Report on Customer Segmentation Using KMeans Clustering

The purpose of this analysis was to perform customer segmentation by leveraging transactional and demographic data from the provided datasets—Customers.csv, Products.csv, and Transactions.csv. The ultimate goal was to understand customer behavior, identify distinct groups within the customer base, and use these insights to inform targeted business strategies.

To begin with, the three datasets were merged into a single unified dataset using CustomerID and ProductID as keys. The merged dataset provided a comprehensive view of customer activities, including their transaction history, total spending, and the variety of products purchased. The data was then aggregated on a customer level to create customer profiles, capturing key metrics such as total spending (total_spent), total transactions (total_transactions), and the number of unique products purchased (unique_products).

Choice of Algorithm: Why KMeans?

KMeans clustering was chosen for this analysis because it is a widely used, efficient, and robust algorithm for grouping data into distinct clusters based on similarity. Given the nature of the data—quantitative features representing customer behavior—KMeans is particularly well-suited as it minimizes within-cluster variance, ensuring that customers within the same group exhibit similar traits. Alternative algorithms, such as hierarchical clustering or DBSCAN, were not selected for the following reasons:

- **Hierarchical Clustering**: Computationally expensive for large datasets and less scalable compared to KMeans.
- **DBSCAN**: Sensitive to parameter selection and less effective for datasets with varying cluster densities, which may have been the case here.

Additionally, KMeans is straightforward to interpret and allows for easy application of dimensionality reduction techniques like PCA to visualize results, making it a practical and accessible choice for this problem.

Choice of Number of Clusters (k=4)

The decision to use 4 clusters was driven by both business logic and an evaluation of cluster quality. An initial assessment was performed using the Davies-Bouldin Index, a metric that evaluates clustering validity based on intra-cluster similarity and inter-cluster separation. A lower Davies-Bouldin score indicates better-defined clusters. After experimenting with different values of k, k=4 provided a balance between meaningful segmentation and interpretability, aligning with practical business needs to categorize customers into distinct and manageable groups.

Tools and Libraries Used

The analysis utilized Python's powerful data science libraries:

1. **Pandas**: For data preprocessing, merging, and aggregating customer profiles.
2. **Scikit-learn**: For scaling the data using StandardScaler, clustering using the KMeans algorithm, and evaluating cluster quality with the Davies-Bouldin Index. PCA was also performed to reduce the dimensionality of the data for visualization purposes.

3. **Matplotlib and Seaborn**: For creating an intuitive scatter plot to visualize customer clusters in two dimensions, making it easier to interpret and present results.

Motive and Business Implications

The primary motive behind this clustering was to segment customers into distinct groups based on their behavior. By identifying clusters, businesses can:

- Develop personalized marketing strategies for different customer groups.
- Identify high-value customers for loyalty programs.
- Recognize underperforming segments that may need re-engagement strategies.
- Tailor product offerings and promotions to the preferences of specific customer groups.

Conclusion

KMeans clustering with k=4 enabled meaningful segmentation of the customer base, as evidenced by the visualization and the Davies-Bouldin Index score. The resulting clusters provide actionable insights into customer behavior and pave the way for data-driven decision-making. The final cluster assignments were saved to Customer_Clusters_KMeans.csv for further use by the business or other stakeholders.

This process demonstrates the utility of advanced data analytics in uncovering patterns within transactional data and converting them into strategic opportunities.