Customer Segmentation / Clustering

Customer segmentation is a crucial aspect of marketing, sales, and customer relationship management. It involves grouping customers into distinct segments based on shared characteristics or behaviors, allowing businesses to deliver targeted marketing strategies and improve customer satisfaction.

In this task, we will perform **customer segmentation** using **clustering techniques** based on customer profile information and transaction data. Clustering is an unsupervised machine learning technique that helps identify patterns and group similar entities together without prior labels. It allows businesses to discover inherent groupings in their data.

Clustering Techniques Overview

Clustering techniques are used to group data points into clusters, where data points within the same cluster are more similar to each other than to data points in other clusters. There are several clustering algorithms, and one of the most commonly used is **K-Means Clustering**.

K-Means Clustering Algorithm

- **K-Means** is a centroid-based clustering algorithm where the objective is to minimize the variance within each cluster. It involves the following steps:

    1. **Initialization**: Select **K** initial centroids (can be done randomly or using specific methods like K-Means++ to improve convergence).
    2. **Assignment**: Assign each data point to the nearest centroid based on the distance (typically Euclidean distance).
    3. **Update**: Recalculate the centroids by computing the mean of all the points assigned to each centroid.
    4. **Convergence**: Repeat steps 2 and 3 until the centroids no longer change significantly or a maximum number of iterations is reached.

Clustering Evaluation Metrics

When evaluating clustering models, it's important to assess how well the clusters have been formed. Several metrics can be used to evaluate clustering quality, including:

1. **Davies-Bouldin Index (DBI)**:

- The **Davies-Bouldin Index** is a metric used to evaluate the separation between clusters. It is based on the average similarity ratio between each cluster and its most similar cluster. The formula for DBI is: DBI=1N∑i=1Nmax⁡i≠j(Si+SjDij)DBI = \frac{1}{N} \sum_{i=1}^{N} \max_{i \neq j} \left(\frac{S_i + S_j}{D_{ij}} \right)DBI=N1i=1∑Ni□=jmax(DijSi+Sj) where:
  - SiS_iSi is the average distance of points within cluster iii to the centroid of that cluster.
  - DijD_{ij}Dij is the distance between the centroids of clusters iii and jjj.
  - A **lower DBI score** indicates that the clusters are well-separated and compact.

2. **Silhouette Score**:
   - The **Silhouette Score** measures how similar a data point is to its own cluster compared to other clusters. It takes values between -1 and 1, where:
     - A score closer to **1** means the point is well-clustered.
     - A score closer to **-1** means the point may have been assigned to the wrong cluster.
     - A score closer to **0** means the point lies between two clusters.
3. **Inertia (Within-Cluster Sum of Squares)**:
   - **Inertia** is the sum of squared distances between each point and its corresponding centroid. It measures how compact the clusters are. A **lower inertia value** typically indicates better clustering.
4. **Cluster Size**:
   - The size of the clusters is also an important metric. Ideally, you want a balance where no cluster is too small or too large, unless there's a specific reason for it.

Steps in the Customer Segmentation Task

1. **Data Preprocessing**:
   - **Data cleaning**: Handle missing values, outliers, and irrelevant data.

- **Feature engineering**: Combine customer profile data (e.g., age, region) and transaction data (e.g., total spending, transaction frequency) into meaningful features for clustering.
- **Standardization**: Standardize the data to ensure that features with different units or ranges (e.g., age vs. total spending) are treated equally by the clustering algorithm.

2. **Choosing the Clustering Algorithm**:
   - While **K-Means** is commonly used, other clustering algorithms such as **Agglomerative Hierarchical Clustering**, **DBSCAN**, or **Gaussian Mixture Models** could be explored, depending on the data structure and problem at hand.
3. **Determining the Optimal Number of Clusters**:
   - The number of clusters (KKK) in K-Means can be determined by using the **Elbow Method**, **Silhouette Analysis**, or **Cross-Validation**.
   - **Elbow Method**: Plot the inertia (sum of squared distances to centroids) for different values of KKK and identify the "elbow" point where the inertia starts to decrease more slowly. This is often a good estimate of the optimal number of clusters.
4. **Running the Clustering Algorithm**:
   - Once the number of clusters is chosen, the **K-Means** algorithm is run, and the data points are assigned to one of the KKK clusters.
5. **Evaluating the Clustering**:
   - After clustering, evaluate the clustering results using metrics such as **DBI**, **Silhouette Score**, and **Inertia**. These metrics will help in understanding how well the clustering algorithm has performed.
6. **Visualizing the Clusters**:
   - **PCA (Principal Component Analysis)** and **t-SNE (t-Distributed Stochastic Neighbor Embedding)** are commonly used dimensionality reduction techniques for visualizing high-dimensional data in 2D or 3D.

- **PCA** projects the data into two or three principal components based on variance, while **t-SNE** is particularly effective for visualizing clusters in a lower-dimensional space.

7. **Interpreting the Clusters**:
    - Once the clusters are formed, interpret them by analyzing the average characteristics of each cluster (e.g., average age, average total spending, transaction frequency).
    - Identify patterns in customer behavior that can help in formulating marketing strategies.

Example Clusters Interpretation:

After performing clustering, you may find:

- **Cluster 1**: Young, frequent shoppers who spend relatively less per transaction.
- **Cluster 2**: Older, high-value customers who make fewer transactions.
- **Cluster 3**: Mid-age customers with moderate spending and transaction frequency.

Visualizing the Clusters:

1. **PCA Scatter Plot**: A scatter plot of customer data after applying PCA shows the customer distribution in 2D based on their cluster assignments.
2. **t-SNE Plot**: A t-SNE plot helps visualize how similar customers from different clusters are in a lower-dimensional space.