Segmentation Project

Merrimack College

DSA5100 Data Exploration

Taryn Popplewell

Data Science Program

The Future is Data!

Customer Segmentation Executive Summary

Introduction

Telecommunications Company Inc is committed to providing top tier services to its base of subscribers. In an effort to build customer loyalty and subscriber retention a segmentation exercise was performed to help paint a picture of the typical Telecommunications Company customer and identify those who are of high value.

Customer Background Data

The segmented customer base was pre-tailored to represent all five regions of the United States roughly equally. We see a similar pattern in customer town size and age, which tells us that this review may not represent the true regional, town, or age distribution of all Telecommunications Company customers, but it does present a de facto balanced view of customers by these variables.

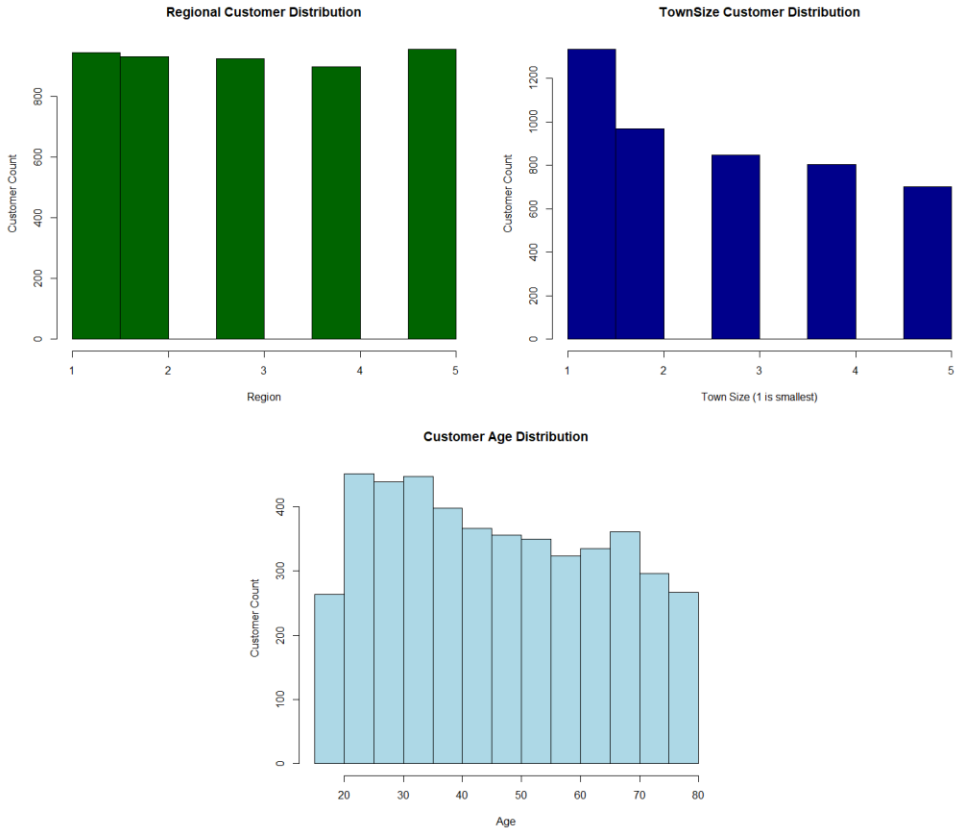

Figure 1. Histograms of Customer Region, Town Size, and Age

Segment Analysis

A k-means segmentation analysis revealed four groups of users separated by tenure with the company in months and the users' tendency to adopt technology. The majority of the survey group is split between Cluster 1 – Low Tenure/Low Tech and Cluster 2 – Low Tenure/High Tech.

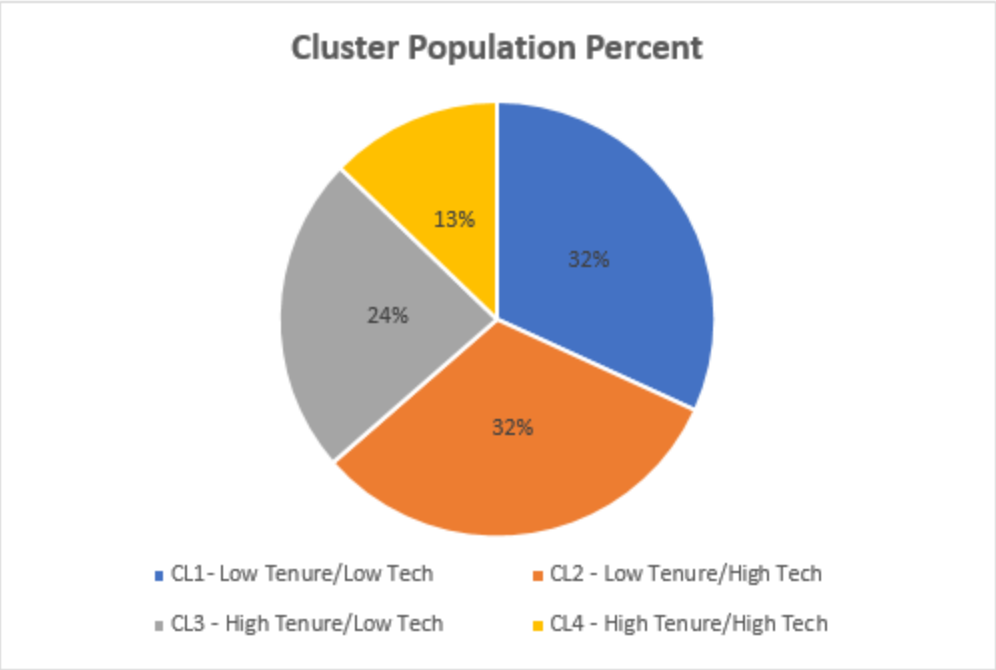

Figure 2. Cluster Percent of Survey Population

Segment Profiles

Using the averages reported for each cluster, the customers within are described as follows.

Cluster 1 is composed of low tenure, low tech adopting customers who are roughly middle aged whose highest level of education is a high school diploma.

Cluster 2 is composed of low tenure, high tech adopting customers who are roughly middle aged who is college educated.

Cluster 3 is composed of high tenure, low tech adopting customers who are approaching retirement age whose highest level of education is a high school diploma.

Cluster 4 is composed of high tenure, high tech adopting customers who are approaching retirement age or already there, who is college educated.

Analysis

The data associated with these clustered customers show a clear pattern of tech-savvy subscribers bringing more value to the company. Figure 3 below shows the average dollars spent over tenure skyrocket into the thousands at Cluster 4 – High Tenure/High Tech. We also see that clusters 3 and 4 spend more on voice than clusters 1 and 2. We'll talk more about that later.

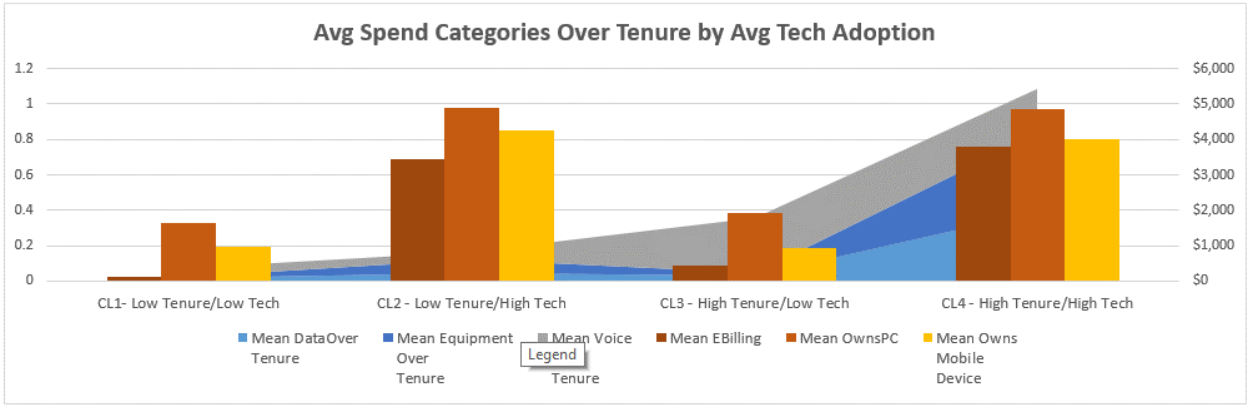

Figure 3. Average Spend Over Tenure by Tech Adoption

In order to better depict the average dollars spent for Cluster 2 – Low Tenure/High Tech, Figure 4 removes Cluster 4. We see that even for Low Tenure customers who are High Tech adopters their data and equipment dollars over tenure outpace those of the less tech-savvy Clusters 1 and 3.

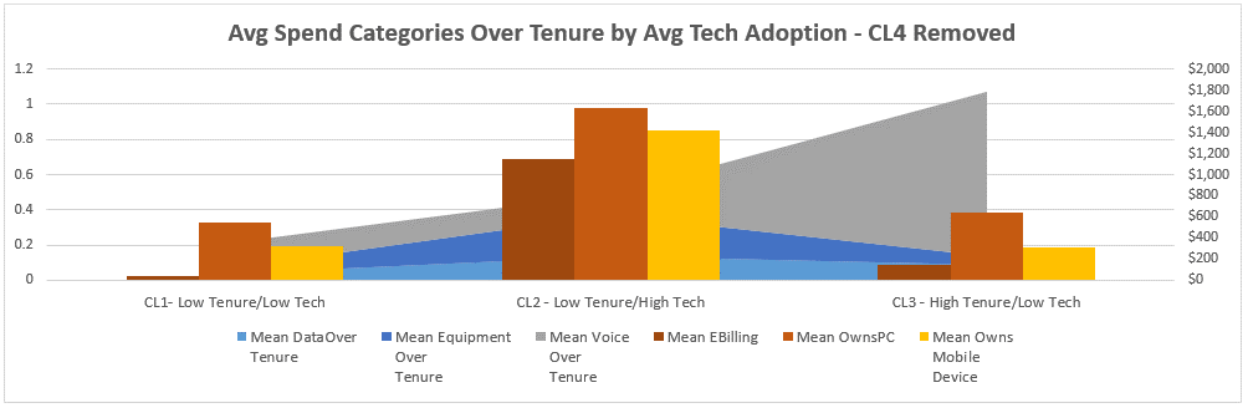

Figure 4. Average Spend Over Tenure by Tech Adoption

Further analysis of the clusters reveal that average data and equipment dollars spent appear to be independent of customer age and their average tenure with the company. Figure 5 shows us that not only is Cluster 4 comprised of high tenure, older customers, but that they also have spent more on data than the other cluster customers. Removing Cluster 4 from Figure 5 would mimic Figure 4, showing that the younger-aged customers of Cluster 2 are spending more on data than the low-tech customers of Clusters 1 and 3. This suggests that technology adoption is a strong indicator of whether a consumer will be a high value Telecommunications Company customer.

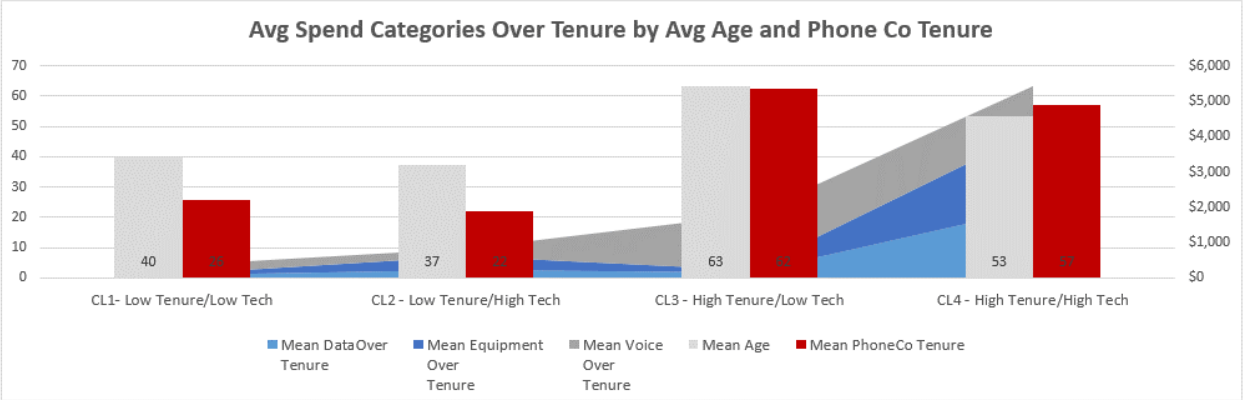

Figure 5. Average Spend Categories Over Tenure by Average Age and Phone Co Tenure

Digging deeper, we see a clear link between education years and technology adoption as illustrated in Figure 6. Our two low-tech clusters, 1 and 3 both have the standard 13 years of education (K-12). The high-tech customers in clusters 2 and 4 have on average 16 years of education, which suggests these are college educated consumers.

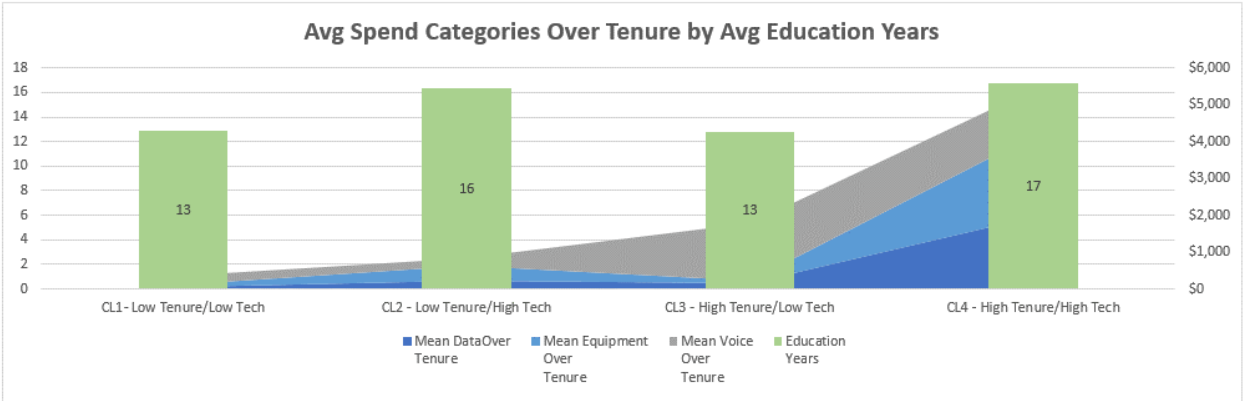

Figure 6. Average Spend Categories Over Tenure by Average Education Years

K-means cluster analysis revealed that the two younger, shorter tenured clusters earn less per household that the older grouped clusters. This is not surprising since usually people's incomes increase as they get older and gain experience. What is interesting though is of clusters 1 and 2, cluster 2 is spending quite a bit more per service on average than cluster 2. This spending average along with cluster 2's high technology adoption against the low increase in household incomes suggest that the customers in cluster 2 place a higher value priority on data and equipment in their monthly budgets.

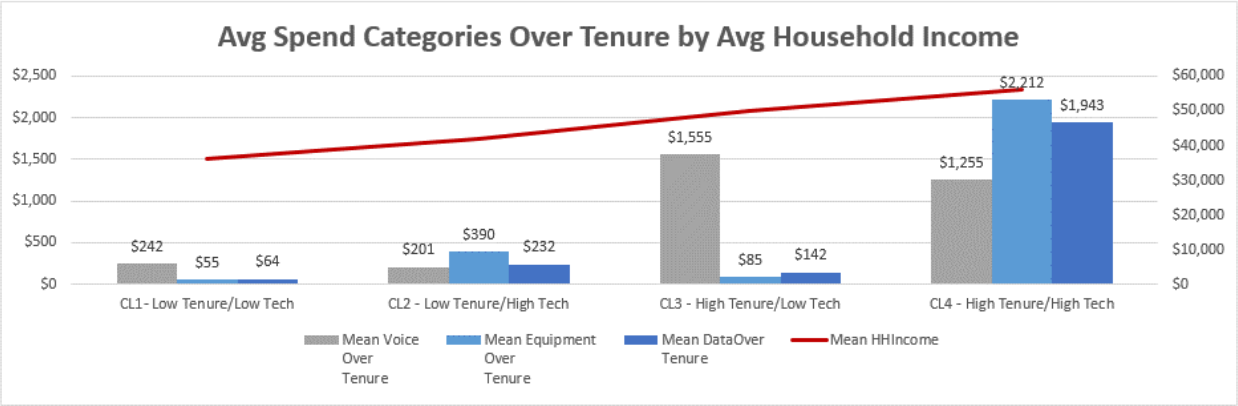

Figure 7. Mean Data Spend Over Tenure Against Household Income

Finally, for the sake of completeness we examine each cluster's mean total over tenure value. Cluster 2 is over double the value of Cluster 1. Cluster 3 has the highest value, but you'll recall from Figure 5 that this is also the highest tenure group. We would expect the overall tenure average dollars to be higher because they have spent more time spending money with the company. Cluster 4 is hot on the heels of Cluster 3 for spending and while this group has a slightly lower average tenure run, they are also an average of 10 years younger than Cluster 3 (again, see Figure 5). If Cluster 4 continues their current spending habits for the next 10 years with the company it's easy to imagine their overall average spend amount eclipsing that of Cluster 3.

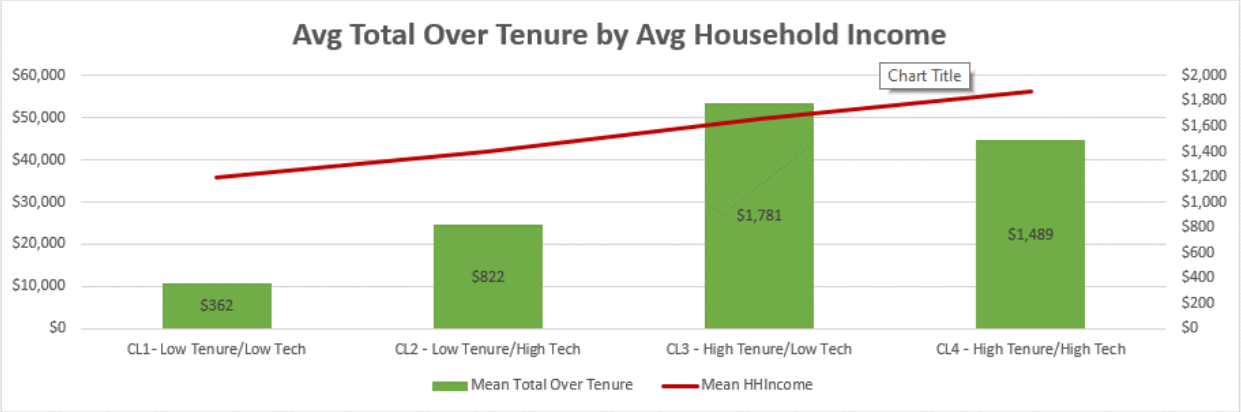

Figure 8. Average Total Over Tenure by Average Household Income

Conclusion and Recommendations

The results of the k-means segmentation exercise show associations with technology adoption, education years, and dollars spent over tenure. Older, high tenured, low tech customers lead in overall average spending (Figure 8), this is a red herring for data detectives hoping to draw conclusions about what makes a valuable customer. College educated customers who are also high technology adopters have been shown to spend more on data and equipment services (Figure 6).

Based on these findings it is recommended that retention efforts focus on customers who have a college education and/or an interest in technology. Outreach work could include selling data and equipment

services to the low tenure/low tech cluster. Partnering with a PC or mobile device manufacturer or retailer could help to get these devices into the hands of customers in Clusters 1 and 3, possibly increasing their likelihood to subscribe and prioritize data and equipment services in their monthly finances.

Detailed Processes and Findings

Data Preparation and Cleansing

The provided data set contained responses from 5,000 customers distributed roughly equally by region, town size, and age (see Figure 1). This immediately tells us that the sample is intended to represent customers equally by these variables. There was a notable right-skew in the household income variable (HHIncome) that was corrected as it could have potential clustering impacts. After the removal of the HHIncome outliers the total data set was dropped to n = 4654.

Data was treated for NAs, zero values, and nonsense value where necessary. Fields with categorical data were recoded to binary 1 = Yes and 0 = No. While it is sometimes customary to replace empty rows of value fields with the overall mean of the field itself, any customer value fields such as VoiceLastMonth or DataOverTenure had its empty rows replaced with a zero. This was done to avoid artificially inflating overall average values in these key value fields.

Some additional fields were created and added to the data set for additional exploratory options. Not all fields may have been used.

New Variable Name	New Variable Description
Age_Buckets	Ages distributed into six groups
AvgVoiceOverTenure	Represents the total amount spent on voice services divided by months of tenure
AvgEquipOverTenure	Represents the total amount spent on equipment rentals divided by months of tenure
AvgWDOverTenure	Represents the total amount spent on wireless data services divided by months of tenure
MonthlyOverTenure	Represents the total amount spent on all services divided by months of tenure
PCoTenure_HighLow	Customers with Phone Co Tenure between 0-36 are 0/Low, those >= 37 are 1/High
MonthlyTotal_HighLow	Customers with MonthlyOverTenure amounts between 0-48 are 0/Low, those >48 are 1/High
TechAdopter	Ebilling = Yes; OwnsPC = Yes; OwnsMobileDevice = Yes
VoiceOnly	EquipmentRental = No; WirelessData = No

Figure 9. Scaled Means of K-means Clustering Technique

AvgTotal_HighLow			
ID	Desc	n	% of n
0	Low	3481	75%
1	High	1173	25%

Age Range	Bucket #
18-24	1
25-34	2
35-44	3
45-54	4
55-64	5
65+	6

Figure 9. Extra Detail on Created Variables

Segmentation Techniques

The two segmentation techniques employed in this exercise were k-means clustering and decision trees. They were selected because of their differing input requirements; the former is unsupervised and does not require a dependent variable, whereas the latter is supervised and does require a dependent variable.

K-means was chosen as the primary tool for analysis because of its versatility. Once a reliable group of variables was identified for clustering, it was possible to merge the cluster variable means with the means of variables excluded from the cluster run itself. This allowed us to broaden our view of the customers in each cluster beyond the variables used in the clustering application and take all data set variables into account.

Dissimilarly, the decision trees produced using the customer data required a more selective approach in terms of variables, and produced a more restrictive set of results that could not be merged with the existing data set. Additionally, due to the nature of the R decision tree algorithm, attempts to build larger trees were often stymied by the function's 'best path' determination, which ignores variables it finds to be of little consequence to the final tree structure. Details on the results produced by decision trees are available later in this document.

K-Means Process and Findings

The K-means segmentation technique was selected as the primary method of analyzation due to the detailed data it was able to return. Decision tree output was less thorough and will be examined later in this document.

The final population was analyzed in R using the factoextra library's kmeans solution. The variables depicted below in Figure 9 were selected for the final accepted cluster. Their produced scaled means are also presented. We see a good variation across means and recall from Figure 2 that all cluster sizes made up 13% or more of the final population.

Cluster	Education Years	Employment Length	HHIncome	Marital Status	Household Size	Home Owner	Card Tenure	Card Spend Month	PhoneCo Tenure	Voice Over Tenure	Equipment Over Tenure	Data Over Tenure	EBilling	OwnsPC	Owns Mobile Device
1	-0.4784	-0.3707	-0.2831	-0.0319	0.1876	-0.0957	-0.5600	-0.0912	-0.5139	-0.4498	-0.4470	-0.3344	-0.6680	-0.6207	-0.5721
2	0.5808	-0.5201	-0.0666	0.2127	0.0201	-0.0346	-0.6552	-0.0027	-0.6721	-0.4932	-0.0625	-0.1551	0.7235	0.7200	0.7479
3	-0.5115	1.0709	0.2252	-0.0909	-0.2913	0.0754	1.2220	0.0128	1.1149	0.9341	-0.4123	-0.2513	-0.5416	-0.5138	-0.5863
4	0.6953	0.2347	0.4539	-0.2805	0.0213	0.1847	0.7621	0.2103	0.8873	0.6177	2.0308	1.6825	0.8690	0.7086	0.6524

Figure 9. Scaled Means of K-means Clustering Technique

The algorithm produced the cluster map you see below in Figure 10. There is some clear overlap across clusters, but all in all, the clusters are acceptably dense, well sized, and have varied means. This grouping is therefore acceptable for further analysis.

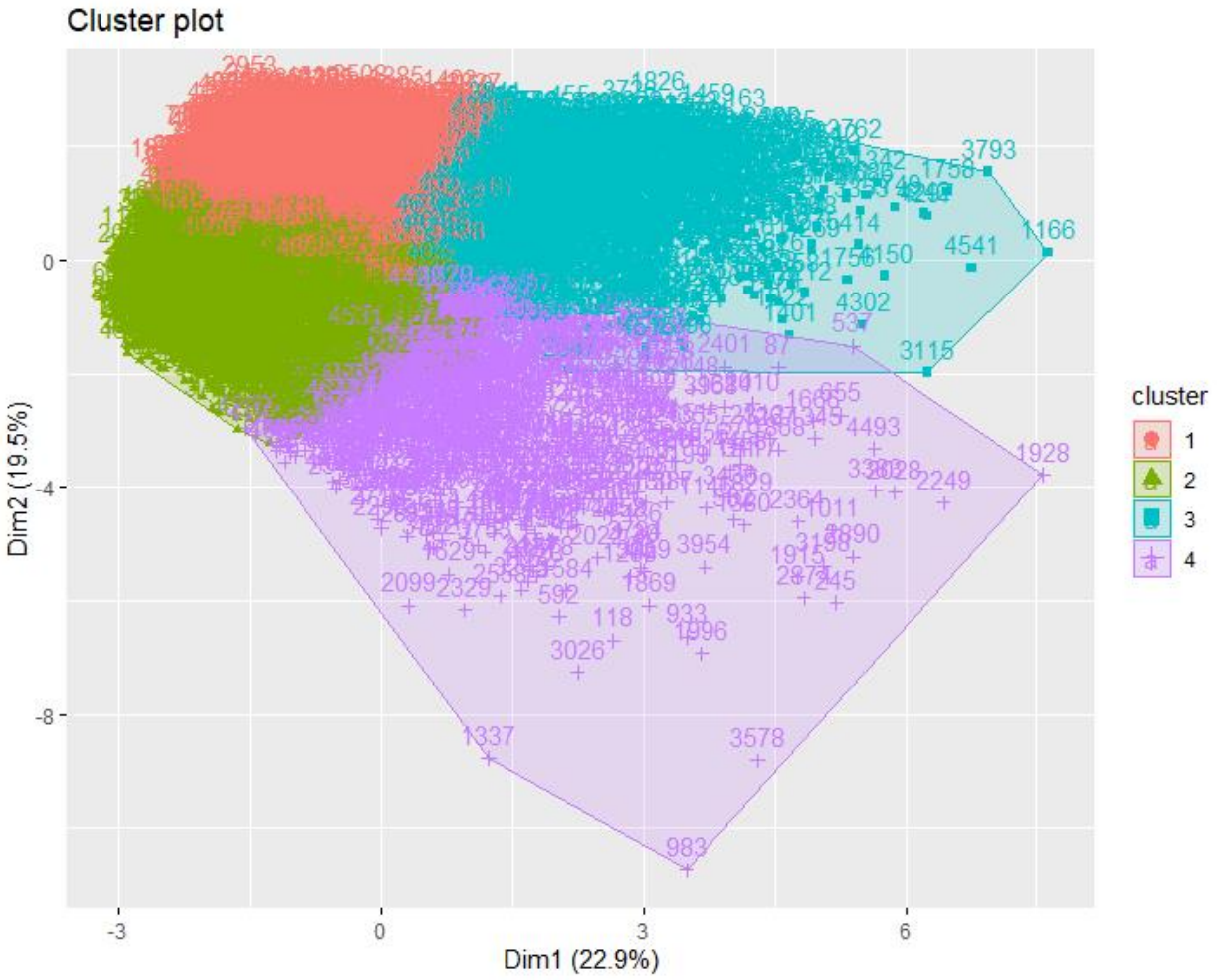

Figure 10. K-means Cluster Visualization

To view the un-scaled means of the rows associated with each cluster, the original customer data set was merged with the cluster assignment data set and summary data was extracted. Below is the unscaled means for Cluster 4 – High Tenure/High Tech. To review the unscaled means for the remaining three clusters please refer to this report's companion Excel Workbook.

	Cluster	Education Years	Employment Length	HHIncome	Marital Status	Household Size	Home Owner	Card Tenure	Card Spend Month	PhoneCo Tenure	Voice Over Tenure	Equipment Over Tenure	Data Over Tenure	EBilling	OwnsPC	Owns Mobile Device
Min.	4	9	0	$9,000	0	1	0	3	$70	26	$103	$0	$0	0	0	0
1stQu.	4	15	5	$30,000	0	1	0	18	$2,084	49	$552	$1,740	$0	1	1	1
Median	4	17	9	$52,000	0	2	1	25	$2,973	59	$912	$2,286	$2,032	1	1	1
Mean	4	16.71	11.17	$56,171	0.3798	2.249	0.7059	25.07	$3,682	57.2	$1,255	$2,212	$1,943	0.7563	0.9714	0.8034
3rdQu.	4	19	16	$77,000	1	3	1	33	$4,456	67	$1,529	$2,813	$2,833	1	1	1
Max.	4	23	42	$131,000	1	8	1	40	$19,781	72	$13,047	$6,525	$12,859	1	1	1

Figure 11. Unscaled Means for Cluster 4 – High Tenure/High Tech

Since the cluster assignments had been merged to the cleansed data set, additional variable means were selected to view alongside the cluster variables. Below is the unscaled additional variable means associated with Cluster 4 – High Tenure/High Tech rows. To review the additional variable means

associated with the remaining three clusters please refer to this report's companion Excel Workbook. Please note that not all cluster-external variables were selected for additional evaluation.

	cluster	Age	Gender	DebtTo Income Ratio	Cars Owned	Active Lifestyle	VoiceLast Month	Equipment LastMonth	Data LastMonth	Monthly Over Tenure	Total Over Tenure
Min.	4	21	0	0.3	0	0	$7	$0	$0	$22	$0
1stQu.	4	41	0	5.1	1	0	$33	$34	$0	$64	$198
Median	4	53	1	8.6	2	0	$49	$42	$39	$91	$698
Mean	4	53.21	0.5025	10.06	2.076	0.4218	$61	$40	$36	$95	$1,489
3rdQu.	4	66	1	14.2	3	1	$71	$50	$53	$117	$1,942
Max.	4	79	1	35.8	7	1	$540	$106	$186	$297	$21,057

Figure 12. Selected Additional Variable Means to View

There was a noticeable jump in Cluster 4's mean dollars over tenure for each value category. These figures were validated in the cluster output to ensure accuracy. As depicted in Figures 3, 4, and 5 of this document we see a strong tendency for High Tech adopters to also be big data spenders over their tenure with the company, even if it's relatively short. High tech adoption was categorized by customers who participate in E-billing, own a personal computer, and also own a mobile device. This categorization was made at the discretion of the analyst.

The means of descriptive demographic variables like marital status, household size, home ownership, and employment length did not exhibit any unusual behavior. Averages were consistent with one another across all clusters and did not seem to have any kind of individual effect on phone company tenure or average spend over tenure. Similar patterns were seen with the credit card tenure and monthly spend amounts. While all these variables contributed to customer cluster assignments, they don't answer any questions individually about a customer value.

The most remarkable patterns across clusters were found in the Education Years, Age, PhoneCoTenure, Ebilling, OwnsPC, OwnsMobileDevice, DataOverTenure, EquipmentOverTenure, MonthlyOverTenure, and TotalOverTenure variables. In an effort to maintain brevity, please refer to the Executive Summary at the top of this document for findings relevant to the goal of identifying high value customers for retention focus.

Key take away – The more tech-savvy and educated the customer, the more likely they are to spend more with the company.

Decision Tree Process and Findings

While decision trees were not selected as the primary clustering technique, they did provide some extra insight to the customer base, and we see decision themes similar to the results produced by k-means clustering. For this reason, a few decision trees produced from the data will be displayed and interpreted.

The final cleansed data set was analyzed in R using the rpart library's rpart decision tree solution. Several different runs were conducted, the most informative of which are shared below. For all decision

trees presented here the MonthlyTotal_HighLow variable was used as the dependent variable. Recall that a high average spend amount over tenure was decided to be greater than $48.

Decision Tree 1 examines TechAdopter and Age_Buckets. We see straight away that customers who are not tech adopters (use Ebilling, own a PC and a mobile device) shoot straight down to the 0/Low bucket. Age doesn't even come into play in that decision. This is very much in line with what our clustering segments found!

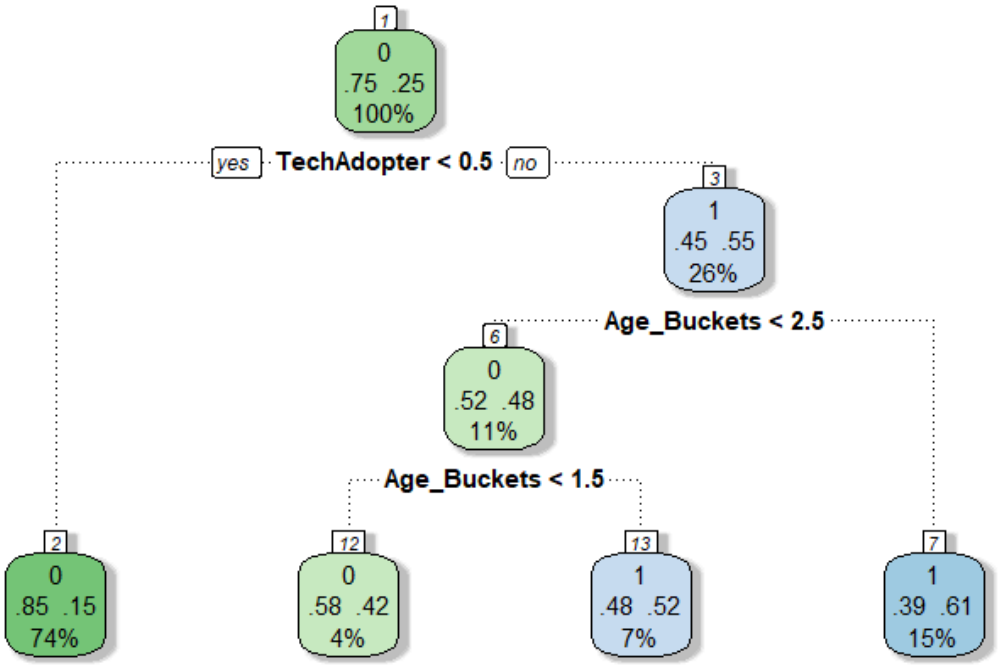

Figure 13. Decision Tree 1

Decision Tree 2 depicted below considers the Tech Adoption variables separately and alongside the Age variable. This tree tells the same story as Decision Tree 1 in greater detail. We see that technology adoption and age play a key role in determining whether a customer is high value.

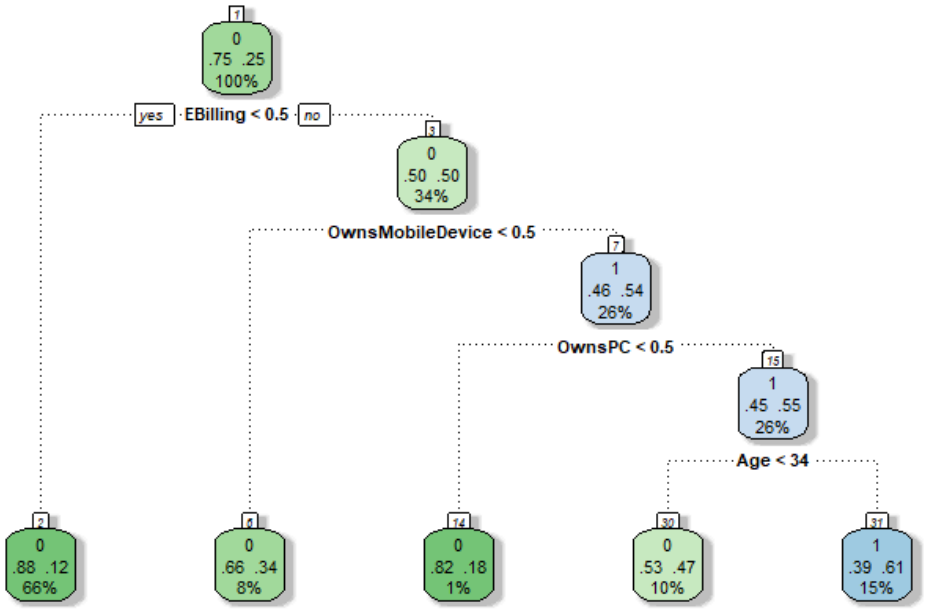

Figure 14. Decision Tree 2

Decision Tree 3 shows us that customers with an average data spend over tenure greater than $17 are likely to be high value customers. Voice customers have to spend more than $48 just on voice over their customer tenure to be considered high value. Money is not where the customer's mouth is!

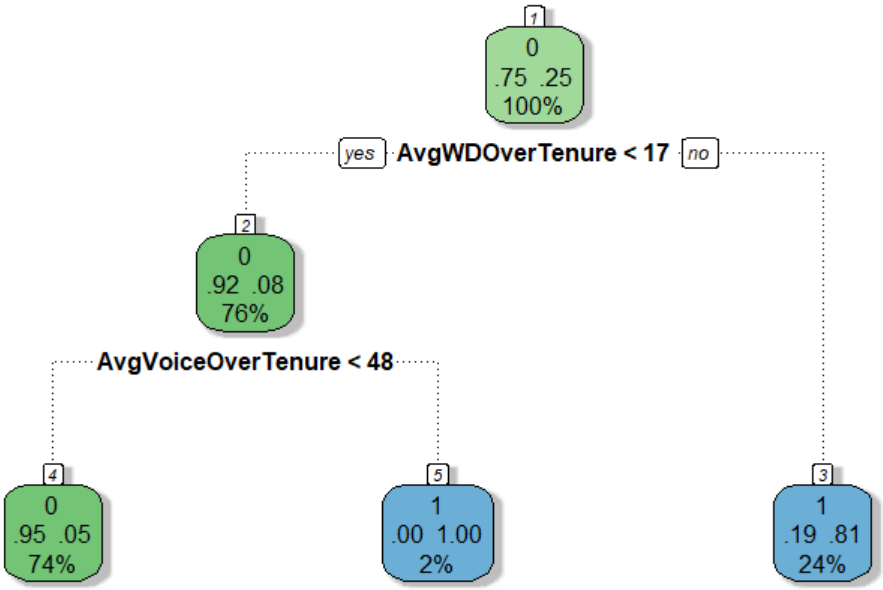

Figure 15. Decision Tree 3

Our fourth decision tree to examine evaluates whether a customer is subscribed only to voice services, tech adoption and age buckets. We see that customers who subscribe only to voice are immediately branched into the low value group. This is consistent with decision tree 3 and with our k-means results. Examining further we see that technology adoption is branched by age. We see greater volatility in high value assignment in the lower age-range buckets. This could be due to younger customers generally earning less than older customers and either not being able to afford technology devices or prioritizing saving and 'starting out' type expenses (such as a home down payment).

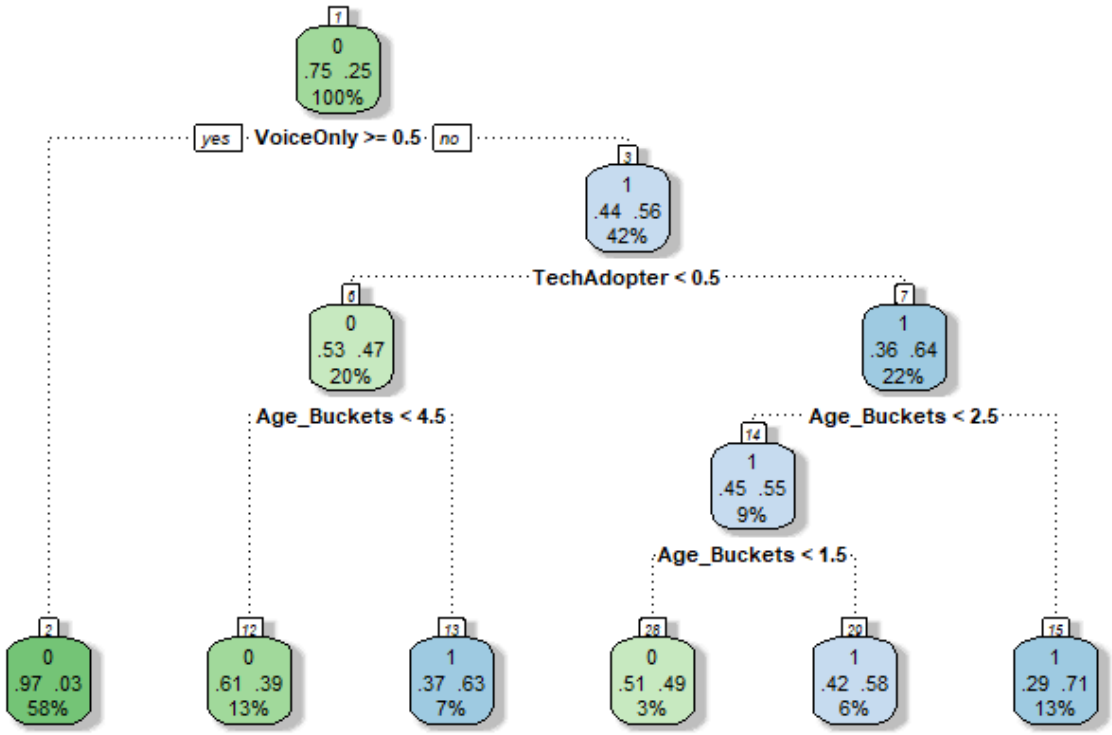

Figure 16. Decision Tree 4

Our fifth and final decision tree examines the means of equipment, data, and voice over tenure. This tree is interesting because it places average equipment over tenure at the top. This is in line with our k-means results but better illustrates the value of rental equipment to the business. Moving down the branches we see the same patters from the k-means clustering repeat. A higher average data over tenure generally leads to a high value assignment.

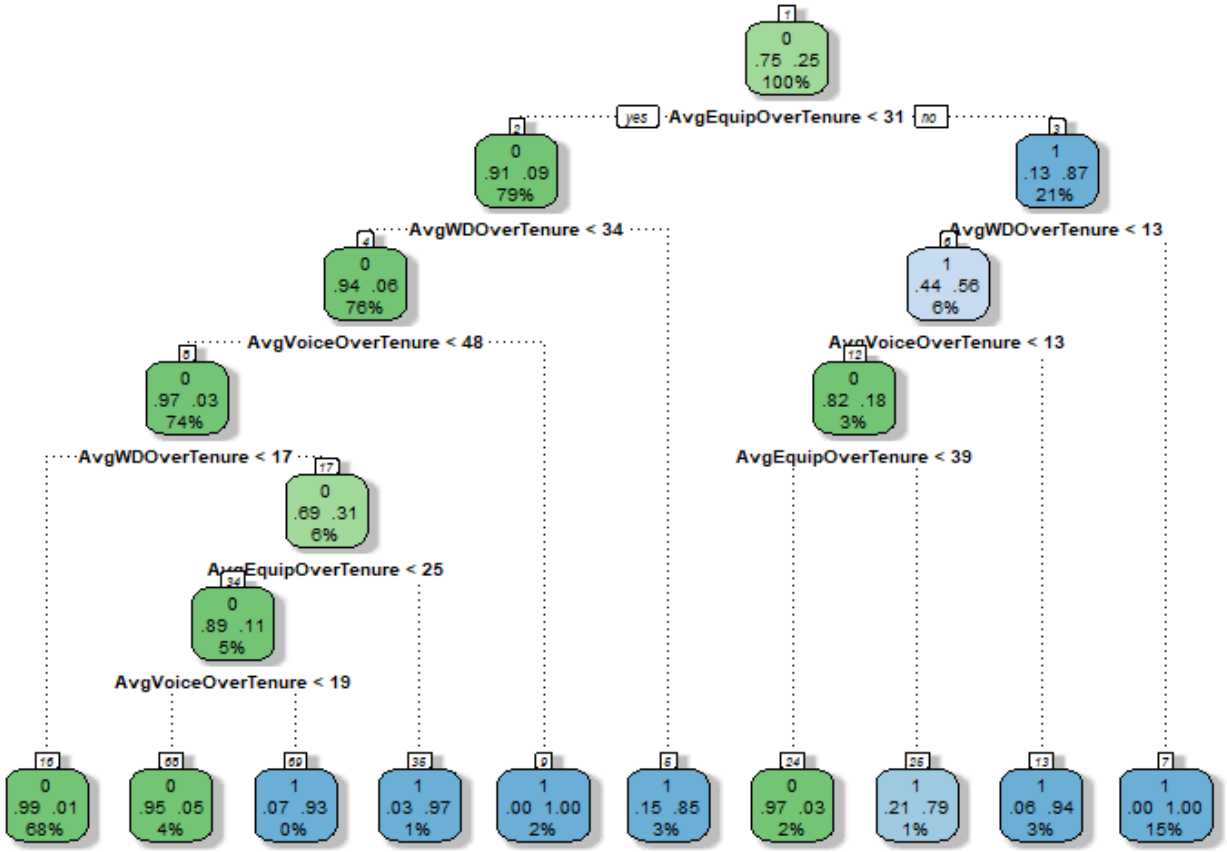

Figure 17. Decision Tree 5

Evaluating decision tree results alongside k-means results may initially seem like an exercise in redundancy, but don't be fooled! This extra technique provides us with two important things. First, it validates our k-means results and instills extra confidence in the findings presented in this report. Second, the presented decision trees provide an easier visual way to examine our key variables in more granular groupings.

Conclusions and Recommendations

As noted in the Executive Summary of this report, the results of the k-means segmentation exercise show associations with technology adoption, education years, and dollars spent over tenure. Older, high tenured, low tech customers lead in overall average spending (Figure 8), this is a red herring for data detectives hoping to draw conclusions about what makes a valuable customer. College educated

customers who are also high technology adopters have been shown to spend more on data and equipment services (Figure 6).

Additional decision tree analysis has provided some extra groupings within key variables that could be useful for marketers looking to add an edge to their ad targeting.

Based on these findings it is recommended that retention efforts focus on customers who have a college education and/or an interest in technology. These customers have shown to be high value, and the longer they are with the company, the more value they will generate.

These customer tendencies have been discovered on a customer base that has not experienced any attempts from the company to affect change. That is to say, the company has not employed any marketing tactics targeted to the customer clusters identified in this report. This means that this is natural customer behavior. The company should consider marketing techniques not only to build its base of college-educated, technology-interested customers, but also to encourage customers who are not technology-inclined to become so. An example of this could be partnering with a PC or mobile device manufacturer/retailer to help get these devices into the hands of customers in Clusters 1 and 3. The goal being to get these customers accustomed to using these devices and then possibly increasing their likelihood to subscribe and prioritize data and equipment services in their monthly finances.

One final recommendation is that this analysis be conducted on a truly random sample of customers. As noted in the Data Preparation and Cleansing portion of this document, age, region, and town size were all represented equally enough in the dataset to suggest the results were tailored or pared down. Since the Age and Age_Buckets variables did come somewhat into play in both the k-means and decision tree results, it might be wise to conduct an additional survey that shows a more accurate distribution of the customer base by age, region, and town size.

Additional Resources

Only snippets of the actual data were included in this report as a mercy to the reader. The companion spreadsheet to this document contains all data and results that the reader may review at their leisure.