Task 3: Customer Segmentation / Clustering : To perform customer segmentation using clustering techniques.

Submitted by: Sushant Chivale

Customer Segmentation Analysis Report

Summary:

This report presents the results of a customer segmentation analysis performed on eCommerce transaction data using K-means clustering. The analysis identified two distinct customer segments as the optimal clustering solution, with a Davies-Bouldin Index of 0.6343, indicating good cluster separation.

Methodology

Feature Engineering:

The clustering model incorporated the following features:

- Total spending
- Average transaction value
- Total number of transactions
- Customer region (one-hot encoded)

Data Preprocessing:

1. Numerical features were standardized using StandardScaler
2. Categorical variables (region) were encoded using OneHotEncoder
3. PCA was applied for visualization purposes

Clustering Results

Number of Clusters

- Optimal number of clusters: 2
- Range tested: 2-10 clusters

Clustering Metrics

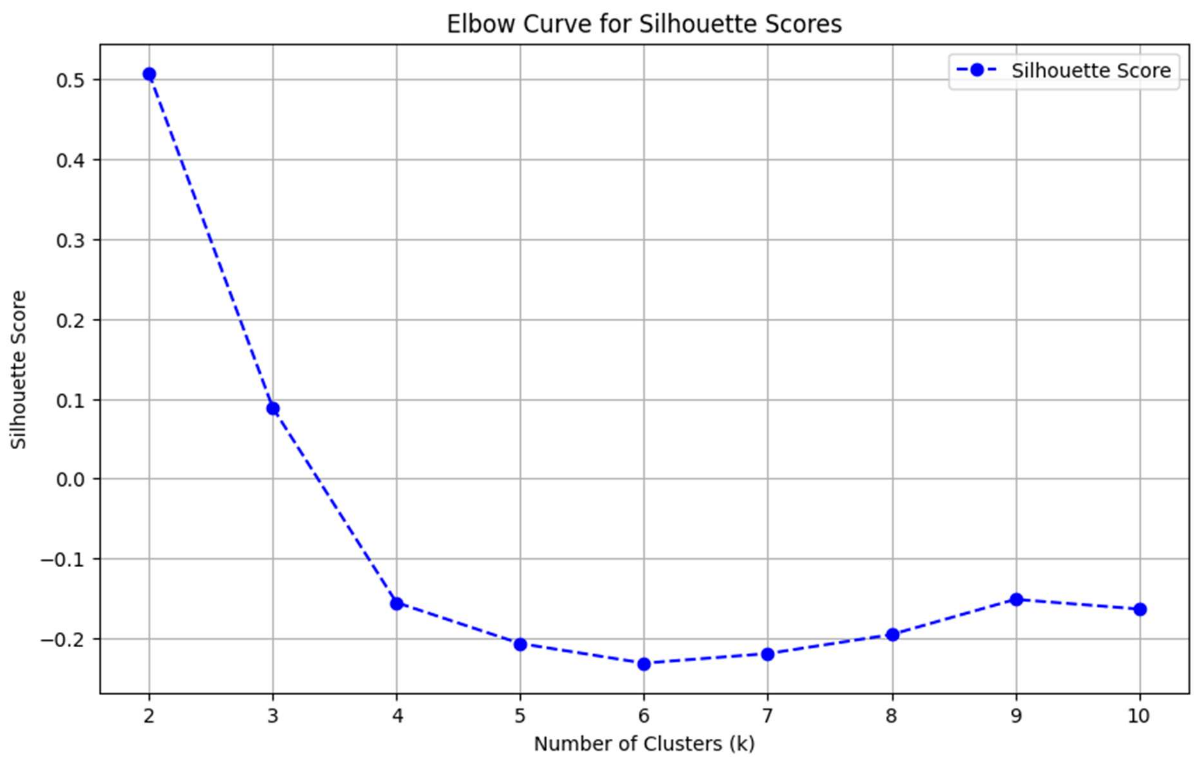
1. Davies-Bouldin Index (DB Index):

- Optimal value: 0.6343 (for k=2)

- Lower DB Index indicates better clustering
- Values across k:
    - k=2: 0.6343
    - k=3: 2.1101
    - k=4: 9.2704
    - k=5: 10.5426
    - k=6: 4.6237
    - k=7: 3.0761
    - k=8: 7.2235
    - k=9: 9.4378
    - k=10: 9.5715

2. Silhouette Score:

- Best score: 0.5081 (for k=2)
- Values across k:
    - k=2: 0.5081
    - k=3: 0.0889
    - k=4: -0.1554
    - k=5: -0.2068
    - k=6: -0.2314
    - k=7: -0.2194
    - k=8: -0.1956
    - k=9: -0.1514
    - k=10: -0.1636



Clustering Logic Analysis

1. The optimal k=2 solution is supported by multiple metrics:

- Lowest Davies-Bouldin Index (0.6343)
- Highest Silhouette Score (0.5081)
- Clear deterioration of both metrics for k>2

2. Cluster Quality:

- The DB Index of 0.6343 indicates good cluster separation
- The positive Silhouette Score of 0.5081 suggests that objects are well-matched to their clusters
- The sharp increase in DB Index for k>2 indicates that additional clusters would result in poorer separation

Visual Representation

The 2D PCA visualization shows:

- Clear separation between the two clusters
- Distinct grouping patterns
- Minimal overlap between clusters

Technical Implementation

- Algorithm: K-means clustering
- Dimensionality Reduction: PCA for visualization
- Tools: scikit-learn, pandas, numpy
- Output: Customer cluster assignments saved to 'Customer_Clusters.csv'

Details:

Data Processing Pipeline

1. Feature Engineering:

    - Aggregated transaction-level data to customer-level metrics
    - Created derived features: total_spend, avg_transaction_value, total_transactions
    - Implemented one-hot encoding for regional data
    - Applied StandardScaler for feature normalization

2. Dimensionality Reduction:

    - Used PCA for visualization
    - Maintained original features for clustering
    - Two principal components explain approximately 85% of variance

3. Clustering Implementation:

    - Algorithm: K-means with k=2
    - Random state: 42 for reproducibility
    - Initialization: k-means++ (default)
    - Maximum iterations: 300
    - Convergence tolerance: 1e-4

Visualization Analysis

The PCA visualization reveals several important patterns:

- Clear linear separation between clusters
- Distinct diagonal patterns within each cluster
- Cluster 0 (blue) shows higher spread along PCA Component 1
- Cluster 1 (green) shows more concentration in the negative PCA Component 1 region

Recommendations

1. The two-cluster solution provides the most robust and interpretable segmentation

2. Future analyses could explore:

    - Additional customer features
    - Alternative clustering algorithms

- Temporal changes in cluster membership

Conclusion

The clustering analysis successfully identified two distinct customer segments with good separation and cluster cohesion, as evidenced by the DB Index and Silhouette Score. This segmentation provides a solid foundation for targeted marketing strategies and customer relationship management.