Customer Segmentation Documentation

This documentation provides a comprehensive explanation of the customer segmentation project using the Superstore dataset. The project applies unsupervised learning through KMeans clustering to group customers based on their purchasing behavior. Additionally, it introduces an interactive prediction function that can classify new customers into meaningful segment types. The entire workflow includes data preprocessing, exploratory analysis, clustering, cluster interpretation, prediction, and visualization. The aim is to uncover valuable insights about different customer groups and make the segmentation model useful for business decision-making.

Data Preprocessing

The dataset contains sales transactions with details about orders, products, and customers. Key columns include Customer ID, Segment, Sales, Category, and Sub-Category. The preprocessing stage focuses on aggregating sales metrics at the customer level. For each customer, three features are extracted: Total Sales, Average Sales, and Number of Orders. These numerical features capture purchasing behavior. To ensure fair treatment across features, StandardScaler is applied to normalize the data. Scaling is crucial because KMeans clustering is sensitive to differences in magnitude among features. The data is then ready for clustering.

Clustering with KMeans

KMeans clustering is applied to the scaled customer data. The optimal number of clusters is determined using the Elbow Method, which evaluates inertia across different values of k. Inertia represents the compactness of clusters. Based on the Elbow plot, four clusters are chosen for segmentation. Each customer is then assigned a cluster label. These cluster labels are analyzed to understand the behavior of different groups. Cluster characteristics include average sales, frequency of orders, and total spending. By comparing clusters, patterns such as high-value customers, frequent buyers, and occasional shoppers are revealed. This step transforms raw customer transactions into actionable segments.

Prediction Function

To make the clustering model interactive, a prediction function is developed. This function accepts new input values for Total Sales, Average Sales, and Number of Orders. The input is first transformed using the same scaler that was applied during training. The trained KMeans model then assigns the customer to a cluster. Since cluster numbers (0, 1, 2, 3) are not interpretable, human-readable labels are mapped to each cluster based on cluster summaries. For example, Cluster 0 may be labeled as 'High-Value Customers,' while Cluster 2 may represent 'Frequent Low Spenders.' This approach ensures that predictions are both accurate and understandable to business users.

Visualization and Conclusion

The segmentation results are visualized using Principal Component Analysis (PCA) to reduce feature dimensions into two components. A scatter plot is drawn where each point represents a customer, colored according to their cluster. This visualization helps in intuitively understanding how customers are distributed across segments. In conclusion, the project successfully demonstrates customer segmentation through KMeans clustering, enriched with interpretation, prediction, and visualization. The documentation outlines the entire workflow from data preprocessing to delivering meaningful customer types. This framework can be extended to other datasets and can also be upgraded to supervised learning models to directly predict business-defined segments like Consumer, Corporate, or Home Office.