Capstone Project: Customer Segmentation

Final Submission

Executive Summary

- Customers can be segmented into 3 groups
- The key attributes for the 3 groups are income and family size, which customer profiles can be based on
- The more meat products and wine products a customer purchases, the more likely they are to be in the high income group
- Customers with the least amount of kids (family size) have the highest incomes
- Higher income groups tend to make more in store and catalog purchases, while lower income tend to make more website visits/purchases.
- The company marketing team can use these three groups to target their ads to the correct channels and the correct market.
- The marketing team can also personalize their communication efforts a lot better based of the group characteristics and behaviors, leading to higher ROI
- The final proposed model is K-Means at K=3 as it had the most clearly distinguishable clusters, can handle a large amount of data, and is cost-effective

Problem and Solution Summary

- The problem is we do not have segmented customers. This problem is crucial to solve as customer segmentation is a critical part of marketing operations with direct consequences on sales and marketing strategy.
    - Customer segmentations help marketers develop personalized communications and offerings that cater to each segment's interest.
    - In today's world, personalized customer interaction is highly preferred by customers. Overall, customer segmentation plays a vital role in optimizing ROI for marketing efforts.

- The goal is to figure out the best possible customer segments using the given customer dataset. This can be achieved using Unsupervised Learning techniques like Dimensionality Reduction and Clustering.

- After visualizing the data by applying T-SNE and PCA, and running clustering algorithms like Gaussian Mixture Model, DBSCAN, Hierarchical Clustering, K-Medoids, K-Means to cluster the data, we found K-Means to be the best clustering model in which customers can be segmented as it showed clear clusters.

- Grouping customers into these segments is a valid solution that will solve the problem of not having segmented customers by giving the Marketing team a lot more information, like group characteristics, customer profiles, and group purchase behaviors, to be able to optimize ROI.

Recommendations for Implementation

- I would suggest that the marketing team segment customers into 3 segments with income and family size being the key attributes.
  - For example, high income, middle income, low income, for income. For family size, 0 kids, 1 kid, 2+ kids.
  - Income and Family Size were the most notable attributes that impacted customer behaviors.
- I would suggest the marketing team runs ads through online/website campaigns for low income customers, and potentially mid as well as these customer visit the website that most
- For high income customers, in store campaigns/promotional ads and clear in store signage might improve purchase behavior the most as these customers tend to make more in store purchases
- High income customers also tend to make more catalog purchases. And with meat and wine being some of the biggest purchases for high income customers, I would recommend to make sure the catalog lists these products in the beginning of the catalog, and continues to call out meats and wines through the catalog
  - The catalog could potentially have deals for meat products and wine products, or bundle offerings.
- I would also recommend an online campaign/promotional efforts that help bring customers with larger family sizes into the store. Perhaps a call out on the website that says something like “Families of 4 or more get half off meat products every Sunday” or something to that nature.
  - This could help lower income customers to purchase products they otherwise wouldn’t, which can help drive up overall company revenue.
- Some key benefits of this solution would be higher customer engagement, less complaints, and more purchases, thereby enabling the marketing team to have a higher ROI.

- By segmenting customers into these 3 groups, the company will not only save money by advertising to the potential wrong channels and markets, but it will increase revenue by personalizing marketing campaigns to customers and encouraging them to make more purchases through clearer and more relatable ads, e-blasts, and overall online and in store communications.
- The potential risks involved with this solution are that the segments should be a different number, and that the grouping should be focused on other attributes in the data, or other attributes of data not collected.
    - However, this shouldn't be considered a heavy risk, as the customers are now more segmented than they were before and therefore revenue and customer experience should be higher due to personalized ad campaigns and a better overall understanding of the customers.
- If anything, we can continue to collect data, and run an analysis and segmenting models again to see if anything has changed and we can refine our marketing efforts again.
- Further analysis needs to be done on future customer behavior, and responses to ad campaigns to make sure things are running smoothly. Therefore, collection of data remains crucial for confirmation of the correct customer segments and for future company benefits.

Capstone Project: Customer Segmentation

Milestone Submission

Problem Definition

The problem is we do not have segmented customers. This problem is crucial to solve as customer segmentation is a critical part of marketing operations with direct consequences on sales and marketing strategy. Research proves that customer segmentation often has a huge impact on people's email engagement as segmented campaigns see over 100% more clicks than non-segmented campaigns. Email marketers also have reported 6-7 times growth in their overall revenue. Another key note is that customer segmentations help marketers develop personalized communications and offerings that cater to each segment's interest. In today's world, this type of customer interaction is highly preferred by customers. Overall, customer segmentation plays a vital role in optimizing ROI for marketing efforts.

The goal is to figure out the best possible customer segments using the given customer dataset. This can be achieved using Unsupervised Learning techniques like Dimensionality Reduction and Clustering.

There are some key questions that need to be answered. Which features are best applicable in order to segment customers effectively, or what is the feature data telling us? Are there any missing values in the data? Which features are positively or negatively correlated? What do these correlations mean in terms of solving our problem? Do the features need to be scaled? Which clustering algorithm is the best and why? What is the optimum amount of customer segments? What are the important characteristics of these segments, or clusters, that will detail the customer profiles?

Given the customer dataset, data science can help us create customer profiles from visualizing and analyzing the customer data set into segments. After visualizing the data by applying T-SNE and PCA, we can use a combination of clustering algorithms like Gaussian Mixture Model, DBSCAN, Hierarchical Clustering, K-Medoids, K-Means to cluster the data. After running each algorithm we can analyze which one segments the data best by weighing out the feasibility and challenges verse the likely benefits and solutions.

Data Exploration

The data provided contains 26 features (or columns) [this is after dropping the ID column], and 2,240 customers (or rows) . It includes several features about each customer like year of birth, education level, marital status, number of kids at home, income, etc. It also includes information on the customer's engagements and interactions with the company. For example, number of days since last purchase, amount spent on different food products in the last two years, different channels of purchases like instore or through the company website, and direct feedback from how the customer has engaged with the first through the fifth and last campaign. Lastly it also includes if the customer complained in the last two years. One important thing to note is the timeframe only goes back to the last 2 years so the data is recent. Of these features almost all data types are integers, with Education, Marital Status, Date of customer's enrollment with company being objects, and Income being a float. Only income is missing 24 values.

After exploring the summary statistics of numerical values it was interesting to see how each customer has 2 kids max at home, with many having none or 1. These customers spend the most amount of money on wines, and meat products. Instore purchases are higher than catalog purchases, with an average of around 5 website visits per month, which is a decent amount for a customer. It also looks like campaign 2 was least effective, with campaigns 3-5 being strongest.

After exploring the summary statistics of categorical values, and fixing some values, there are 4 unique Education values, 5 unique Marital Statuses, 3 unique Kids home values (0,1,2), 3 unique Teens home values (0,1,2), and 2 unique Complain values.

Univariate Analysis

After performing univariate analysis on the numerical data…

- Income (after outliers were dropped, and missing values were filled in with median incomes) has a somewhat normal distribution with peaks at $40,000 and around $65,000
- The amount spent on wines, fruits, meat products, fish products, sweet products, and gold products are all heavily skewed right. Once again overall customers spend more on wines and meat products

After performing univariate analysis on the categorical data…

- The majority of customers are married or together (around 1400 customers)

- The majority of customers are graduates, I'm assuming college, as basic education seems likely to be a high school graduate. There are also a good amount of customers with Masters and PhDs.
- The bar plots for Kidhome, Teenhome, and Complain generated strange bar graphs that were all uniform, although looking at the data that doesn't seem to be the case. I'll have to reassess this later.

In general the univariate analysis might hint at where customers could likely be segmented. For example, perhaps customers might be clustered with different incomes, education levels, and marital statuses, as these were much clearer to see differences in.

Bivariate Analysis

After performing Bivariate analysis on the numerical data…

Negatively Correlated Features:

- Income and number of website visits
- Income and number of kids at home
- Amount spend on wines and number of kids home
- Number of Catalog Purchases and number of kids home
- Number of Store Purchases and number of kids home

Positively Correlated Features:

- Income and amount spent on wines
- Income and amount spent on meat products
- Income and number of Catalog Purchases
- Income and number of Store Purchases
- Amount spent on Wines and (1) Number of Catalog Purchases (2) Number of Store Purchases
- Amount spent on Meat products and number of catalog purchases

After performing Bivariate analysis on the categorical data…

- The more education a customer has the more income they earn which makes sense
- Marital Status didn't impact income much. Only noteworthy observation being that widows have more income.
- Customers with 0 kids have more income, which makes sense

In general, the bivariate analysis shows income plays a huge role in customer behavior. And amount spent on Wines and Meat Products play a big role as well, impacting catalog and store purchases greatly.

Data Treatments / Pre-Processing Steps

Feature Engineering

Dropped improbable age data
After dropping some potentially false Age data, a histogram shows us the age ranges from 20 to around 75, in a somewhat normal distribution with peaks at around 40 and 45.

Created a new variable "Family Size"
After doing some work to combine Kidhome and Teenhome ("Kids"), and Marital Status ("Status"), I was able to derive a new variable Family Size.

Created a new variable "Expenses"
I combined all the amounts spent on different food products into one new variable named Expenses

Created a new variables "NumbTotalPurchases"
After combining to number of purchases on different channels, I was able to create a variable for the total number of purchases

Created a new feature "Engaged in Days" to indicate how long a customer has been with the company.

Created a new feature called "TotalAcceptedCmp" that shows how many offers customers have accepted.

Created a new feature called "AmountPerPurchase" indicating the amount spent per purchase.
The mean amount spent per purchase was $33 dollars. This graph was also heavily right-skewed.

Overall, after reviewing a scatter plot, not surprisingly the more income a customer has, the more they spend. Also the bigger a family size the lower income a customer has, with a slight upward trend around 5. In general, age, family size, expenses, and income, not surprisingly all play an important role in customer behavior.

Lastly, before building the model, I dropped all unnecessary columns - demographic attributes - like Year of Birth, Day, Complain, Status, Marital_status, etc. This is to focus the customer segmentation solely on behavioral attributes. And then I scaled the data to ensure the features all have a similar range in order to better analyze and develop models. This is especially important for models that compute a distance metric like K-Means.

Dimensionality Reduction

T-SNE

I fit the data with T-SNE with the number of components equal to 2 to visualize how the data is distributed.There weren’t very clear clusters, but I could definitely see a pattern of potentially a lot of different clusters developing.

PCA

Next I applied PCA to reduce the amount of multicollinearity between variables, as high multicollinearity results in poor models.

Building Models

I used the algorithms K-Means, K-Medoids, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Model to see which model performed the best on the data.

K-Means
3 received the best silhouette score at .27, and therefore went with K=3.
Of the 3 clusters, 0 had a count of 1035, 1 had 569, and 2 had 622 values. This lets me know 0 is the largest cluster, and maybe K-Means isn't the best clustering model for this data as it's unevenly distributed. It could also mean that potentially the 0 cluster simply will have more customers. However, after visualizing K-Means with K=3 with PCA, it had very clearly distinguished clusters. This could potentially be the best model.

After cluster profiling with K-Means, cluster 1 seems like the highest income cluster, cluster 2 is in the middle, and cluster 0 has the lowest income - based on their purchase and overall behavior. After viewing boxplots it not only confirmed this, but also showed cluster 1 has the least amount of kids, highest expenses, and purchases in store the most. It's mainly based on income and not much more, so I tried K=5.

With K=5, the data was a little more evenly balanced, but not much better, as cluster 0 still had a lot more values than clusters 1-4, with 984 values and no other cluster reaching 500. After plotting these clusters, this doesn't seem like the best option either as the visualization didn't show the clusters existing in completely separated areas, but perhaps better than K=3 as the data was more interesting and spread out. For example, after highlighting the max values between clusters, it looks like clusters 1, 3, and 4 have the highest income, while cluster 0 has the least, and 1 just above 0. The income groups did have different age groups and preferences not evident in K=3. Therefore, for K-Means we'll choose K=5

K-Medoids
With K=5 the silhouette score of 0.12 shows a weak structure or overlapping clusters. When checking the distribution it's slightly better than K-Means, but still cluster 0 has by far the most value, with nearly 200 more values than the next closest cluster - cluster 4.

After creating a PCA plot the clusters 1, 2, and 3 are poorly clustered and are all overlapping each other.

After highlighting the highest average values throughout the clusters it looks like cluster 0 has the highest income, while cluster 2 has the lowest income. What's reassuring is that the location of the highest income cluster (cluster 0) in K Medoids is similar to the location of the highest cluster in K-Means with K=5, and mainly with K=3. What's concerning is all the overlap in the clustering. After visualizing boxplots of all the variables, it confirmed cluster 0 had the highest income, the most expenses, and the smallest family size - which has been a trend. In store purchases continued to be the highest purchase channel for all clusters, with cluster 0 leading the way.

List the most meaningful insights from the model relevant to the problem
A meaningful insight has 3 components

- Good interpretation of the output from the model
- Potential reason for that output
- What is means for the problem/business

Hierarchical Clustering
When running Hierarchical Clustering, first I found the cophenetic correlation for different distances with different distance methods. The highest cophenetic correlation was was 0.83

Cityblock
While Complete linkage had the most compact clusters, Average Linkage had the highest cophenetic correlation score with 0.83. After locating the best space for a horizontal line, it looks like 3 clusters is best for this method

Chebychev
Single Linkage, Complete Linkage, and Average Linkage all look like 2 clusters are the optimum amount.

Mahalanobis
The Single Linkage, Complete Linkage, and Average Linkage all looked very different from the Chebychev method. This time giving us 3 clusters as the best result for all three linkages. Single Linkage and Average Linkage looked very close to each other.

Euclidean
The clusters from the Single Linkage, Complete, Linkage, and Average Linkage all were not very clear.

After visualizing with PCA the clusters 1, 2, 3 still look jumbled up with no clear separated clusters. Clusters 0 and 4 are separated well.

Cluster 0 came out to be the highest income with cluster 2 having the least income. Overall it seems the clusters are slightly more balanced than K-Means and K-Medoids. Cluster 4 had the second highest income. In store purchases continued to be the highest channel of purchases. Something interesting of note was that cluster 0 had the highest amount of in store purchases as well as Catalog purchases, however it had the lowest amount of website visits per month.

DBSCAN

When running DSCAN it seems like it is the most evenly distributed clustering method so far. While cluster 0 still has a large amount of more values than the rest of the clusters 1-4, 1-4 are much more evenly distributed. At this point, it seems like potentially cluster 0 might simply cover a wider amount of customers.

After visualizing the clusters with PCA, it's still frustrating to find the clusters not clearly separated. 0 and 4 are clearly separated, but 1, 2, and 3 are still overlapping. The only slight positive is it seems with DBSCAN they're slightly less overlapping.

Gaussian Mixture Model

When running a Gaussian Mixture model, the silhouette score was low at .09. However it was distributed exactly the same as DBSCAN with fairly well distributed clusters, other than 0 with around 200+ more values than the rest.

Comparison of Techniques & Their Performances

After going through all the different clustering techniques, K-Means with k=3 seemed to be the best technique. It had the highest Silhouette score, which is the best metric for clusters by providing a measurement of how similar each data point is to its own cluster.

K-Means also had the most clearly defined clusters after running PCA. It was the only technique that didn't produce overlapping clusters.

K-Means is also computationally fast and therefore a cheaper option. It can also handle large data sets fast because of its simplicity.

Furthermore K-Means is easy to explain to stakeholders and easy to interpret. While it can be sensitive to outliers and scaling issues, overall it achieved the best results.

-

There is definitely opportunity to improve the performance however. Perhaps with stronger parameters as the Silhouette score could be a lot higher, and the clusters could be separated better.

Proposal for the Final Design Solution

My proposal for the Final Design Solution is to adopt a K-Means model based on its results compared to the others. As stated above, while it can be improved, it had the highest silhouette score and the most clearly defined clusters. It’s also very easy to interpret and performs well with large data sets due to its simplicity.

K-Means will find the best clusters from the data. These clusters will allow the marketing team to clearly segment customers, which will improve marketing and advertising efforts and result in a better marketing ROI. Customer profiles can be clearly defined from the clusters, giving the marketing team clear characteristics for each customer profile. Because of this, communication efforts and campaigns can be highly personalized, leading to much better results.