Executive Summary

Situation: 'Hoyt Phone Company' is a nationwide phone company, they have drawn 5000 customers from their customer base at random in hopes of identifying customer values and increasing retention with their product. This segmentation of customers is being done in the hopes of giving the company a competitive advantage in the phone industry. Our goal is to segment our customers based on their value to the company as well as predict the likelihood of retaining various groups of customers. Having this information will help us produce different marketing campaigns and offer services and rewards to our customers.

Processes: The data science team received the data frame of raw data, which consisted of 5000 customers (rows) and 60 variables (columns) this made for 300,000 data points. The first step was cleaning the data and identifying useful variables for the segmentation we would be conducting later on in the workflow. Empty and NA values were turned into 0's, numbers that were labeled characters were transformed into numeric values, and unnecessary special characters were removed from the data. The variables we identified as useful for segmenting this data were; ID, Age, Gender, Education, Phone Co Tenure, Voice Over Tenure, Equipment Over Tenure, Data Over Tenure, TV Watching Hours, and Household Income. New variables were created with the average of all the existing variables for later use.

For the segmentation of this data, we used two different forms of segmentation, a ruled-based technique and an unsupervised technique. The first technique used was rule-based segmentation "Rule-based systems are a basic type of model that uses a set of prewritten rules to make decisions and solve problems. Developers create rules based on human expert knowledge that enable the system to process input data and produce a result."(Foster, 2023). Using our previous knowledge of the Phone industry we identified rules in order to segment our customer data set and see if we could find trends. Using customer age, gender, and voice-over tenure we segmented this data set. Voice data is our most leveraged product and also has the highest mean (use) relative to other predictors(Data, Equipment, TV Hours, etc.). Knowing this we felt this would be a good gauge to measure customers' value and predict the likelihood of retention.

The second technique was unsupervised learning, where we leveraged K-means clustering to cluster and identify segments within our customer data. "divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups." (*K means clustering - introduction* 2023). For this segmentation, we used voice, equipment, and data tenure which we averaged into monthly averages for the three and calculated the monthly average for all three for every customer. Using the average monthly value with the customer tenure with the phone company we segment the customers into 6 different groups with varying level of value and retention probability.

Outcomes: After conducting segmentation and visualizations on our data set using rule-based learning and unsupervised learning it became apparent to us that the unsupervised learning technique yielded a much better result in segmenting our customer. I believe the fact that the unsupervised learning method was able to incorporate so many variables into its segmentation that it just made for a better picture of our customer groups.

The rule-based outcome resulted in six different segments. Segment 1 had customers with low engagement/value for the company and also were males with the lowest average age. Segment 2 were also males but on average older who had higher value to the company but kinda sat at the medium level. Segment 3 which according to rule-based learning is our most valued customer, these individuals were the oldest on average older, and brought the most value to our company. Segment 4 was the same as segment 1 but consisted of only females, segment 5 matched segment 2, and segment 6 matched segment 3. For ease of this paper, we will combine 1 and 4, 2 and 5, 3 and 6 together into Low, Medium, and High valued customers.

The unsupervised method of segmentation yielded a much better outcome for segmenting customers into value groups. Using voice-over tenure, equipment-over tenure, data-over tenure, TV-watching hours, household income, and phone co-tenure. We calculated monthly averages for every customer's data, equipment, TV, and voice using this information. With those averages, we calculated the monthly value each customer held and then used a scatter plot to plot our customers and applied K-means clustering to our data with 6 k-means segments. We found this technique yielded a much better picture of the value of our customer base. Being able to apply multiple variables to segmenting provides a much clearer picture of each customer's contribution to the company through data, equipment, voice, and TV. Segment 3 was our lowest tenure and lowest monthly value customers. These customers on average were the youngest (35) and had significantly lower incomes than those of other segments. Segment 6 was medium lengthened tenure and low monthly value customers. This segment was mid-aged customers which had an average age of 47 which is interesting as there are two more segments with a younger average age that outperform them, which disproves my initial suspicion of age having a direct correlation with value and tenure. Segment 1 is our high-tenure and low-value customers. These customers use us for the bare minimum but also have a very high likelihood of retention. They also are our oldest on average segment at 59. Segment 2 are our medium values and low to mid-tenured customers, these are our second youngest segment and second lowest paid. Segment 4 is our high tenure and medium value customers our second oldest segment and second highest income as well. Finally, Segment 5 our high value and high tenure customers not the oldest segment but by far the highest average income customers. These are our core customers who should receive the platinum treatment.

Recommendations: My professional opinion is the company should leverage the unsupervised machine learning technique of K-means clustering to segment our customers going forward. It is a cheap, easy-to-implement but also very sophisticated and efficient method that fits our use case perfectly. Rule-based learning in this situation is not a good fit, considering all

the additional time and rules needed to match the K-means method. For clarity purposes, all segments referenced going forward will be the k-means segments. Segment 3: our low-value and low-tenure customers. I feel it would be best to target these customers with a simple promotional campaign of potential products that may interest them and could be used to increase their likelihood of retention and value. Segment 6: we are more confident in their retention as they have been with the company for at least 2 years, it is the value that they are low on. I think pushing a promotional campaign for the different products we offer could be useful. Segment 1: these are our long-tenured and oldest customers, It might be useful to cater a marketing campaign fit for an older generation, no need to worry about retention with these customers. Segment 2: our low to mid-tenured customers with medium value. I think the increase in spending with the company helps tame retention concerns but it still would be useful to introduce popular products to these individuals. Segment 4: our long-tenured medium valued customers, I think it'd be useful to push new products with discounts and promotional deals to these customers as they are long-tenured and likely to stay with our company and they have the second highest average income so the profit potential is there. Finally, segment 5: our high-tenured and high-valued customers. These individuals should be rewarded for their commitment and also have exclusive offers such as sporting events, theme parks, paid subscriptions, etc. These are the customers that will rave about us to their friends and family and potentially gain us more customers.

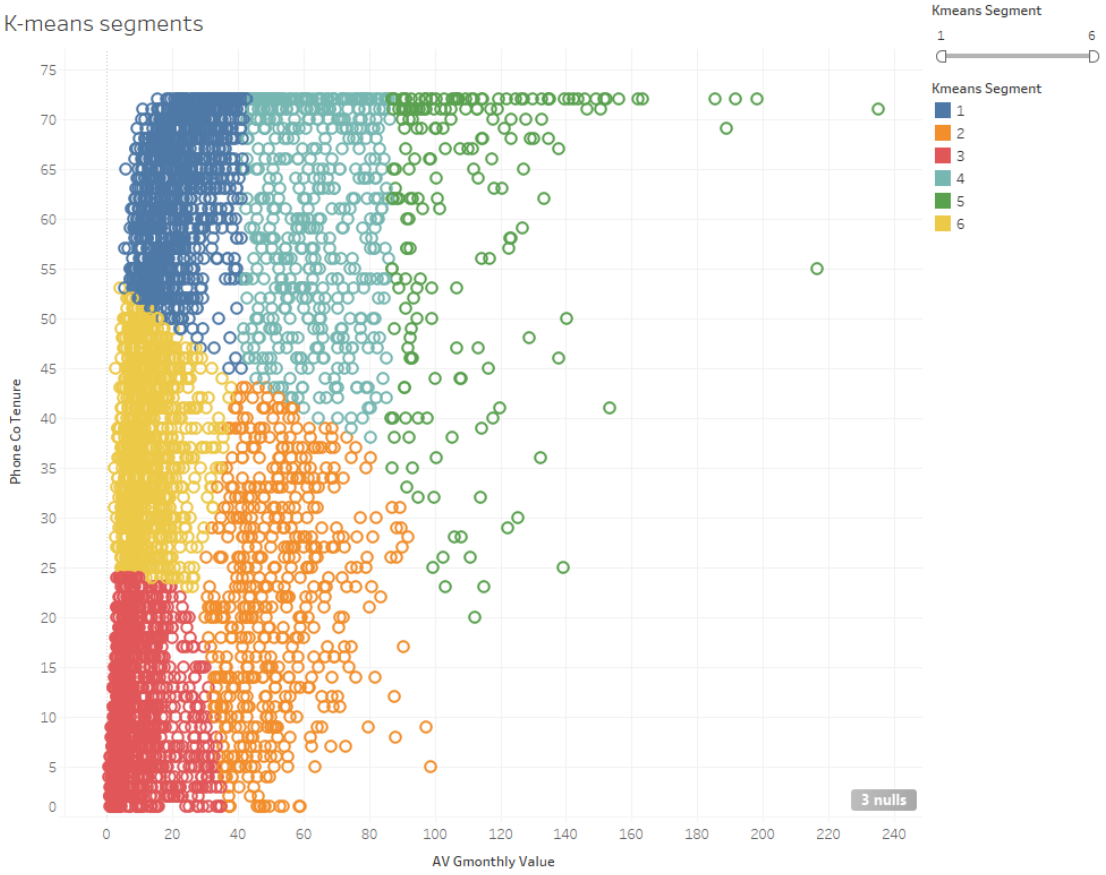
Technical Report:

Unsupervised learning - K-means Clustering:

Customer Segment Metric

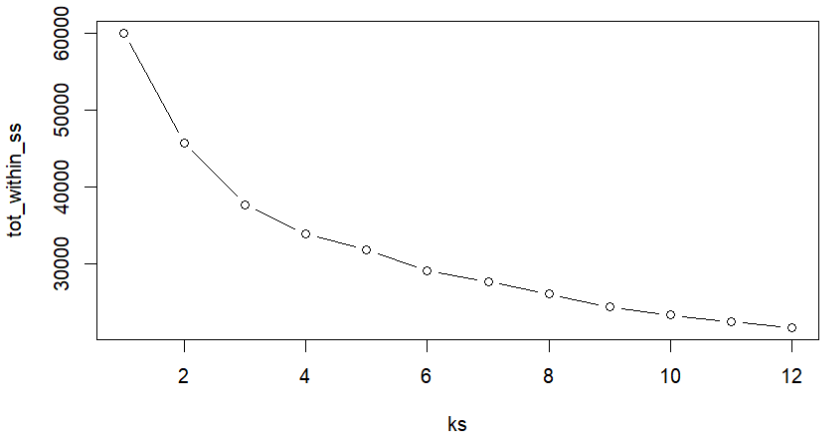
Kmeans ..	Avg. Age	Avg. HH Income	Avg. Phone Co Tenure	Avg. TV Watching Hours	Avg. Data Over Tenure	Avg. Equipment Over Tenure	Avg. Voice Over Tenure
1	59	58,808	63	20	17	360	1,259
2	38	56,458	22	20	753	623	188
3	35	37,711	11	19	14	74	63
4	56	76,697	61	20	1,618	1,228	1,740
5	56	105,049	60	20	3,587	2,601	2,239
6	47	49,822	37	20	23	228	365

Here is a brief overview of the metrics for the different value segments, as you can see segment 5 has the highest averages across the board except for age and phone co-tenure. Segment 1 has the highest average phone co-tenure and age but has a low customer value. Segment 1 holds a lot of value potential if it were to be targeted correctly.



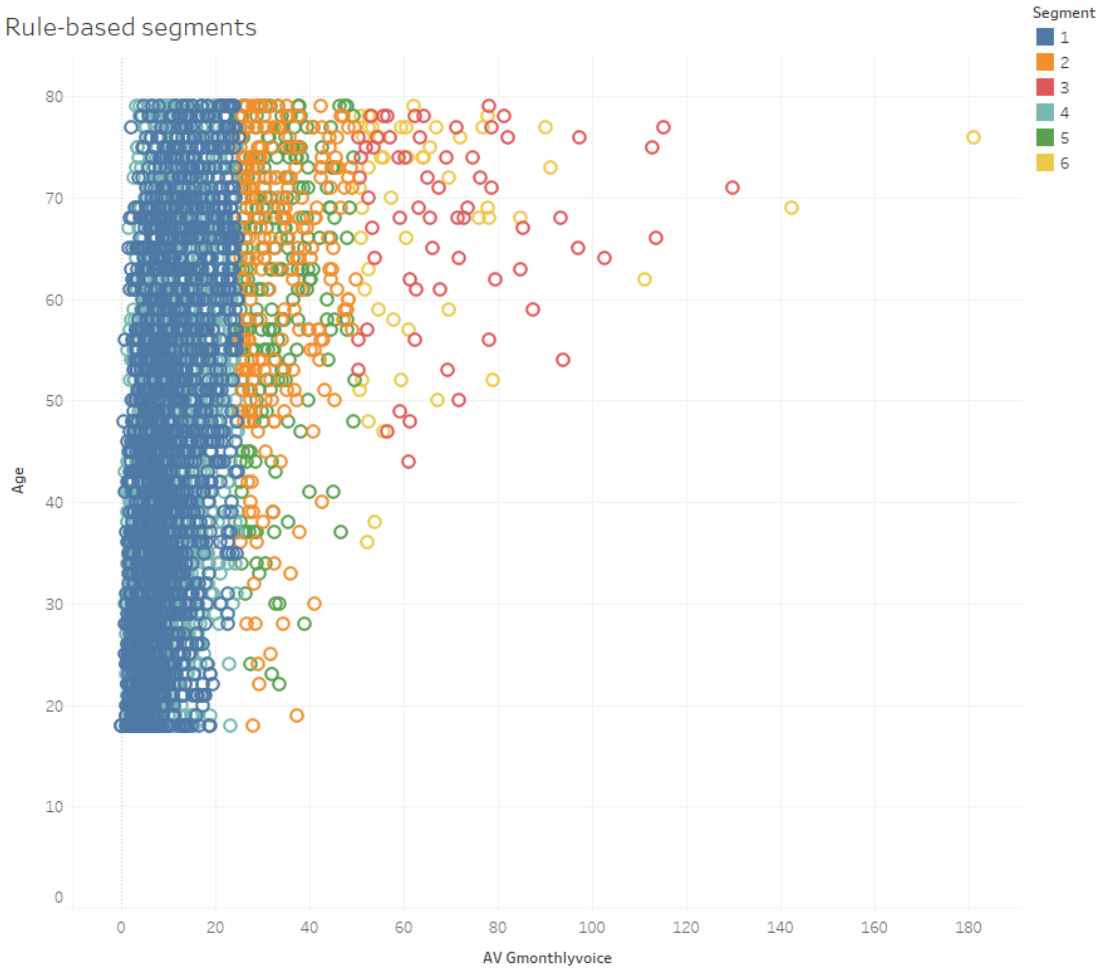
Here we can see the clusters more clearly, segments 1 and 4 hold value as they are long-term customers but haven't necessarily brought a lot of value to the company yet. Segments 2, 3, and 6 are segments we want to try and increase retention rate in through promotional opportunities and targeted ads.

Here is a breakdown of household incomes by segment, unsurprisingly segment 5 who are our core customers has the highest annual income. Segment 4 has the next highest income and they are our long-tenured lower-valued customers which is another reason to target these customers with marketing and product campaigns. The same goes for Segment 1.

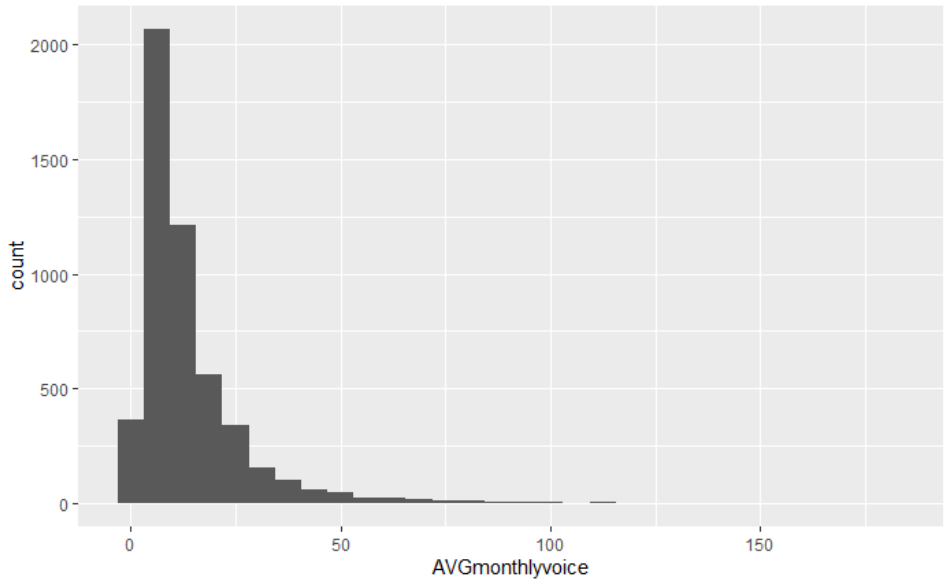


We decided on using 6 segments from using the 'Elbow Method' "The elbow method is a graphical method for finding the optimal K value in a k-means clustering algorithm. The elbow graph shows the within-cluster-sum-of-square (WCSS) values on the y-axis corresponding to the different values of K (on the x-axis). The optimal K value is the point at which the graph forms an elbow." (elbow-method, builtin.com)

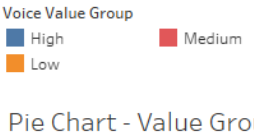
Rule Based Learning:



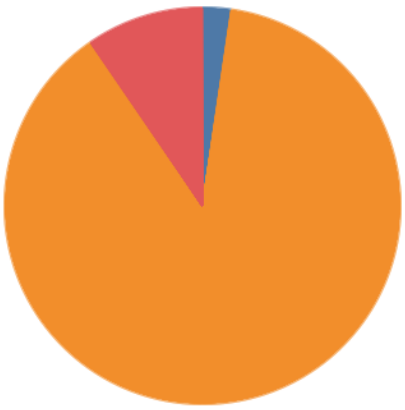
This was the clustering outcome for our rule-based learning method that I feel did not yield a good result a large majority of our customers are in groups 1 and 4 which based on the metrics we used of gender, age, and voice usage doesn't necessarily represent their value. There are many other variables to consider that are not easily added compared to the K-means method.



We created our segments based on the distribution of average monthly voice use, but as you can see under 25 is the bulk of our customers so with these segments we are unable to get a good understanding of customers' value.



Pie Chart - Value Groups



Seen in this pie chart you can see that the low segment greatly outweighs the other two segments thus making it impossible to make intelligent business decisions from this data.

Things I would do differently: The biggest thing I would do differently is incorporate different types of data. I primarily used numeric data for both of my methods, I realize there are a lot of potential reads in variables like ‘VM’, ‘Pager’, ‘Internet’, etc. that I did not leverage.

Source:

1. https://www.techtarget.com/searchenterpriseai/feature/How-to-choose-between-a-rules-based-vs-machine-learning-system

Foster, E. (2023, August 2). *Choosing between a rule-based vs. Machine Learning System: TechTarget*. Enterprise AI. https://www.techtarget.com/searchenterpriseai/feature/How-to-choose-between-a-rules-based-vs-machine-learning-system

2. https://www.geeksforgeeks.org/k-means-clustering-introduction/

GeeksforGeeks. (2023, August 25). *K means clustering - introduction*. GeeksforGeeks. https://www.geeksforgeeks.org/k-means-clustering-introduction/

3. https://builtin.com/data-science/elbow-method

Stop using elbow method in k-means clustering. Built In. (n.d.). https://builtin.com/data-science/elbow-method