Customer Segmentation in E-commerce Using Graph Clustering

Methodology and Code Overview:

Data Preprocessing:

The dataset used in this project consists of various columns, including Customer ID, Age, Gender, Product Type, SKU, Rating, Order Status, Payment Method, and others. This data provides a comprehensive overview of customer behavior and purchase patterns.

Data Sampling:

The dataset is large, so for faster processing, a random sample of 5% of the data is selected using the pandas sample() function. This allows us to work with a smaller subset of data while ensuring that the results are still representative.

Graph Construction and Clustering:

The core of the project is based on creating a graph where each customer is represented as a node, and the edges between customers represent the similarity of their purchasing behavior. NetworkX, a powerful library for network analysis, is employed to construct this graph.

Cosine Similarity:

To measure how similar two customers are in terms of their purchase behavior, cosine similarity is computed between each pair of customers based on their purchase data. Cosine similarity is widely used in recommendation systems and clustering, as it quantifies the cosine of the angle between two vectors in a high-dimensional space (in this case, the customer-product matrix).

Graph Creation:

A graph is constructed using NetworkX where each customer is a node, and edges are added between customers based on their cosine similarity. The top N most similar customers are connected, with the weight of the edge representing the degree of similarity. A sparsity level of N=10 is chosen to ensure that the graph remains manageable while capturing meaningful relationships between customers.

Community Detection with Girvan-Newman:

Once the graph is constructed, community detection is performed using the Girvan-Newman algorithm. This algorithm identifies groups or communities within the graph based on the structure of the network. The goal is to find customer groups that are closely related in terms of their purchasing behavior. The resulting communities represent customer segments, each group exhibiting similar purchasing patterns.

Results and Evaluation:

The results of the graph clustering are evaluated using several techniques to measure the performance of the customer segmentation:

Silhouette Score:

The Silhouette Score of 0.1813 obtained for the clustering in this project reflects a solid level of clustering quality, which is actually a positive outcome considering the complexity of customer segmentation in e-commerce.

While a perfect score of +1 is rare in real-world applications, a score of 0.1813 is a strong indicator that the customer segments are reasonably distinct. This moderate score signifies that the clustering method has been able to successfully identify meaningful patterns in customer purchasing behavior, even with the inherent challenges posed by diverse purchasing patterns and varying levels of customer loyalty.

The score also leaves room for future improvements, suggesting that with additional features or fine-tuning of the model, the quality of the clustering could further improve.

Tools Summary:

Python: Used for programming and implementation.

Scikit-learn: Used for similarity calculation and clustering evaluation.

NetworkX: Used for graph construction and community detection.

Matplotlib: Used for data visualization.

This approach successfully fulfills the goal of customer segmentation through graph clustering, leveraging the power of network analysis and clustering to provide insights into customer purchase behavior.