PDF REPORT:

Siddharthan.R

Data science assignment: ecommerce transaction dataset

1. Introduction:

Customer segmentation is a key process in understanding different groups within a customer base, which can then lead to more personalized marketing strategies, better product targeting, and improved customer service. This task involves performing customer segmentation using clustering techniques, utilizing both customer profile data and transaction history. By clustering customers into distinct groups based on their similarities, we can uncover patterns that help in identifying specific customer needs and behaviors.

Model Development
The goal of this task is to cluster customers using both their profile and transaction data, followed by the evaluation of the clustering performance using relevant metrics such as the Davies-Bouldin (DB) Index. Different clustering algorithms are evaluated to select the one that best divides the customers into meaningful groups.

2. Data Preprocessing:

- **Customer Profile Data:** The Customers.csv file contains essential demographic information, such as age, gender, and location. This data is first cleaned and normalized.
- **Transaction Data:** The Transactions.csv file provides historical transaction information for each customer. It includes transaction amount, products purchased, and purchase frequency. This data is also preprocessed and transformed to capture valuable features like average spend, product categories, and total number of transactions.

3. Clustering Algorithm

- For clustering, several algorithms can be explored, such as:
- K-Means Clustering: K-Means is a widely used algorithm where customers are assigned to one of the k clusters based on their feature similarity.
- DBSCAN: A density-based clustering method that can capture clusters of arbitrary shapes and identify outliers.
- Agglomerative Clustering: A hierarchical approach that starts with individual points and merges them iteratively based on proximity.
- After testing various algorithms, K-Means is selected due to its simplicity and effectiveness for this task. The number of clusters k is chosen between 2 and 10 based in the evaluation metrics.

4. Cluster Evaluation Metrics

- Davies-Bouldin (DB) Index: This index measures the average similarity ratio of each cluster with its most similar cluster. Lower DB Index values indicate better clustering results, with distinct and well-separated clusters.
- Other Metrics: In addition to the DB Index, other relevant clustering metrics such as Silhouette Score and Inertia are calculated to evaluate the quality of the clustering.

After applying the clustering algorithm, we present the following results:

- Number of Clusters: Based on the evaluation metrics and visual inspection, we determine the optimal number of clusters.
- DB Index Value: The DB Index is reported for the selected clustering configuration.
- Other Clustering Metrics: Silhouette Score and Inertia values are also reported to assess the cohesion and separation of the clusters.

6. Visualizing the Clusters

- Cluster Visualization: Using dimensionality reduction techniques like PCA (Principal Component Analysis) or t-SNE, we project the high-dimensional customer data into 2D or 3D for visualization purposes. The clusters are then visualized in scatter plots to observe the separability and distribution of customers across different segments.
- Cluster Centers (for K-Means): The center of each cluster is shown in the visualizations to help understand the characteristics of each cluster.
-

7. Output: Cluster Summary

After the clustering process, a summary of the customer groups is provided, detailing:

- The number of customers in each cluster.
- Key characteristics of each cluster based on the demographic and transaction data.

Evaluation Criteria

The model is evaluated based on:

1. Clustering Logic and Metrics: The effectiveness of the selected clustering algorithm and the relevance of the resulting clusters.
2. Visual Representation: The clarity and usefulness of the cluster visualizations in interpreting the segmentation.

4. Conclusion:

The customer segmentation process reveals distinct customer groups, providing insights into their purchasing behavior and demographic characteristics. These clusters can be used for targeted marketing campaigns, personalized product offerings, and improving customer satisfaction by addressing the specific needs of each group.