Segmentation & Profiling Project

Merrimack College

Data Exploration

Amol Gote

(Data science student)

July 3rd, 2020

**Executive Summary**

This reports provides the customer segmentation for customers on telecommunications company which will help the organization in making effective economical customer retention strategies. Based on the data analytics exercise customers have been categorized on four categories Platinum, Gold, Silver and Bronze. These four categories have been primarily driven by the revenue generated by the customers over the tenure of their association with Telecommunications Company. Key value factor is revenue generated over the tenure and average monthly revenue generated. Apart from the value generation factors other attributes related to demographics line age, gender, education, employment length, marital status and household size have been factored in. Financial features like household income, debt to income ratio, total debt has been factored in as well. Distribution categories is as shown below

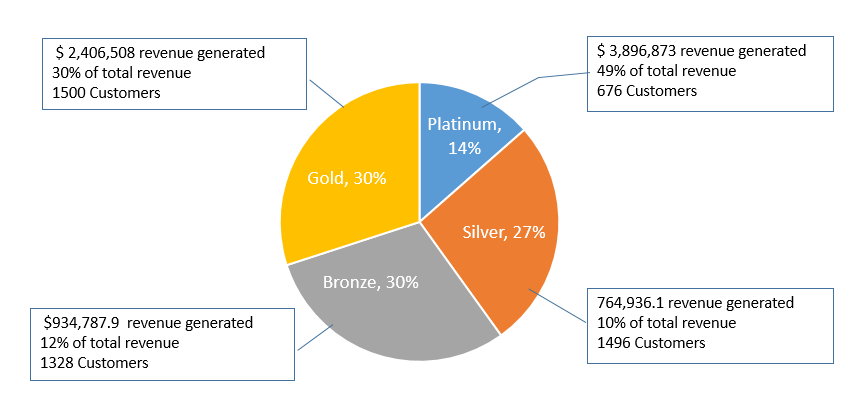


Figure 1: Customer segmentation distribution

Platinum customers who count for only 14% of the total customer base generate almost 50% of the total revenue, followed by gold with 30% of total revenue, silver and bronze generate very minimal revenue with over all percentage hovering around 10-12%. Average monthly revenue generated by platinum customers is $103, by gold customer is $27.50, for silver and bronze it is almost same which is $24.

As far as demographics is concerned, for platinum customer is dominated by customers aged in between 40 to 60, with average age of 53, gold customers are dominated by the customers with age greater than 60, with average age of 63, while silver and bronze is dominated by customers aged between 18 to 40 years of age, with average age of 35 and 37 respectively. Customers below age of 18 years are either in silver of bronze category. Gender distribution is equal across all customer segments, there is no bias. For platinum marital status is not of significance as there are nearly equal married and unmarried customers, with married ones on higher side a bit. Silver customer segment comprises of married customer and with very less unmarried customers. Bronze customer hardly contains any married customers. Median household size for platinum and gold customers is 2, while that for silver is 4, for bronze household size is hovers around 1 which is self-evident from the fact that majority of bronze customer is dominated by unmarried customers.

For household income average income for platinum is $90,000, for gold customers it is $65,000 for bronze and silver the average household income is around $40,000. Debt to income ratio for platinum, gold and silver customers is around 10% while that for bronze customers it shoots up to 37%. Nearly all the customers in the platinum segment are multiline customers, while the 65% of the gold customer are multiline, silver and bronze customers are dominated by non-multiline customers. Platinum and gold customer have been associated with an average of 56 months, followed by silver for 28 months and bronze by 20 months.

**Segmentation solution**

There were 2 type of clustering techniques that were utilized to identify the customer segments firstly K-mean and secondly Hierarchical Clustering. For both clustering techniques following variables were used Gender, Age, EducationYears, EmploymentLength, HHIncome, DebtToIncomeRatio, MaritalStatus, HouseholdSize, HomeOwner, PhoneCoTenure, Multiline, OverTenureTotalValue, AverageMonthlyRevenue. For statistical analysis added following variables OverTenureTotalValue, AverageMonthlyRevenue which could help in finding high-value customers segment.

K-Mean

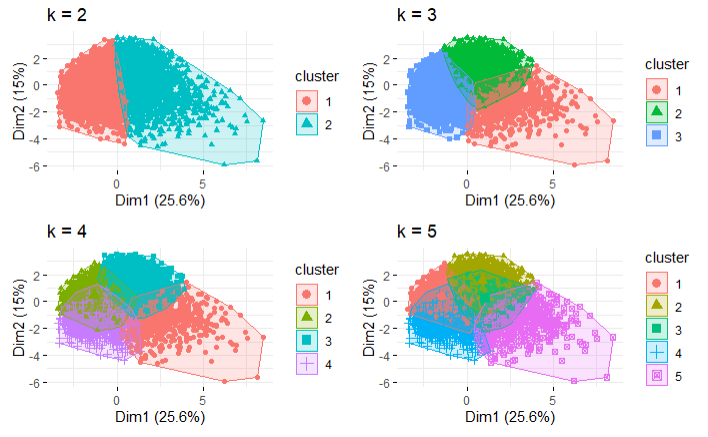
Firstly K-mean technique was performed, k-mean only accepts numerical variables so all categorical variables were converted to numeric variables, For k-means the number of cluster (k) must be set before starting the algorithm. So ran the k-mean analysis with varying cluster size from 2 to 9. Determining the exact cluster size is challenge and this is where graphical methods of Elbow method and Silhoutte method came to rescue. Elbow method clearly indicated the bend at cluster number 4. Based on Figure 2 & 3 is evident that cluster size of 2, 3 and 4 give clear differentiating customer segments other cluster sizes there is considerable amount of overlap especially with higher cluster sizes line 6, 7, 8 and 9. Figure 4 which is output of the elbows method shows bend at cluster size of 4, 5 and 6, if you compare the cluster distribution in Figure 1, then cluster with size of 4 is more prominent with less overlap. Cluster size 5 and 6 all clusters are not clearly visible. 

Figure 2: K-mean distribution with cluster size of 2, 3, 4 and 5

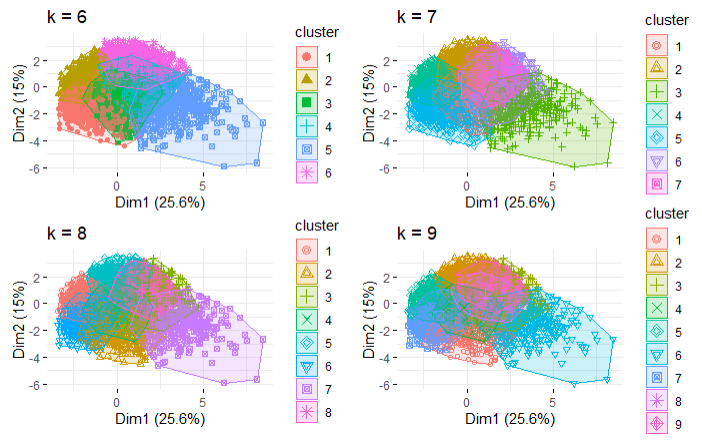


Figure 3: K-mean distribution with cluster size of 6, 7, 8 and 9

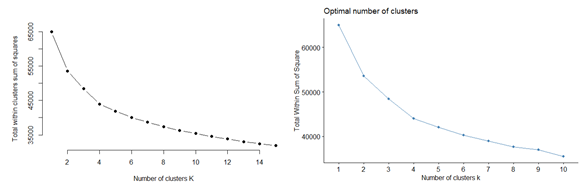


Figure : Elbow method output

Hierarchical Clustering

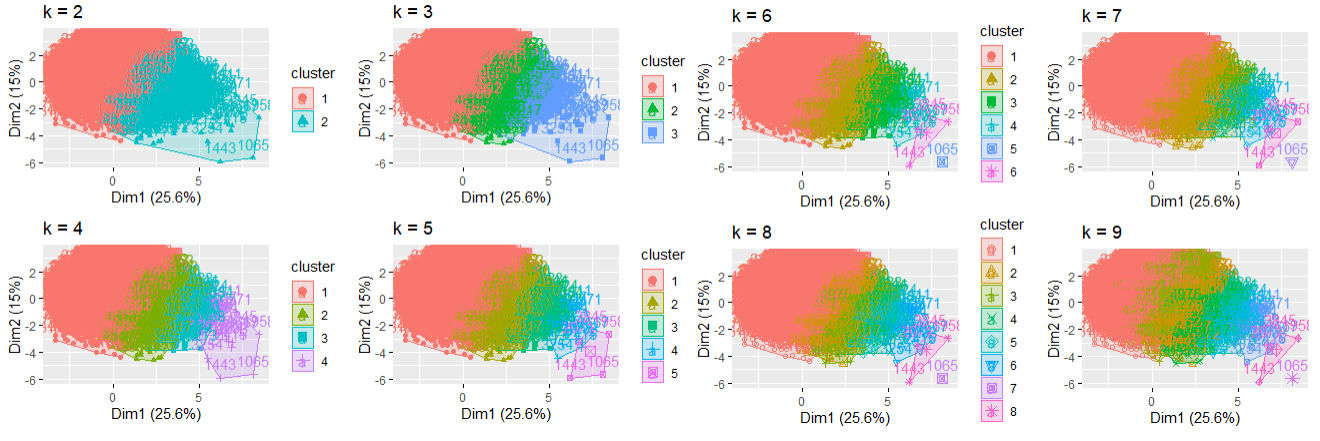
In this analysis as well ran the technique with cluster sizes ranging from 2 to 9, below are the cluster distribution. One major distinguishing factor between K-mean and hierarchical is that one of the cluster in hierarchical is consistent large (cluster with red color in Figure: 5) and there is no even distribution of points across the clusters. Also elbow method clearly indicated that optimal cluster size is 4, but based on the points distribution analysis there are very less number of points in cluster 3 and 4.

Figure 5: Hierarchical Clustering with cluster size from 2 to 9

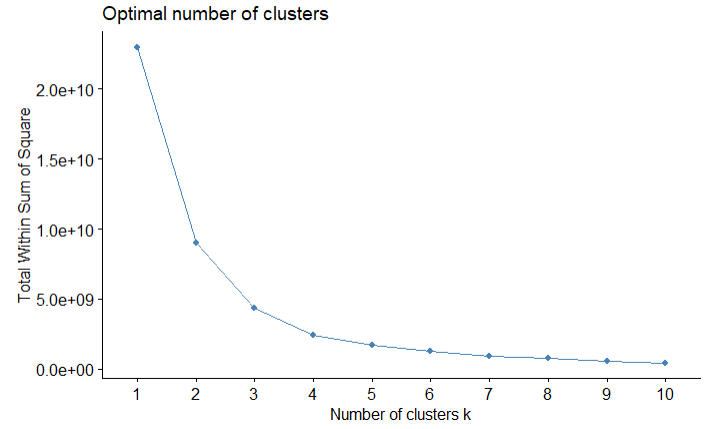


Figure 6: Elbow method output.

**Detailed Findings**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| # of Customers | 676  14% | 1500  30% | 1496  12% | 1328  10% |
| $ Revenue | $3,896,873  49% | $2,406,508  30% | $934,788  12% | $764,936  10% |
| $ Revenue (Average Monthly) | $103 | $28 | $25 | $24 |
| House hold Income (Avg) | $90,000 | $65,000 | $41,000 | $40,000 |
| Household size (Avg) | 2 | 2 | 4 | 1 |
|  |  |  |  |  |
| Education (Avg) | 17 Years | 13 Year | 14 Years | 14 years |
| Age (Avg) | 54 years | 65 years | 35 years | 37 years |
| Association with company (Avg) | 56 months | 57 months | 28 months | 20 months |
| Debt to income ratio | 10% | 10.5% | 9.5% | 37% |
| Gender | Male: 355  Female: 321 | Male: 752  Female: 748 | Male: 577  Female: 751 | Male: 805  Female: 691 |
| Multiline |  |  |  |  |
| Loan Default |  |  |  |  |
| Car Ownership |  |  |  |  |
| Voice Revenue $ | $1,035,413 | $1,811,989 | $420,014 | $ 274,883 |
| Equipment Revenue $ | $1,400,841 | $338,197 | $310,291 | $301,552 |
| Data Revenue $ | $1,460,619 | $256,321 | $204,482 | $188,501 |
| Marital status |  |  |  |  |

# Bibliography

*K-means Cluster Analysis.* Retrieved from University Of Cincinnati Business Analytics R

Programming Guide: <https://uc-r.github.io/kmeans_clustering>

*Hierarchical Cluster Analysis.* Retrieved from University Of Cincinnati Business Analytics R

Programming Guide: <https://uc-r.github.io/hc_clustering>

Sarfin, R (2019, Aug 7). *Data Quality Dimensions: How Do You Measure Up? (+*

*Downloadable Scorecard).* Retrieved from syncsort: [https://blog.syncsort.com/](https://blog.syncsort.com/2019/08/data-quality/data-quality-dimensions-measure/)

[2019/08/data-quality/data-quality-dimensions-measure/](https://blog.syncsort.com/2019/08/data-quality/data-quality-dimensions-measure/)