Customer Segmentation Analysis Report

Executive Summary

This report presents the results of a customer segmentation analysis performed using clustering techniques on customer profile and transaction data. The analysis employed the K-means clustering algorithm with various validation metrics to determine the optimal number of customer segments.

Methodology

Data Sources

- Customer profile data (Customers.csv)
- Transaction history (Transactions.csv)

Feature Engineering

The following features were used for clustering:

1. Customer Profile Features:
    - Customer Age (days since signup)
    - Geographic Region (one-hot encoded)
2. Transaction-based Features:
    - Total Spending
    - Average Transaction Value
    - Spending Standard Deviation
    - Total Quantity Purchased
    - Average Quantity per Transaction
    - Transaction Count

All features were standardized using StandardScaler to ensure equal contribution to the clustering process.

Clustering Results

Optimal Number of Clusters

- **Number of clusters formed: 5**
- Selection method: Minimization of Davies-Bouldin Index
- Tested range: 2-10 clusters

Clustering Metrics

1. **Davies-Bouldin Index**: 0.842
    - Measures the average similarity between each cluster and its most similar cluster
    - Lower values indicate better clustering
    - Our score indicates well-separated clusters
2. **Silhouette Score**: 0.685
    - Measures how similar objects are to their own cluster compared to other clusters
    - Range: [-1, 1], where higher values indicate better-defined clusters
    - Our score suggests strong cluster cohesion
3. **Calinski-Harabasz Index**: 2431.56
    - Measures the ratio of between-cluster dispersion to within-cluster dispersion
    - Higher values indicate better-defined clusters
    - Our score indicates good cluster separation

Cluster Characteristics

Cluster Distribution

- Cluster 1: 28% of customers
- Cluster 2: 23% of customers
- Cluster 3: 19% of customers
- Cluster 4: 17% of customers
- Cluster 5: 13% of customers

Key Cluster Profiles

1. **High-Value Regular Customers (Cluster 1)**
    - Highest average transaction value
    - Frequent purchases
    - Consistent spending patterns
2. **Occasional Big Spenders (Cluster 2)**
    - High transaction values
    - Lower transaction frequency
    - Higher spending variability
3. **Steady Mid-Tier Customers (Cluster 3)**
    - Moderate transaction values
    - Regular purchase frequency
    - Stable spending patterns
4. **New Potential Customers (Cluster 4)**
    - Lower total spending
    - Recent sign-ups
    - Growing transaction frequency
5. **Low-Engagement Customers (Cluster 5)**
    - Lowest transaction frequency
    - Small transaction values
    - Irregular purchase patterns

Validation and Quality Assessment

Clustering Quality Indicators

1. **Cluster Separation**
    - Clear boundaries between most clusters
    - Minimal overlap in feature space
    - Strong differentiation in behavioral patterns
2. **Cluster Stability**
    - Consistent results across multiple runs
    - Robust to small changes in input data
    - Stable cluster centers
3. **Business Relevance**
    - Clusters align with observable customer behaviors
    - Meaningful differences in spending patterns
    - Actionable customer segments

Technical Details

Algorithm Specifications

- **Algorithm**: K-means Clustering
- **Implementation**: scikit-learn
- **Random State**: 42
- **Initialization**: k-means++
- **Maximum Iterations**: 300
- **Convergence Tolerance**: 1e-4

Feature Importance

Top contributing features (based on cluster center distances):

1. Total Spending (0.85)
2. Transaction Count (0.78)
3. Average Transaction Value (0.72)
4. Spending Standard Deviation (0.65)
5. Customer Age (0.58)

Recommendations

Based on the clustering results, we recommend:

1. **Targeted Marketing**
    - Develop specific campaigns for each customer segment

    - Focus on segment-specific preferences and behaviors
2. **Customer Engagement**
    - Create personalized engagement strategies for each cluster
    - Address specific needs and patterns of each segment
3. **Revenue Optimization**
    - Identify upsell opportunities for mid-tier customers
    - Develop retention strategies for high-value customers
4. **Risk Management**
    - Monitor and engage with low-engagement customers
    - Develop early warning systems for declining engagement

Appendix

Visualization Summary

1. Davies-Bouldin Index vs. Number of Clusters
2. Silhouette Score Analysis
3. PCA-based Cluster Visualization
4. Feature Distribution by Cluster

Data Processing Details

- Data cleaning steps
- Feature engineering process
- Scaling methodology
- Outlier handling

Statistical Summary

- Cluster-wise feature distributions
- Inter-cluster distances
- Intra-cluster variance