MIT Applied Data Science Program

Marketing Campaign Customer Segmentation

September 30, 2022

Kristi Yamashita

Refined Insights

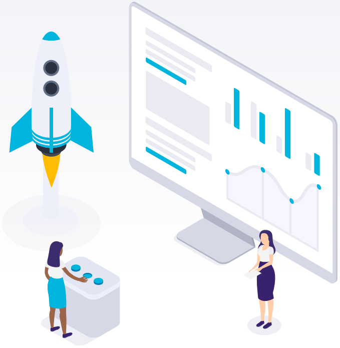
- Drop demographic attributes and other variables that won't help with segmentation
  - Year_Birth, "Dt_Customer, day, Complain, Response, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Marital_Status, Status, Kids, Education, Kidhome, Teenhome, Income, Age, Family_Size
- There is a strong positive correlation between expenses and amount spent on wine (0.90) and meat products (0.85) in the past 2 years. This may suggest that these are more expensive products that greatly contributes to a customer's overall expenses.
- There is a strong positive correlation between number of total purchases and number of purchases made through the company's website (0.79) or directly in stores (0.83).
- There is also a strong positive correlation between number of purchases made using a catalog and expenses (0.80).
- The negative correlation between number of visits to the company's website in the last month and number of purchases made using a catalog or directly in stores makes sense because if customers are using these channels to make purchases, they probably don't have to visit the company's website.
- In the elbow plot, the elbow is seen for K=3 and K=5 as there is some drop in distortion at K=3 and K=5. We tested the K-Means algorithm using both of these k values to see which provides better results.
- Using the K-Means algorithm with K=3, we seem to get 3 distinct clusters including low-income, middle-income, and high-income customers. We get deeper insights into different types of customers by using K=5.
  - Cluster 0: Middle/Low-Income Customers with Kids
  - Cluster 1: High-Income Customers who don't have kids and spend a more modestly
  - Cluster 2: Middle-Income Customers
  - Cluster 3: Low-Income Customers
  - Cluster 4: High-Income Customers who don't have kids and spend a lot

Comparison of Various Techniques

Algorithm	Best Solution	Silhouette Score	Remarks
K-Means	K = 5	0.13021374284739756	In the elbow plot, the elbow is seen for K=3 and K=5. We used both of these K values to see which provides better results. We get deeper insights into different types of customers using K=5.
K-Medoids	K = 5	0.107528069592116	Gives us similar clusters to K-Means algorithm.

Algorithm	Best Solution	Silhouette Score	Remarks
Hierarchical Clustering	● Cityblock distance and average linkage ● K = 3	n/a	The cophenetic coefficient has an average of 0.736922421650532 and highest coefficient of 0.8671371105092277, which means it is a pretty good measure of how faithful the dendrogram preserves the pairwise distance between the original unmodeled data points.
DBSCAN	Eps = 3 and min sample = 20	0.3398851568849134	Highest silhouette average is 0.3398851568849134 for eps = 3 and min sample = 20.
Gaussian Mixture Model	K = 5	0.14344403792681099	Gives us similar clusters to K-Means and K-Medoids algorithm.

Proposal for the Final Solution Design



DBSCAN

- This algorithm has the highest silhouette score compared to the other algorithms used.

Algorithm	Best Solution	Silhouette Score	Remarks
DBSCAN	Eps = 3 and min sample = 20	0.3398851568849134	Highest silhouette average is 0.3398851568849134 for eps = 3 and min sample = 20.

THANKS!

Any questions?

You can find me at:

- kristi.yamashita@yum.com
- www.linkedin.com/in/kristiyamashita