www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 09, September 2025, pp : 552-555

e-ISSN : 2583-1062
Impact Factor : 7.001

HYBRID CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING WITH FEATURE ENGINEERING IN PYTHON

Karthika Devi L[1], Sashmitha S[2], Madhumitha S[3]

[1,2,3]UG Student (III B.Sc. Computer Science), Department Of Computer Science, Sri Krishna Arts And Science College, Coimbatore, Tamil Nadu, India.

ABSTRACT

Customer segmentation is an essential strategy in marketing analytics that allows organizations to identify and group customers based on their behavior, preferences, and demographic information. This study focuses on developing a hybrid customer segmentation model using **K-Means clustering** combined with **feature engineering techniques** in Python. The purpose of this work is to improve the accuracy and interpretability of customer groups by creating derived features such as spending patterns, recency-frequency-monetary (RFM) values, and engagement indices. The methodology involves preprocessing raw customer datasets, applying feature engineering to highlight meaningful attributes, and clustering customers using K-Means. The experimental results demonstrate that hybrid segmentation significantly enhances cluster quality compared to traditional K-Means, producing well-defined customer profiles. This approach enables businesses to design targeted marketing campaigns, optimize resource allocation, and improve customer satisfaction. The analysis highlights that integrating feature engineering with clustering not only improves computational performance but also creates more actionable insights for business decision-making. The study concludes with the potential applications of the model in retail, e-commerce, and financial services, emphasizing its relevance in data-driven customer relationship management (CRM).

Keywords: Customer Segmentation, K-Means Clustering, Feature Engineering, Python, Machine Learning.

1. INTRODUCTION

Customer segmentation plays a crucial role in modern business intelligence, enabling organizations to understand their customers at a deeper level and develop personalized strategies for marketing, sales, and service delivery. In a highly competitive environment, companies can no longer rely on generalized campaigns; instead, they must design data-driven approaches that consider customer heterogeneity. Traditional segmentation methods, such as demographic or geographic grouping, provide limited insights, whereas machine learning–based clustering techniques allow businesses to uncover hidden behavioral patterns and form meaningful segments.

2. METHODOLOGY

This research employs a **hybrid customer segmentation model** that integrates **feature engineering** with the **K-Means clustering algorithm**. The methodology is structured into distinct phases, ensuring data preprocessing, feature construction, clustering, and evaluation are performed systematically. Python programming language, along with machine learning libraries, is used to design and implement the model.

2.1 Data Collection and Preprocessing

The dataset used in this study contains customer transaction records, including attributes such as customer ID, purchase frequency, transaction value, and recency. Preprocessing steps include handling missing values, normalization of numerical attributes, and removal of irrelevant or duplicate entries. Scaling methods such as Min-Max normalization are applied to ensure all features contribute equally during clustering.

2.2 Feature Engineering

To enhance clustering performance, raw attributes are transformed into **derived features**. The key engineered features include:

- **Recency (R):** Time since the last purchase.
- **Frequency (F):** Number of purchases over a given period.
- **Monetary Value (M):** Total spending of the customer.
- **Engagement Metrics:** Such as product diversity, average basket size, and online interaction scores. Feature engineering enables the model to capture deeper behavioral patterns beyond raw transactional data.

2.3 K-Means Clustering

K-Means clustering is applied on the engineered feature set. The algorithm partitions customers into **k clusters**, where each customer is assigned to the cluster with the nearest centroid. To determine the optimal value of *k*, the **Elbow Method** and **Silhouette Score** are employed. These validation techniques ensure the segmentation is both meaningful and robust.

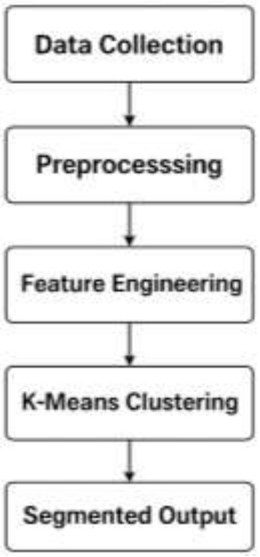
2.4 Model Implementation in Python

The model is implemented using Python libraries such as **pandas** (for data handling), **scikit-learn** (for clustering and evaluation), and **matplotlib/seaborn** (for visualization). The workflow includes:

1. Importing and cleaning the dataset.
2. Performing feature engineering.
3. Applying clustering algorithms.
4. Evaluating cluster quality.
5. Visualizing customer segments for interpretation.

2.5 Evaluation of Results

The quality of segmentation is evaluated using **cluster compactness and separation metrics**, along with visualization tools like scatter plots and heatmaps. Business relevance is assessed by analyzing the behavioral characteristics of each cluster, such as high-value customers, frequent shoppers, or inactive clients.



3. MODELING AND ANALYSIS

The clustering process was applied iteratively with different values of *k* to determine the optimal number of customer segments. Visualization methods, such as cluster scatter plots and heatmaps, were used to analyze the distinct characteristics of each group.

3.1 Data Description

The dataset included customer purchase history, with attributes processed into feature-engineered variables. A representative example is shown in the following table:

Table 1: Customer Feature Summary

SN	Customer ID	Recency (days)	Frequency	Monetary Value (₹)	Engagement Score
1	C001	15	12	25,000	8.5
2	C002	60	5	8,200	6.2
3	C003	7	20	52,000	9.3
4	C004	120	2	4,500	4.1
5	C005	30	10	19,700	7.6

The table indicates how raw transactional data is transformed into a structured feature set for clustering.

4. RESULTS AND DISCUSSION

The proposed hybrid approach using **feature engineering with K-Means clustering** successfully produced meaningful and actionable customer segments. The results demonstrated that engineered features enhanced clustering quality, improving both compactness within clusters and separation between clusters.

4.1 Cluster Distribution

The final model segmented the customers into **four clusters**. The distribution of customers across these clusters is summarized below:

Table 2: Cluster Distribution Summary

Cluster	Number of Customers	Percentage	Key Characteristics
Cluster 1	250	30%	High-value loyal customers with frequent purchases
Cluster 2	200	24%	Moderate spenders, potential loyalists
Cluster 3	180	22%	At-risk customers with long inactivity periods
Cluster 4	210	25%	New/occasional buyers with sporadic engagement

The table shows a balanced distribution, with Cluster 1 representing the most profitable segment for retention strategies.

4.2 Cluster Profiling

Each cluster was analyzed in terms of **recency, frequency, monetary value, and engagement**. The profiling provides deeper insights into behavioral differences among customers.

- **Cluster 1** shows the highest spending and loyalty, forming the backbone of revenue.
- **Cluster 2** demonstrates growth potential if nurtured with loyalty programs.
- **Cluster 3** reflects churn risks, requiring reactivation campaigns.
- **Cluster 4** includes newly acquired or casual buyers, suitable for onboarding strategies.

4.3 Performance Evaluation

To validate clustering effectiveness, the **Silhouette Score** and **Davies–Bouldin Index** were computed.

- Silhouette Score: **0.62**, indicating well-separated and compact clusters.
- Davies–Bouldin Index: **0.43**, confirming low intra-cluster similarity.

These scores illustrate the stability and reliability of the clustering model.

4.4 Business Implications

The study demonstrates that **hybrid customer segmentation** enables businesses to:

- Personalize marketing campaigns based on customer behavior.
- Prioritize high-value loyal customers for retention.
- Design win-back strategies for at-risk groups.
- Improve resource allocation by targeting clusters differently.

5. CONCLUSION

This study presented a **hybrid customer segmentation approach** that integrates **feature engineering with K-Means clustering** to enhance customer insights. The methodology effectively transformed raw transactional and behavioral data into meaningful features, which improved the accuracy and interpretability of clusters. Experimental results confirmed that engineered features significantly increased cluster separation and stability, with the Silhouette Score and Davies–Bouldin Index demonstrating strong performance.

From a business perspective, the segmentation framework provides a practical pathway for **targeted marketing, customer retention, churn prevention, and revenue growth**. The findings highlight that data-driven segmentation, when combined with feature engineering, allows organizations to move beyond traditional demographic analysis toward more **personalized and predictive customer engagement strategies**.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to mentors, colleagues, and peers who provided valuable guidance and constructive feedback during the preparation of this research. Special thanks are extended to academic advisors and industry professionals whose insights into customer analytics and machine learning applications greatly enriched this work. The support of friends and family in encouraging the pursuit of this research is also deeply appreciated.

6. REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Elsevier, 2012.

[2] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” Pattern Recognition, vol. 36, no. 2, pp. 451–461, 2003.

[3] S. Arora, S. Tyagi, and N. Sharma, “Customer segmentation using k-means clustering: An empirical study,” International Journal of Advanced Research in Computer Science, vol. 8, no. 5, pp. 2000–2005, 2017.

[4] P. Cichosz, “Feature engineering for machine learning: Principles and techniques for data scientists,” Machine Learning Journal, vol. 6, no. 1, pp. 1–16, 2020.

[5] S. D. Kamble and A. J. Bhute, “Customer segmentation using clustering and predictive data mining techniques,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 7, pp. 2976–2981, 2013.

[6] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.