. (2003). Mining for profitable customers. In *International Conference on Information Technology: Coding and Computing*. Las Vegas, NV, USA, USA: IEEE. Retrieved from https://ieeexplore.ieee.org/document/1197605

[4] Arumawadu, H., Rathnayaka, R., & Illangarathne, S. (2015). Mining Profitability of Telecommunication Customers Using K-Means Clustering. *Journal Of Data Analysis And Information Processing*, *03*(03), 63-71. doi: 10.4236/jdaip.2015.33008

[5] Tselios, D., Messina, F., Chaikalis, C., & Savvas, I. (2017). Understanding customers' behaviour of telecommunication companies increasing the efficiency of clustering techniques. In *25th Telecommunication Forum (TELFOR)*. Belgrade, Serbia: IEEE. Retrieved from https://ieeexplore.ieee.org/document/8249274

[6] Jianing, Q., & Changchun, G. (2011). The application of Data Mining in CRM. In *2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*. Dengleng, China: IEEE. Retrieved from https://ieeexplore.ieee.org/document/6010697

[7] Danqin, W., & Xiaolong, Z. (2014). Mobile user stability prediction with Random Forest model. In *International Conference on Data Science and Advanced Analytics (DSAA)*. Shanghai, China: IEEE. Retrieved from https://ieeexplore.ieee.org/document/7058108

[8] Jonathan, M., & Tor, K. (2012). 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* (pp. 77-84). Beijing, China. Retrieved from https://dl.acm.org/citation.cfm?id=2351327

[9] Alhilman, J., Rian, M., Marina, W., & Margono, K. (2014). Predicting and clustering customer to improve customer loyalty and company profit. In *2nd International Conference on Information and Communication Technology (ICoICT)*. Bandung, Indonesia: IEEE. Retrieved from https://ieeexplore.ieee.org/document/6914087/authors#authors

[10] Panuš, J., Jonášová, H., Kantorová, K., Doležalová, M., & Horáčková, K. (2016). Customer segmentation utilization for differentiated approach. In *International Conference on Information and Digital Technologies (IDT)*. Rzeszow, Poland: IEEE. Retrieved from https://ieeexplore.ieee.org/document/7557178

[11] Shin, H., & Sohn, S. (2004). Multi-attribute scoring method for mobile telecommunication subscribers. *Expert Systems With Applications*, *26*(3), 363-368. Retrieved from https://doi.org/10.1016/j.eswa.2003.09.013

[12] Mohamad Amin, S., Ungku Ahmad, U., & Lim, S. (2012). Factors Contributing to Customer Loyalty Towards Telecommunication Service Provider. *Procedia - Social And Behavioral Sciences*, *40*, 282-286. Retrieved from https://doi.org/10.1016/j.sbspro.2012.03.192