Segmentation and Profiling – Final Project
DSA5100- Data Exploration

Owen R. Evans
Data Science Student

Executive Summary

It is essential that the management team of NewCo Telecom (our Company) understand that our large customer base is not monolithic and is comprised of highly diverse groups/segments with disparate needs and varying revenue potential. Generic marketing strategies, mass social media campaigns and singular pricing strategies may only cater to a small proportion of our customer base. By not understanding the characteristics of key customer segments it becomes very difficult for us to drive future business growth. Additionally, those telecom companies that are not able to properly perform and act upon customer segmentation may also suffer from revenue loss by not immediately identifying customer groups most prone to churn.

There are a variety of machine learning techniques that can automate and provide for effective customer segmentation. The two techniques explored in this analysis (k-means clustering, decision tree classification) are examples of unsupervised and supervised learning techniques. They are both capable of handling clustering/classification tasks of multi-dimensional data sets. The limitations and advantages of both approaches will be comprehensively detailed in the analysis below.

For the purposes of this segmentation analysis, an engineered data set containing 67 variables for nearly 5000 customers was utilized after significant pre-processing and exploratory analysis. This engineered data set contains rich details concerning the demographic, psychographic, financial and behavioral characteristics of our customers. The proceeding segmentation analysis was based primarily on customer value, based upon the average revenue attributed to each customer (ARPU). For both segmentation approaches, a vector of feature variables focused primarily on education, employment status, age and individual income of our customers was used for segmentation.

Of the two approaches, it was determined that clustering via k-means afforded customer value segments with high intra-cluster similarity and good separation across clusters. This technique effectively separated our customer base into three distinct value segments. High value segments were typified by those customers with high ARPU values. Customers from this high value segment tended to be highly educated, middle-aged professionals with significant work experience and moderate income. In comparison to other segments, high value customers also tended to utilize services other than voice, with a fairly significant proportion of revenue attributed to both data and equipment transactions. High value customers were also more likely to own cars/homes and had a higher tendency to be unmarried.

A mid-value customer segment of equal size was also identified. This mid-value segment was comprised mainly of older retirees with significant employment experience and relatively high individual incomes. This segment primarily utilizes our voice service and is fairly lucrative with respect to revenue generation. This segment is also comprised of our most loyal customers with the majority of these customers exhibiting a company tenure in excess of 36 months.

A low-value customer segment comprised of young working professionals with limited experience and education was also identified. The poor revenue performance of this segment appeared to be driven primarily by low individual income. The results suggest that this segment may be the most responsive to aggressive pricing strategies. A tiered pricing strategy that encourages customer

loyalty in this segment today may prove to be more lucrative in the future as customers within this segment improve their financial standing.

A similar segmentation analysis was attempted via decision tree classification. The results from this analysis were directionally equivalent to that afforded through k-means clustering, but the segments obtained were subject to high error rates, misclassification and overfitting. While the decision tree approach may allow for greater flexibility via the use of both numerical and categorical feature variables, it was evident in this case that the k-means algorithm tended to afford better results.

The customer segmentation results obtained through k-means clustering successfully identified three customer segments based upon overall value. The exercise has provided a deeper understanding of these segments that will eventually allow for future revenue growth by specifically targeting to these segments. It has successfully identified the key characteristics and the service utilization for high value customers. Targeted products and marketing campaigns will maximize the revenue growth in this segment in the future. Low values segments will also benefit from more targeted business activities via revised pricing strategies. The end result of the following analysis is a marketing roadmap to future revenue growth.

Introduction.

The primary intent of the following customer segmentation analysis is to gain a more comprehensive understanding of our customer base. The goal is to segment customers according to overall value, defined primarily as potential to gain higher revenue. This analysis will allow for further business growth by achieving the following: (1) more targeted marketing campaigns to higher value segments, (2) identification of low value segments more prone to churn, (3) more targeted product offerings for maximum revenue growth and (4) personalize customer service. It can be shown that customer segmentation is an effective way to drive revenue growth through more personalized marketing campaigns. The following analysis will evaluate the effectiveness of both unsupervised (K-means clustering) and supervised (decision tree classification) machine learning algorithms. A comprehensive comparison of the accuracy and quality of the outputs from the two techniques will allow one to choose the best approach. A comprehensive profile of the of the derived customer value segments will then be derived using the most optimal approach.

Customer Segmentation via K-Means Clustering.

As indicated above, the primary goal of this data analysis is to effectively segment our customer based upon on overall customer value. For the purposes of this data mining exercise, customer value was assessed according to an average monthly revenue per user (ARPU) metric, calculated as the quotient of an individual customer's total revenue and their respective tenure as customer of our business. To facilitate initial clustering of our customer base via key feature variables, a K-means clustering algorithm was utilized. Feature input variables for this clustering exercise were judiciously chosen based upon the hypothesis that customer value is driven primarily by disposable income, education, age and employment status. As such, the following feature variables were utilized for clustering: (1) *Age*, (2) *EducationYears*, (3) *IncomeIndLog*, (4) *EmploymentLength* and (5) *ARPU_Log*. The engineered variable (*IncomeIndLog*) was derived as the quotient of household income and household size and was intended to represent a meaningful income metric for an individual customer. It is important to note that both *IncomeIndLog* and *ARPU_Log* were derived from a log-transformation due to the presence of a significant right-handed skew in their respective data distributions. Data distributions with a significant skew can affect the results of clustering via K-means, as a few data points can significantly distort the centroids of each cluster formed through the K-means algorithm. Finally, since all selected feature variables have disparate units, they were all subject to scaling prior to scaling. This step is necessary to ensure that all variables are treated equally during the clustering exercise and are not affected by a singular variable with inherently high values.

There are a few more limitations of the K-means algorithm that may also affect the efficacy of our segmentation. First, since the K-means algorithm is based upon the iterative minimization of intra-cluster distances it will only work with numeric/continuous variables where a Euclidean distance can be reliably determined. It is possible to perform k-means clustering on categorical variables via hot encoding (i.e. conversion to 1,0), but this is not recommended. Euclidean distances for data that has an arbitrary scale and/or order have little meaning. Additionally, the outcomes of an encoded categorical variable are mutually exclusive; they are either 1 or 0. This leads to a scaled data distribution with high variance which will overly influence the clustering results. For the purposes of this analysis, only numeric variables were utilized for clustering via

K-means. It is important to note that there are other clustering algorithms (K-modes) that may be more appropriate for categorical data.

Another limitation of the K-means algorithm is that the number of clusters need to be specified in advance. This imparts an element of subjectivity that could ultimately influence the results of clustering by inadvertently specifying a non-optimal number of clusters. Fortunately, there are a variety of heuristic and statistical methods to determine the optimal number of clusters. For the purposes of this analysis, the optimal number of clusters for the chosen feature variables were determined via three different methods: (1) silhouette analysis, (2) elbow analysis and (3) gap statistic. The results of these analysis (Figure 1) were largely mixed, with both the gap statistic and WSS method indicating an optimal number of clusters of 3 and 4, respectively. Results obtained from the Silhouette method indicate an optimal *k*-value of 2, which likely indicates a slight overlap of clusters beyond this value. However, a two-cluster model would likely not afford an informative customer segmentation outcome and that a trade-off between cluster purity and utility must be made. This fact highlights one of key limitations of K-means, that the choice of the *k*-hyperparameter is not always unambiguous. The rest of this k-means analysis focused primarily on validating a 3 and 4-cluster model.

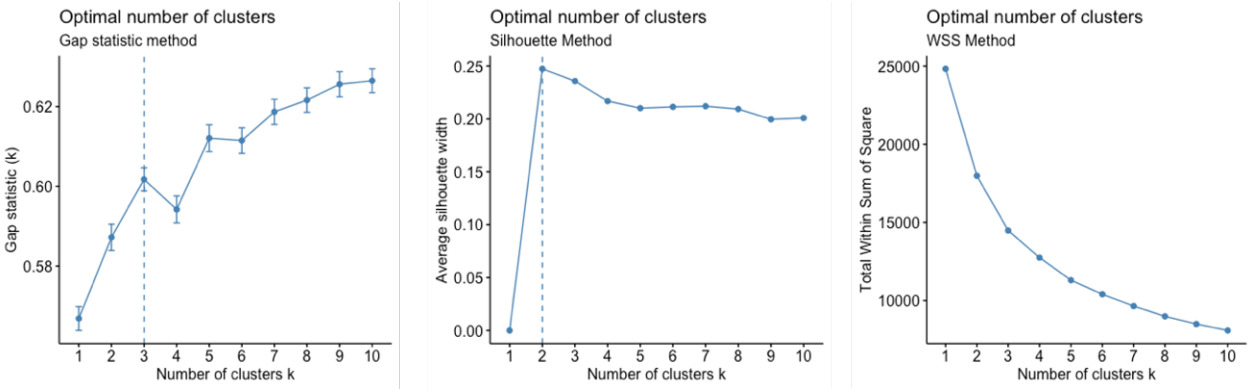

Figure 1. Optimization of the Number of Clusters

In order to further define the optimal number of clusters to utilize for this k-means clustering exercise, it is recommended to visualize the clustering results in two-dimensional space using the *fviz_cluster* function from the factoextra package. This function, once applied to the k-means results for *k*=3 and *k*=4, will perform a dimensional reduction of the feature variables via principal component analysis and plot the results of each k-means models as a function of the first two principal components. Shown in Figure 2 are the visualizations for k=3 and k=4 clustering models. These plots indicate that diminished overlap and clusters with more equal sizing are obtained for the k=3 model, which likely suggests it is a more optimal solution than k=4.

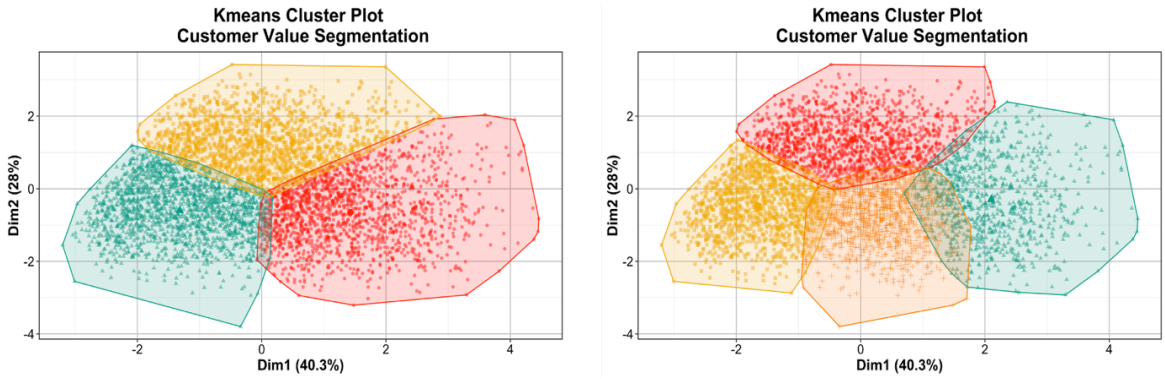

Figure 2. Clustered data points as a function of the first two principal components as derived from a K-means clustering algorithm using *k*=3 (left) and *k*=4 (right).

Validation of the effectiveness of the K-means clustering algorithm for *k*=3 and *k*=4 can be derived through a variety of statistical techniques, all of which are intended to ensure a high degree of intercluster similarity and that each individual cluster is distinct. A silhouette analysis affords a silhouette width metric that effectively gives a quantitative measure of how close a point in a particular cluster is to data points in neighboring clusters. Silhouette width values that are greater than zero indicates tight clustering with clearly distinct data points within each cluster. Conversely, silhouette coefficients equal to or less than zero indicate ineffective clustering. Shown in Figure 3 are the results from silhouette analysis of the k-means clustering model using *k*=3 and *k*=4, respectively. The plot affords modest and positive silhouette width values for most data points within each of the specified clusters for both models. Although the presence of a few data points with negative silhouette widths indicates a possible misclassification of a few select data points, the derived k-means model has done an effective job in clustering for both *k*=3 and *k*=4 models. However, for the three-cluster model we observe a slightly higher average silhouette width and clusters with equal sizes. For the 4-cluster model, lower silhouette widths and the presence of clusters with differing sizes are observed. These results coupled with the visualizations above confirm that the optimal value of *k* for this clustering exercise is 3.

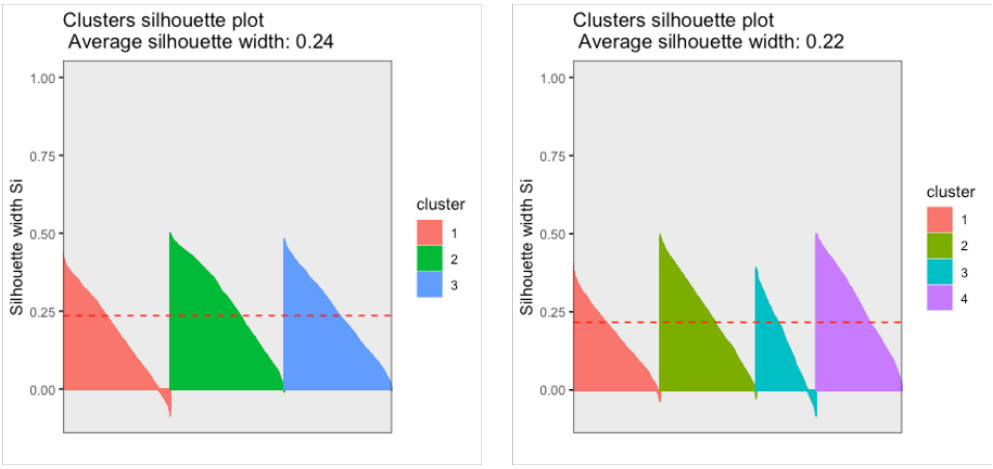

Figure 3. Validation of K-means Clustering Model via Silhouette Analysis for k=3 (left) and k=4 (right).

Characterization of the specific clusters identified via the K-means algorithm can serve as an additional means to validate results. Specifically, an inspection of the cluster medians (k=3) indicates disparate and distinct values for the feature variable of interest across the three clusters. The clustering has successfully divided the customer space into high, mid and low-value segments with key characteristics (Table 1). The high-value customer segment is comprised of highly educated middle-aged professionals, whereas the low-value segment is comprised of young, unexperienced folks with minimal education. Lastly, the mid-value segment appears to be comprised of older retirees with extensive work experience.

Table 1. Cluster medians and segmentation labels obtained from K-means clustering.

Cluster	Size (n)	Segment Label	Age	Education Years	Individual Income ($/year)	Employment Length (years)	Average Revenue per User, ARPU ($/month)
1	1617	Mid Value - Older, Experienced, Low	66	13	32000	19.0	21.87
2	1725	Low Value - Young, Unexperienced, Low	31	13	12500	3.0	8.13
3	1625	High Value - Middle Aged and Well Educated	43	17	29500	5.0	54.42

Visualization of the range of values observed for the feature variables of interest as a function of customer segment is afforded by a side-by-side boxplot (
Figure *4*). The boxplot visualization indicates the range of values observed for each of the variables within each segment. In general, the interquartile ranges observed are numerical distinct across the three segments, indicating that the k-means algorithm was reasonably successful in segmenting the customer base into value segments.

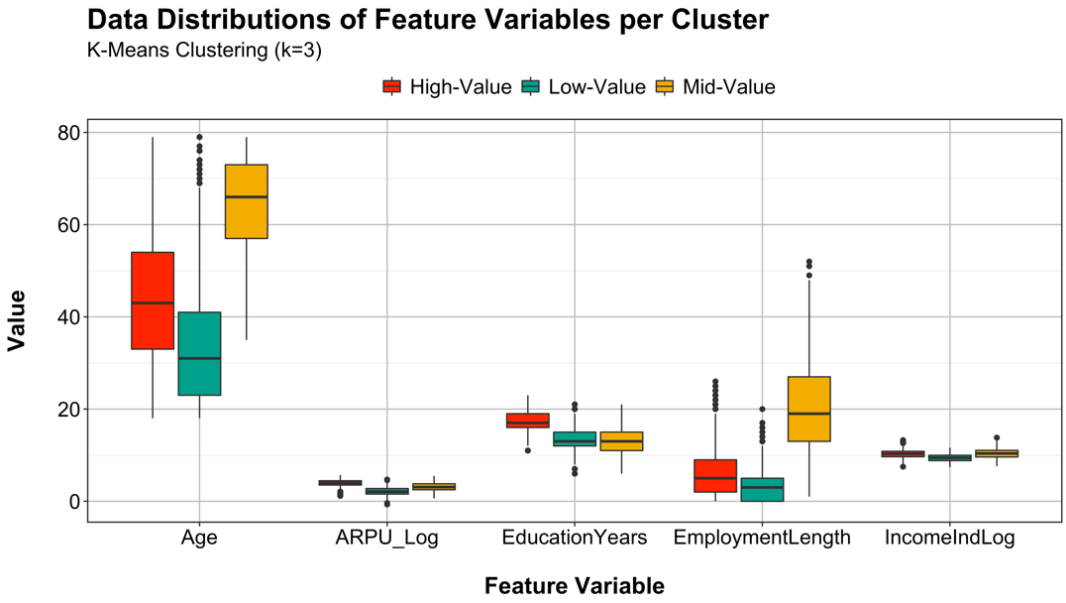

Figure 4. Data distribution of the feature variables as a function of customer segment.

Customer Segmentation via Decision Tree Classification.

To contrast and potentially confirm the segmentation results obtained from k-means clustering, a similar analysis was conducted using decision tree classification. Decision tree classification is a supervised learning technique that will use pre-defined labels (or classes) to repeatedly subset a data set into homogenous groups based upon optimal information gain. In order to facilitate this analysis, each observation was assigned a customer value label based upon ARPU value. The basis for this labeling was largely based upon the range of ARPU values for each customer segment from k-means clustering analysis. Specifically, low value customers tended have ARPU values of <12, high value customers have ARPU values of >35 and mid-value customers had ARPU values of 12-35. These labels were utilized to initially perform a decision tree analysis using the feature variables of interest: age, individual income, education, employment status. Attempts to perform a multi-class classification (mid, low and high value) were plagued with computational complexity, long run times and outputs with a high error rate. As such, individual analyses for each value segment were instead subject to binary classification using feature variables of interest.

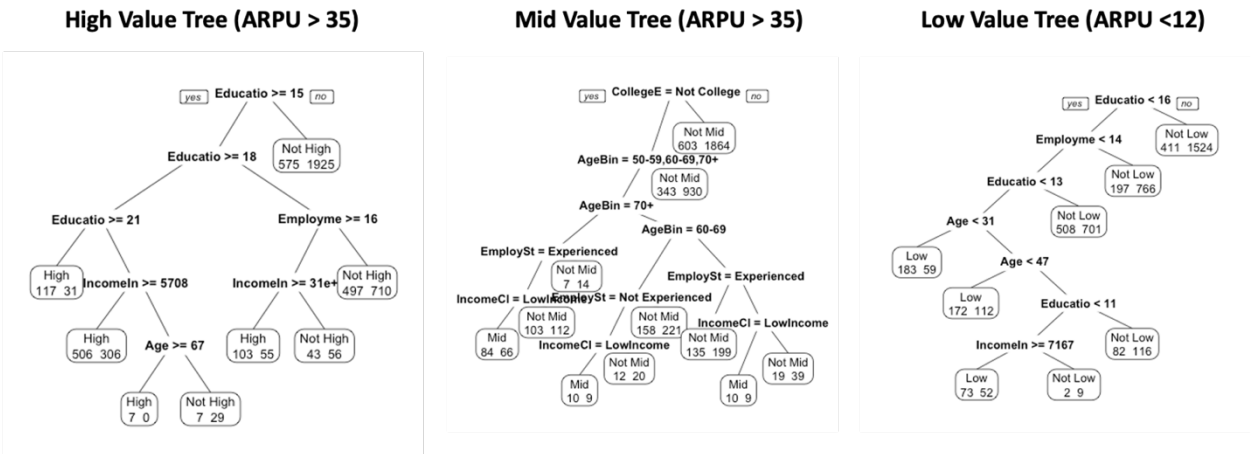

Figure 5. Classification decision tree outcomes for high, mid and low value segments.

Shown in Figure 5 are the pruned individual decision tree outcomes for all customer value classes. Similar to the k-means outcomes, the high value segment is typified by customers that are well-educated (*EducationYears* >15) with modest individual incomes (<$31,000). Similarly, the low-value segment was typified by young, inexperienced customers with low incomes. Unfortunately, classification of the mid-value segment using the same feature variables tended to afford complex trees, high cross validation errors and poor-quality outcomes. An improved decision tree model was only obtained when the feature variables were binned/categorized. In this case, education status, employment status and income were converted to binary variables. Utilizing this approach, an acceptable decision tree was obtained for the mid-value customer segment which was typified by middle aged folks with significant employment experience.

In comparison to the clustering results obtained from the K-means algorithm, the segmentation results via decision tree classification resulted in very high error rates, increased complexity and over-fitting. This effect seems to be driven by the fact that a decision tree is a greedy algorithm

and will make splits in the data to satisfy a local minimum and not afford a globally optimal outcome. While the two techniques afforded the same results directionally, the segmentation derived by k-means clustering did a profoundly better job in segmenting the customer base with high accuracy and minimal misclassification. As such, the following customer profiling exercise will solely be based upon the results from k-means clustering.

Customer Segmentation – Profiling.

As indicated above the results from k-means clustering were based upon a select number of key feature variables. However, the data set in question contains a rich source of other variables that can serve to further profile the customer value segments. For instance, additional demographic information such as gender, marital status and job category are key aspects that can further characterize the three customer value segments.

Table 2. Cross-tabulation results of demographic customer attributes and customer value segment.

		Customer Segment					
Variable	Factor	Low Value		Mid Value		High Value	
Gender	Female	864	50.1%	804	49.7%	811	49.9%
	Male	844	48.9%	803	49.7%	809	49.8%
	Undeclared	17	1.0%	10	0.6%	5	0.3%
Marital Status	Married	925	53.6%	726	44.9%	737	45.4%
	Unmarried	800	46.4%	891	55.1%	888	54.6%
Job Category	Agriculture	71	4.1%	89	6%	51	3.1%
	Crafts	134	7.8%	192	12%	123	7.6%
	Labor	173	10.0%	347	21%	160	9.8%
	Other	3	0.2%	7	0%	5	0.3%
	Professional	433	25.1%	419	26%	523	32.2%
	Sales	732	42.4%	286	18%	600	36.9%
	Service	179	10.4%	277	17%	163	10.0%
Retired	No	1688	98%	1007	62%	1537	95%
	Yes	37	2%	610	38%	88	5%

Shown in Table 2 are cross-tabulation results of customer demographic attributes and customer value segment. The analysis is intended to highlight key similarities and differences across the three identified segments. All segments appear to have equal distributions of genders, indicating that the telecom services supplied by our company cater equally to all people regardless of gender. Marital status appears to differ somewhat across all three segments. Primarily, that the high value segment, which is comprised of middle-aged and well-educated customers, has a higher proportion of unmarried customers. This contrasts with the low-value customer segment which has a comparatively lower proportion of unmarried customers.

Customers within the high value segment appear to possess jobs within the professional and/or sales category, with smaller proportions in the services or labor category. Mid-value customers have a higher proportion of customers with jobs in the labor and service category. Interestingly, the breakdown of job categories for low value customers appears to be heavily weighted in the

sales category. Finally, an inspection of the retirement status across the three segments confirms that the mid-value segment is comprised primarily of older retirees.

Table 3. Cross-tabulation results of psychographic customer attributes and customer value segment.

Variable	Factor	Customer Segment					
		Low Value		Mid Value		High Value	
Owns Pets	Yes	555	32.2%	486	30.1%	483	29.7%
	No	1170	67.8%	1131	69.9%	1142	70.3%
Car Ownership	Lease	276	16.0%	340	21.0%	178	11.0%
	None	165	9.6%	152	9.4%	179	11.0%
	Own	1284	74.4%	1125	69.6%	1268	78.0%
Owns Mobile Device	No	1011	58.6%	1141	71%	435	26.8%
	Yes	714	41.4%	476	29%	1190	73.2%
Phone Co Tenure	1-12months	601	35%	51	3%	269	17%
	12-24months	365	21%	78	5%	267	16%
	24-36months	311	18%	161	10%	283	17%
	>36months	448	26%	1327	82%	806	50%

Shown in Table 3 are cross-tabulation results of customer psychographic attributes and customer value segment. The intent behind this analysis is to determine if any association exists with regard to customer behavior and value. For instance, would attributes that infer a high level of disposable income (such as car ownership) predominate in certain customer value segments? The cross tabulation indicates that high value customers are indeed more likely to own cars and mobile devices. Customers within the high value segment also appear to be somewhat loyal, with an appreciable portion having a company tenure in excess of 36 months. Interestingly, the mid-value segment has the greater proportion of folks with increased company tenure. This is likely due to the fact that this particular segment is comprised mainly of older retirees.

Table 4. Utilization rate (last month) of customers within each segment for voice, data and equipment services.

Service	% Utilization Over Last Month	Low Value		Mid Value		High Value	
Equip Utilization	<25%	1482	86%	1394	86%	478	29%
	25-50%	80	5%	178	11%	799	49%
	50-75%	140	8%	43	3%	321	20%
	>75%	23	1%	2	0%	26	2%
Voice Utilization	<25%	69	4%	31	2%	441	27%
	25-50%	212	12%	155	10%	654	40%
	50-75%	56	3%	190	12%	275	17%
	>75%	1388	80%	1240	77%	253	16%
Data Utilization	<25%	1576	91%	1398	87%	761	47%
	25-50%	72	4%	164	10%	723	45%
	50-75%	69	4%	52	3%	132	8%
	>75%	8	0%	2	0%	8	0%

Finally, in order to effectively target future marketing campaigns towards our high value segment, it is imperative to understand what services and products these particular customers are primarily using. Shown in Table 4 is the utilization rates (calculated as the fraction of total revenue for each service) of customers within each segment for the separate services/products offered - voice, data and equipment. Interestingly, customers from the low and mid value segments primarily use voice services. Whereas customers from the high value segment have a larger percentage of revenue attributed to data and equipment transactions. As such, to maximize revenue it would be advisable to target and market these specific services to this pre-determined customer segment.

Summary and Conclusions.

The preceding affords a comprehensive customer segmentation analysis of Newco Telecom with the primary intent in defining segments based upon overall customer value. Both unsupervised clustering and supervised classification were utilized for segmentation using key feature variables, including age, education/employment status, individual income and average user revenue. It was determined that clustering via k-means afforded a more accurate and reliable segmentation outcome. Three distinct segments were identified with key characteristics. Specifically, a high value segment comprised mostly of highly educated, working professionals is responsible for a large amount of revenue generation for the company. These high value customers appear to heavily utilize our data and equipment services and are not solely voice customers. A low value customer segment seems to be comprised of inexperienced young folks with minimal education. Finally, a mid-value customer segment was identified that was primarily comprised of affluent, older retirees. By successfully identifying the key customer value segments it is now possible to drive future revenue by targeting/marketing to those most lucrative customers.

#END

Appendix – R Code for K-means Clustering

```
# Load libraries and data
library(data.table)
library(lubridate)
library(bit64)
library(dplyr)
library(stringr)
library(anytime)
library(tidyverse)
library(ggplot2)
library(broom)
library(janitor)
library(dlookr)
library(purrr)
library(psych)
library(openxlsx)
library(expss)
library(cluster)    # clustering algorithms
library(factoextra)
library(wesanderson)
library(klustR)
library(ggally)

setwd("~/Desktop/Segmentation Project/Working Folder")
df <- read.xlsx(file.choose(), 1)

# First Trial - Good Clusters
# Feature Selection,   would like to find high value customer segments based upon ARPU or Reve
nue
# Choose Select Variables
# Needs to be numeric,  no outliers.  Scale numerical data.
df$HouseholdSize<-as.numeric(df$HouseholdSize)
df$IncomeIndLog <- log(df$HHIncome/df$HouseholdSize)
df1 <- select(df, Age, EducationYears, IncomeIndLog, EmploymentLength, ARPU_Log)

# Scale/Normalize Data
df1.scaled <- scale(df1)

# Determine best number of clusters
fviz_nbclust(df1.scaled, kmeans, nstart = 25, method = "wss")+
  labs(subtitle = "WSS Method")

fviz_nbclust(df1.scaled, kmeans, nstart = 25, method='silhouette')+
  labs(subtitle = "Silhouette Method")

fviz_nbclust(df1.scaled, kmeans, nstart = 25,  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")

nc <- NbClust(df1.scaled, min.nc=3, max.nc=5, method="kmeans") # too long
table(nc$Best.nc[1,]) #
barplot(table(nc$Best.nc[1,]))

# Determine Optimal # Clusters then run Kmeans Algorithm
# Four clusters
set.seed(1234)
km1 <- kmeans(df1.scaled, 4, nstart=25) # Kmeans results
kcluster=km1$cluster
```

```
kmeans.med1 <- aggregate(as.data.frame(df1.scaled),
                         by=list(cluster=kcluster), median)


kmeans.med2 <- aggregate(as.data.frame(df1),
                        by=list(cluster=kcluster),


df.w_clusters <- cbind(df, cluster = km1$cluster)  # add clusters to original DF

cluster.plot <- fviz_cluster(km1, data = df1, geom="point", alpha=0.5,
              main = paste ("Kmeans Cluster Plot \n Customer Value Segmentation")) +
  scale_fill_manual(values = wes_palette("Darjeeling1"))+
  scale_color_manual(values = wes_palette("Darjeeling1"))+
  theme_bw()+
  theme(legend.position = "none",
    plot.title = element_text(hjust=0.5, face="bold", size=20),
    axis.text.y = element_text(color="black", size=14),
    axis.text.x = element_text(color="black", size=14),
    axis.title = element_text(color="black", face="bold", size=16),
    panel.border = element_rect(fill=NA, size=1),
    panel.grid.major = element_line(color="grey75", size=0.5))

ggsave(filename="KMeansClusterPlot.png",
       plot=cluster.plot, device="png",
       height=5.5, width=9.0, units='in', dpi=600)

sil <- silhouette(km1$cluster, dist(df1.scaled))
fviz_silhouette(sil,ggtheme=theme_bw(), print.summary = TRUE)

ggplot(df.w_clusters, aes(x=cluster, y=ARPU_Log, group=cluster))+geom_boxplot()

km1$size  # identification of size of clusters
# END for k=4


# Three Cluster Results
set.seed(1234)
km2 <- kmeans(df1.scaled, 3, nstart=25) # Kmeans results
kcluster2=km2$cluster

# aggregate data, median of each cluster
kmeans2.med1 <- aggregate(as.data.frame(df1.scaled),
                         by=list(cluster=kcluster2), median)
write.xlsx(kmeans2.med1, "ClusterMedian2.xlsx")

# aggregate data, median of each cluster
kmeans2.med2 <- aggregate(as.data.frame(df1),
                         by=list(cluster=kcluster2), median)
write.xlsx(kmeans2.med2, "ClusterMe1.xlsx")

# add clusters to original DF
df2.w_clusters <- cbind(df, cluster2 = km2$cluster)
write.xlsx(df2.w_clusters, "df2.w_clusters.xlsx")

cluster.plot2 <- fviz_cluster(km2, data = df1, geom="point", alpha=0.5,
                              main = paste ("Kmeans Cluster Plot \n Customer Value Segmentation
")) +
  scale_fill_manual(values = wes_palette("Darjeeling1"))+
  scale_color_manual(values = wes_palette("Darjeeling1"))+
```

```
  theme_bw()+
  theme(legend.position = "none",
        plot.title = element_text(hjust=0.5, face="bold", size=20),
        axis.text.y = element_text(color="black", size=14),
        axis.text.x = element_text(color="black", size=14),
        axis.title = element_text(color="black", face="bold", size=16),
        panel.border = element_rect(fill=NA, size=1),
        panel.grid.major = element_line(color="grey75", size=0.5))

ggsave(filename="KMeansClusterPlot2.png",
       plot=cluster.plot2, device="png",
       height=5.5, width=9.0, units='in', dpi=600)

sil <- silhouette(km2$cluster, dist(df1.scaled))
fviz_silhouette(sil,ggtheme=theme_bw(), print.summary = TRUE)

ggplot(df2.w_clusters, aes(x=cluster2, y=ARPU_Log, group=cluster2))+geom_boxplot()

km2$size

# Visualizations of Feature Variables as a Function of Cluster

features1 <- c("cluster2", "Age", "EmploymentLength", "EducationYears")
df.select1 <- df2.w_clusters[,features1] %>%
  pivot_longer(-1, names_to = "FeatureVariable") %>% mutate(cluster2 = case_when(
    cluster2 == "1" ~ "Mid-Value",
    cluster2 == "2" ~ "Low-Value",
    cluster2 == "3" ~ "High-Value"
  ))

# FeatureVar1 - Visualizations

FeatureVar1 <- ggplot(df.select1, aes(x=value, y=FeatureVariable, fill=factor(cluster2)))+
  geom_boxplot() +
  coord_flip() +
  scale_fill_manual(values = wes_palette("Darjeeling1"))+
  scale_color_manual(values = wes_palette("Darjeeling1"))+
  theme_bw()+
  theme(plot.title = element_text(hjust=0, face="bold", size=22),
        axis.text.y = element_text(color="black", size=16),
        axis.text.x = element_text(color="black", size=16),
        axis.title = element_text(color="black", face="bold", size=18),
        plot.subtitle = element_text(hjust=0, size=16),
        panel.border = element_rect(fill=NA, size=1),
        legend.text = element_text(size=16),
        legend.position = "top",
        legend.title = element_blank(),
        panel.grid.major = element_line(color="grey75", size=0.5)) +
  labs(title="Data Distributions of Feature Variables per Cluster",
       subtitle="K-Means Clustering (k=3)\n", x="Value\n",
       y="\nFeature Variable")

ggsave(filename="FeatureVar1.png",
       plot=FeatureVar1, device="png",
       height=6.5, width=9.5, units='in', dpi=600)

# FeatureVar2

features2 <- c("cluster2", "ARPU_Log", "IncomeIndLog")
df.select2 <- df2.w_clusters[,features2] %>%
```

```
  pivot_longer(-1, names_to = "FeatureVariable") %>% mutate(cluster2 = case_when(
    cluster2 == "1" ~ "Mid-Value",
    cluster2 == "2" ~ "Low-Value",
    cluster2 == "3" ~ "High-Value"
  ))

FeatureVar2 <- ggplot(df.select2, aes(x=value, y=FeatureVariable, fill=factor(cluster2)))+
  geom_boxplot() +
  coord_flip() +
  scale_fill_manual(values = wes_palette("Darjeeling1"))+
  scale_color_manual(values = wes_palette("Darjeeling1"))+
  scale_x_continuous(sec.axis = sec_axis(~ .), name = "Value")+
  theme_bw()+
  theme(plot.title = element_text(hjust=0, face="bold", size=22),
        axis.text.x = element_text(color="black", size=16),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.ticks.y.right = element_line(),
        axis.text.y.right = element_text(color="black", size=16),
        axis.title.y = element_blank(),
        axis.title.x = element_text(color="black", face="bold", size=18),
        axis.title.y.right = element_text(color="black", face="bold", size=18),
        plot.subtitle = element_text(hjust=0, size=16),
        panel.border = element_rect(fill=NA, size=1),
        legend.text = element_blank(),
        legend.position = NULL,
        legend.title = element_blank(),
        panel.grid.major = element_line(color="grey75", size=0.5)) +
  labs(title=NULL,
       subtitle=NULL,
       y="\nFeature Variable")

ggsave(filename="FeatureVar2.png",
       plot=FeatureVar2, device="png",
       height=4.5, width=3.5, units='in', dpi=600)
```