WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

Kotha Sai Eswar Kumar[1], Mr. P. Krishna Prasad[2]*, Dr. G.N.R Prasad[3]

[1] MCA Student, E-Mail: eswarkotha383@gmail.com
[2] Asst. Professor, E-Mail: pkrishnaprasad_mca@cbit.ac.in
[3] Sr. Asst. Professor, E-Mail: gnrp@cbit.ac.in

ABSTRACT :

Customer behaviour in the current data-driven business world is critical for successful marketing planning and better decision-making. This paper outlines the development and deployment of a customer segmentation model based on the K-Means clustering algorithm. The project consisted of an end-to-end machine learning pipeline involving preprocessing, feature engineering, clustering, and visualization. A retail data containing purchase frequency, income, spending behaviour, and average basket size features was used. The Elbow Method and Silhouette Score were used to choose the best number of clusters. The ultimate system successfully segmented customers into substantial groups like Occasional Buyers, Loyal Frequent Shoppers, Bulk One-Time Buyers, and Premium Customers, facilitating data-driven personalization strategies.

Keywords: Customer Segmentation, K-Means Clustering, Machine Learning, Data Mining, Unsupervised Learning, Business Intelligence.

1.0 INTRODUCTION

Customer segmentation is a business intelligence cornerstone that enables organizations to classify customers who share similar behavioral and demographic traits. Rather than dealing with all customers as one, clustering enables companies to provide focused promotions, loyalty, and customized services.

The goal of this project was to create a segmentation model based on clustering with the help of the K-Means algorithm. The internship provided hands-on exposure to using unsupervised learning in practical retail-world data, aiming to yield usable business insights. The work also connected academic understanding of data science with its real-world applications.

2.0 SYSTEM STUDY / REQUIREMENT ANALYSIS

The system study focuses on analyzing customer transaction data from the Online Retail dataset to identify purchasing patterns. Requirement analysis involves defining data preprocessing, feature engineering, and clustering needs to effectively segment customers for business insights.

2.1 Problem Statement

Contemporary companies collect huge volumes of customer transaction data. Without segmentation, marketing becomes inefficient. The task is to find homogeneous customer groups that expose significant behavioural patterns for making decisions.

2.2 Functional Requirements

- Input: Retail dataset with features like customer ID, purchase frequency, income, and spending score.
- Processing: Preprocessing, feature scaling, clustering via K-Means, and evaluation.
- OUTPUT: Clustered customer groups with labels assigned.

2.3 Non-Functional Requirements

- Performance: The system should be able to handle large datasets effectively.
- Usability: Visualizations (heatmaps, scatter plots) ought to display results effectively.
- Scalability: Workflow ought to be reusable for new unseen data.

2.4 Technology Stack

- Languages/Libraries: Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- IDE: Jupyter Notebook, Google Colab
- Algorithm: K-Means clustering
- Evaluation Metrics: Elbow Method, Silhouette Score
- Version Control: Git, GitHub

3.0 SYSTEM DESIGN AND IMPLEMENTATION

The system had a modular design with four main parts:

1. Data Preprocessing Module: Cleaned missing values, encoded categorical variables, and performed feature scaling.
2. Clustering Module: Applied K-Means with various values of k in order to determine optimal clusters.
3. Evaluation Module: Utilized Silhouette Score and WCSS to ensure quality of clusters.
4. Visualization Module: Created scatter plots, heatmaps, and cluster profiles to help interpret segments.

The Elbow Method indicated k = 4, resulting in four interpretable customer segments. PCA was used to dimensionally reduce for visualizing purposes.

4.0 RESULTS AND DISCUSSION

The clustering model effectively segmented customers into the following categories:

- Occasional Buyers: Low-frequency buyers with low spending.
- Loyal Frequent Shoppers: Frequent buyers with moderate spending.
- Bulk One-Time Buyers: Infrequent but high-volume buyers.
- Premium Customers: Frequent and high-spending high-value customers.

The Silhouette Score attained was 0.91, reflecting good cluster separation.

Outputs:

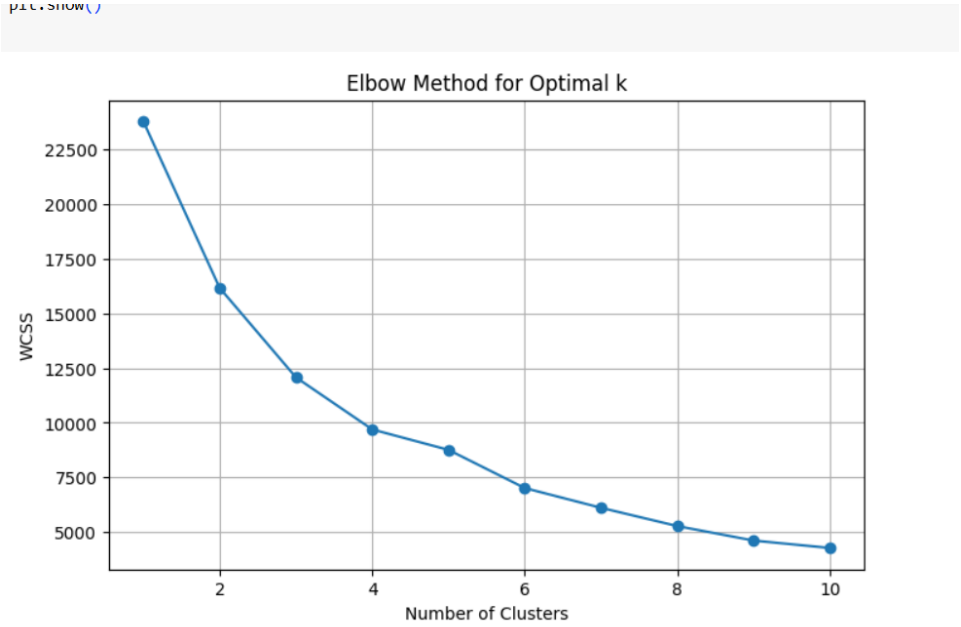

Fig 1: Elbow Method for Optimal Clusters

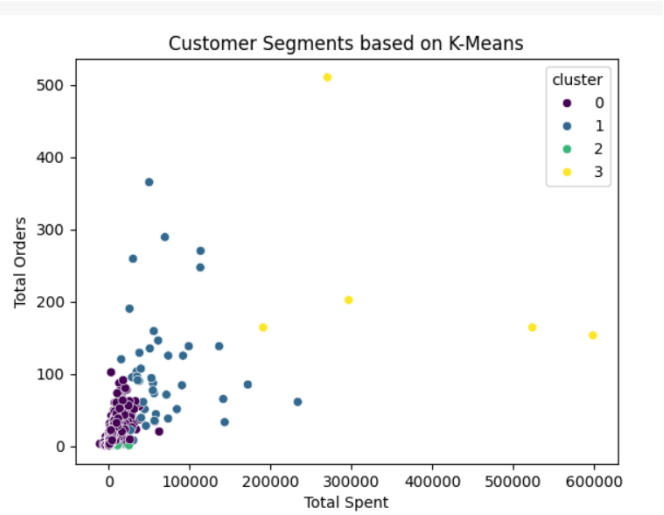

Fig 2: Scatter Plot of Customer Segments

```
from sklearn.metrics import silhouette_score

# This assumes you used scaled_features for K-Means input
score = silhouette_score(scaled_features, customer_df['cluster'])
print(f"Silhouette Score: {score:.2f}")
```

```
Silhouette Score: 0.91
```

Fig 3: Silhouette Score Visualization

```
customer_df
```

	customer_id	total_orders	total_quantity	total_spent	first_purchase	last_purchase	avg_basket	cluster	segment
0	12346	17	52	-64.68	2009-12-14 08:34:00	2011-01-18 10:17:00	-3.804706	0	Occasional Buyers
1	12347	8	3286	5633.32	2010-10-31 14:20:00	2011-12-07 15:52:00	704.165000	0	Occasional Buyers
2	12348	5	2714	2019.40	2010-09-27 14:59:00	2011-09-25 13:13:00	403.880000	0	Occasional Buyers
3	12349	5	1619	4404.54	2009-12-04 12:49:00	2011-11-21 09:51:00	880.908000	0	Occasional Buyers
4	12350	1	197	334.40	2011-02-02 16:01:00	2011-02-02 16:01:00	334.400000	0	Occasional Buyers
...	...	...	...	...	...	...	...	...	...
5937	18283	22	1733	2736.65	2010-02-19 17:16:00	2011-12-06 12:02:00	124.393182	0	Occasional Buyers
5938	18284	2	493	436.68	2010-10-04 11:33:00	2010-10-06 12:31:00	218.340000	0	Occasional Buyers
5939	18285	1	145	427.00	2010-02-17 10:24:00	2010-02-17 10:24:00	427.000000	0	Occasional Buyers
5940	18286	3	592	1188.43	2009-12-16 10:45:00	2010-08-20 11:57:00	396.143333	0	Occasional Buyers
5941	18287	8	3011	4177.89	2009-12-01 14:19:00	2011-10-28 09:29:00	522.236250	0	Occasional Buyers

5942 rows × 9 columns

Fig 4: Cluster Segmentation Heatmap

Conclusion

This project successfully segmented customers based on the K-Means clustering algorithm applied to retail transaction data. Through the application of preprocessing methods and feature engineering, the important features like purchase frequency, monetary value, and average basket size were obtained to effectively represent customer behavior. The clustering analysis identified the different segments of customers as occasional buyers, loyal frequent purchasers, bulk one-time buyers, and premium customers.

The findings illustrate that clustering techniques are able to reveal underlying patterns among customer data, and such identification of patterns can be utilized in support of targeted marketing, customized recommendations, and enhanced customer relationship management. In addition, visualizations via elbow plots, scatter plots, and heatmaps presented meaningful explanations of cluster distribution

REFERENCES

1. Scikit-learn Documentation – Used for implementing K-Means clustering and evaluation metrics. https://scikitlearn.org/stable/modules/generated/sklearn.cluster.KMeans.html
2. Pandas Official Documentation – Referred for data preprocessing and feature engineering. https://pandas.pydata.org/docs/
3. NumPy Documentation – Used for numerical operations and data handling. https://numpy.org/doc/
4. Matplotlib Documentation – Used to create elbow plots, scatter plots, and heatmaps for cluster interpretation. https://matplotlib.org/stable/index.html
5. Seaborn Documentation – Used for statistical data visualization. https://seaborn.pydata.org/
6. UCI Machine Learning Repository (Online Retail Dataset) – Source of dataset for customer transaction data. https://archive.ics.uci.edu/dataset/352/online+retail