2022

Clustering – Graded Assignment

NGO HUMANITARIAN AID – CLUSTERING - ASSIGNMENT

BHAVESHKUMAR THAKER

UPGRAD – CHANDIGARH UNIVERSITY – M.SC. IN DATA SCIENCE

Assignment based Subjective Questions

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)

Answer 1:

Problem Statement

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. NGO have been able to raise around $10 million. The CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues CEO is facing is related to choosing the countries that are in the direst need of aid.

Business Objective

CEO of HELP International NGO wants to know the top 5 or 10 countries which are in direst need of aid. For that they want to build a Model which identifies the countries in need of Aid.

Analysis Approach

Following analysis approach was followed:

- Load and Clean the Data
  - o Check for Null values and convert percentage values to normal values
- Perform Exploratory Data Analysis (EDA)
  - o Perform Univariate analysis and Bivariate analysis to understand the data
- Data Preparation
  - o Standardize the data
- Assess Clustering Tendency
  - o Perform Hopkins' Statistic test on data to find clustering tendency of the data provided.
  - o **The score received on data was 90.92%. The data supports clustering.**
- Apply Principal Component Analysis (PCA)
  - o Make scree plot
    - ▪ **4 components were good enough to get 94% of variance in the data.**
  - o Apply PCA and identify number of components which doesn't have correlation between them
    - ▪ **4 Incremental PCA components were taken.**
    - ▪ **The Hopkins' Statistic score received on PCA data was 87.13%. The data supports clustering.**
- Perform Clustering using Kmeans

  - Find optimal number of clusters using Elbow Curve, and Silhouette Analysis
    - **Number of K (clusters) was taken as 3**
  - Perform KMeans clustering with optimal number of clusters
  - Identify Countries which are in the direst need of aid.
- Perform Clustering using Hierarchical Clustering
  - Perform complete linkage Hierarchical Clustering and cut the tree with optimal number of clusters
  - Identify Countries which are in the direst need of aid.

Question 2a: Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer 2a: Following are the differences between KMeans Clustering algorithm and Hierarchical Clustering algorithm.

- KMeans Clustering algorithm needs a prior knowledge of number of centroid (K) whereas Hierarchical Clustering algorithm do not need these kinds of parameters. cut_tree() function is used to create the number of clusters of any choice under hierarchical clustering.
- In KMeans clustering the algorithm will calculate the centroid each time at each iteration.
- KMeans clustering algorithm is fast compared to Hierarchical clustering algorithm.

Question 2b: Briefly explain the steps of the K-means clustering algorithm.

Answer 2b: Following are the steps following for KMeans Clustering algorithm:

- Perform Principal Component Analysis (PCA) if required.
    - PCA is useful in reducing the dimensionality.
- Identity number of clusters
    - Perform Elbow Curve analysis to find optimal number of clusters.
    - Perform Silhouette Score analysis to find optimal number of clusters.
- Execute KMeans algorithm
    - Provided identified number of clusters as input and perform fit to get the cluster labels assigned to the data.
    - Visualize the identified clusters with data.

Question 2c: How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Answer 2c: Following steps were following to get the optimal number of clusters:

- Identity number of clusters
    - Perform Elbow Curve analysis to find optimal number of clusters.
        - The Elbow Curve showed 3 optimal clusters.
    - Perform Silhouette Score analysis to find optimal number of clusters.
        - The Silhouette Score showed 3 optimal clusters.
- Silhouette Analysis

$$silhouette\ score = \frac{p - q}{\max(p, q)}$$

    - **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.
    - 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.
- Elbow Curve Analysis
    - The elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

Question 2d: Explain the necessity for scaling/standardization before performing Clustering.

Answer 2d: In statistics, standardization or data normalization or feature scaling refers to the process of rescaling the values of the variables in data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When working with data where each variable means something different the fields are not directly comparable. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

Question 2e: Explain the different linkages used in Hierarchical Clustering.

Answer 2e: Following linkages are used in Hierarchical Clustering:

- **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.
- **Complete Linkage:** For two clusters R and S, the complete linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.
- **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.
- **Centroid linkage:** It returns the distance between centroid of Clusters.
- **Ward linkage:** The linkage function specifying the distance between two clusters is computed as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster. Ward's Method seeks to choose the successive clustering steps so as to minimize the increase in ESS at each step.