© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716518

Customer Segmentation Using Unsupervised Machine Learning Techniques: A Data Driven Approach

AYUSH NIGAM[1], AVINASH MAURYA[2], SURESH KUMAR TIWARI[3], DR. SANJAY PACHAURI[4]
[1,2,3,4]*Department of Data Science (DDCS), GNIOT College, Greater Noida, India*

Abstract- ***Customer segmentation is a critical component of modern marketing analytics and business intelligence. This paper presents a comparative analysis of three widely used unsupervised machine learning algorithms — K-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Agglomerative Hierarchical Clustering — for the purpose of customer segmentation in the retail domain. The UCI Online Retail Dataset is preprocessed and transformed using RFM (Recency, Frequency, Monetary) analysis to extract meaningful behavioral features. Each algorithm is applied to the RFM feature space and evaluated using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Cluster visualizations are generated using Principal Component Analysis (PCA) for dimensionality reduction. Experimental results demonstrate that K-Means achieves superior segmentation quality in terms of Silhouette Score, while DBSCAN effectively identifies outlier customers. This study provides a structured benchmark for selecting appropriate clustering methods in customer analytics and offers actionable insights for personalized marketing strategy formulation.***

Index Terms- ***Customer Segmentation, Machine Learning, K-Means Clustering, DBSCAN, Hierarchical Clustering, RFM Analysis, Unsupervised Learning, Retail Analytics.***

I. INTRODUCTION

Customer segmentation is the process of dividing a heterogeneous customer base into homogeneous groups based on shared characteristics such as purchasing behavior, demographics, and engagement patterns. It enables businesses to identify high-value customers, design targeted marketing campaigns, optimize resource allocation, and improve customer retention rates.

Traditional rule-based segmentation methods, such as demographic categorization, are increasingly inadequate for handling the complexity of modern transactional data. With the exponential growth in e-commerce and digital retail, machine learning-based clustering approaches have gained significant traction due to their ability to discover hidden patterns in large-scale datasets without requiring labeled training data.

Among unsupervised learning techniques, clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering are the most commonly applied for customer segmentation tasks. However, each algorithm exhibits distinct characteristics in terms of cluster shape assumptions, sensitivity to noise, scalability, and interpretability. A systematic comparison of these methods on a standardized feature representation is essential for guiding practitioners in selecting the most appropriate algorithm for their specific use case.

The RFM (Recency, Frequency, Monetary) model is a well-established behavioral framework that quantifies customer engagement through three dimensions: how recently a customer made a purchase, how frequently they purchase, and how much monetary value they contribute. RFM-transformed features serve as a compact, business-interpretable representation of customer behavior, making them highly suitable as input features for clustering algorithms.

This study makes the following contributions: (1) applies RFM feature engineering to the UCI Online Retail Dataset to derive customer behavioral profiles; (2) implements and evaluates K-Means, DBSCAN, and Agglomerative Hierarchical Clustering on the RFM feature space; (3) assesses performance using multiple cluster validity indices; (4) visualizes clusters using PCA-based 2D projection; and (5) provides business-level interpretation of the resulting customer segments.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716518

II. LITERATURE REVIEW

Extensive research has been conducted on machine learning-based customer segmentation. Several key studies inform the methodology of the present work.

Ngai et al. [1] provided a comprehensive review of CRM applications using data mining, establishing the theoretical foundation for behavioral customer analysis. The study emphasized the utility of RFM as a key feature extraction framework for segmentation tasks.

Khajvand et al. [2] proposed an RFM-weighted customer value model for segmentation in the banking domain, demonstrating that RFM-transformed features yield more interpretable clusters than raw transactional features.

Chen et al. [3] applied K-Means clustering on the UCI Online Retail Dataset and demonstrated its effectiveness in identifying distinct purchasing behavior groups, establishing a benchmark for comparative studies on this dataset.

Seret et al. [4] performed a comparative evaluation of K-Means and Hierarchical Clustering on retail customer data, finding that K-Means outperforms hierarchical approaches in terms of computational efficiency, while hierarchical clustering provides richer dendrogrammatic representations.

Ankerst et al. [5] introduced OPTICS as an extension of DBSCAN for varying density clusters, while Ester et al. [6] originally proposed DBSCAN, demonstrating its robustness to noise and ability to detect arbitrary cluster shapes — properties not shared by K-Means.

Wei et al. [7] applied PCA-based dimensionality reduction prior to clustering on customer transactional data, demonstrating that PCA preprocessing improves cluster quality metrics and visualization clarity.

Huang [8] proposed K-Modes and K-Prototypes as extensions of K-Means for mixed data types. While the present study focuses on numerical RFM features, this work informs the preprocessing design choices.

Sinaga and Yang [9] proposed an unsupervised K-Means clustering algorithm with an automatic determination of cluster count, addressing the limitation of manual K selection in standard K-Means.

Despite extensive existing literature, a systematic three-way comparison of K-Means, DBSCAN, and Hierarchical Clustering on RFM-transformed retail customer data with multi-index evaluation remains underexplored. This paper addresses that gap.

III. DATASET DESCRIPTION

The UCI Online Retail Dataset [10] is employed in this study. It contains transactional records of a UK-based online retail company spanning the period from December 2010 to December 2011. The dataset comprises 541,909 records with 8 attributes: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country.

Table 1 summarizes the key statistical properties of the dataset prior to preprocessing:

Table I. Dataset Statistical Summary

Attribute	Type	Non-Null Count	Description
InvoiceNo	Categorical	541,909	Unique invoice identifier
CustomerID	Numerical	406,829	Unique customer identifier
Quantity	Numerical	541,909	Units purchased per transaction
UnitPrice	Numerical	541,909	Price per unit (GBP)
InvoiceDate	DateTime	541,909	Date and time of transaction
Country	Categorical	541,909	Country of the customer

After removing records with missing CustomerID values and filtering out cancelled transactions (InvoiceNo beginning with 'C'), the dataset is reduced

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716518

to 392,692 clean transactional records representing 4,372 unique customers. All transactions are denominated in GBP and represent B2C retail purchases.

IV. METHODOLOGY

A. RFM Feature Engineering

RFM analysis is applied to transform raw transactional data into three customer-level behavioral metrics:

- Recency (R): Number of days since the customer's most recent purchase, computed relative to a reference date of 2011-12-10.
- Frequency (F): Total number of distinct invoices (transactions) made by the customer during the observation period.
- Monetary (M): Total revenue generated by the customer, computed as the sum of (Quantity × UnitPrice) across all transactions.

To mitigate the influence of outliers and skewness in the RFM distributions, log transformation is applied to Frequency and Monetary values. All three RFM features are subsequently standardized using StandardScaler (zero mean, unit variance) prior to clustering.

B. Clustering Algorithms

Three clustering algorithms are implemented using Python's scikit-learn library (v1.3):

K-Means Clustering partitions n observations into K clusters by minimizing the within-cluster sum of squared distances to the cluster centroid. The optimal number of clusters K is determined using the Elbow Method on inertia values for $K \in \{2, 3, 4, 5, 6, 7, 8\}$ and validated using the Silhouette Score. K=4 is selected as optimal.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups together points that are closely packed in the feature space and marks as outliers the points that lie alone in low-density regions. The hyperparameters epsilon (ε) and minimum samples (MinPts) are tuned using k-distance graphs. Optimal values of ε=0.5 and MinPts=5 are identified.

Agglomerative Hierarchical Clustering builds a hierarchy of clusters by successively merging the closest pairs of clusters using Ward's linkage criterion, which minimizes the total within-cluster variance. A dendrogram is plotted to determine the optimal cut yielding 4 clusters, consistent with K-Means.

C. Evaluation Metrics

Three internal cluster validity indices are employed for performance evaluation:

Silhouette Score (SS) measures how similar each point is to its own cluster compared to other clusters. The score ranges from -1 to +1, where higher values indicate well-defined, compact clusters. Davies-Bouldin Index (DBI) computes the average similarity between each cluster and its most similar cluster; lower values indicate better separation. Calinski-Harabasz Index (CHI) evaluates cluster compactness and separation based on the ratio of between-cluster dispersion to within-cluster dispersion; higher values indicate better-defined clusters.

D. Dimensionality Reduction and Visualization

Principal Component Analysis (PCA) is applied to reduce the 3-dimensional RFM feature space to 2 principal components for cluster visualization purposes. The two principal components explain approximately 94.7% of the total variance in the RFM feature space.

V. EXPERIMENTAL RESULTS

A. Cluster Validity Comparison

Table 2 presents the comparative evaluation of the three clustering algorithms across all three validity indices:

Table II. Clustering Algorithm Performance Comparison

Algorithm	Silhouette Score ↑	Davies-Bouldin Index ↓	Calinski-Harabasz Index ↑	No. of Clusters
K-Means	0.412	0.891	1842.3	4

DBSCAN	0.318	1.204	1103.7	3 + noise
Hierarchical	0.387	0.943	1698.5	4

K-Means achieves the highest Silhouette Score (0.412) and Calinski-Harabasz Index (1842.3), indicating the most compact and well-separated clusters. Hierarchical Clustering performs comparably on the Davies-Bouldin Index (0.943 vs 0.891 for K-Means). DBSCAN identifies 3 dense clusters along with a noise cluster (approximately 8.2% of customers classified as outliers), reflecting its sensitivity to parameter selection.

B. Customer Segment Profiles (K-Means)

The four clusters identified by K-Means correspond to distinct customer behavioral profiles as described in Table 3:

Table III. Customer Segment Profiles (K-Means, K=4)

Cluster	Segment Label	Avg. Recency (days)	Avg. Frequency	Avg. Monetary (£)	Business Strategy
0	Champions	15	12.4	£2,840	Loyalty rewards, VIP programs
1	At-Risk Customers	142	3.1	£480	Re-engagement campaigns
2	Potential Loyalists	38	6.8	£1,120	Upsell & cross-sell offers
3	Lost Customers	287	1.4	£180	Win-back discounts

Cluster 0 (Champions) represents the most valuable customers with recent, frequent, and high-value purchases. Cluster 1 (At-Risk) comprises customers who were once active but have not purchased recently. Cluster 2 (Potential Loyalists) includes moderately engaged customers with growth potential. Cluster 3 (Lost Customers) contains dormant customers with low engagement across all RFM dimensions.

VI. DISCUSSION

The experimental results indicate that K-Means is the most effective algorithm for RFM-based customer segmentation on the UCI Online Retail Dataset, outperforming DBSCAN and Hierarchical Clustering on all three-evaluation metrics. The relatively spherical and well-separated nature of RFM clusters in the standardized feature space favors K-Means, whose centroid-based optimization is well-suited to compact, convex cluster shapes.

DBSCAN, while less competitive on global validity metrics, offers a unique advantage in identifying noise customers — a segment of potential interest for anomaly detection, fraud identification, or targeted re-activation strategies. Its sensitivity to the ε and MinPts hyperparameters, however, requires careful tuning and domain expertise.

Hierarchical Clustering provides competitive performance and the additional interpretability advantage of a dendrogram, which enables stakeholders to visually inspect the cluster merging process and adjust granularity without rerunning the algorithm. This property is particularly valuable in exploratory marketing analytics.

The RFM preprocessing step is identified as a key contributor to clustering quality. The log transformation and standardization of RFM features reduce the impact of high-value outliers and normalize the feature distributions, resulting in more coherent clustering outcomes across all three algorithms.

VII. CONCLUSION

This paper presented a systematic comparative study of K-Means, DBSCAN, and Agglomerative

Hierarchical Clustering algorithms for retail customer segmentation using RFM feature engineering on the UCI Online Retail Dataset. The study evaluated algorithm performance using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, and generated PCA-based cluster visualizations for interpretability.

K-Means demonstrated superior overall performance, yielding four interpretable customer segments — Champions, Potential Loyalists, At-Risk Customers, and Lost Customers — each with actionable marketing implications. DBSCAN contributed noise-based outlier detection capability, while Hierarchical Clustering offered dendrogram-based interpretability.

Future work will explore ensemble clustering approaches, deep learning-based autoencoders for feature extraction, incorporation of demographic and psychographic features alongside RFM, and real-time streaming segmentation pipelines for dynamic customer analytics. Extending the proposed framework to domain-specific datasets in banking, healthcare, and telecom sectors is also planned.

REFERENCES

[1] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," Expert Systems with Applications, vol. 36, no. 2, pp. 2592–2602, 2009.

[2] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," Procedia Computer Science, vol. 3, pp. 57–63, 2011.

[3] D. Chen, S. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing & Customer Strategy Management, vol. 19, no. 3, pp. 197–208, 2012.

[4] A. Seret, B. Baesens, and J. Vanthienen, "A new knowledge-based system for using multiple customer segmentation approaches in retail datasets," Expert Systems with Applications, vol. 41, no. 8, pp. 3883–3892, 2014.

[5] M. Ankerst, M. Breunig, H. P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," ACM SIGMOD Record, vol. 28, no. 2, pp. 49–60, 1999.

[6] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. KDD, 1996, pp. 226–231.

[7] J. Wei, H. Lin, and H. Wu, "A review of the application of RFM model," in Proc. African Journal of Business Management, vol. 4, no. 19, pp. 4199–4206, 2010.

[8] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283–304, 1998.

[9] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020.

[10] D. Chen, "Online Retail Data Set," UCI Machine Learning Repository, 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/online+retail