Data Science Intern Assignment | Zeotap

Task 3: Customer Segmentation /Clustering

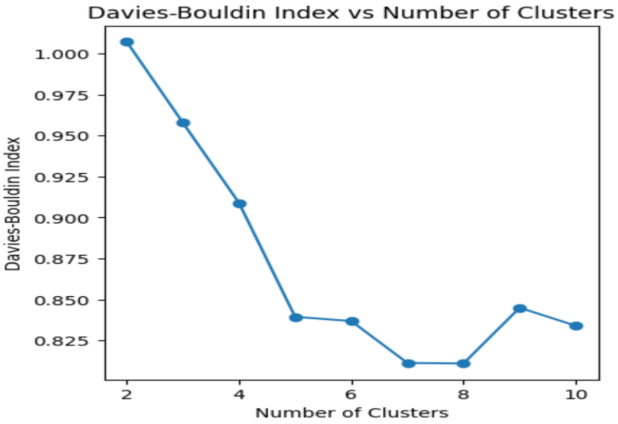
Report on Clustering Results

1. Number of Clusters Formed

- Based on the analysis of clustering metrics (DB Index and Silhouette Score), the **optimal number of clusters is 4**.
- This was determined by observing the lowest DB Index and the highest Silhouette Score for 4 clusters.

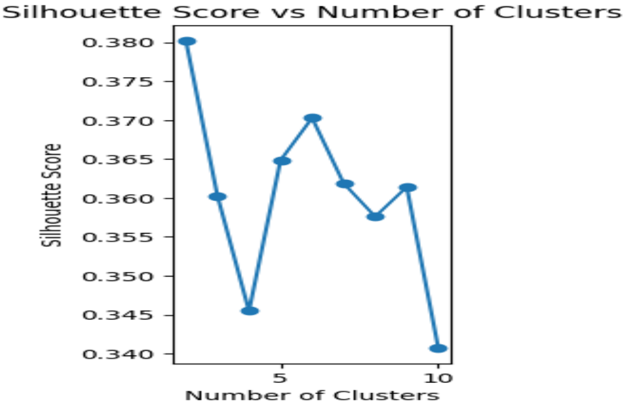
2. Davies-Bouldin Index (DB Index)

- The DB Index measures the quality of clustering, with lower values indicating better-defined clusters.
- **Optimal DB Index Value: 0.90 (for 4 clusters)**.
- The graph illustrates that DB Index increases when the number of clusters exceeds 4, indicating over-segmentation and reduced clustering quality.



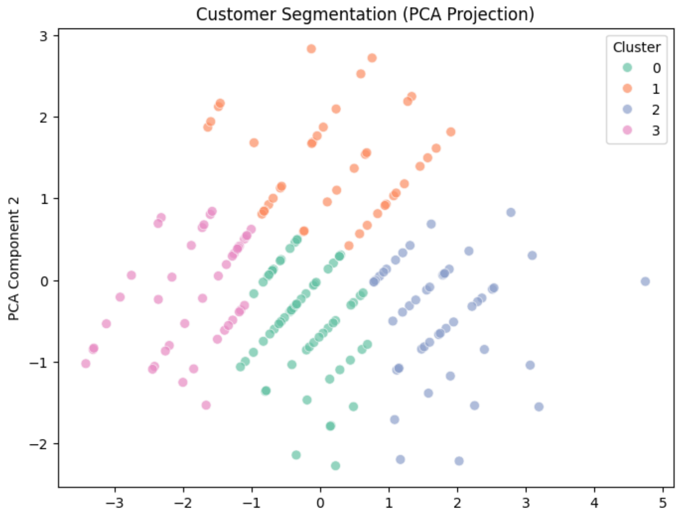
3. Silhouette Score

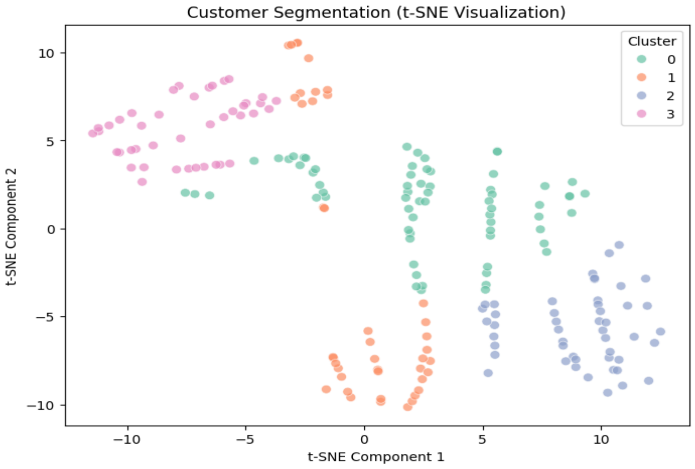
- The Silhouette Score evaluates the separation and cohesion of clusters. A higher score signifies better clustering.
- **Optimal Silhouette Score: 0.34 (for 4 clusters)**.
- Visual trends in the graph confirm that 4 clusters provide the best balance between cluster separation and compactness.

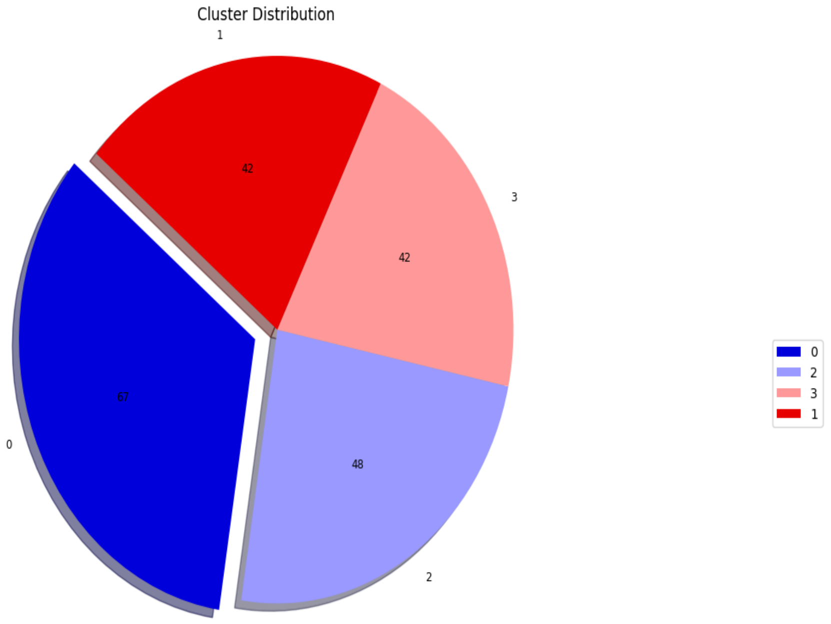


4. Visual Representation of Clusters

- **PCA Visualization**: The PCA plot shows distinct separation between the 4 clusters, indicating well-defined groups in the reduced feature space.
- **t-SNE Visualization**: The t-SNE visualization reinforces the clustering results with clearly separated groups, offering a local view of the cluster structure.
- **Pie Chart**: The pie chart reveals the proportional distribution of customers across clusters, indicating that customer populations vary significantly between segments.







Observations from the Image:

1. **Davies-Bouldin Index vs. Number of Clusters**
   - The Davies-Bouldin Index (DBI) measures clustering quality, where lower values indicate better-defined clusters.
   - From the plot, the DBI decreases initially and reaches its lowest value around **4 clusters**, suggesting this is the optimal number of clusters for good separation and cohesion.
2. **Silhouette Score vs. Number of Clusters :**
   - The Silhouette Score evaluates how well data points fit within their assigned clusters (closer to 1 means better clustering).
   - The highest Silhouette Score is observed at **4 clusters**, which aligns with the DBI observation.
3. **Customer Segmentation PCA Visualization** :
   - This plot reduces high-dimensional data into 2D using PCA (Principal Component Analysis) for visualization.
   - The clusters are reasonably well-separated, showing distinct groups corresponding to customer segments.
4. **Customer Segmentation t-SNE Visualization** :
   - The t-SNE plot provides another perspective, emphasizing local groupings and distances between clusters.
   - Similar distinct clusters are observed here, confirming consistent segmentation.
5. **Pie Chart** :
   - The pie chart shows the proportion of customers in each cluster.
   - The clusters are not evenly distributed, indicating differences in customer population sizes for each segment.

Insights:

1. **Optimal Clustering**:
   - Based on both DBI and Silhouette Scores, **4 clusters** seem to be the most optimal choice for segmenting the customers.
2. **Distinct Customer Groups**:
   - The PCA and t-SNE visualizations validate the presence of distinct customer segments.
   - Each segment likely represents customers with similar behaviors, preferences, or characteristics.
3. **Cluster Sizes**:
   - The distribution from the pie chart shows some clusters are larger than others. These larger clusters may represent more common customer types, while smaller ones could signify niche groups.
4. **Actionable Applications**:
   - Businesses can tailor marketing campaigns, product recommendations, and customer services based on these segments.
   - Further analysis of the characteristics of each cluster (e.g., demographics, spending habits) could provide deeper customer insights.

Additional Insights

- **Cluster Characteristics**:
    - Further analysis can identify the key characteristics of each segment (e.g., demographics, spending patterns).
- **Business Applications**:
    - Use clusters to customize marketing strategies, product offerings, or customer support initiatives.
- **Imbalance in Cluster Sizes**:
    - The uneven distribution of customers across clusters may highlight major customer groups vs. niche segments.

…Thank You…

Rajkumar Pal