Customer Segmentation Using Clustering Techniques

By: Nishant Dixit

Introduction:

Clustering is an unsupervised learning technique that identifies natural groups or clusters that exist within the data. Thus, unlabelled data is used to train the model for unsupervised learning. One of the major applications of clustering is customer segmentation. Customer segmentation helps in understanding customers and provide personalized marketing aiming for improved customer satisfaction and increased revenue.

Objective:

To build an unsupervised learning model for customer segmentation.

Methodology:

- Data Wrangling
- Data Preprocessing
- Hyperparameter Selection
- Model Building
- Evaluating the models

Data Wrangling:

The datasets Customers.csv and Transactions.csv were utilized for clustering. Data aggregation is done on Transactions data and merged with Customers data. Data cleaning including missing values removal and data type conversions were made.

Data Preprocessing:

The numerical columns in the were checked for skewness and transformation was done to reduce the skewness. Encoding for categorical columns was performed. As clustering involves distance calculation, all the numerical columns were scaled where the mean =0 and standard deviation=1. This ensures all the columns are given equal importance and to avoid any bias to the columns.

Hyperparameter Selection:

Both KMeans clustering and Agglomerative clustering was performed. The optimal number of clusters (k) for KMeans clustering was found using an elbow plot and silhouette score. Elbow plot showing the elbow point and the inertia/within cluster sum of squares is low was found to be at 3. The silhouette score for k=3 was closest to 1. Hence, the optimal k was considered as 3.

In case of Agglomerative clustering, the hyperparameters linkage method and optimal k was selected using cophenetic correlation coefficient and dendrogram. The cophenetic correlation coefficient that is close to 1 was observed with average linkage method and based on the dendrogram 2 clusters was chosen optimal.

Model Building:

1. Kmeans Clustering

KMeans Clustering model was trained with the pre-processed data and optimal k value of 3. The three clusters resulted were visualized to understand the patterns.

Cluster 0: Nearly 40% of the customers, majorly from North America. These customers are newly signed up exhibiting higher purchase frequency and higher revenue contribution.

Cluster 1: 26% of the total customers, long-term customers from South America who generate consistent revenue through frequent transactions.

Cluster2: 33% of customers, least engaged and are from include a diverse range of signup.

The clusters 0 and 1 from KMeans clustering are not distinct and are elliptical.

2. Agglomerative Clustering

Agglomerative clustering with average linkage method and optimal number of clusters as 2, resulted in cluster0 and cluster1with a highly imbalanced proportion of 98% and 2% respectively.

Cluster 0: Predominantly from South America with very low average purchase frequency but with a longer tenure.

Cluster 1: Customers from North America, high purchase frequency and higher spending pattern.

Model Evaluation:

Metrics used for evaluating the models are Silhouette Score and Davis-Bouldin Index (DB Index).

Silhouette score measures the closeness of each data point within its cluster when compared with neighbouring clusters. The value ranges from +1 to -1. Higher silhouette score indicates well – assigned clusters. Davis-Bouldin Index measures the average similarity between each cluster and its most similar cluster. The DB index value ranges form 0 to infinity. Lower score indicates better clustering.

The silhouette score and DB index was calculated for both the models. The silhouette score is higher and DB index is lower for Agglomerative Clustering, suggesting it as a better clustering method.

Evaluation

In [120]:
```
print('Silhouette Score for KMeans Clustering: ', silhouette_score(x, df_new['km_label']))
print('Silhouette Score for Agglomerative Clustering: ', silhouette_score(x, df_new['agg_label']))
```

```
Silhouette Score for KMeans Clustering:  0.25351291152874794
Silhouette Score for Agglomerative Clustering:  0.26433637989294534
```

In [122]:
```
print('DB Index for KMeans Clustering: ', davies_bouldin_score(x, df_new['km_label']))
print('DB Index for Agglomerative Clustering: ', davies_bouldin_score(x, df_new['agg_label']))
```

```
DB Index for KMeans Clustering:  1.3285843612586048
DB Index:  0.9557015499699533
```

In [ ]:
```
# Lower DB index and Higher Silhouette Score with Agglomerative Clustering suggests better clustering method.
```

Conclusion:

Based on the agglomerative clustering method, the customer data inherently shows 2 clusters with a highly imbalanced proportion of customers within these clusters. However, the clusters are well established with distinct purchase behaviour and regional variation.