GMMA 869
Machine Learning and AI

Dr. Stephen W. Thomas

Individual Assignment 1
July 5, 2020

William Chak Lim Chan

How Lovely!

Preamble

Download the "customer" dataset: *jewelry_customers.csv*.

Scenario

You work at a local jewelry store. You've recently been promoted and the store owner asked you to better understand your customers. Using some sneaky magic (and the help of Environics!), you've managed to collect some useful features for a subset of your customers: age, income, spending score, and savings (i.e., how much savings they have in their bank account). Use these features to segment your customers and create customer *personas*.

Tasks

1. [Code] Perform a clustering analysis of the dataset.
    a. Load, clean, and preprocess the data as you find necessary.
    b. Cluster the data using any clustering algorithm discussed in class. Measure goodness-of-fit. Try different values of hyper parameters to see how they affect goodness-of-fit.
    c. Print summary statistics for each cluster.
2. [Text] What do you think the best hyper parameter values are? Why?

The best hyper parameters for the K-Means algorithm on this dataset appear to be when K = 5. This is evidenced by a low WCSS and high silhouette score, suggesting a tight and compact cluster size that has instances that are closer to its own cluster. The K = n values below 5 produce a high WCSS (clusters that are spread out) and acceptable silhouette scores. The K = n values above 5 produce low WCSS (tight clusters) but low silhouette scores (instances that are closer to another cluster than its own).
K=3, WCSS=384.81, Sil=0.70
K=4, WCSS=189.70, Sil=0.76
K=5, WCSS=66.51, Sil=0.80
K=6, WCSS=61.13, Sil=0.63
K=7, WCSS=56.83, Sil=0.44
K=8, WCSS=52.91, Sil=0.30

3. [Text] Describe and interpret the clusters with words. That is, create personas.

Cluster	Age	Income	SpendingScore	Savings
0	33	$105,266	0.31	$14,963
1	60	$72,448	0.77	$6,890
2	88	$27,866	0.33	$16,659
3	24	$128,029	0.90	$4,087
4	86	$119,994	0.06	$14,809

Cluster 0 are one of the prime target customer groups. They're young, have high earning power, high savings, and moderate spend habits. They are in a stage of life where gifting expensive jewelry for a special occasion is within their means and on the cards. Given their spending habits but despite their large earning power, they are selective with their purchases, focusing on quality.

Cluster 1 are customers who are entering the retirement phase, or who have recently retired. They have moderate earning power and low savings, driven by their lofty spending habits. This group could treat themselves to some jewelry to celebrate a successful career. Given the excitement of retirement around this age, a reward purchase is not out of the question for this group.

Cluster 2 are elderlies who have saved up enough for a retirement within their means. They are the oldest of the five cohorts and are nearing the end stages of life. This cluster is a non-target given their income, spending habits and savings.

Cluster 3 represents the second of our prime target customer groups. The youngest of our clusters, these young adults have just entered the workforce and command huge salaries given their age (probably investment bankers or MBB consultants....). Their low savings are marked by their lavish spending habits which make them prime targets for expensive jewelry. Not only are they impulsive purchasers, they impulse purchase expensive things, which translates to $$$ for us.

Cluster 4 are the elderly frugal spenders. With high balances in their chequing and savings accounts, these folks, despite having all the money in the world, are hunting for a more than bargain. Nearing the 100 year mark, an expensive jewelry purchase is the last thing on their mind.

4. [Text] How good are your results? What could you do to make them better?

The results given by K-Means from a technical score standpoint is the best given the number of K's available. I could also try running this algorithm using DBSCAN to cross-validate the number of clusters and pick the algorithm and cluster number based on business knowledge. I would have to tune the MinPts and Eps to generate the optimal scores and derive the correct number of clusters.