Telecom Customer Segmentation Using K-Means and Two-Step Clustering Algorithm

Salar Masood[1], Moaz Ali[1], Faryal Arshad[1], Ali Mustafa Qamar[1], Aatif Kamal[1],
Ahsan Rehman[2], Summaya Mumtaz[1], Khurram Javed[1]
[1]Department of Computing, School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST)
[2]Business Analytic Consultant, IBM - Global Business Services
Islamabad, Pakistan
{09bitsmasood, 09bitmoaza, 09bitfarshad, mustafa.qamar, aatif.kamal,12mscssmumtaz, 12mscskjaved}@seecs.edu.pk
ahsanr@ibm.com

ABSTRACT: *Pakistan hosts a competitive and fluid telecommunication market and for a company to sustain, create customer value and increase economic efficiency, it needs to better understand its customers. The purpose of clustering or customer segmentation is to deliver actionable results for marketing, product development and business planning. In this paper, we focus on customer segmentation using clustering algorithms on real data of a telecommunication company in Pakistan. After choosing appropriate attributes for clustering, we used two-step clustering algorithm and k-Means algorithm for creating different customer segments. The results of both algorithms are compared and the insights obtained from each segment were analyzed before suggesting marketing strategies.*

Keywords: Customer Segmentation, Two-Step Clustering Algorithm, Business Intelligence, K-Means Clustering Algorithm, Clustering, Segmentation

Received: 17 August 2013, Revised 21 September 2013, Accepted 26 September 2013

1. Introduction

Pakistan hosts the world’s largest and the most experienced telecommunication companies. The number of mobile subscribers has reached around 123 million, with more than 90% of the country having cellular services along with a tele-density of more than 62% as of January 2013. In such a competitive environment, in order to increase profits and create a loyal customer base, a company first needs to better understand its customers. For this purpose, different clustering algorithms are used in this research work.

Segmentation is applied in order to classify the customers into different groups according to one or more attributes. The customers within the same group have greater similarity, and the ones in a different group have greater differences. It further helps in better understanding the customers, producing optimal price plans, creating tailored products and ultimately helps in reducing churn. Moreover, it helps in providing a multidimensional view of the customer for better marketing. Cross and Thompson [2] have applied supervised (decision trees) as well as unsupervised (K-means clustering) machine learning algorithms to identify customer segments and to predict different risks.

Another related work is by McCarty and Hastak [7] who compared different segmentation approaches such as RFM, CHAID (CHi-squared Automatic Interaction Detection) and logistic regression in data mining.

Li [5] has identified various customer segments belonging to a retail supermarket before using association rules in order to perform customer characteristics analysis. Zhang et al. [9] clustered telecom customers of Liaoning, China based on consumer's behavior. Like the work of Li [5], they also described the characteristics of different cluster groups. Furthermore, they also put forward a marketing strategy based on their study.

In this research, we first applied the two step clustering algorithm on real customer data of a telecommunications company of Pakistan pertaining to the call usage, revenue and recharge analysis, The two-step algorithm was selected since it can do clustering even if the categorical attributes are used which was the case here. Moreover, K-means clustering algorithm was also applied using the same attributes that were used for two-step clustering algorithm. The results obtained from both algorithms were compared. Although many related research works have been performed earlier, we know of no such work in the Pakistan telecommunication market.

The paper is organized as follows: Section II describes the research methodology including pre-processing, data transformation, correlation analysis, the use and comparison of two-step clustering algorithm as well as k-means algorithm along with their results. Section III discusses various marketing strategies which could be developed based on the performed analysis whereas the paper is in concluded in Section IV along with shedding light on future perspectives.

2. Research Methodology

This section describes in detail the research methodology including the preprocessing, data transformation, correlation analysis, application of the two-step clustering algorithm and the K-means algorithm. In the first step, the data was preprocessed and transformed into categorical variable types. Next, correlation analysis was performed so as to select the attributes for clustering. This is followed by performing clustering on customer call usage data, revenue data and recharge analysis data resulting in five revenue segments. Each of the revenue segments were further divided into five sub-segments based on usage and recharge data.

2.1 Preprocessing and Transformation of data

Preprocessing is applied to fill in missing values, removing the outliers and resolving the inconsistencies in the data. The data set comprised of nine days of 5, 109 customers' daily call usage, SMS(Short Message Service) usage and revenue generation. Table 1 and 2 shows the descriptive statistics about the revenue generation, call usage and SMS usage attributes.

	Final_Revenue	Calls_Revenue	SMS_Revenue	VAS_Revenue
Count	5109	5109	5109	5109
Mean	4.019	3.680	3.278	3.245
Sum	20534	18800	16748	16577
Min	1.000	2.000	2.000	2.000
Max	7.000	6.000	5.000	5.000
Range	6.000	4.000	3.000	3.000
Variance	3.958	2.207	1.368	1.378
Standard Deviation	1.989	1.486	1.170	1.174
Standard Error of Mean	0.028	0.021	0.016	0.016
Median	4.000	4.000	3.000	3.000
Mode	3.000	2.000	2.000	2.000

Table 1. Descriptive Statistics of Revenue generation attribute

	SMS _Usage	**On-Net_ SMS_Usage**	**Off-Net_ SMS_Usage**	**Calls_ Usage**	**On-Net_ Calls_Usage**	**Off-Net Calls_Usage**	**Peak_Call _Usage**	**Off-Peak_ Calls _Usage**
Count	5109	5109	5109	5109	5109	5109	5109	5109
Mean	3.225	3.246	2.211	5.630	4.147	3.275	3.530	2.278
Sum	16479	16583	1129	28765	21188	16730	18036	11638
Min	2.000	2.000	2.000	2.000	2.000	2.000	1.000	2.000
Max	5.000	5.000	3.000	10.000	7.000	5.000	6.000	3.000
Range	3.000	3.000	1.000	8.000	5.000	3.000	5.000	1.000
Variance	1.338	1.384	0.167	7.333	3.271	1.299	2.889	0.201
Standard Deviation	1.157	1.176	0.408	2.708	1.809	1.140	1.700	0.448
Standard Error of Mean	0.016	0.016	0.006	0.038	0.025	0.016	0.024	0.006
Median	3.000	3.000	1.000	6.000	4.000	3.000	4.000	2.000
Mode	2.000	2.000	2.000	2.000	2.000	2.000	3.000	2.000

Table 2. Descriptive Statistics of calls and SMS usage attributes

Using box-plot (as shown in Figure 1-4), the outliers are identified and replaced by the mean of the attributes. The missing values are also replaced by the mean of the attributes. This daily data was aggregated using pivot tables so as to get good insights from different cluster.

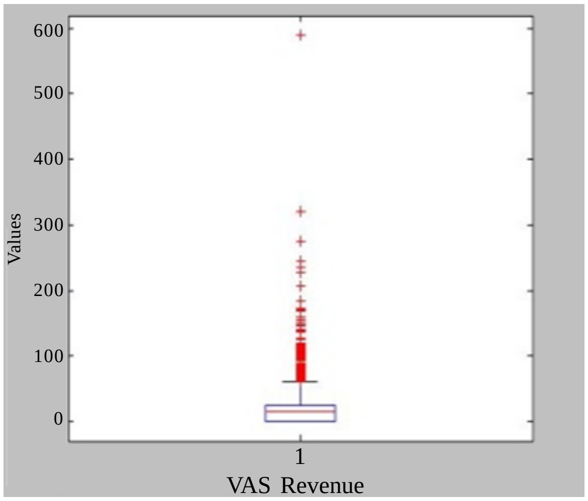

Figure 1. Boxplot for VAS Revenue

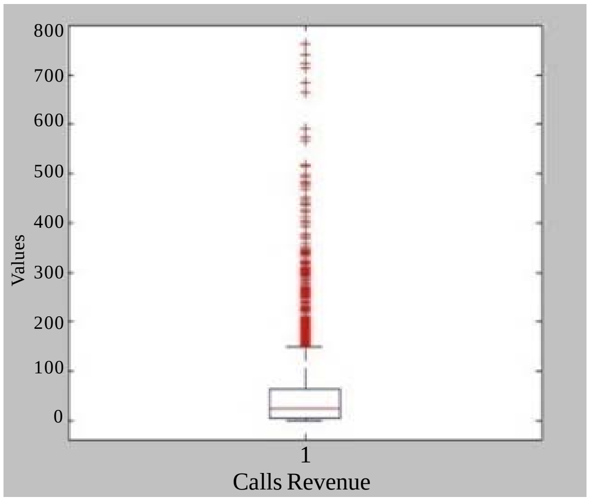

Figure 2. Boxplot for Calls Revenue

This was followed by performing correlation analysis in order to see the relationship between two attributes on the basis of Pearson correlation value. A positive value of correlation means that if one variable increases, the other also increases and vice versa. If the value is negative, it means that if one variable decreases the other increases and vice versa [1]. The values of Pearson correlation fall in the range of [–1, 1]. Table 3 shows the various attributes used in correlation analysis.

Table 4 shows the correlation results for the revenue attributes. The Pearson correlation values for revenue related to calls, SMS and VAS with the *Final Revenue* attribute are 0.925, 0.592 and 0.568 respectively. These values are quite high (more than 0.5), which shows that these attributes have a strong relationship with the *Final Revenue* attribute. Therefore, these attributes could be selected for clustering purposes.

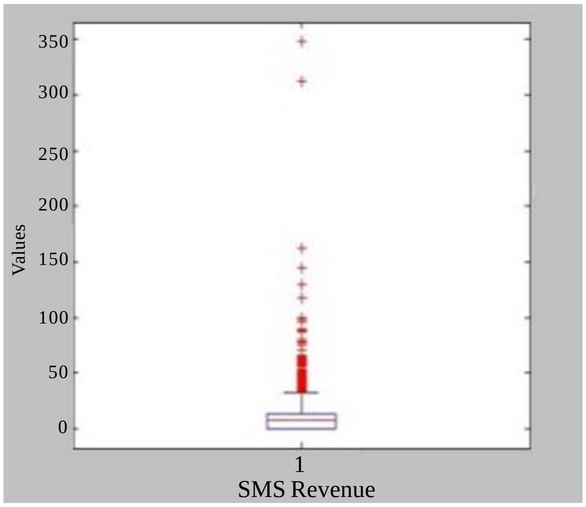

Figure 3. Boxplot for SMS Revenue

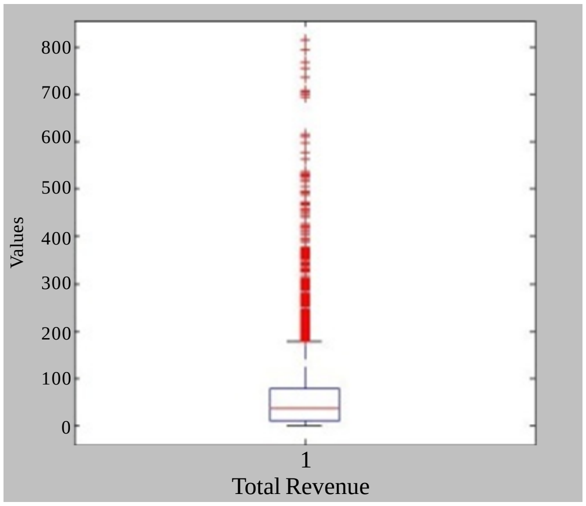

Figure 4. Boxplot for Total Revenue

Final Revenue (Pakistani Rupees-PKR)
Calls Revenue
SMS Revenue
VAS (Value Added Services) Revenue
Call Usage (Minutes of Use-MOU)
Peak Call Usage (MOU)
Off Peak Call Usage (MOU)
On-Net Call Usage (MOU)
SMS Usage (Count)
On-Net SMS Usage (Count)
Off-Net SMS Usage (Count)

Table 3. Attributes used for Correlation Analysis

	Final Rev	**Calls Rev**	**SMS Rev**	**VAS Rev**
Final Rev	1	0.925	0.592	0.568
Calls Rev	0.925	1	0.429	0.405
SMS Rev	0.592	0.420	1	0.912
VAS Rev	0.568	0.405	0.912	1

Table 4. Correlation Matrix for Revenue Attribute

	Calls Usage	**On-Net**	**Off-Net**	**Peak**	**Off-Peak**
Calls Usage	1	0.951	0.493	0.910	0.455
On-Net	0.951	1	0.335	0.854	0.432
Off-Net	0.493	0.335	1	0.594	0.261
Peak	0.910	0.854	0.594	1	0.424
Off-Peak	0.455	0.432	0.261	0.424	1

Table 5. Correlation Matrix for Call Usage Attribute

Table 5 shows the results for different attributes related to the call usage. The call usage parameter for On-Net, Off-Net, Peak and Off-Peak had correlation values of 0.951, 0.493, 0.910 and 0.455 respectively with the Calls Usage attribute. However, only On-Net calls usage and Peak calls usage were selected for clustering purpose since their values are more than 0.5.

Table 6 shows the values for Pearson correlation related to the SMS usage attributes. In this case, both On-Net SMS as well as Off-Net SMS have high values (0.862 and 0.832 respectively) and hence are selected for clustering purposes.

	Total SMS	On-Net	Off-Net	Peak	Off-Peak
Total SMS	1	0.862	0.832	0.964	0.645
On-Net	0.862	1	0.437	0.877	0.425
Off-Net	0.832	0.437	1	0.752	0.680

Table 6. Correlation Matrix for SMS Usage Attribute

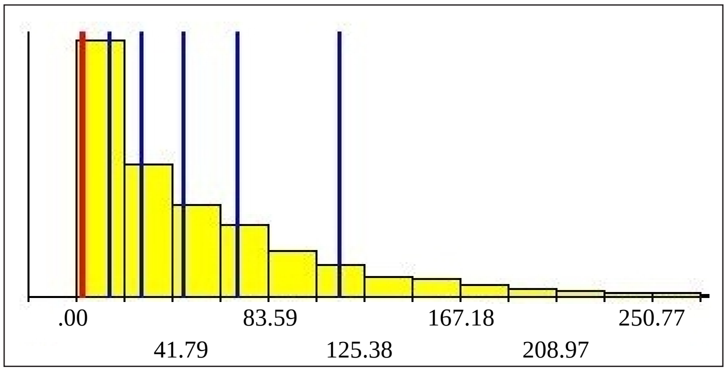

Figure 5. Binning *Call Usage* Attribute with Ten Bins

2.2 Discretization of Continuous Attributes

It is easier for a company to divide its customers as low, medium and high valued customers and analyze them rather than assigning some numeric value to them. In our case, all of the aforementioned attributes were continuous and were discretized with an equal number of cases in each bin. Figure 5 and Table 7 show the bins formed for *Final Revenue* attribute. There are seven bins in total which are: less than 2, 2-13, 14-27, 28-45, 46-69, 70-113, 114+. We used visual binning method with equal percentiles to discretize our data for each attribute. The minimum number of bins formed by SPSS against total revenue was 7 and more bins could not be formed since whole of the customer data was equally distributed in these bins.

No.	Label
1	< 2
2	2-13
3	14-27
4	28-45
5	46-69
6	70-113
7	114 +

Table 7. Bins for *Final Revenue*

No	Value	Label
1	Less than 0	< 0
2	4	0 -3
3	11	4 - 10
4	20	11 -19
5	33	20 - 32
6	57	33 -56
7	96	57 - 95
8	165	96 -164
9	317	165 - 316
10	High	317 +

Table 8. Binning *Call Usage* Attribute

Similarly the *Calls Usage* attribute was also discretized to 10 bins in total which are: less than 0, 0-3, 4-10, 11-19, 20-32, 33-56, 57-95, 96-164,165-316, 317+ as shown in Figure 6 and Table 8. Here also, the visual binning method was employed with equal percentiles to discretize our data for each attribute. The minimum number of bins formed by SPSS against *Calls Usage* was 10.

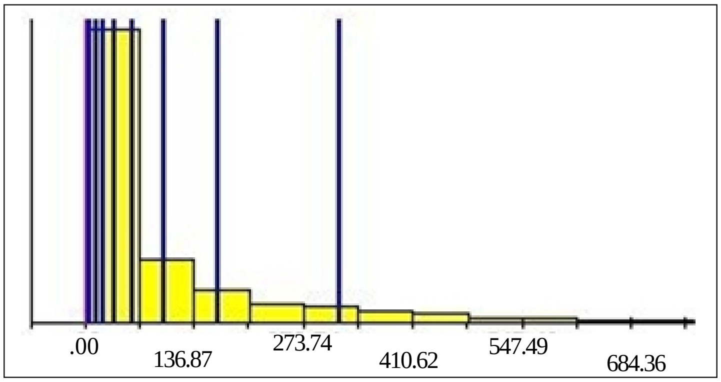

Figure 6. Binning *Call Usage* Attribute with Ten Bins

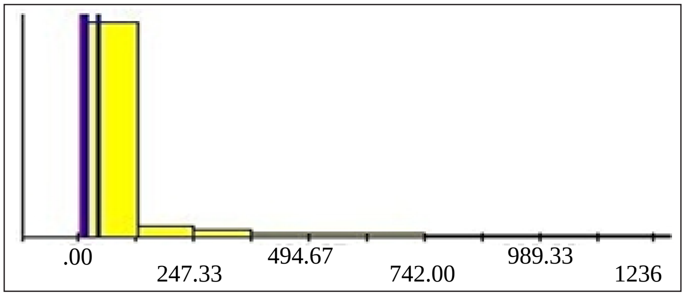

Figure 7. Binning *SMS Usage* attribute with Five Bins

No	Value	Label
1	Less than and Equal to	0 <= 0
2	7	1 - 7
3	11	8 - 11
4	36	12 - 36
5	High	3/+

Table 9.Binning SMS Usage Attribute

No.	Label
1	0 - 9
2	10 - 24
3	25 - 46
4	47 - 84
5	85+

Table 10 . Calls Revenue

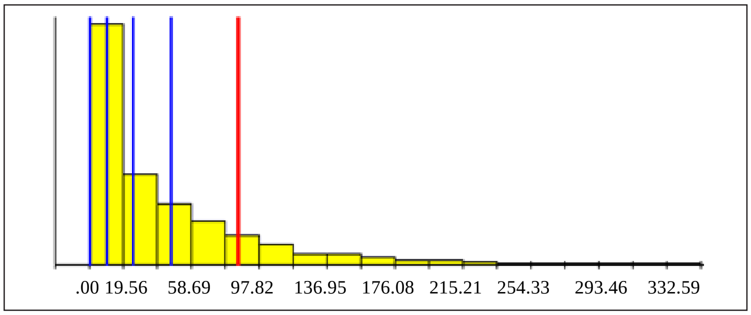

Figure 8. Binning *Calls Revenue* Attribute with Five Bins

Moreover, the *SMS Usage* attribute was also discretized into 5 bins which are less than 0, 1-7, 8-11, 12-36, and 37+ as shown in Figure 7 as well as Table 9.

Similarly, *Calls Revenue* attribute was discretized into five bins which are less than 0, 0-9, 10-24, 25-46, 47-84 and 85+ as shown in Figure 8 and Table 10.

In the similar manner, *SMS Revenue* attribute was discretized into 5 bins which are less than 0, 0-3, 4-9, 10-15 and 16+ as shown in Figure 9 and Table 11.

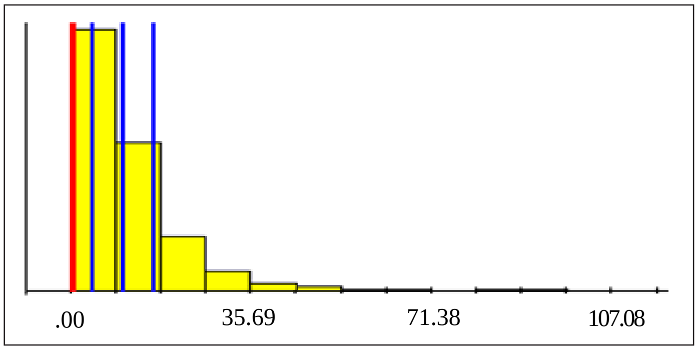

Figure 9. Binning *SMS revenue* attribute with Five Bins

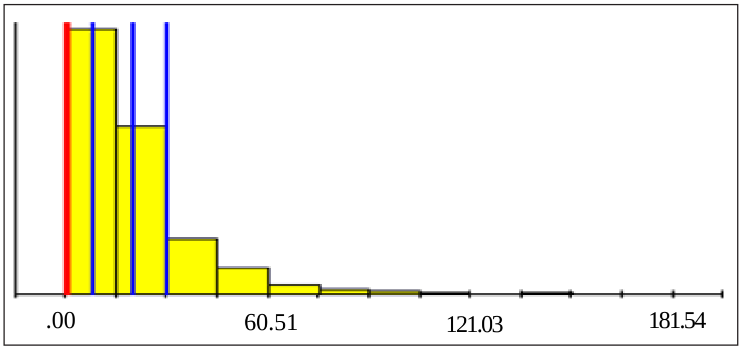

Figure. 10. Binning *VAS Revenue* attribute with Five

VAS Revenue attributes was also discretized into five bins which are less than and equal to 0, 1-8, 9-20, 21-30 and 31+ as shown in Figure 10 and Table 12.

No.	Label
1	< 0
2	0 - 3
3	4 - 9
4	10 - 15
5	16 +

Table 11. SMS Revenue

No.	Label
1	< .00
2	1.00 - 8.00
3	9.00 - 20.00
4	21.00 - 30.00
5	31.00+

Table 12. VAS Revenue

2.3 Two Phase/Step Clustering Algorithm

Namvar et al. [8] developed a two phase clustering algorithm for intelligent customer segmentation. In the first step, they used K-means clustering in order to cluster the customers into different segments using their RFM (Recency, Frequency, Monetary) value. This was followed by further clustering of each cluster based on the demographic data. The same algorithm is used in this research since this algorithm can handle large data sets having both categorical and continuous attributes at the same time. This algorithm works in two steps:

1. In the first step, all of the instances are assigned to pre-clusters.

2. During the second step, these pre-clusters are handled as individual cases and are further clustered using the hierarchical clustering algorithm. The two step algorithm can determine itself the number of clusters or else they could also be specified by the user.good clustering has been done by the algorithm as visible from Figure 11.

A rather classical approach is to cluster customers based on their Call Detail Records (CDRs) [6]. However in our case, the customers were segmented based on their revenue attributes such as *Final Revenue*, *Calls Revenue*, *SMS Revenue* as well as *VAS Revenue*. The number of clusters was manually selected as five because the value for silhouette measure of cluster cohesion and separation (which indicates the results as poor, fair or good) was found to be more than 0.5 which indicates that

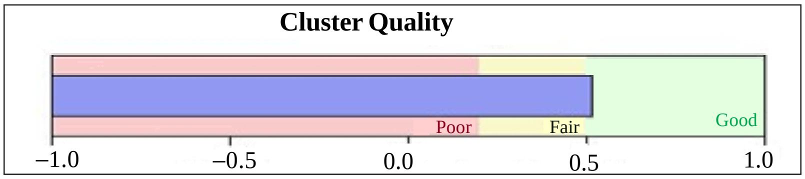

Figure 11. Silhouette measure of cohesion and Separation

The classification of results into categories such as poor, fair or good are based on the work of Kaufman and Rousseeuw [4] regarding interpretation of cluster structures. The silhouette measure averages over all of the instances as shown in equation below:

$$\frac{B - A}{max\,(A, B)}$$

Where **A** is the distance between an instance and the center of the cluster to which it belongs and **B** is the distance between an instance and the nearest center from nearby clusters. A silhouette coefficient of **1** would mean that all cases are located directly on their cluster centers. A value of **−1** would mean that all cases are located on the cluster centers of some other cluster. A value of **0** shows that on average, cases are equi-distant between their own cluster center and the nearest different cluster.

The noise handling was kept to 3%. This meant that the algorithm ran on 4517 cases after removing 592 outlier cases from a total of 5109 customers. With this value, we got a silhouette value more than 0.5 which classifies our result as good. With noise handling less than 3%, the results were fair.

2.4 Results of Two Step Clustering Algorithm

The clusters obtained as a result of using two-step clustering algorithm are given in Table 13 and shown in Figure 12. One can observe that the biggest cluster is cluster 2. This cluster contains 22.9% (1035 cases) of the total customers. The customers in this cluster have total revenue more than Rs. 21–30. One can also note that although cluster 2 is the biggest one and has got the largest total revenue, it does not have the largest SMS revenue as well as the VAS revenue.

No.	Total Cases	Total Rev	Calls Rev	SMS Rev	VAS Rev
1	922	> 114	> 84	> 16	> 31
2	1035	> 114	> 84	10 – 15	21 – 30
3	836	14 – 27	10 – 24	0 – 3	1 – 8
4	1006	< 2	0 – 9	0 – 3	1 – 8
5	718	2 –13	0 – 9	4 – 9	9 – 20

Table 13. Results of Two Step Clustering Algorithm

Similarly, cluster 4 is the second largest cluster having 22.3% (1006 instances) of the total instances. The customers in this cluster have the lowest total revenue(less than Rs.2) where as the calls revenue is between Rs. 0-9, calls revenue between Rs. 0-9, SMS revenue between Rs. 0-3 and VAS Revenue between Rs. 1-8.

Similarly, cluster 1 has got 20.4% (922 cases) of the total cases. The customers in this cluster have total revenue more than Rs 114, calls revenue more than Rs. 84, SMS revenue more than Rs. 16 and VAS revenue more than Rs. 31. This cluster has got the largest revenue with calls, SMS as well as VAS.

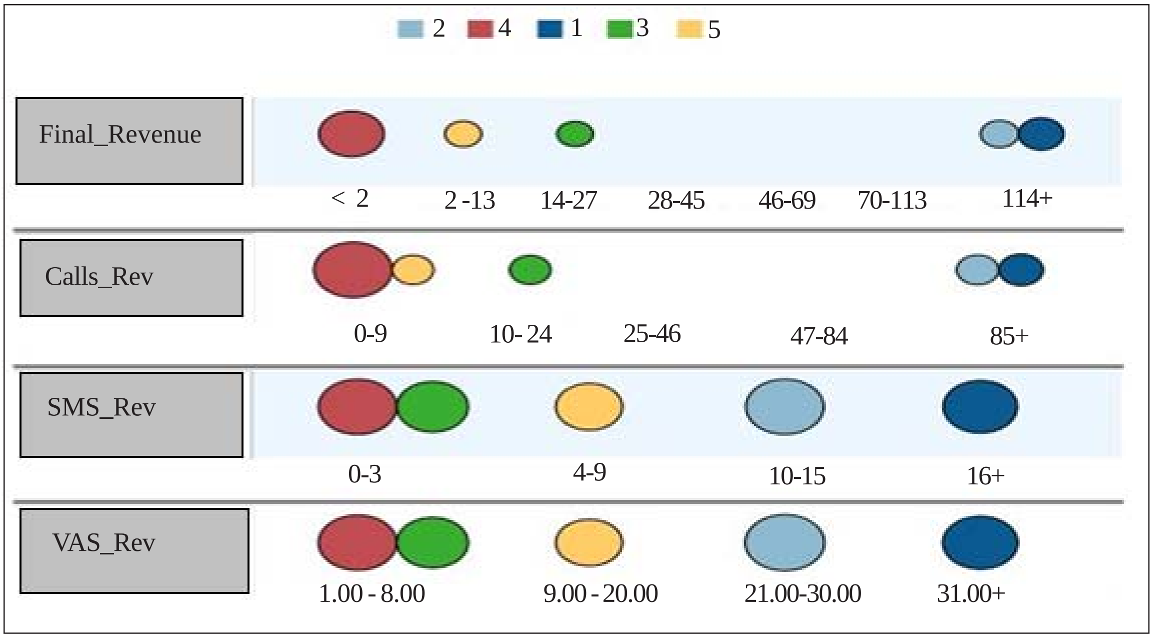

Figure 12. Revenue Clusters using Two-Step Clustering Algorithm

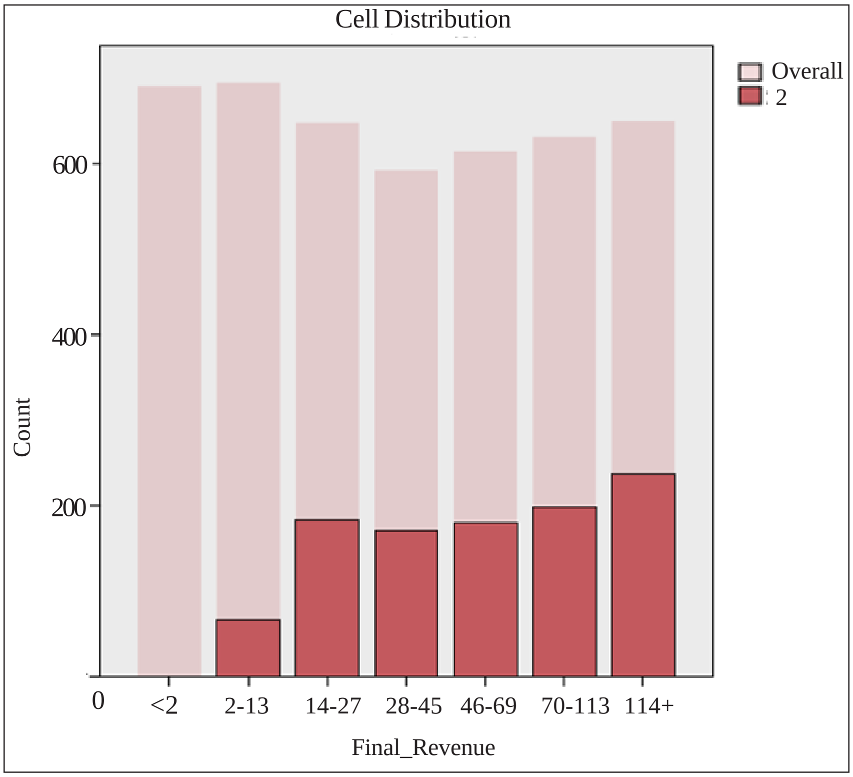

Figure 13. Final Revenue Count

Cluster 3 is relatively a smaller cluster, consisting of just 18.5% (836 cases) of the total instances. The customers in this cluster have total revenue between Rs. 14-27, calls revenue between Rs.10-24, SMS revenue between Rs.0-3 and VAS revenue between Rs. 1-8. The smallest cluster is cluster 5, having only 15.9% (718 cases) of the total cases. The customers in this cluster have total

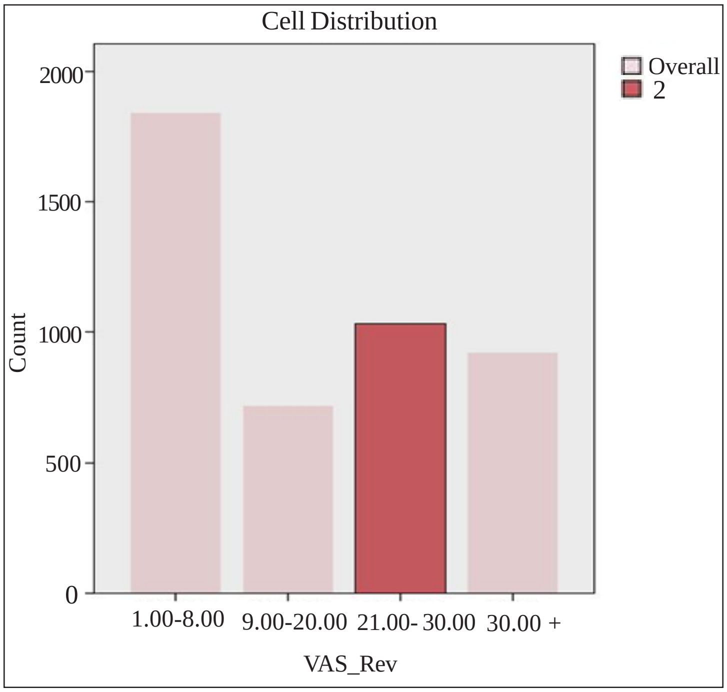

Figure 14. VAS Revenue Count

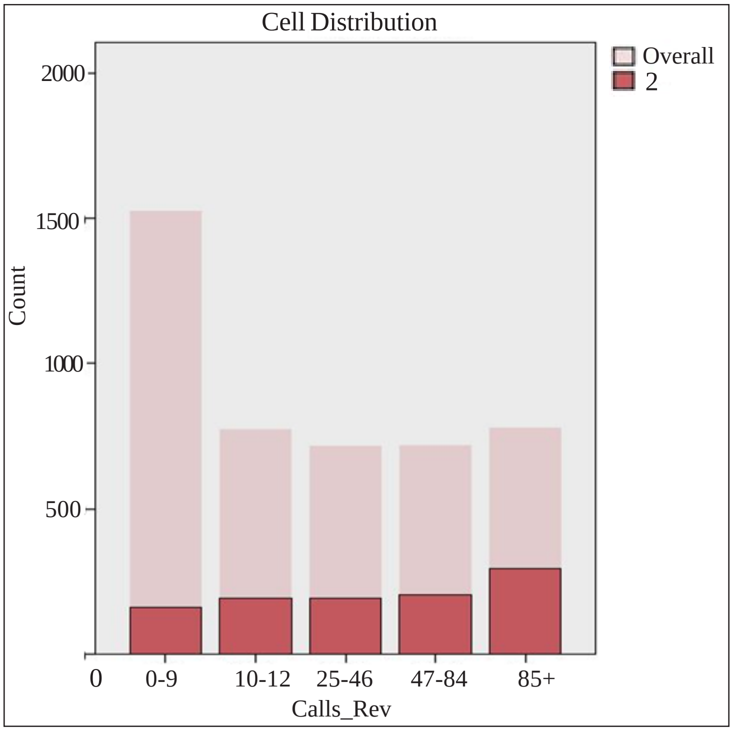

Figure 15. Calls Revenue Count

revenue between Rs. 2-13, calls revenue between Rs. 0-9, SMS revenue between Rs. 4-9 and VAS revenue between Rs. 9-20. In all of the clusters, the VAS revenue is greater than the SMS revenue. Similarly, calls revenue is also greater than SMS revenue for all of the clusters, a phenomenon which shows that the people have a higher tendency to make a call as compared to sending a SMS.

The next step was to perform further segmentation of each revenue segment based on customer's call and SMS usage. One of

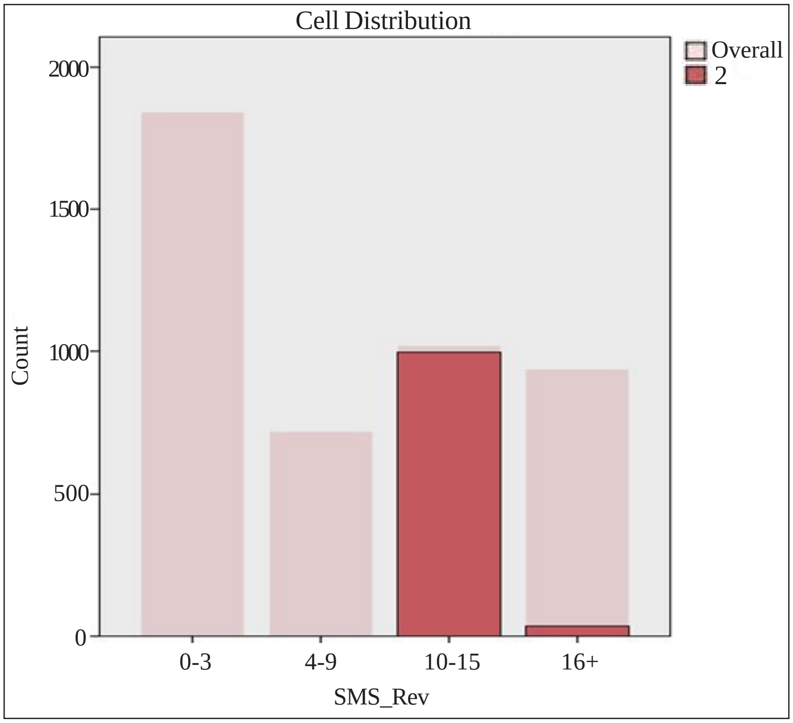

Figure 16. SMS Revenue Count

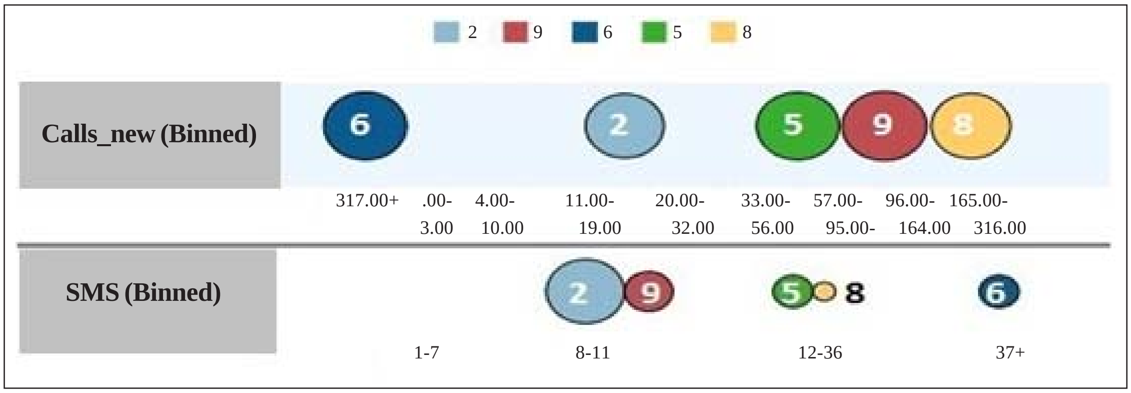

Figure17. Segmentation of Revenue Cluster

the challenges was that we did not know whether a customer is subscribed to free minutes and other such subscriptions. Because of these subscriptions, a segment might show more or less usage than the amount of revenue generated against it.

As a first step, the revenue cluster 2 was segmented into a number of clusters but only top five segments were analyzed and compared. In Figure 6, one can observe that the four clusters i.e. 5, 6, 8 and 9 which are of almost the same size fall in categories 33-56, 317+, 96-164 and 57-95 minutes of use respectively. Moreover, this justifies the high revenue of Rs. 114+ for cluster 2 as shown in Figure 17. Jansen [3] performed customer segmentation based on usage call behavior (incoming or outgoing). Moreover, he estimated the customer segment using Support Vector Machine (SVM) based on the customer profile. He got an accuracy of 80.3%.

Similarly, the cluster comparison on the basis of SMS (Figure 17) shows that this segment has relatively medium SMS usage. Most of the clusters falling in categories have a SMS count between 8-36 in total. In short, this cluster shows a higher call usage and medium SMS usage.

Cluster 4 of revenue was similarly analyzed and compared. Figure 19 shows that the most of the cases related to the calls fall within only 0-3 minutes of use, thereby justifying revenue of less than 2 for this segment of revenue. On analyzing the division along SMS usage, one can observe that a very big cluster falls in the 1st category i.e.1-7 SMS. In general, this cluster has very low call as well as SMS usage.

Figure 21 shows the segmentation of cluster 1. It can be observed that both call as well as SMS usage is very high in this segment as all segments fall in higher categories. Figure18 shows the cluster 3 of revenue further segmented on the basis of call and SMS usage. It can be seen that this segment has high call usage and low to medium SMS usage. Figure 20 shows cluster 5 segmented on the basis of call and SMS usage. This segment has medium call usage and very low to low SMS usage.

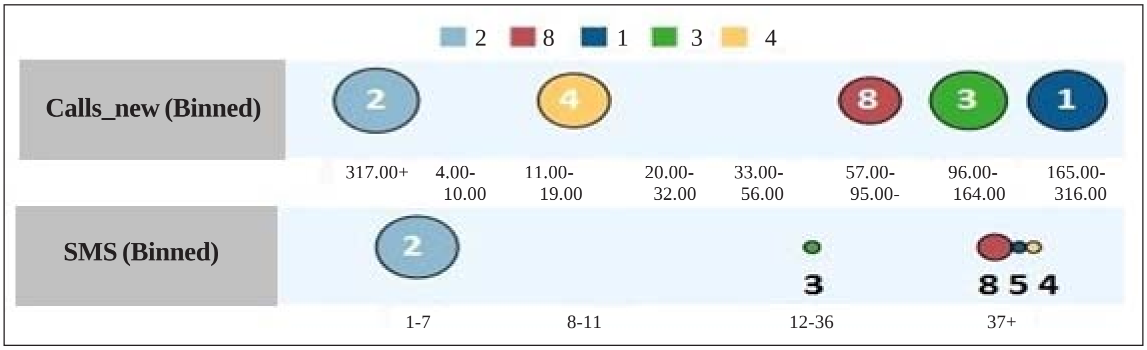

Figure 18. Segmentation of Revenue Cluster 3

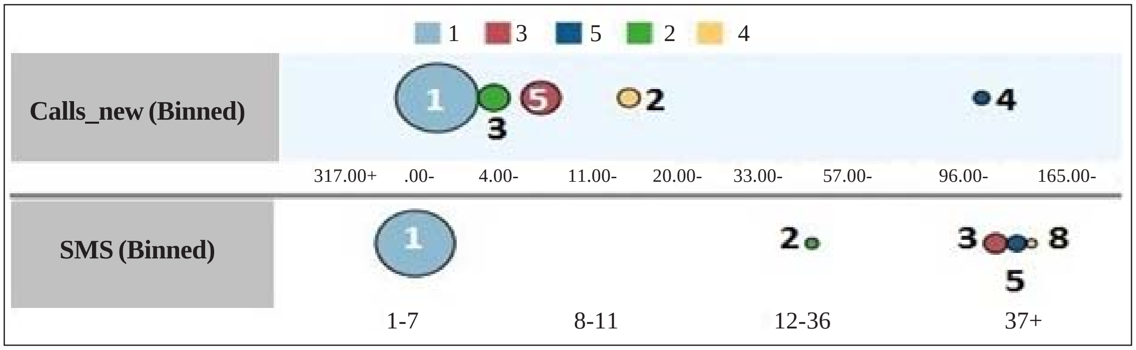

Figure 19. Segmentation of Revenue Cluster 4

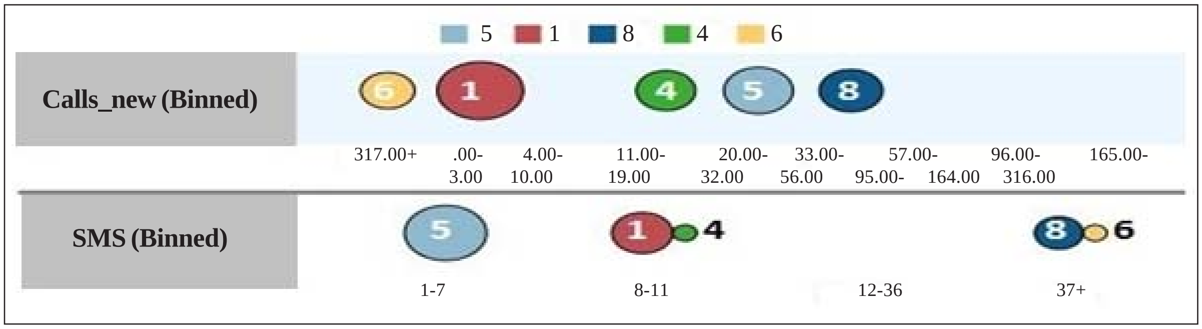

Figure 20. Segmentation of Revenue Cluster 5

2.5 K-Means Clustering Algorithm

K-Means is one of the most widely used algorithms in data mining and machine learning techniques. It is the simplest unsupervised learning algorithm that is used to solve the problem of clustering. K-centroids are used in order to represent clusters. The placement of centroids is important as change in position may result in different clusters. The first step is to assign each data

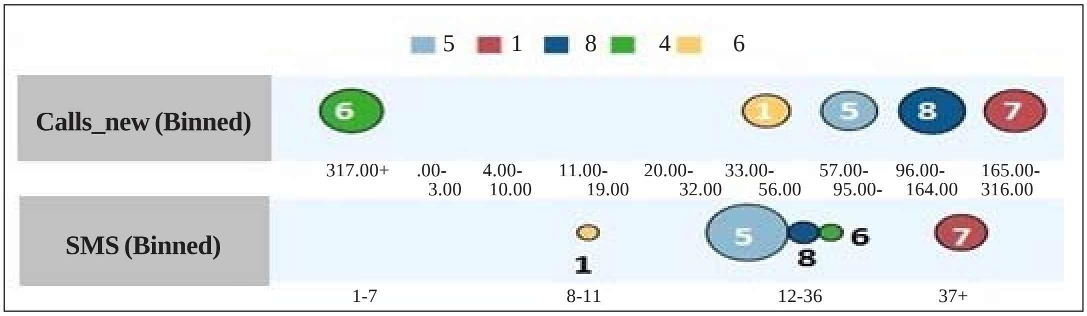

Figure 21. Segmentation of Revenue Cluster 5

point to the nearest centroid. In the second step, new centroids are calculated based on the points previously assigned to each cluster. The distance between new assigned points and the centroids is calculated and clusters are adjusted accordingly. In each iteration, the centroids change their location and iterations are performed until there is no change in centroids. The main objective is to reduce the mean squared distance of each data point to its nearest centroid. This technique is known as squared-error distortion [10]. The squared error function is mentioned below:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^j - c_j\|^2$$

The algorithm works in the following manner:

1. Choose any *k*-points from the data set. These points represent initial centroids.
2. Assign each data point to its closest centroid.
3. After assigning all data points to *k*-centroids, recalculate the centroids.
4. Repeat step 2 and 3 until centroids don’t change their position any more.

2.6 Results of K-Means Clustering Algorithm

The clusters obtained after applying K-Means algorithm on the data set are given in Table 14. The biggest cluster obtained is cluster 2 which contains 29.4% (1503 cases) of the total customers. The customers in this cluster have total revenue between Rs. 70-113, calls revenue between Rs.47-84, SMS revenue more than Rs. 16 and VAS revenue between Rs. 21-30. It can be observed that cluster 2 contains the largest total revenue and calls revenue. This cluster also contains largest SMS and VAS revenue

No	Total Cases	Total Rev	Calls Rev	SMS Rev	VAS Rev
1	569	70 - 113	47 - 84	0 - 3	1 - 8
2	1503	70 - 113	47 - 84	> 16	21 - 30
3	783	28 - 45	10 - 24	10 -15	21 - 30
4	1367	< 2	0 - 9	0 - 3	1 - 8
5	887	28 - 45	2 - 3	4 - 9	9 - 20

Table 14. Results of K-Means Clustering Algorithm

Cluster 4 is the second largest cluster having 26.7% (1367) customers of the total cases. The customers belonging to this cluster have quite low total revenue i.e. less than Rs. 2. The calls revenue is also less in this cluster which is between Rs. 0-9. SMS revenue is between Rs. 0-3 and VAS revenue is between Rs. 1-8.

Cluster 1 has got 11.1% (569) customers of the total cases which is the smallest cluster. The customers in this cluster have total revenue between Rs. 70-113, calls revenue between Rs. 47-84, SMS revenue between Rs. 0-3 and VAS revenue between Rs. 1-8. This cluster hasgot a greater total revenue and calls revenue but SMS and VAS revenue generation is less in this group.

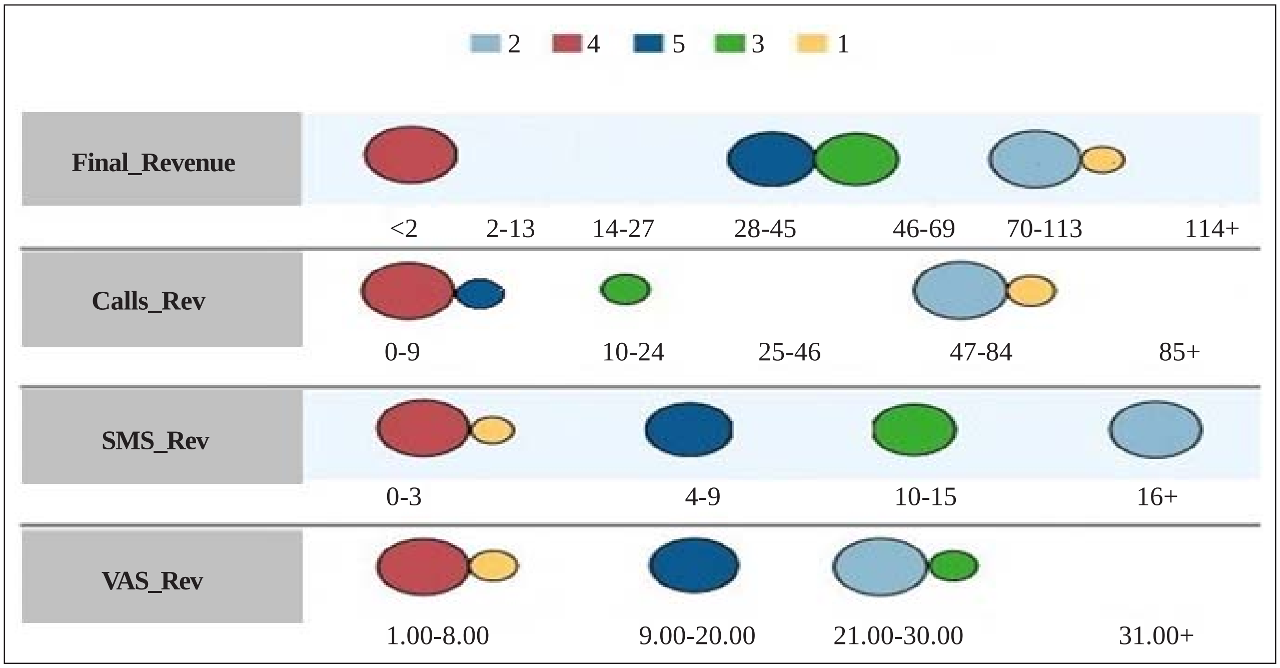

Figure 22. Revenue Clusters using K-Means Clustering

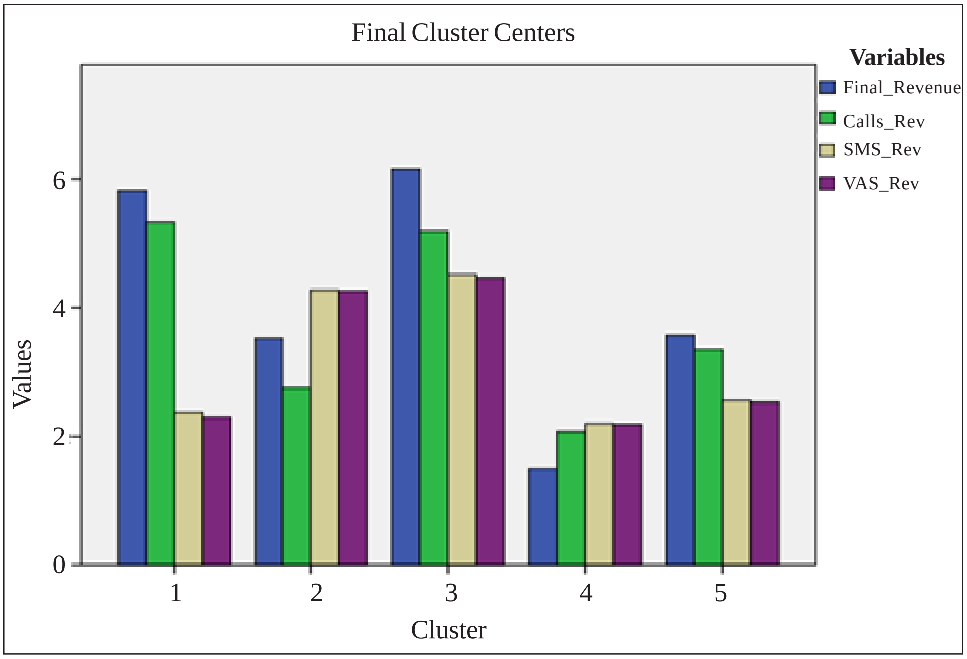

Figure 23. Bar Graph showing Distribution of Clusters

Cluster 3 consists of 15.3% (783) of the total cases. The total revenue in this cluster is between Rs. 28-46, calls revenue between Rs. 10-24, SMS revenue between Rs. 10-15 and VAS revenue lying between Rs. 21-30. Cluster 5 comprises of 17.3% of the total cases. The total revenue is between Rs. 28-45, calls revenue between Rs. 2-3, SMS revenue between Rs. 4-9 and VAS revenue is between Rs. 9-20.

In all of the clusters, calls revenue is more as compared to SMS and VAS revenue which shows that people prefer calls over SMS. Similarly it can be observed that VAS revenue is more as compared to SMS revenue in all clusters, showing that customers are more interested in value-added services as compared to sending SMS.

2.7 Comparison of Two-Step Clustering Algorithm and K-Means Algorithm

K-Means assigned 1503 cases to cluster 2 (the largest cluster) with total revenue between Rs. 70-113, calls revenue between Rs. 47-84, SMS revenue more than Rs. 16 and VAS revenue between Rs. 21-30 while two step clustering algorithm assigned 1035 cases to cluster 2 (the largest cluster) with total revenue more than 114, calls revenue more than Rs. 84, SMS revenue between Rs. 10-15 and VAS revenue more than Rs. 31. Both algorithms show that these are the loyal customers and they generate large revenue.

K-means assigned 1367 cases to cluster4 which is the 2nd largest cluster with total revenue less than 2, calls revenue between Rs. 0-9, SMS revenue between Rs. 0-3 and VAS revenue between Rs. 1-8 while two step clustering algorithm assigned 1006 cases to cluster 4 which is the 2nd largest with all other values exactly the same as in K-means. Both algorithms represent that these customers are not very loyal to the company and they do not generate good amount of revenue for the company.

K-means assigned 887 cases to the cluster 5 with total revenue between Rs. 28-45, calls revenue between Rs. 2-3, SMS revenue between Rs. 4-9 and VAS revenue between Rs. 9-20 while two step clustering algorithm assigned 718 cases to cluster 5 with total revenue between Rs. 2-13, calls revenue between Rs. 0-9, SMS revenue between Rs. 4-9 and VAS revenue between Rs. 9-20. In case of total revenue, K-means shows more revenue generation than two step clustering algorithm.

K-means assigned 783 cases to the cluster 3 with total revenue between Rs. 28-45, calls revenue between Rs, 10-24, SMS revenue between Rs. 10-15 and VAS revenue between Rs. 21-30 while two step clustering algorithm assigned 836 cases to cluster 3 with total revenue Rs. 14-27, calls revenue between Rs. 10-24, SMS revenue between Rs. 0-3 and VAS revenue between Rs. 1-8. 569 cases were assigned to cluster 1 by K-means with total revenue between Rs. 70-113, calls revenue between Rs. 47-84, SMS revenue between Rs. 0-3 and VAS revenue between Rs. 1-8 while two step clustering algorithm assigned 922 cases to cluster 1 with total revenue more than Rs. 114, calls revenue more than Rs. 84, SMS revenue more than Rs. 16 and VAS revenue more than Rs. 31. Both algorithms represent that these customers generate reasonable amount of revenue mostly coming from calls.

3. Marketing Strategies

Since our segmentation is done on nine days of customer data, it would be better to discuss weekly and daily packages which could be offered to these segments. Different tailored packages, bundles and offers can be given to these segments based on the type of revenue they generate and their usage behavior. Cluster 1 is indeed the segment of loyal customers as their revenue from all categories is the highest. Such customers need to be retained as they generate most of the company revenue. They must be given special offers, rewards and discounts so that they feel more loyal to the company. When we analyze cluster 2 from Table V, we can see that they are more inclined towards making calls which shows that they are more comfortable to communicate through calls rather than sending SMS. There is a need to provide them with optimum price plans for making calls. They could be offered packages such as:

- Talk Monthly 24 hours a day throughout the month on same network and all land-line numbers for Rs. 300 + tax per month.
- Talk Weekly 24 hours a day on same network for Rs. 80 _ tax per week.
- Talk Daily 24 hours a day on same network and friends and family numbers for Rs. 12 + tax daily.

Cluster 3 shows good call usage and some signs of VAS usage as well. They are least interested in communicating through SMS. Hence they should be offered good price plans for calls and good VAS offers such as:

- Talk Daily 24 hours a day on same network and friends and family numbers for Rs.12 + tax daily.
- To enhance their VAS usage, we need to offer them free Internet buckets of 3-10 MB etc. on weekly basis.

Analyzing cluster 4 of revenue from Table V, we can see that they generate the least revenue in all categories. We need to offer them different packages and bundles related to call and SMS only so that their usage of the core services could increase. They should not be offered VAS promotions as their core services usage is already very low and they would not bother for extra services offer.

They could be offered packages such as:

- On recharge of Rs. 50, get free 10 min + 100 SMS on all networks and on recharge of Rs. 100, get 30min + 300 SMS

- For only Rs. 10 + Tax per day, all calls on same network are free. Moreover, get 100 min + 300 SMS free as a bonus.

Cluster 5 is of technology lovers because their VAS (Value Added Services) revenue is good. Their SMS revenue is also fairly good as compared to their call revenue as shown in Table V. We should give incentives in using VAS. Following offers could be given:

- For Rs. 250, get 5000 SMS and MMS each with 1.5 GB Internet for the whole month. Note that such an offer will also cater for their SMS use.
- For Rs. 30, send unlimited messages through WhatsApp (a free on-line chatting application on cell phone) monthly.

4. Conclusion And Future Work

In this paper, we have shown that through the use of customer segmentation, a telecommunication company could easily market its customers with right products and services. Moreover, this also helps in offering tailored packages, offers and bundles for customers. In this way, it becomes easier for company officials to create marketing campaigns from scratch for speciûc customer segments instead of the whole customer base. The data set consisted of 5,109 customers' daily call and SMS usage as well as revenue generation data spanning over 9days. The continuous attributes were discretized using binning method. The two-step clustering algorithm, as well as K-Means algorithm was applied and the results were thoroughly analyzed. It was shown that VAS (Value Added Services) usage was greater than the SMS usage for all customers' segments. Every cluster was analyzed so as to uncover its call as well as SMS usage behavior. The results for the two algorithms were in accordance for most of the clusters.

In future, a revenue prediction model could be built which predicts total revenue based on call, SMS and VAS usage for each segment using different machine learning algorithms such as multinomial logistic regression. With data of different services used by customer, each segment could further be analyzed on the basis of the services used. This would eventually allow offering better targeted promotions for increasing the sales.

We would like to thank IBM for rendering help under IBM Academic Initiative.

References

[1] Bhatia, N., Vandana (2010). Survey of nearest neighbor techniques. *International Journal of Computer Science and Information Security,* 8 (2).

[2] Cross, G., Thompson, W. (2008). Understanding your customer: Segmentation techniques for gaining customer insight and predicting risk in the telecom industry: Data mining and predictive modeling, SAS Global Forum.

[3] Jansen, S. M. H. (2007). Customer segmentation and customer profiling for a mobile telecommunications company based on usage behavior, A Vodafone Case Study, Master thesis, University of Maastricht, July.

[4] Kaufman, L., Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Inter science, New York (Series in Applied Probability and Statistics).

[5] Li, Z. (2011). Research on customer segmentation in retailing based on clustering model, *In:* Computer Science and Service System (CSSS), 2011 International Conference on, p. 3437–3440.

[6] Lin, Q., Wan, Y. (2009). Mobile customer clustering based on call detail records for marketing campaigns, *In:* Management and Service Science, 2009. MASS '09. International Conference on, p. 1–4.

[7] McCarty, J. A., Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression. *Journal of Business Research,* 60 (6) 656 – 662.

[8] Namvar, M., Gholamian, M. R., KhakAbi, S. (2010). A two phase clustering method for intelligent customer segmentation. *In:* Intelligent Systems, Modeling and Simulation (ISMS), 2010 International Conference on, p. 215–219.

[9] Tang, J. F., Zhang, T. J., Huang, X. H., Luo, X. G. (2011. Case study on cluster analysis of the telecom customers based on consumers' behavior. In Industrial Engineering and Engineering Management (IE EM), 2011 IEEE 18*th* International Conference on, 2, p. 1358–1362.

[10] Gersho, A., Gray, R.M. (1992). Vector Quantization and Signal Compression. Boston: Kluwer Academic.