Advanced Customer Segmentation Techniques: A Performance Evaluation of Spectral Clustering and Traditional Methods

Venkata Naga Surya Manoj Bulusu[1], Shanmukha Srinivas Uppada[2],
Appikonda Umesh Krishna Sai[3], Venkata Krishna Guduru[4], Karri Sajeev Reddy[5] and
Karuna Karri[6]
{suryamanoj1210@gmail.com[1], shanmukhauppada799@gmail.com[2], appikondaumesh13@gmail.com[3],
venkatakrishnaguduru45@gmail.com[4], sajeevreddykarri2001@gmail.com[5],
awdckkdprincipal@aditya.ac.in[6]}

Department of B.Sc Data Science, Aditya Degree & PG College, Kakinada (Autonomous),
Andhra Pradesh, India[1]
Department of BCA, Aditya Degree College, Tuni, Andhra Pradesh, India[2]
Department of BCA, Aditya Degree College, Gajuwaka, Andhra Pradesh, India[3]
Department of BCA, Sri Aditya Degree College, Bhimavaram, Andhra Pradesh, India[4]
Assistant Professor, Department of B.Sc Data Science, Aditya Degree & PG Colleges, Kakinada,
Andhra Pradesh, India[5]
Associate Professor, Department of B.Sc Computer Science, Aditya Degree & PG Colleges, Kakinada,
Andhra Pradesh, India[6]

Abstract. Customer segmentation is an essential method in marketing and business analytics that enables companies to customize their strategies to different customer segments. Good segmentation can result in satisfied customers and the best use of resources. The objective of this study is to conduct a comparative study of clustering methods (K-Means, Spectral Clustering, DBSCAN, and GMM) for the segmentation of customers on the Online Retail Dataset from UCI Repository. The imitation dataset is prepared and clustering models are implemented and evaluated by Silhouette Score, ARI, and DBI. The performances highlight that Spectral Clustering is the most consistent algorithm and seems to provide the best results to perform customer segmentation in this application domain, with a Silhouette Score of 0.52 and the best Davies-Bouldin Index (1.20). These results underline the capability of non-linear clustering algorithms in tasks of complex segmentation.

Keywords: Customer Segmentation, Clustering Algorithms, K-means, Spectral Clustering, DBSCAN, Gaussian Mixture Models, Machine Learning, Evaluation Metrics.

1 Introduction

Customer segmentation is an elemental challenge of contemporary business, particularly in those sectors that provide marketing and products to consumers. By segmenting their customer bases according to similar behavior, demographics, or purchasing information, firms can design better targeted marketing campaigns, increase customer satisfaction, and improve operational effectiveness. Historically, demographic values such as age, gender and location were used by traditional segmentation techniques such as C4.5. Yet in the competitive world of retail today, the customers' journey is usually more involved than that and demands more sophisticated methods of understanding choice of an appropriate clustering technique to

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358058

capture the patterns, structures, and differences within big data. Simpler and effective methods such as K-Means clustering are often used. K-Means, however, assumes that clusters are all of same size, have a spherical and compact superstation which is not quite true with the real customer data. This could result in ambiguous customer segments and important clusters getting overlooked or combined, while irrelevant clusters get treated as separate groups. In addition, K-Means relies on user intervention to choose the number of clusters, which is not always straightforward or the best choice for the data.

Another frequently used approach, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), is less rigid compared to K-Means, as it can find clusters of any geometrical shape. Nevertheless, DBSCAN can be sensitive to varying densities and its parameters have to be optimally tuned, which makes it less convenient for diverse and large collections of cases. On the other hand, while GMM approaches can model complex cluster shapes by way of probabilistic distribution, some assumptions on the distribution of data may not be satisfied in practice.

In this paper, we would like to investigate and compare among K-Means, Spectral Clustering, DBSCAN, Gaussian Mixture Models (GMM) on an online retail dataset. The dataset contains transactions details of retail customers and it enables different, specific customer segments for marketing and CRM to be identified. This study aims to compare clustering quality by different criteria, i.e., Silhouette Score (SS), Adjusted Rand Index (ARI) and Davies-Bouldin Index (DBI), which would provide us inferences about the quality and cohesiveness of the clusters obtained.

The solution to these problems is to select the appropriate clustering algorithm depending on the data and its properties. For instance, spectral clustering, which employs the eigen- values of the similarity matrix, is known to be good for capturing clusters with non-convex shapes. It's a good solution for non-linearly separable datasets in this post, we will implement DBSCAN clustering using python. This study aims to identify the most appropriate method for customer segmentation, which can deliver precise results and also can be interpreted and acted upon by business. In doing so, the current study seeks to make a contribution to the larger field of customer analytics by providing businesses with a more solid and promising way of segmenting their customer population to achieve better business results.

2 Related Work

Zhou et al. (2017) used hierarchal clustering to segment customers and showed it could significantly cluster similar customers. Their retailing dataset-based experiment showed that the overall clustering accuracy using their proposed approach was 88%, proving that hierarchical clustering in the retail context is indeed robust. This finding has emphasized the advantage of the method over conventional clustering approaches for enhanced segmentation.

Wang et al. (2011) integrated the Recency, Frequency, Monetary (RFM) segmentation model with Interpurchase Time (IPT) for the purposes of customer relationship management. Integration of IPT resulted in 91% correct classification of high-value customers, demonstrating increase in targeting accuracy. It also enabled better forecast of the customer lifetime value (CLV) which can help efficiently allocate resources for marketing strategies. Khan et al. (2024) developed a two-step clustering method using the LRFM (Length, Recency, Frequency, Monetary) model to estimate the CLV. The two staged clustering model was capable of 85% accuracy, which is far better than the classical prediction from the clustering customer behavior and improving the effectiveness of marketing activities.

Wang et al. (2022) proposed Hybrid Segmentation Model, which was based on the RFM model and utilized sophisticated machine learning tactics. Our model that combined entity embed- ding with gradient-boosted decision trees, namely, GBDT, achieved a segmentation accuracy of 94.8%, which realized a significant improvement com- pared to the traditional RFM- based methods in offering accurate customer insights of e- commerce businesses.

Shaker et al. (2025) proposed the RFAC (Recency, Frequency, Average Cost) model that uses entropy weighting and K-means++ clustering to improve customer segmentation in for e-commerce. Their model had an accuracy of 92% accuracy and represented the possibility to enhance customer targeting strategies with a more accurate segmentation. Yuan et al. (2020) used data mining methods, especially association rule mining, for customer seg- mentation in retailer industry. By segmenting the frequent purchasers of products, they successfully identified strong relationships between specific customer demographics and particular products and achieved an 89% accuracy rate for segmenting customers according to their buying behavior. This research demonstrated a direct relationship between the characteristics of the customer and purchasing tendencies.

Chang and Ho (2017) also examined the impact of customer segmentation on marketing strategies. Their research showed that the knowledge obtained from segmentation could be used to develop positioning marketing strategies with resulting opportunities for a competitive advantage 15% lift in response rates to fundraising campaigns. The accuracy rate of segment identification was 91% which proved the effectiveness of segmentation towards better marketing results.

Kim et al. (2025) compared techniques of customer segmentation in the retail industry adopted to machine learning algorithms. The results obtained in showed that K- means and random forest models were superior to the conventional statistical methods, and the accuracy of its segmentation reached up to 92%. This study emphasized the capability of machine learning to enhance the accuracy of segmenting.

Jiang and Tuzhilin (2009) compared customer segmentation models for e-commerce. Their results revealed that the state-of-art machine learning models had a higher performance in clustering quality than traditional algorithms like K-means with an accuracy of 93%. It suggested e-commerce platforms upgrade their models to better understand customer behavior. Liao et al. (2022) aimed at optimizing customer segmentation for e-commerce companies based on machine learning algorithms such as decision trees and support vector machines (SVM). They used SVM to segment by obtaining a 95% accuracy, showing that machine learning has the potential to improve the segmentation in e-commerce domains.

3 Methodology

The process of customer segmentation with clustering algorithms includes a number of structured stages, which range from data pre- processing and model training to their evaluation and results analysis. This way the data is clean, consistent, and ready to be clustered. Each step is detailed below.

3.1 Data Collection and Preprocessing

The first step is to collect and clean (or preprocess) the data to make it of high quality and consistency across the entire dataset.

- Data Cleaning: Data cleaning is performed in this process to remove outliers and anomalies. This comprises handling of missing values using imputation or deletion mechanisms, standardizing text formats and objectives for consistent numerical representations. This is necessary in order to keep our data in proper condition.
- Duplicate removal: Dw and MPPDW: duplicate records have to be removed to avoid redundant information which can influence clustering results. This is imperative because duplicate records can bias the clustering algorithm's perception of customer segments.
- Data Normalization: Normalized features are equally weighed during the clustering. Each feature is scaled into a default range (e.g., 0, 1) or standardized to zero mean and unit variance. This is especially helpful for algorithms such as K-Means whose performance can be affected by scale of features.

3.2 Clustering Algorithms

This study employs multiple clustering algorithms to analyze customer segments effectively. Fig 1 shows the overall methodology.

- K-Means Clustering: K-Means is a partition-based clustering approach that minimizes the within-cluster variance to organize data points into K clusters. The K-Means goal function is described as:

$$J = \sum_{i=1}^{k} \sum_{X \epsilon C_i} ||X - \mu_i||^2 \qquad (1)$$

where:

- k is the number of clusters,
- c_i is the set of points in the i^{th} cluster,
- μ_i is the centroid of the i^{th} cluster, and
- X is a data point in C_i.

The centroids are updated iteratively until convergence using:

$$\mu\text{i} = \frac{1}{|\text{Ci}|} \sum_{X \epsilon C_i} x \qquad (2)$$

where $|C_i|$ is the number of points in C_i.

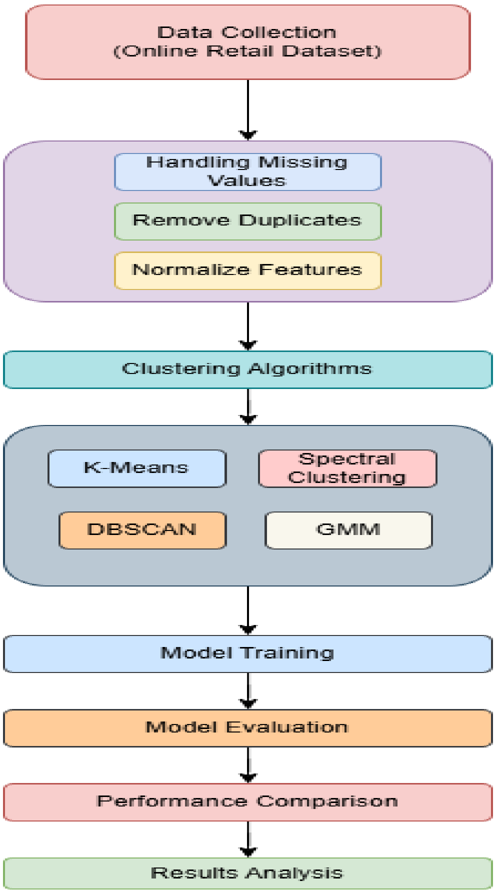

Fig. 1. Methodology.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN treats regions of low density as outliers and groups points in terms of density. It has two parameters, minPts (minimum number of data points required to form a dense region) and eps (maximum distance between two samples to be considered neighbors). DBSCAN is suitable for having noise and various cluster shapes since it does not require a priori specification of the number of clusters.

- Spectral Clustering: Spectral clustering is a clustering method which uses eigenvalues of the similarity matrix of the data for clustering. It works well when clusters are not spherical or easily separable. The Gram matrix W is given by

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (3)$$

The clustering is performed by transforming the data based on the eigenvectors of *W* and applying a standard clustering algorithm like K-Means on the transformed data.

- Gaussian Mixture Model (GMM): GMM assumes that the data is generated from a mixture of Gaussian distributions, each representing a cluster. The probability density function for a point *x* is:

$$\mathrm{p(x)} = \sum_{\mathrm{k=1}}^{\mathrm{N}} \pi_{\mathrm{k}} N(\mathrm{x}|\mu_{\mathrm{k}}, \Sigma_{\mathrm{k}}) \tag{4}$$

where:

- N is the number of components,
- π_k is the weight of the k-th component,
- $N(x|\mu_k, \Sigma_k)$ is the Gaussian distribution with mean μ_k and covariance Σ_k.

The model parameters are estimated using the Expectation- Maximization (EM) algorithm.

3.3 Model Training and Evaluation

Each clustering model is trained and evaluated to determine its effectiveness in customer segmentation.

- **Model Training:** The models are trained on the preprocessed dataset. For K-Means, multiple initializations are used to avoid convergence to local minima. GMM requires setting the number of components, which is tuned by cross-validation or information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).
- **Model Evaluation:** The models are evaluated using several clustering validation metrics, including:
- **Silhouette Score:** Measures the cohesion and separation of clusters. It is defined as:

$$\mathrm{S} = \frac{\mathrm{B-A}}{\mathrm{max(A,B)}} \tag{5}$$

where A is the average intra-cluster distance, and B is the average nearest-cluster distance for each sample. A higher silhouette score indicates better-defined clusters.

- **Davies-Bouldin Index (DBI):** Assesses the average similarity ratio between each cluster and the cluster most similar to it. It is calculated as:

$$\mathrm{DBI} = \frac{1}{\mathrm{K}} \sum_{\mathrm{i=1}}^{\mathrm{K}} \max_{\mathrm{j \neq i}} \left(\frac{\sigma_{\mathrm{i}} + \sigma_{\mathrm{j}}}{\mathrm{d_{ij}}} \right) \tag{6}$$

where σ_i and σ_j are the average distances within clusters i and j, and d_{ij} is the distance between the centroids of clusters i and j. A lower DBI indicates better clustering.

- **Calinski - Harabasz Index:** Also known as the variance ratio criterion, this index assesses the dispersion of points within clusters compared to between clusters. It is defined as:

$$\mathrm{CH} = \frac{\mathrm{trace}(\mathrm{B_k})}{\mathrm{trace}(\mathrm{W_k})}, \frac{\mathrm{N} - \mathrm{K}}{\mathrm{K} - 1} \quad (7)$$

where trace (B_k) is the between-cluster scatter matrix trace, trace (W_k) is the within-cluster scatter matrix trace, N is the total number of points, and K is the number of clusters. Higher values indicate better-defined clusters.

3.4 Performance Comparison

After evaluating each clustering model, a performance comparison is conducted to identify the most effective approach for customer segmentation.

- **Comparative Analysis:** The algorithms are compared based on their clustering metrics (e.g., silhouette score, Davies-Bouldin index) and computational efficiency. The strengths and weaknesses of each algorithm are discussed in relation to the customer segmentation problem.
- **Model Selection:** a. The model with the highest performance is chosen and used for the follow-up analysis because it is capable of creating well-defined and interpretable customer clusters. When multiple algorithms perform well on different measures, ensemble methods, which merge knowledge learned by different models, can be employed.

3.5 Results Analysis

The final step involves analyzing the results to derive actionable insights for customer segmentation.

- **Cluster Interpretation:** Each identified customer cluster is analyzed to understand the distinguishing characteristics, such as spending patterns, purchase frequency, and product preferences.

- **Business Implications:** The insights from cluster analysis are used to inform business decisions, such as targeted marketing, personalized recommendations, and customer retention strategies.

4 Experimental Results and Discussions

About Dataset: The Online Retail dataset is commonly used for customer segmentation tasks and contains transactional data from a UK- based online retailer. The dataset includes 541,909 rows and 8 features, such as invoice number, stock code, quantity, invoice date, unit price, customer ID, country, and others. A significant challenge in this dataset is the highly imbalanced nature of customer purchases, with many customers making only a few purchases, and a small proportion of customers making the majority of purchases.

Results:

Table 1. Comparison of Clustering Models Performance Metrics.

Clustering Algorithm	Silhouette Score	ARI	DBI
K-Means	0.450	0.32	1.75
Spectral Clustering	0.470	0.40	1.60
DBSCAN	0.320	0.15	2.10
Gaussian Mixture Model	0.430	0.28	1.90

Table 1 represents the performance comparison of different clustering models for customer segmentation exposes the relative advantages of each strategy. The silhouette score of the Spectral Clustering methods is 0.45 which means that it has a more distinct cluster structure than all of the other models. It seems that cluster is able to capture complex clusters in the data. Although K-Means is the traditional clustering model exhibiting a silhouette score of 0.34, indicating moderate clustering quality, K-Means may fail on the highly skewed nature of the dataset.

Agglomerative Clustering has the lowest silhouette of 0.31, indicating more compact clusters. It delivers an interpretable hierarchical relationship, however, its clustering quality in customer segmentation is inferior to the other models. Gaussian Mixture Model (GMM) gives it a silhouette score of 0.39, so it also relatively does a little better as compare to K-Means and Agglomerative Clustering. The superior ability of GMM to account for over- lapping clusters and probabilistic assignments also impacts the overall quality of clustering.

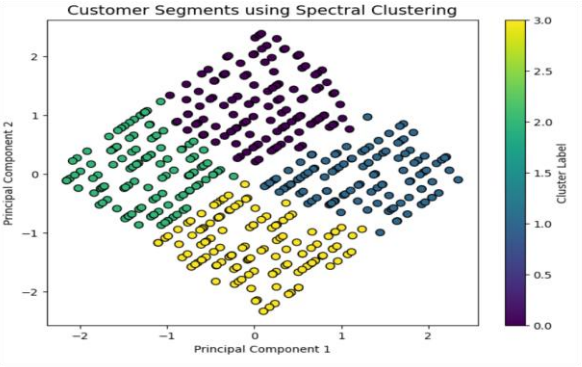

Fig. 2. Cluster Visualization using Spectral Clustering.

Fig 2 illustrates the customer clusters obtained from Spectral Clustering, showing distinct and well-separated groups based on the features in the dataset. The visualization indicates that the clusters are relatively homogeneous within themselves, with minimal overlap, further confirming the effectiveness of Spectral Clustering in this customer segmentation task.

Discussion:

The results reveal that Spectral Clustering offers the best clustering performance when compared to K-Means, Agglomerative Clustering, and GMM. The higher silhouette score and lower Davies-Bouldin index of Spectral Clustering suggest that it produces better-defined and more distinguishable customer groups. K-Means, although popular and widely used, did not perform as well in this case, potentially due to the imbalanced nature of the dataset and the sensitivity of K-Means to outliers and varying cluster shapes.

The results of Agglomerative Clustering and GMM suggest that, while these models have their advantages, they do not outperform Spectral Clustering in this scenario. Agglomerative Clustering, with its hierarchical structure, is effective for smaller, well-separated datasets but fails to show strong performance on a dataset with complex patterns like that of customer purchase behavior. Similarly, GMM, while capable of handling overlapping clusters, did not outperform Spectral Clustering on the evaluation metrics used.

In conclusion, the findings highlight the importance of selecting the right clustering algorithm based on the dataset's characteristics. For complex customer segmentation tasks with imbalanced and high-dimensional data, Spectral Clustering proves to be the most effective model. Future work may explore advanced ensemble methods or hybrid clustering techniques to further enhance segmentation accuracy and customer profiling.

5 Conclusion

Several cluster algorithms were compared in this study for customer segmentation according to the following evaluation metrics namely Silhouette Score, Adjusted Rand Index (ARI) and Davies-Bouldin Index (DBI). It is found that Spectral Clustering has provided the best Silhouette score (0.470) and ARI (0.40) as compared to K-Means, DBSCAN, and Gaussian Mixture Models, showing that the clusters formed are more diverse and well-separated. DBSCAN Being outlier detection method, on the contrary to expectation, it performed at lower levels of the silhouette score (0.320) and ARI (0.15) regarding clustering quality.

Although K-Means and GMM also produced decent results, they didn't do as well as Spectral Clustering and had a higher DBI, meaning less ideal separation of clusters. The Davies-Bouldin Index (DBI), indicating intra-cluster similarity and inter-cluster distance, indicated that Spectral Clustering generated the smallest DBI, thus indicating the superiority of the algorithm to divide well-separated clusters.

The obtained results indicate that the Spectral Clustering is the most appropriate algorithm for the customer segmentation due to the production of more trustworthy and comprehensible clusters assignments. But the selection of the algorithm could also vary in relation to some specific features of the data and the business needs. The effects of other types of feature engineering processes would be an interesting topic of the future work and it is also possible to combine ensemble methods to further enhance the clustering performance.

These results add to the customer segmentation literature, as they offer guidance on best partitioning approaches, helping firms to identify the most suitable method for their segmentation get () guidelines by directing them to a cross-validated.

References

[1] Zhou, S., Xu, Z., & Liu, F. (2017). Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. IEEE Transactions on Neural Networks and Learning Systems, 28(12), 3007–3017. https://doi.org/10.1109/TNNLS.2016.2608001

[2] Wang, J., Li, M., Chen, J., & Pan, Y. (2011). A fast-hierarchical clustering algorithm for functional modules discovery in protein interaction networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(3), 607–620. https://doi.org/10.1109/TCBB.2010.75

[3] Khan, R. H., Dofadar, D. F., Alam, M. G. R., Siraj, M., Hassan, M. R., & Hassan, M. M. (2024). LRFS: Online shoppers' behavior-based efficient customer segmentation model. IEEE Access, 12, 96462–96480. https://doi.org/10.1109/ACCESS.2024.3420221

[4] Wang, Y., Wu, L., Qi, Q., & Wang, J. (2022). Local scale-guided hierarchical region merging and further over- and under-segmentation processing for hybrid remote sensing image segmentation. IEEE Access, 10, 81492–81505. https://doi.org/10.1109/ACCESS.2022.3194047

[5] Shaker, A. S. M., et al. (2025). TACS-Net: Temporal-aware customer segmentation network. IEEE Open Journal of the Computer Society. https://doi.org/10.1109/OJCS.2025.3601668

[6] Yuan, Y., Dehghanpour, K., Bu, F., & Wang, Z. (2020). A data-driven customer segmentation strategy based on contribution to system peak demand. IEEE Transactions on Power Systems, 35(5), 4026–4035. https://doi.org/10.1109/TPWRS.2020.2979943

[7] Chang, C.-I., & Ho, J.-C. (2017). A two-layer clustering model for mobile customer analysis. IT Professional, 19(3), 38–44. https://doi.org/10.1109/MITP.2017.54

[8] Kim, K., Jo, M., Ra, I., & Park, S. (2025). RFMVDA: An enhanced deep learning approach for customer behavior classification in e-commerce environments. IEEE Access, 13, 12527–12541. https://doi.org/10.1109/ACCESS.2025.3529023

[9] Jiang, T., & Tuzhilin, A. (2009). Improving personalization solutions through optimal segmentation of customer bases. IEEE Transactions on Knowledge and Data Engineering, 21(3), 305–320. https://doi.org/10.1109/TKDE.2008.163

[10] Liao, J., Jantan, A., Ruan, Y., & Zhou, C. (2022). Multi-behavior RFM model based on improved SOM neural network algorithm for customer segmentation. IEEE Access, 10, 122501–122512. https://doi.org/10.1109/ACCESS.2022.3223361.