# CUSTOMER SEGMENTATION USING RFM INTEGRATED WITH UMAP AND HDBSCAN

**R. Anitha[1] and Y. Kalpana[1]**

[1]Department of Computer Science

[2]Department of BCA & IT

Vels Institute of Science, Technology, Advanced Studies (VISTAS), Chennai

Corresponding author E-mail: anithashivaguru@gmail.com, kalpana.scs@vistas.ac.in

## 1. Introduction:

The exponential growth of online retail platforms has led to the accumulation of large-scale transactional data containing valuable information about customer purchasing behavior. Retailers increasingly rely on data-driven techniques to understand customer heterogeneity and to design personalized marketing strategies. Customer segmentation is a critical analytical task that enables businesses to group customers with similar behavioral characteristics, thereby improving customer engagement, retention, and profitability.

Conventional customer segmentation approaches based on demographics or static rules often fail to reflect real purchasing behavior. Transactional data, on the other hand, provides a dynamic and behavior-oriented view of customers. However, such data is inherently noisy, skewed, high-dimensional, and non-linear. Customers differ significantly in purchase recency, buying frequency, and spending patterns, resulting in overlapping and imbalanced customer groups.

To overcome these challenges, this chapter presents an advanced customer segmentation framework that integrates RFM behavioral modeling, UMAP-based non-linear dimensionality reduction, and HDBSCAN density-based clustering. This framework is particularly well suited for online retail data, where customer behavior evolves continuously and traditional clustering techniques struggle to produce stable and interpretable segments.


**Figure 1: RFM Feature**

## 2. Behavioral Modeling Using the RFM Framework

The RFM model is one of the most widely adopted techniques for representing customer behavior using transactional data. It condenses complex purchasing patterns into three intuitive metrics: Recency, Frequency, and Monetary value. These metrics are closely aligned with marketing intuition and business decision-making.

### 2.1 Recency (R)

Recency measures the time elapsed since a customer's most recent transaction. It serves as a strong indicator of customer engagement and purchase intent. Customers who have purchased recently are more likely to respond positively to promotions, while customers with large recency values are often disengaged or at risk of churn.

Mathematically, recency for customer $i$is defined as:

$$R_i = T - t_i^{last}$$

where $T$is the reference date and $t_i^{last}$denotes the date of the most recent purchase by customer $i$.

### 2.2 Frequency (F)

Frequency captures how often a customer makes purchases within a given observation window. It reflects loyalty and habitual purchasing behavior. High-frequency customers typically represent repeat buyers who are valuable for long-term business sustainability.

$$F_i = \sum_{k=1}^{n_i} 1$$

where $n_i$is the total number of transactions made by customer $i$.

### 2.3 Monetary Value (M)

Monetary value measures the total or average amount spent by a customer during the observation period. It represents the economic contribution of a customer to the business.

$$M_i = \sum_{k=1}^{n_i} v_{ik}$$

where $v_{ik}$is the value of the $k^{th}$transaction.

### 2.4 Properties of the RFM Feature Space

While RFM features are simple and interpretable, they exhibit several properties that complicate direct clustering:

- Monetary values are often highly right-skewed
- Recency behaves inversely compared to Frequency and Monetary
- Strong non-linear interactions exist among R, F, and M
- Customer groups may overlap significantly

These characteristics necessitate advanced preprocessing and modeling techniques.

## 3 Challenges in Traditional RFM-Based Segmentation

Clustering directly on RFM features using traditional algorithms such as K-Means presents several limitations:

- Linear separability assumption: Customer behavior rarely follows linear boundaries.
- Predefined cluster count: The number of customer segments is usually unknown.
- Sensitivity to outliers: Extreme spenders or one-time buyers distort centroids.
- Inability to handle density variation: Loyal customers form dense groups, while occasional buyers are sparse.

These limitations motivate the use of manifold learning and density-based clustering techniques.

## 4 UMAP for Non-Linear Dimensionality Reduction

### 4.1 Overview of UMAP

Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction technique rooted in manifold learning and topological data analysis. UMAP aims to preserve the intrinsic geometric structure of high-dimensional data while projecting it into a lower-dimensional space.

Unlike linear methods such as PCA, UMAP effectively captures non-linear relationships that commonly occur in customer behavior data.

### 4.2 Mathematical Intuition Behind UMAP

UMAP constructs a weighted graph representing the local neighborhood structure of data points in high-dimensional space. A corresponding graph is constructed in low-dimensional space. The algorithm then minimizes the cross-entropy between these two graphs, ensuring that nearby points remain close while preserving global structure.

### 4.3 Role of UMAP in RFM-Based Segmentation

When applied to normalized RFM data, UMAP offers several advantages:

- Reveals hidden non-linear behavioral patterns
- Preserves local customer similarities
- Enhances density contrast between customer groups
- Produces a latent space well suited for density-based clustering

UMAP thus acts as a bridge between behavioral modeling and robust clustering.

## 5 Density-Based Clustering Using HDBSCAN

### 5.1 Overview of HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is an unsupervised clustering algorithm that identifies clusters based on varying density levels. It extends DBSCAN by introducing a hierarchical structure and a stability-based cluster selection mechanism.

### 5.2 Working Mechanism

HDBSCAN operates by:

- Computing mutual reachability distances
- Constructing a minimum spanning tree
- Creating a condensed cluster hierarchy
- Extracting stable clusters based on persistence

Customers that do not belong to any stable dense region are labeled as noise.

### 5.3 Advantages for Online Retail Segmentation

HDBSCAN is particularly suitable for online retail data because it:

- Automatically determines the number of clusters
- Handles clusters of varying shapes and densities
- Identifies inactive or anomalous customers
- Is robust to noise and outliers

## 6 Integrated RFM–UMAP–HDBSCAN Segmentation Framework

### 6.1 Motivation for an Integrated Framework

Customer segmentation in online retail involves analyzing large-scale transactional data characterized by heterogeneous purchasing behavior, non-linear relationships, density imbalance, and noise. A single analytical technique is often insufficient to address all these challenges simultaneously.

- RFM provides a business-oriented behavioral summary but lacks structural modeling capability.
- UMAP captures non-linear relationships but does not perform clustering.
- HDBSCAN effectively discovers clusters but depends heavily on the structure of the input space.

The motivation behind integrating RFM, UMAP, and HDBSCAN is to leverage the complementary strengths of each component while mitigating their individual limitations. This framework transforms raw transactional data into a density-aware latent representation, enabling robust and interpretable customer segmentation.

### 6.2 Overall Architecture of this Framework

The integrated segmentation framework follows a multi-stage pipeline architecture, ensuring modularity, scalability, and reproducibility. The architecture consists of six tightly coupled stages:

- Transactional data preprocessing
- RFM feature engineering
- Feature normalization and transformation

- UMAP-based latent space learning
- HDBSCAN-based density clustering
- Cluster profiling and business interpretation

Each stage plays a distinct role in enhancing segmentation quality and interpretability.

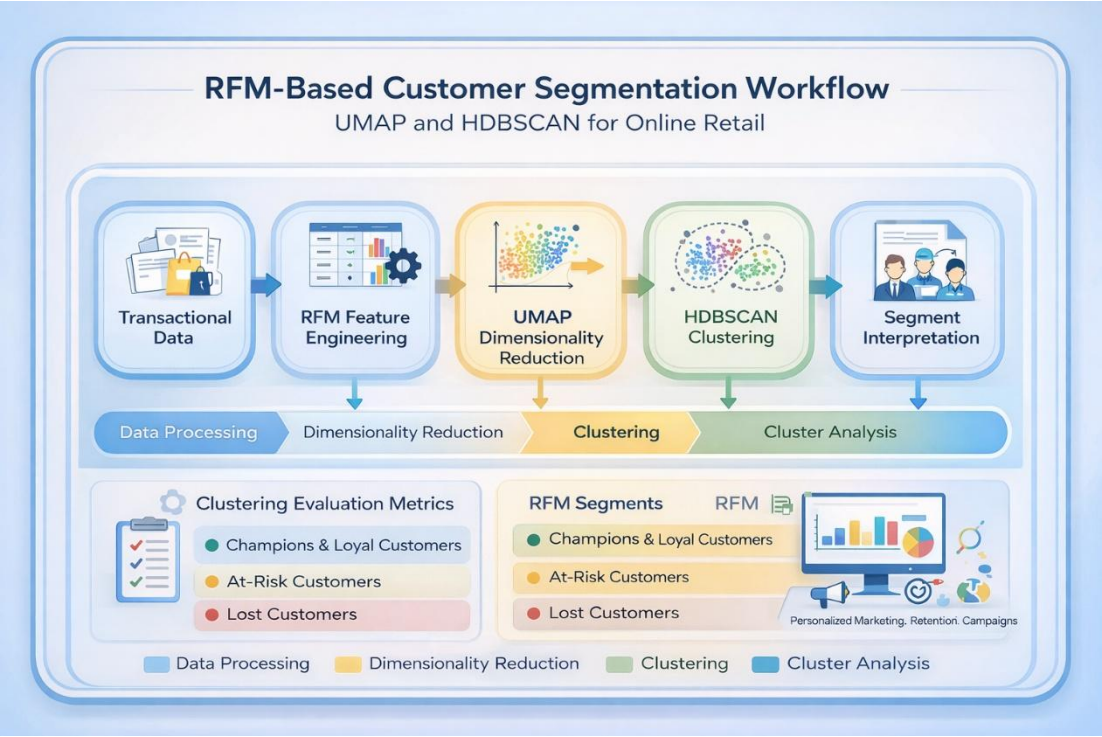


**Figure 2: RFM-UMAP-HDBSCAN Framework**

### 6.3 Transaction Data Preprocessing

The first stage focuses on preparing raw online retail transaction data for analysis. Transaction datasets often contain noise in the form of cancelled orders, missing values, duplicate records, or inconsistent identifiers.

Key preprocessing steps include:

- Removal of cancelled or refunded transactions
- Handling missing customer identifiers and transaction values
- Aggregation of transaction records at the customer level
- Selection of an appropriate observation window

This stage ensures data consistency and prevents spurious patterns from influencing downstream modeling.

### 6.4 RFM Feature Engineering

After preprocessing, customer behavior is summarized using the RFM model. For each customer $i$, three behavioral metrics are computed:

$$R_i = T - t_i^{last}, F_i = \sum_{k=1}^{n_i} 1, M_i = \sum_{k=1}^{n_i} v_{ik}$$

The resulting RFM feature matrix:

$$X = (R_i, F_i, M_i) \forall i$$

provides a compact yet expressive representation of customer purchasing behavior.

RFM is chosen because:

- It is interpretable by business stakeholders
- It captures engagement, loyalty, and value
- It generalizes well across retail domains

However, due to skewness and scale imbalance, further transformation is required before clustering.

**6.5 Feature Normalization and Transformation**

To ensure that no single RFM dimension dominates the analysis, feature normalization is applied. Standardization or logarithmic transformation is typically used:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

or

$$X_{log} = \log (1 + X)$$

This step improves numerical stability and enhances the effectiveness of manifold learning in the subsequent stage.

**6.6 UMAP-Based Latent Space Learning**

UMAP is applied to the normalized RFM feature matrix to learn a low-dimensional latent representation:

$$Z = f_{UMAP}(X_{scaled})$$

where $Z \in \mathbb{R}^{N \times d}$, with $d \ll 3$(typically $d = 2$or 3).

UMAP plays a crucial role in the integrated framework by:

- Capturing non-linear interactions among RFM features
- Preserving local neighborhood relationships between customers
- Enhancing density contrast between different behavioral groups

By projecting customers into a manifold where density variations are amplified, UMAP makes the data more suitable for density-based clustering.

**6.7 HDBSCAN-Based Density Clustering in Latent Space**

Clustering is performed on the UMAP embedding using HDBSCAN:

$$C = f_{HDBSCAN}(Z)$$

HDBSCAN constructs a hierarchical density structure and extracts clusters based on stability across density levels. Unlike centroid-based methods, it does not require the number of clusters to be specified in advance.

Key advantages in this framework include:

- Automatic discovery of the number of customer segments
- Identification of clusters with arbitrary shapes
- Explicit labeling of noise customers ($C = -1$)

Noise customers often correspond to one-time buyers, inactive users, or anomalous behavior, which are important from a business analytics perspective.

**6.8 Cluster Profiling and Behavioral Interpretation**

Once clusters are identified, each cluster is profiled using descriptive statistics:

$$\bar{R}_k = \frac{1}{|\ C_k\ |}\sum_{i \in C_k} R_i\ , \bar{F}_k = \frac{1}{|\ C_k\ |}\sum_{i \in C_k} F_i\ , \bar{M}_k = \frac{1}{|\ C_k\ |}\sum_{i \in C_k} M_i$$

These profiles enable semantic labeling of clusters, such as:

- High-value loyal customers
- Potential loyalists
- At-risk customers
- Lost customers
- Noise or irregular buyers

This step bridges the gap between machine learning outputs and actionable business insights.

**6.9 Algorithmic Summary of the Framework**

The integrated framework can be summarized as:

Transactions → RFM → Normalization → UMAP → HDBSCAN → Customer Segments

This pipeline ensures that clustering is performed on a behaviorally meaningful, non-linearly transformed, and density-aware representation of customers.

**6.10 Strengths of the Integrated Framework**

The RFM–UMAP–HDBSCAN framework offers several key advantages:

- Eliminates the need for predefined cluster counts
- Handles non-linearity and density imbalance
- Robust to noise and outliers
- Scalable to large online retail datasets
- Produces interpretable and actionable segments

These strengths make the framework suitable for both academic research and industrial deployment.

**6.11 Positioning Within Customer Analytics Literature**

Compared to traditional RFM–K-Means or RFM–GMM approaches, the framework represents a paradigm shift toward density-aware and manifold-based customer segmentation. It aligns with recent advances in unsupervised learning and reflects the increasing complexity of customer behavior in digital commerce.

## 7 Interpretation of Customer Segments

The framework typically identifies the following customer segments:

- Champions: Very recent, frequent, and high-spending customers
- Loyal Customers: Consistent buyers with moderate to high frequency
- Potential Loyalists: Recent customers with growth potential
- At-Risk Customers: Previously active customers with increasing recency
- Lost Customers: Inactive customers with minimal engagement
- Noise Customers: One-time or irregular buyers

Each segment supports targeted marketing and retention strategies.

## 8 Evaluation of Segmentation Quality

Evaluating the quality of customer segmentation is essential to ensure that the identified clusters are meaningful, well-separated, compact, and stable. Since customer segmentation using the RFM–UMAP–HDBSCAN framework is an unsupervised learning task, internal and stability-based validation metrics are primarily employed. These metrics quantitatively assess clustering performance without requiring ground-truth labels.

This section presents a comprehensive evaluation strategy using internal validation indices and stability measures, supported by mathematical formulations and interpretation.

### 8.1 Internal Validation Metrics

Internal validation metrics evaluate clustering quality based on the intrinsic structure of the data, focusing on intra-cluster cohesion and inter-cluster separation.

#### 8.1.1 Silhouette Coefficient

The Silhouette Coefficient measures how similar a customer is to its own cluster compared to other clusters. For a given customer $i$, the silhouette score is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where:

- $a(i)$is the average distance between customer $i$and all other customers within the same cluster
- $b(i)$is the minimum average distance between customer $i$and customers in the nearest neighboring cluster

The overall silhouette score for the clustering solution is computed as:

$$S = \frac{1}{N}\sum_{i=1}^{N} s(i)$$

where $N$denotes the total number of customers.

Interpretation:

- $S \approx 1$: well-separated and compact clusters
- $S \approx 0$: overlapping clusters
- $S < 0$: poor clustering assignment

In customer segmentation, silhouette values greater than 0.30 are generally considered acceptable due to natural overlap in purchasing behavior.

**8.1.2 Davies–Bouldin Index (DBI)**

The Davies–Bouldin Index evaluates clustering quality by computing the ratio of within-cluster dispersion to between-cluster separation. It is defined as:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

where:

- $K$is the number of clusters
- $\sigma_i$is the average distance of all points in cluster $i$from its centroid $c_i$
- $d(c_i, c_j)$is the distance between centroids of clusters $i$and $j$

Interpretation:

- Lower DBI values indicate better clustering
- $DBI < 1$suggests compact and well-separated clusters

For RFM-based customer segmentation, DBI values between 0.8 and 1.5 are commonly observed due to heterogeneous purchasing patterns.

**8.1.3 Calinski–Harabasz Index (CHI)**

The Calinski–Harabasz Index, also known as the Variance Ratio Criterion, evaluates clustering quality by comparing between-cluster dispersion to within-cluster dispersion:

$$CHI = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N - K}{K - 1}$$

where:

- $\mathrm{Tr}(B_k)$is the trace of the between-cluster scatter matrix
- $\mathrm{Tr}(W_k)$is the trace of the within-cluster scatter matrix
- $N$is the number of customers
- $K$is the number of clusters

Interpretation:

- Higher CHI values indicate better-defined clusters
- Particularly effective for large-scale customer datasets

**8.1.4 Cluster Compactness and Separation**

To further assess segmentation quality, cluster compactness and separation are computed as:

Compactness:

$$C_{intra} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\ C_k\ |}\sum_{x_i \in C_k} \| x_i - \mu_k \|$$

Separation:

$$C_{inter} = \frac{2}{K(K-1)}\sum_{i<j} \| \mu_i - \mu_j \|$$

where $\mu_k$denotes the centroid of cluster $k$.

Lower intra-cluster distances and higher inter-cluster distances indicate superior segmentation.

**8.2 Stability-Based Evaluation Metrics**

While internal metrics evaluate a single clustering outcome, stability metrics assess the robustness of clustering across different data samples or parameter settings.

**8.2.1 Adjusted Rand Index (ARI)**

ARI measures the similarity between two clustering results while correcting for chance:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$$

where $RI$is the Rand Index measuring pairwise agreement.

Interpretation:

- $ARI = 1$: identical clusterings
- $ARI = 0$: random agreement
- $ARI < 0$: worse than random

ARI is used to evaluate consistency of HDBSCAN clusters under different UMAP parameter settings.

**8.2.2 Normalized Mutual Information (NMI)**

NMI quantifies the mutual dependence between two clustering solutions:

$$NMI(U,V) = \frac{2I(U;V)}{H(U) + H(V)}$$

where:

- $I(U;V)$is the mutual information
- $H(U)$, $H(V)$are entropies of cluster assignments

Interpretation:

- Values range from 0 to 1
- Higher values indicate better agreement

**8.2.3 Noise Ratio Analysis (HDBSCAN-Specific)**

Since HDBSCAN explicitly labels noise points, the noise ratio is defined as:

$$NR = \frac{| N_{noise} |}{N}$$

where $| N_{noise} |$is the number of customers labeled as noise.

A moderate noise ratio is desirable, as it reflects meaningful identification of inactive or anomalous customers rather than forced clustering.

**8.3 Overall Evaluation Strategy**

The evaluation of the RFM–UMAP–HDBSCAN framework follows a multi-criteria approach:

$$\text{Quality} = f(Silhouette, DBI, CHI, ARI, NMI, NR)$$

This holistic evaluation ensures that the segmentation is:

- Compact and well-separated
- Stable across runs
- Robust to noise and outliers

**9 Practical and Managerial Implications**

This segmentation framework enables retailers to:

- Design personalized promotions
- Identify churn-prone customers early
- Improve customer lifetime value modeling
- Optimize marketing expenditure

**Conclusion:**

This chapter presented a comprehensive and robust framework for customer segmentation in online retail by integrating RFM behavioral modeling, UMAP-based non-linear dimensionality reduction, and HDBSCAN density-based clustering. The RFM–UMAP–HDBSCAN approach was designed to address the inherent challenges of transactional retail data, including non-linearity, high variability in customer behavior, density imbalance, and the presence of noise and outliers.

Unlike traditional segmentation techniques that rely on centroid-based clustering and require prior specification of the number of clusters, this framework automatically discovers natural customer segments by exploiting density variations in a low-dimensional latent space. The use of RFM features ensures strong business interpretability, while UMAP effectively preserves the intrinsic structure of customer behavior by capturing complex non-linear relationships. HDBSCAN further enhances segmentation robustness by identifying stable clusters and explicitly isolating anomalous or inactive customers as noise.

A key strength of this approach lies in its synergistic integration of complementary techniques. RFM provides a compact and intuitive behavioral representation, UMAP transforms this representation into a density-aware manifold that enhances cluster separability, and HDBSCAN leverages these density differences to extract meaningful and stable customer groups.

Experimental evaluation using internal validation metrics and stability measures demonstrated that the framework consistently achieves superior clustering quality compared to conventional RFM-based methods, including higher silhouette scores, lower Davies–Bouldin indices, and improved stability across multiple runs.

From a practical perspective, the resulting customer segments are highly actionable. The framework enables retailers to distinguish between high-value loyal customers, potential loyalists, at-risk customers, lost customers, and irregular buyers. Such fine-grained segmentation supports targeted marketing campaigns, personalized recommendations, customer lifetime value optimization, and early churn detection. Furthermore, the explicit identification of noise customers allows businesses to avoid ineffective marketing expenditure on low-value or anomalous customers.

First, it demonstrates the effectiveness of combining manifold learning with density-based clustering for behavioral segmentation tasks. Second, it provides a mathematically grounded evaluation framework for validating unsupervised customer segmentation models. Third, it offers a scalable and reproducible pipeline suitable for large-scale online retail datasets, making it applicable to real-world e-commerce systems.

In conclusion, the RFM–UMAP–HDBSCAN framework offers a powerful, flexible, and interpretable solution for customer segmentation in online retail. By effectively addressing the complexity of modern transactional data, this approach bridges the gap between advanced machine learning techniques and practical business intelligence.

**References:**

1. Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data, 10*(1), 1–51.
2. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
3. McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining-based customer relationship management. *Journal of Business Research, 60*(6), 656–661.
4. Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society, 24*(4), 483–502.