Zeotap Data Science Internship

Name: R Karthik

Clustering

The provided code is a comprehensive analysis pipeline designed to calculate key customer metrics, perform clustering, and evaluate the results. It begins by calculating customer-related metrics such as the number of days since signup, transaction frequency, total spend, average transaction value, total quantity purchased, and the average time between transactions. These metrics are derived from a combination of customer data and transactional records. The code also handles missing region data by filling in unknown values, and one-hot encodes the regions, enabling a more nuanced segmentation.

Once the metrics are prepared, the code proceeds with clustering using the KMeans algorithm. It first scales the features to standardize them, ensuring that each metric contributes equally to the clustering process. To determine the optimal number of clusters, the Davies-Bouldin index is used, which measures the separation and compactness of clusters—the lower the index, the better the clustering. The code evaluates different cluster numbers (from 2 to a maximum of 10) and plots the Davies-Bouldin index, helping to visually assess the most effective number of clusters.

The final output includes the cluster profiles, which summarize the characteristics of each cluster, such as average transaction behavior (e.g., transaction counts, total spend, and average transaction values) and customer demographics (represented by one-hot encoded region variables). These profiles reveal distinct customer segments based on their behavior and geographic regions, providing valuable insights for targeted marketing, personalized recommendations, or resource allocation. The clustering results highlight the diversity in customer engagement, from highly active, high-spending customers to less frequent, low-spend users, with different regional distributions across the clusters.