BA 820: Unsupervised Machine Learning

Aditya Sinha

U18474952

Assignment #1 – Executive Summary Clustering Forum message

**Overview:**

Upon receiving the dataset, we first conducted data exploration. During this time, we found empty space in text field and duplicated values for features that could be important to our analysis. Therefore, we took several approaches to clean the dataset before proceeding to the next step. Then we performed cluster analysis on the clean dataset and discovered that the data could potentially be segmented into around 4 like-groups. Lastly, we provided the client with some insights on the clusters and provided recommendations based on our findings.

**Exploratory Data Analysis, Data Cleaning, and Preprocessing:**

Our dataset consists of 2362 rows and 301 columns. Each row represents an individual message, and the columns are mostly details about the message. 300 columns are float values and 1 of column is the textual data- text.

We ran a check to see if there are any identical columns and we didn’t find any identical columns. After that we stripped off any terminal spaces in the text column and then replaced empty string with null value. After that we dropped the rows with null value in the text column. There were 58 duplicate values which we also dropped. Since there are no duplicate and null values left in text column, we set the unique left column as the indices.

Next, we used ProfileReport from pandas\_profiling package to generate a summary report of all columns to find any missing values.

When we completed the data cleaning process, we managed to keep over 97% of the data. There are now 2303 messages, 301 columns, and no missing values. The last step to take before we conduct cluster analysis is to standardize the data. All the columns of this data record have similar measurement and the summary statistics between the columns (such as min, max and mean) are similar to one another, therefore we don’t need to scale the data.

**Hierarchical Clustering and K-Means Clustering:**

Since the goal is to segment stocks into like-groups, we decided to try out two clustering techniques-hierarchical clustering and K-Means clustering – and examine which one yields better results.

1. We used correlation between the distance matrix and the cophenetic distance to decide the clustering method and metric. Then used a dendrogram to visualize the hierarchical relationship and determine the number of clusters.
2. We recorded the inertia and the average silhouette score for different numbers of clusters before conducting K-Means clustering.

The graph on the left demonstrates how the inertia decreases as the number of clusters K increases, while the graph on the right shows how the average silhouette score decreases as K increases. From these two graphs, we could see that the optimal number of clusters is also around 3 or 4, where the inertia is low enough (but does not converge yet), and the average silhouette score is still high.

1. We further inspected the silhouette score for each message.

We could see from the graphs below that the majority of messages have a positive silhouette score, indicating that they are similar to the other messages in their assigned clusters.

1. We examined the distribution of messages in clusters based on different clustering methods.

From the heatmap below, we could see that the distribution of messages is more evenly on hierarchy for 3 cluster and K-Mean for 4 cluster. Overall, the most even distribution is for 4-K-Mean clustering. Therefore, for this dataset, we would compare and 3- hierarchical clustering and 4-K-Mean clustering as the method to segment observations.

**Conclusion and Recommendations:**