Segmentation and Profiling

Nelson Tran

Master of Data Science, Merrimack College

DSE5004 Visual Data Exploration

Dr. Michael Dupin

May 3rd, 2024

Summary

In this report, ruled-based segmentation and unsupervised modeling using K-means clustering will be compared to develop customer segmentation that can support effective, and economically sound customer retention efforts. The comparison between the two techniques will come from the number of segments produced, cross-segment differences, segment-specific profiles, and segment performance and value. There will also be recommendations and considerations after the comparison of the two. Before anything was done for segmentation, the data had to be reviewed and cleaned up for ease of use with clustering, scaling, etc. Cleaning the data includes encoding variables to numeric values or removing the dollar signs which would make those numbers "characters."

The variable used for ruled-based segmentation is the "Total_Debt" variable I created by adding credit card debt and other debts into one variable. After I previewed the data, I thought that the best way to separate this variable would be using quartiles that were divided into "low_debt," "med_debt," and "high_debt." Customers that are more engaged with the news would be more likely to be retained by telecommunications companies which became a factor when using rule-based segmentation. A customer would be of higher value if they are a news subscriber and lower value if they aren't a subscriber. Defaulting on a loan is also another rule used to segment the three groups more. This resulted in 6 segments that include "defaulted_high_value," "high_value," "defaulted_medium_value," "medium_value," "defaulted_low_value," and "low_value" with "def" indicating if a customer had defaulted on a loan before. If a customer cannot pay for a loan, then the chances of retaining them could be lower and potentially targeting these customers might be wasted efforts from the company. For unsupervised modeling using K-means clustering, they work on real-valued and non-binary variables and tend to be better with normalized or scaled variables. First, any categorical variables will be encoded so that they can be used in K-means clustering. For this clustering to work, the number of clusters had to be specified before and the "set.seed" function was used for reproducibility. Once the clustering is complete, the optimal number of clusters needs to be defined by investigating the variation changes as the number of segments increases. The optimal number of clusters is 5 with very minimal improvement afterward. Lastly, once the segments are complete from both ruled-based segmentation and unsupervised modeling, those segments are then added to the original data to be compared.

The 3 recommendations that are chosen are "defaulted_high_value," "high_value," and medium_value." These segments were picked because they all either have potential or overall value. None of the segments from the k-means clustering model were effective because when reviewing the data cluster 3 was the only segment that was noteworthy with the household income and total debt being particularly high for this cluster. One interesting thing I noticed with rule-based segmentation is that the higher the recency, frequency, or monetary (RFM) value a customer is, the older they are. I also noticed that customers who have defaulted on a loan before are on average much younger than their counterparts that have never defaulted on a loan. Total debt was the variable used with news subscribers as a segmentation because these customers tend to have more income available while being engaged.

This data can be used to target these customers as the recommended segments tend to be older individuals, who have more available income, engaged in the news. This allows the telecommunication company to tailor their marketing towards this demographic specified from the segments mentioned. There is one data limitation that may help improve these findings, which is the length of time a customer has had their debt. Additional information like ethnicity or social and cultural variables with the recommended segments could help get a better grasp of customer retention efforts.

1. Data Cleaning and Segmentation Techniques

For ruled-based segmentation, a variable "Total_Debt" was created from the sum of a customer's credit card debt and their other debt together. For k-means clustering, works on real-valued or non-binary variables. The data was then cleaned by removing the "X" variable that did not serve any purpose, replacing or removing any N/a values that may be in the data frame, and removing the "$" from any values that had it which would change this variable into an integer or numeric which would help in k-means clustering.

1.1 Ruled-Based Segmentation

The “Total_Debt” variable was created to aid in telecommunication companies’ retention efforts if they have enough available income to pay for such services. The variable was segmented in the following steps.

- Created a quantile for intervals .25, .50, and .75.
- Customers were segmented by these quantiles.
    - <.25 = low_debt
    - >.25 and <.75 = med_debt
    - >.75 = high_debt
- The lower the debt the higher value a customer would be:
    - Low_debt = high value
    - Med_debt = medium value
    - High_debt = low value
- Customers' values were affected by whether they were new subscribers or not. If they are engaged, then their potential/overall value increases while their value decreases if they are not.
- Lastly, the 3 segments are then split up by if a customer has ever defaulted on a loan. This helps determine if a client can pay their bills, which can help in identifying certain demographics that the company can target in their market to retain them.
- The resulting segments were then added back to the cleaned data for further data analysis.

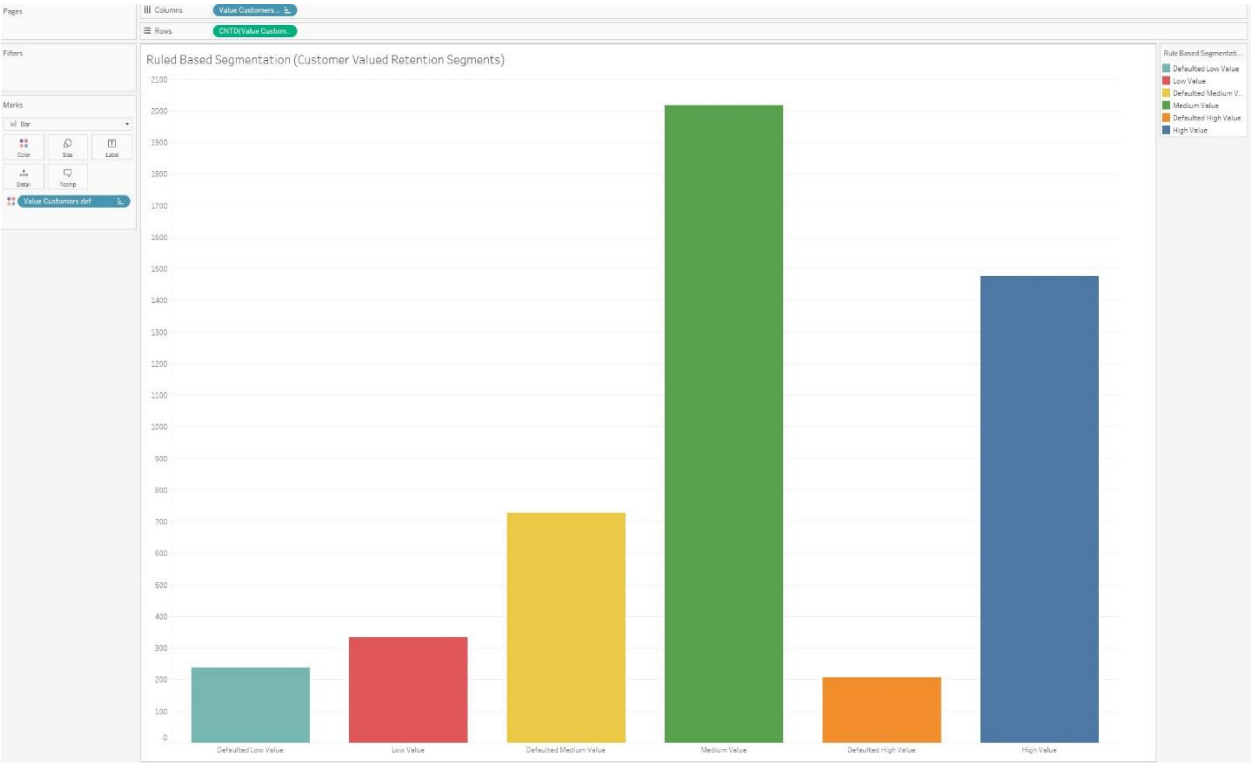

Figure 1: Distribution of the 6 resulting segments from ruled-based segmentation. From left to right are the segments. "Defaulted

Many of the customers would be in the "medium value" segments because this creates another value to base the results from. This segment can be an intermediary between high and low value to customers that a company can focus its attention on to retain. Patrons who are engaged in telecommunications are more likely to be retained, which increases their overall value.

1.2 Unsupervised Modeling (K-Means Clustering)

This methodology doesn't include choosing a variable that could eliminate bias and allow for the data to speak on its behalf. This technique typically works much better when using real-valued or non-binary variables. The steps taken to do this are:

- Encoding categorical variables "Votes," "UnionMember," "Gender," "JobCategory," "Retired," "LoanDefault," "MaritalStatus," "CarOwnership," "CarBrand," "PoliticalPartyMem," "CreditCard," "ActiveLiftstyle," "EquipmentaRental", "CallingCard," "WirelessData," "Multiline," "VM," "Pager," "Internet," "CallerID," "CallWait," "Callfoward," "ThreeWayCalling," "EBilling," "OwnsPC," "OwnsMobileDevice," "OwnsGameSystem," "OwnsFax," and "NewsSubcriber."
- Removing "$" from variables to change them into a numerical variable.
    - "HHIncome," "CarValue," "CardSpendMonth," "VoiceLastMonth," "VoiceOverTenure," "EquipmentLastMonth," "EquipmentOverTenure," "DataLastMonth," and "DataOverTenure."
- Any "N/A" values were converted to 0 or omitted entirely when appropriate.
- Data was scaled and used for k-means clustering.

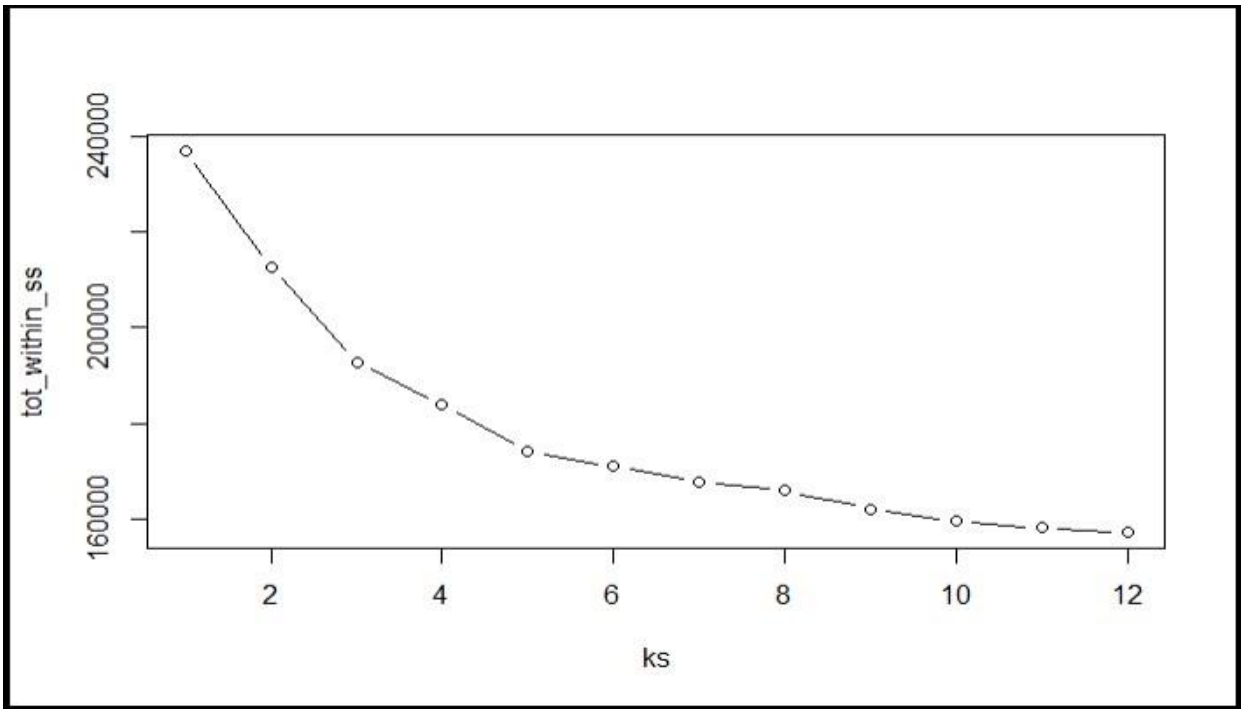

Figure 2: Finding the most optimal number of clusters. 5 is the most optimal number of clusters because as the number of clusters increases there is minimal improvement to the model.

- 5 is the most optimal number of clusters shown in the graph above.
- The model was run once more with 5 clusters specified and the resulting segments were added back to the non-scaled data for further data analysis.

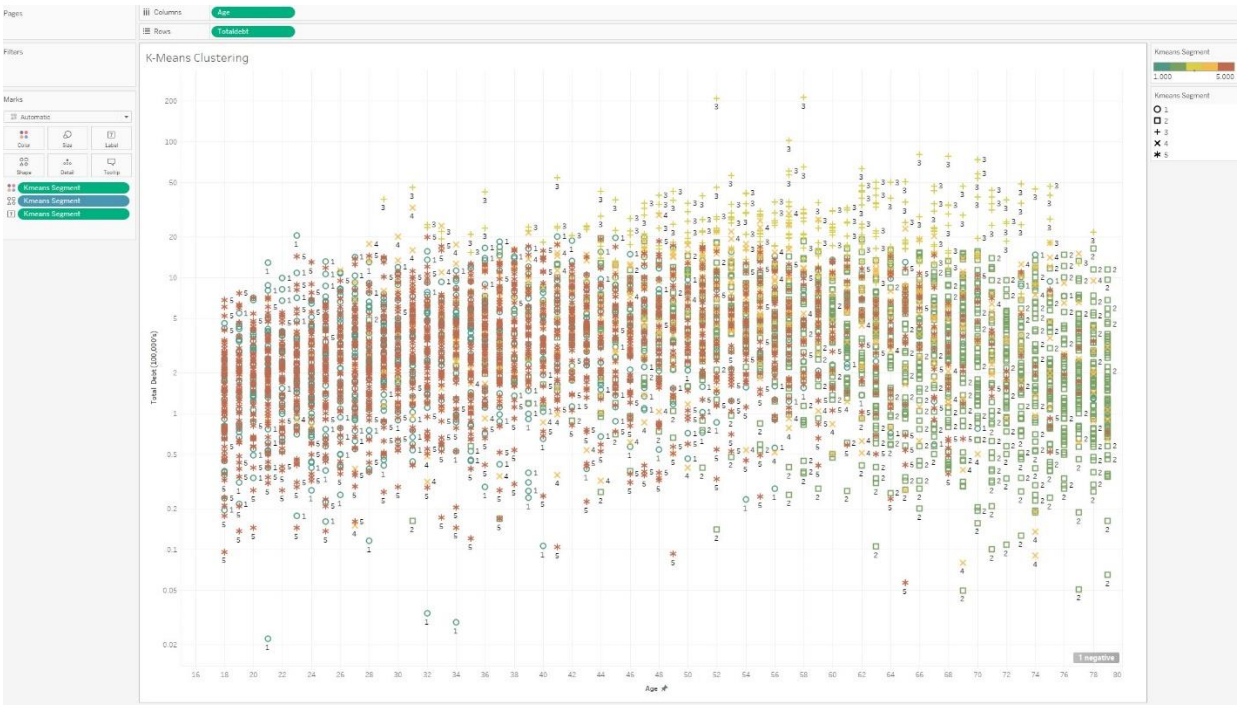

Figure 3: K-means clustering. In an attempt to better view this graph, the log was taken. Cluster 1 is circle-shaped and tends to be on the left side of the graph. Cluster 2 is square-shaped and is on the bottom right side of the

graph. Cluster 3 is a plus-sign which tends to be near the top portion of the graph. Cluster 4 is an "x" that is spread all over the bottom of the graph. Cluster 5 is an "*" that tends to be just about everywhere on the graph.

1.3 Comparisons and Recommendations

K-means Clustering
There are a few things that immediately jump out when looking at the 5 segments from k-means clustering. Segment 3 appears to have clusters that contain older individuals that have a higher total debt. This segment also has the highest household income which could contribute to the higher debt of a customer. Segment 1 tends to be younger customers that have less debt than other segments. The third thing that stands out is that aside from segment 3, there doesn't seem to be any segment that can be seen from the graph.

Ruled-Based Segmentation
When high, medium, and low-value customers were segmented by whether they defaulted on the loan or not created 3 interesting segments. For each tier of valued customers, the average age of patrons who have defaulted on a loan before is much younger than individuals who have not defaulted on a loan. This shows that older folks tend to have a better record for paying bills which improves their RFM value. This one piece of information can also help companies differentiate between customers focus their efforts on the retention of customers and keep their profits up. The higher valued customers also seemed on average older which could indicate that these customers are more reliable and could be the key when focusing on retention efforts.

The differences can be seen when you compare the two methods resulting in one technique that will help the company with customer retention efforts.

- Number of Segments:
  - *Ruled-based*: 6 segments.
    - "Defaulted high value," "high value," "defaulted medium value," "medium value," "defaulted low value," and "low value."
  - *K-means Clustering*: 5 segments.
- Cross-segment differences
  - *Ruled-based*: segment on total debt, news subscription, and whether a customer had defaulted on a loan before.
  - K-means Clustering: This model uses all variables that are scaled for a better fit for clustering.
- Segment Value and Potential
  - *Ruled-based*: Customers with lower debt have more money at their disposal which increases the chance of retention. A news subscription shows an engagement level with telecommunication companies that provide a higher retention rate. Defaulting on a loan can also have a big impact on high retention rates. This shows a history of missing payments that would lower the retention rate of a customer. This lowers the potential value of a customer but can lead to foresight in offering promotions that could hypothetically increase the retention rate.
  - *K*-means Clustering: Segments that cluster towards a lower total debt could have a higher retention rate. Some clusters that fit this description are clusters 1 and 2. On average these clusters show that their total debt is relatively low compared to other customers. The only difference between them is that Cluster 1 is younger while Cluster 2 is older. Retirement could play a big role showing that clusters 1 and 2 are high retention segments while the other segments are naturally low retention segments based on the total debt of the clusters. Based on the figure below, the other clusters can have high potential value too non-retired customers may want to keep up with the news leading to higher engagement for a telecommunications company.

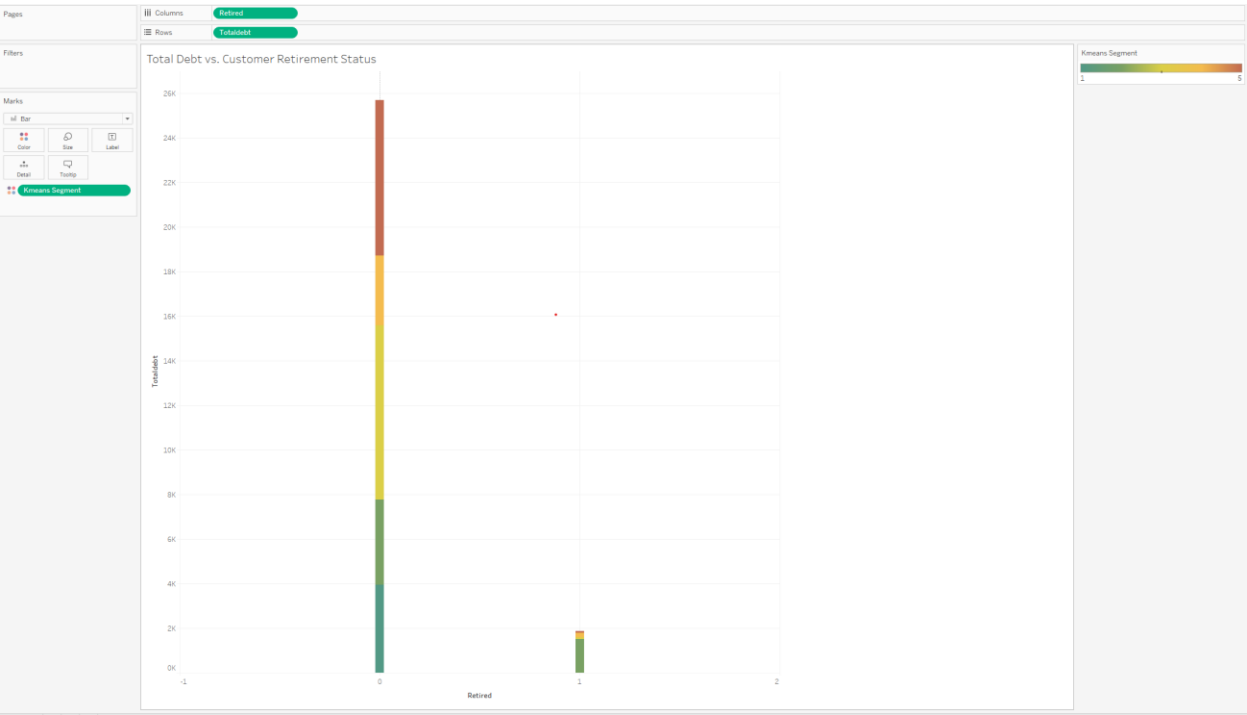

Figure 4: Total Debt vs. Customer Retirement Status. Almost if not all customers in the data are not retired which could increase the potential value of retention for each cluster.

Recommendations

The recommended solution to developing customer segmentation that can support effective, and economically sound, customer retention efforts is ruled-based segmentation. This technique produced more unique segments than the unsupervised k-means clustering did. This means that the company can concentrate on a certain demographic rather than a broad range of customers. This in turn can increase profitability and reduce costs from marketing and promotions. There is a better likelihood of these customers staying with the telecommunication company when their services are marketed towards them. Typically, after customers you can go after another unique segment that prioritizes other patrons. Once you retain a certain portion of your customer base research and development into advertising can be kept at a minimum, which will also reduce costs for companies.

2. Conclusions and Goals for the Business

Ruled-based segmentation is the most recommended when considering methods for developing customer segmentation that can support effective, and economically sound, customer retention efforts. The goal of the business is to identify segments that have the highest overall or potential value that maximizes profit while minimizing the loss of resources. A few ways that a business can do this are engaging in targeted marketing and communication, focusing on product and service quality, and monitoring customer satisfaction and feedback. Using ruled-based segmentation allows for companies to expand customer retention which is crucial to sustaining business growth and profitability.

There are a few limitations to the data that could be addressed like additional values to segment from like ethnicity or social and cultural differences. One thing that constrained the total debt value created above for ruled-based segmentation is the length of the other debt for each customer. This could provide insight into nuanced segmentation that is not on the surface when looking at the data. Data sets that have more variables and information are always a benefit as they help distinguish customers more than having less data would. Variables that have nothing to do with the goal of a business can always be removed to further the quality of data. The quality of a data set will be a big factor as to how well data will perform in a model.