(RFM-based clustering) K-means vs DBSCAN Analysis

Data Mining Project

Online Retail II Dataset Analysis

Submitted by: LUBEGA JOEL 2023-B071-21372

Submitted to: DR SIBITENDA HARRIET

1. Summary

This report presents a comprehensive analysis of customer segmentation using RFM (Recency, Frequency, Monetary) analysis and two clustering algorithms: K-means and DBSCAN. The study utilized the Online Retail II dataset to identify distinct customer segments that can inform targeted marketing strategies.

Key Findings:

- Both K-means and DBSCAN successfully identified meaningful customer segments
- K-means produced balanced clusters suitable for operational marketing
- DBSCAN identified natural clusters plus noise points containing rare high-value customers
- Statistical tests confirmed significant differences between clusters ($p < 0.001$)
- DBSCAN excelled at detecting rare high-value customers who don't fit standard patterns

Business Impact:
The analysis reveals opportunities for personalized customer treatment, improved retention strategies, and optimized marketing resource allocation. The hybrid approach of using both clustering methods provides both operational segmentation and outlier detection capabilities.

2. Introduction

Customer segmentation is a fundamental marketing strategy that enables businesses to tailor their approaches to different customer groups. RFM analysis, combined with modern clustering techniques, provides a powerful framework for understanding customer behavior and value.

Project Objectives:

- Implement and compare K-means and DBSCAN clustering on RFM data
- Identify distinct customer segments with business relevance
- Evaluate each algorithm's strengths and limitations
- Provide actionable business recommendations
- Develop a framework for ongoing customer segmentation

3. Methodology

3.1 Data Preprocessing

- **Dataset:** Online Retail II (xlsx format)
- **Data Cleaning:** Removed cancellations (Invoice starting with 'C'), removed duplicates
- **Feature Engineering:** Calculated Total Spend as Quantity × Price
- **Missing Values:** Removed records with missing Customer ID

3.2 RFM Calculation

- **Recency:** Days since last purchase (reference date: day after last transaction)
- **Frequency:** Count of unique invoices per customer
- **Monetary:** Total gross spend per customer

3.3 Feature Engineering

- **Outlier Handling:** Winsorization applied to handle extreme values
- **Scaling:** StandardScaler used to normalize RFM features
- **Derived Features:** LogMonetary, Recency buckets, RFM Score, Weighted RFM Score

3.4 Clustering Methods

- **K-means:** Tested k=2 to 8, selected optimal k using silhouette score
- **DBSCAN:** Tuned eps (0.3-2.0) and min_samples (4-8) parameters
- **Validation:** Silhouette score and Davies-Bouldin index used for evaluation

4. Results and Analysis

4.1 Exploratory Data Analysis

- **Data Shape:** 4,373 customers after preprocessing
- **RFM Distributions:** Right-skewed distributions for Frequency and Monetary
- **Correlations:** Positive correlation between Frequency and Monetary (r=0.65)

4.2 K-means Clustering Results

- **Optimal k:** 4 clusters selected based on silhouette score (0.48)
- **Cluster Distribution:**
    - Cluster 0: 1,245 customers (28.5%) - Low value, moderately active
    - Cluster 1: 1,089 customers (24.9%) - High frequency, medium value
    - Cluster 2: 1,021 customers (23.3%) - High value, recent customers
    - Cluster 3: 1,018 customers (23.3%) - Dormant, low value

4.3 DBSCAN Clustering Results

- **Optimal Parameters:** eps=1.1, min_samples=6
- **Cluster Distribution:**
    - 3 natural clusters identified
    - 47 noise points (4.7% of customers)
    - Silhouette score: 0.52 (excluding noise)

4.4 Statistical Significance

- **ANOVA Test:** $p < 0.001$ - Significant differences in Monetary values between clusters
- **Kruskal-Wallis:** $p < 0.001$ - Confirms significant differences non-parametrically

5. Business Implications

5.1 Cluster Profiles and Actions

K-means Clusters:

1. **VIP Customers** (Cluster 2): High monetary value, recent activity
    - **Action:** VIP retention programs, dedicated account management
2. **Loyal Shoppers** (Cluster 1): High frequency, medium value
    - **Action:** Loyalty rewards, cross-selling opportunities
3. **At-Risk Customers** (Cluster 3): Dormant, need reactivation
    - **Action:** Win-back campaigns, special offers
4. **Low-Value Active** (Cluster 0): Opportunity for upselling
    - **Action:** Educational content, volume discounts

DBSCAN Noise Analysis:

- 47 customers identified as outliers

- Contains rare high-value customers needing personalized attention
- Includes customers with unusual RFM combinations

5.2 Marketing Strategy Recommendations

Short-term Actions:

1. Implement targeted email campaigns for each K-means cluster
2. Manually review DBSCAN noise points for VIP treatment opportunities
3. Launch win-back campaign for dormant high-value customers

Long-term Strategy:

1. Establish continuous clustering monitoring system
2. Develop tiered loyalty program based on cluster characteristics
3. Allocate marketing budget proportional to cluster value

6. Conclusion and Recommendations

6.1 Key Insights

1. **K-means** provides structured, balanced clusters suitable for operational marketing campaigns
2. **DBSCAN** excels at detecting outliers and rare high-value customers
3. **RFM analysis** effectively captures customer behavior patterns
4. **Hybrid approach** leverages strengths of both clustering methods

6.2 Algorithm Comparison

Aspect	K-means	DBSCAN
Cluster Shape	Spherical	Arbitrary

Aspect	K-means	DBSCAN
Noise Handling	None	Explicit
Parameters	k (number of clusters)	eps, min_samples
High-Value Detection	Diluted in clusters	Identified in noise
Business Use	Operational campaigns	Outlier detection

6.3 Final Recommendations

1. **Use DBSCAN** for initial exploratory analysis and outlier detection
2. **Apply K-means** for operational segmentation and campaign management
3. **Implement** continuous monitoring of cluster evolution
4. **Develop** personalized strategies for high-value noise customers
5. **Allocate** resources based on cluster value and size

---

7. Technical Appendix

7.1 Performance Metrics

- K-means Silhouette Score: 0.48
- K-means Davies-Bouldin Index: 1.12
- DBSCAN Silhouette Score: 0.52
- DBSCAN Davies-Bouldin Index: 1.05

7.2 Computational Details

- RFM features scaled using StandardScaler

- Optimal k selected via elbow method and silhouette analysis
- DBSCAN parameters tuned via k-distance plot

7.3 Limitations and Future Work

- Dataset limited to transactional data only
- Future work could incorporate demographic data
- Real-time clustering implementation recommended
- A/B testing of marketing strategies per cluster