

AI-Driven Customer Intelligence System

Advanced Customer Segmentation Using Unsupervised Learning

1. Executive Summary

This project presents a complete end-to-end implementation of an unsupervised customer segmentation system designed to extract strategic business intelligence from raw retail transaction data. In the absence of labeled datasets, advanced clustering algorithms were applied to discover hidden customer behavior patterns.

Three distinct customer segments were identified using RFM-based behavioral features. Gaussian Mixture Model achieved the highest Silhouette Score (0.3907), while DBSCAN produced the lowest Davies-Bouldin Index (0.7535), indicating strong internal cluster compactness. These insights provide actionable recommendations for targeted marketing, retention, and revenue optimization strategies.

2. Problem Statement

Organizations frequently accumulate large volumes of transactional data without structured customer labels. This lack of segmentation limits their ability to personalize marketing campaigns, identify high-value customers, predict churn risk, and allocate resources efficiently.

The objective of this project is to apply unsupervised learning techniques to automatically identify meaningful customer groups and convert behavioral patterns into strategic business recommendations.

3. Dataset Description

The Online Retail dataset containing over 5,000+ transaction records was utilized. Key attributes include transaction date, quantity, unit price, product descriptions, and customer identifiers.

RFM (Recency, Frequency, Monetary) analysis was conducted to summarize purchasing behavior at the customer level. Additional derived features such as average order value and unique product count were included to enrich behavioral representation.

4. Data Preprocessing & Feature Engineering

Data Cleaning Steps:

- Removed missing Customer IDs
- Eliminated invalid or negative transaction values
- Filtered cancelled transactions

Feature Engineering:

- Recency: Days since last purchase
- Frequency: Total number of transactions
- Monetary: Total spending per customer
- Average Order Value and Product Diversity metrics

5. Clustering Methodology

Four unsupervised clustering algorithms were implemented and evaluated: KMeans, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Model (GMM).

Optimal cluster selection was performed using the Elbow Method, Silhouette Score, and Davies-Bouldin Index. K=3 demonstrated the best balance between cluster separation and interpretability.

Model Performance Comparison

Algorithm	Silhouette Score	Davies-Bouldin Index
KMeans	0.3825	0.9813
Hierarchical	0.3785	0.9643
DBSCAN	0.3722	0.7535
GMM	0.3907	0.9707

6. Cluster Interpretation & Business Profiling

Cluster 0 – Dormant / At-Risk Customers

High recency, low frequency, and low monetary contribution. These customers demonstrate inactivity and represent churn risk. Targeted re-engagement campaigns and discount strategies are recommended.

Cluster 1 – Premium Loyal Customers

Low recency, high transaction frequency, and highest revenue contribution. These customers form the core revenue base and should receive loyalty rewards, exclusive promotions, and personalized premium offers.

Cluster 2 – Regular Mid-Tier Customers

Moderate purchasing activity and spending. These customers present strong upselling and cross-selling opportunities to increase lifetime value.

7. Business Insights & Strategic Recommendations

- Premium Loyal Customers generate the highest overall revenue.
- Dormant Customers exhibit high churn probability.
- Mid-Tier Customers show potential for revenue growth through targeted engagement.
- Segmentation enables optimized marketing budget allocation.

8. Conclusion

This project demonstrates how unsupervised learning can transform raw transactional data into strategic customer intelligence. The developed modular and scalable pipeline can be deployed in real-world enterprise environments to support data-driven decision-making and personalized marketing initiatives.