Assignment – Part II

Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(what EDA you performed, which type of Clustering produced a better result and so on)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Solution:

- Our main objective for this assignment is to find the countries that are in direst need of aid. Our job is to find those countries using socio-economic and heath factors which will show overall development of the country.
- So, in order to do so first we analyse the dataset – It has total 167 countries with no missing values or duplicate country records. All of the columns except country are numeric continuous variables.
- Then we visualize the data by plotting the pair plot to check if any patterns in the data
- Then we plot the correlation matrix and heatmap between variables to check the multicollinearity and found some of the variables are having multicollinearity. Also we identified some of the redundant feature-pairs from the correlation values.
- Based on business knowledge we were able to remove the redundant features and retain important features for our analysis.
- Since k-means method is affected by outliers, we first identified the features having outliers using boxplot. Almost all features have outliers. Then we capped these values to 5-95% Quartile values for k-means to perform correctly on the other values. We will later use the original values for final analysis.
- Then we identified the optimal no of clusters(k) using several elbow methods. We find that visually k=3 shows better distribution of data, and k=4 defines more correct average values. By calculation k=2 showed the best values, but only 2 segments didn't make any sense. We performed few iterations on of K-Means, we could see that using 3 Clusters provided a better output in terms of a balanced cluster size. So we considered the 'K-Means with 3 Clusters' as our FINAL MODEL
- Similarly, we performed no of iterations on hierarchical tree(using complete and single linkage methods). We found better results with complete linkage method. We identified that optimum results were at k=3.
- In the final cluster by observing the countries we identified the clusters into Developed, Underdeveloped and Developing.
- In the Underdeveloped cluster we got 49 countries.
- We selected top 10 countries in direst need of aid. We did this by comparing all the GDPP, Income and child_mort of all the countries within the Underdeveloped cluster with the mean GDPP, mean Income and mean child_mort rates.

- Since we had initially capped the values, now we manually analysed the other factors for the selected countries on the original data and confirmed that they really require external support and funding according to our criteria mentioned in the problem statement.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

K-Means Clustering	Heirarchical Clustering
We need to have desired number of clusters ahead of time.	We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights
It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.	Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
Works very good in large dataset	Works well in small dataset and not good with large dataset
The main drawback of k-Means is it doesn't evaluate properly outliers.	Outliers are properly explained in hierarchical clustering
K-means only used for numerical.	Hierarchical clustering is used when we have variety of data as it doesn't require to calculate any distance.

b) Briefly explain the steps of the K-means clustering algorithm.

Step 1: Randomly select K points as initial centroids.
Step 2: All the data points closest to the centroid will create a cluster center by finding the Euclidean distance between the datapoints and centroids.
Step 3: Once we assign all the points to each of k clusters, we need to update the cluster centers or centroid of that cluster created.
Step 4: Repeat 2,3 steps until cluster centers reach convergence.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

For any unsupervised algorithm we have to determine the optimal number of clusters into which the data may be grouped. The optimal values of k are subjective to the problem and dataset, and there are many ways to do find optimal k.

Some of them are statistical methods like comparing distances/weights of some kind to group similar/different items, while some compare evidences against null hypothesis. These

are often used as complementary evaluation techniques rather than one being preferred over the other. For eg:

1. SSE

- When you plot SSE as a function of the number of clusters, notice that SSE continues to decrease as you increase k.
- As more centroids are added, the distance from each point to its closest centroid will decrease.
- There's a sweet spot where the SSE curve starts to bend known as the elbow point.
- The x-value of this point is thought to be a reasonable trade-off between error and number of clusters.
- To perform the elbow method, run several k-means, increment k with each iteration, and record the SSE
- In this example, the elbow is located at x=3 or x=4 or x=5:

2. Elbow Curve to get the right number of Clusters

- One of the most popular methods
- Two types :
  - Distortion:
    It is calculated as the average of the squared distances
    from the cluster centers of the respective clusters.
    Typically, the Euclidean distance metric is used.
  - Inertia:
    It is the sum of squared distances of samples to their closest cluster center.
    Inertia is the sum of squared error for each cluster.
    Therefore the smaller the inertia the denser the cluster(closer together all the points are)

- To determine the optimal number of clusters, we have to select the value of k at the "elbow" ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 3, 4 or 5.

3. Silhouette Analysis

- Formula:
silhouette score=(p−q)/max(p,q)
where,
p = is the mean distance to the points in the nearest cluster that the data point is not a part of
q = is the mean intra-cluster distance to all the points in its own cluster.

- A higher silhouette coefficient suggests better clusters
- The silhouette coefficient is a measure of cluster cohesion and separation. It quantifies how well a data point fits into its assigned cluster based on two factors:
  - How close the data point is to other points in the cluster
  - How far away the data point is from points in other clusters
- Silhouette coefficient values range between -1 and 1. Larger numbers indicate that samples are closer to their clusters than they are to other clusters. The value of the silhouette score range lies between -1 to 1.
  - A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
  - A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

- In the scikit-learn implementation of the silhouette coefficient, the average silhouette coefficient of all the samples is summarized into one score.
- The silhouette score() function needs a minimum of two clusters, or it will raise an exception.
- A list holds the silhouette coefficients for each k
- Plotting the average silhouette scores for each k shows that the best choice for k is 3 since it has the maximum score:

Our approach:

- The maximum scores are for k=2.
- However in real-world business it makes no sense to have only 2 clusters to represent the entire data.
- The elbow extends from k=3 to k=5, and then it starts to flatten out.
- We will check k-means for k=3,4,5 values

d) Explain the necessity for scaling/standardisation before performing Clustering.

- If we check the min, max, mean and median values in the columns there is lot of variation in the mean values of the columns.
- This may be because data in each col is measured in different units of measurement.
- However most ML algorithms would consider one feature like income more important than child_mort only because the values for income are larger and have higher variability from country to country.
- As k-means algorithm computes the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. This is because the ML algorithms internally use SVD to compute the principal components and assume that the data is scaled and centred.

- Hence Feature Scaling is required so that all the cols are brought down in same range so that they can be compared with each other.
- There are two common ways of rescaling:

1. Normalizing/Min-Max scaling :
    - - (value- min)/(max-min)
    - - It will convert attribute between range of 0 to 1 like probability.
    - - Scikitlearn provides MinMaxScalar class for this
    - - Normalization is not needed for One-Hot encoded features because they are already in the range between 0 to So, normalization would not affect their value. So we do not do it for dummy variables.

2. Standardizing :
    - - (value-mean)/std
    - - For Gaussian distribution, it gives standard normal random variable
    - - It will converts attribute so that its mean becomes 0 and its standard deviation becomes 1. In other words, it centers the variable at zero and variance at 1.
    - - Standardizing the One-Hot encoded features would mean assigning a distribution to categorical features. So we do not do it for dummy variables.
    - - Standardizing is preferred over Normalizing because if we make a mistake, then all the values will not be affected we are using the mean. But if we by mistake in any value, then min, max and (min-max) may change for Normalizing and it will affect all the values.
    - - Scikitlearn provides StandardScaler class for this

e) Explain the different linkages used in Hierarchical Clustering.

Linkage is a technique used in Agglomerative Clustering.
Linkage helps us to merge two data points into one using below linkage technique.
Single linkage: The distance between two clusters is calculated by the minimum distance between two points from each cluster.
Complete linkage: The distance between two clusters is calculated by the maximum distance between two points from each cluster.
Average linkage: The distance between two clusters is the average distance between every point of one cluster to the another every point of other cluster.
Ward linkage: The distance between clusters is calculated by the sum of squared differences with all clusters.