Customer Segmentation for an Online Store using K-Means Clustering

1st Md Limon Mia
dept. Computer Science and Engineering
Dhaka International University
Dhaka, Bangladesh
mdhlimonmia@gmail.com

2nd Ananna Rani Dash
dept. Computer Science and Engineering
Dhaka International University
Dhaka, Bnagladesh
dristy202008@gmail.com

***Abstract*—Customer segmentation is a powerful strategy used in e-commerce to enhance personalized marketing and increase customer retention. This research focuses on implementing K-Means clustering to identify distinct customer groups based on their behavioral attributes such as annual income and spending patterns. Using a real-world e-commerce dataset, data preprocessing and scaling techniques were applied to prepare the data for clustering. The Elbow Method was employed to determine the optimal number of clusters. The results provided clear segmentation of customers into groups, helping businesses target each group more effectively. This paper demonstrates how machine learning techniques can be applied to customer data to produce actionable business insights.**

***Index Terms*—Customer segmentation, K-Means, Machine Learning, E-Commerce, Clustering, Scikit-learn**

I. Introduction

The rapid growth of e-commerce has significantly increased the volume of customer data collected by online retailers. Understanding and leveraging this data has become essential for businesses to remain competitive. One effective approach is customer segmentation, which involves grouping customers based on shared characteristics or behaviors. This process allows businesses to tailor marketing efforts, personalize user experiences, and improve customer satisfaction.

Customer segmentation can be achieved using various statistical and machine learning techniques. Among them, K-Means clustering is widely adopted due to its simplicity, efficiency, and interpretability.

II. Literature Review

Several studies have highlighted the importance of customer segmentation in improving marketing performance and customer satisfaction. Traditional segmentation approaches rely on demographic data or manual classification. However, with the advancement of machine learning, unsupervised clustering algorithms such as K-Means, DBSCAN, and hierarchical clustering have gained popularity.

K-Means is particularly effective in scenarios where the number of segments is known or can be estimated. The algorithm partitions customers into k groups by minimizing intra-cluster variance. Research has shown that this method yields meaningful insights into customer behavior.

III. Methodology

This study uses the K-Means clustering algorithm to identify customer segments in an online store. The analysis was conducted using Python libraries such as `pandas`, `scikit-learn`, and `matplotlib`. The methodology followed these steps:

1) **Data Collection**: A public e-commerce dataset containing features such as ‘CustomerID‘, ‘Age‘, ‘Annual Income (k$)‘, and ‘Spending Score (1–100)‘ was used.
2) **Dataset Description**: The dataset used in this study contains anonymized records of customers from an e-commerce platform. Key attributes include:
   - Customer ID
   - Gender
   - Age
   - Annual Income
   - Spending Score
3) **Data Preprocessing**:The dataset was cleaned by removing missing values and irrelevant columns. Actually this dataset already perfect but I also checked. Here preprocessing steps included:
   - Handling missing values
   - Selecting relevant features: **Annual Income** and **Spending Score**
   - Standardizing features using **StandardScaler** from **scikit-learn**
4) **Feature Selection**: Only numerical attributes (‘Annual Income‘, ‘Spending Score‘) were selected for clustering.
5) **Feature Scaling**: To ensure all features contribute equally to the distance calculations in **K-Means**, feature scaling was applied:

```
from sklearn.preprocessing import
 StandardScaler
scaler = StandardScaler()
x_scaled = scaler.fit_transform(x)
```

6) **Determining the Number of Clusters**: The Elbow Method was used to determine the best number of clusters (k$):

```
# Elbo Method to find optimal k
arr = []
```

Identify applicable funding agency here. If none, delete this.

```
for k in range(1, 11):
 kmeans = KMeans(n_clusters=k,
  init='k-means++', random_state=42)

 kmeans.fit(x_scaled)
 arr.append(kmeans.inertia_)
 plt.figure(figsize=(10, 5))
 plt.plot(range(1, 11), arr,
  marker='o',color='b',
  linestyle='-', markersize=6)
plt.title('Elbow Method to Determine
  Optimal k')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.grid(True)
plt.show()
```

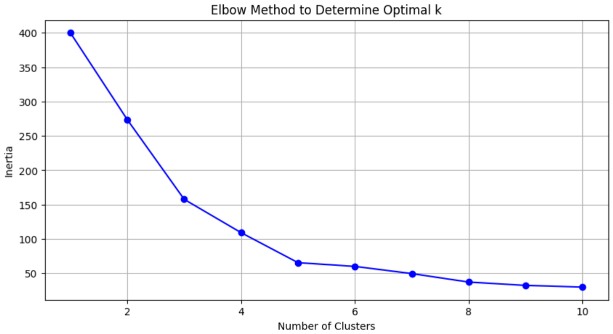

Fig. 1. Elbow Method for optimal number of clusters

7) **Applying K-Means Clustering**: K-Means clustering was applied to the data. Based on the Elbow Method, 5 clusters were selected:

```
optimal_k = 5
kmeans = KMeans(n_clusters= optimal_k,
 init = 'k-means++', random_state= 42)
y = kmeans.fit_predict(x_scaled)
```

8) **Visualization**: Cluster results were visualized using a 2D scatter plot.

```
cluster_labels = {
0: 'Low Income, Low Spending',
1: 'High Income, High Spending',
2: 'Moderate Income, High Spending',
3: 'High Income, Low Spending',
4: 'Moderate Income, Low Spending'
}

plt.figure(figsize=(10, 6))
for lbl, name in cluster_labels
   .items():
 plt.scatter(x_scaled[y == lbl, 0],
 x_scaled[y == lbl, 1],  label=name,
 s = 40)
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.title('Customer Segmentation
 using KMeans')
plt.legend()
plt.grid(True)
plt.show()
```

IV. Results

After applying the Elbow Method, the optimal number of clusters was found to be 5. K-Means was then applied, and customers were grouped based on their income and spending scores.

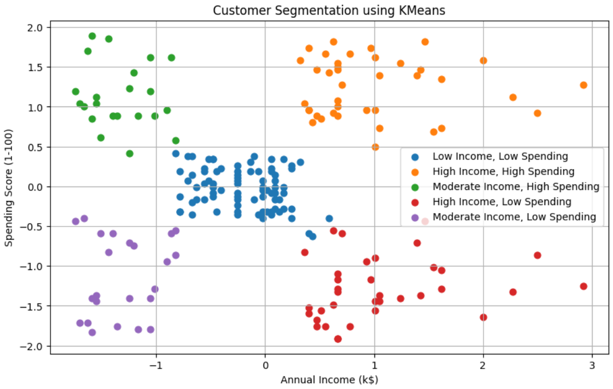

Fig. 2. Customer segmentation using K-Means

TABLE I
Summary of Customer Segments

Segment	**Avg. Income (k$)**	**Avg. Score**
Low Income, Low Spending	55.30	49.25
High Income, High Spending	86.54	82.13
Moderate Income, High Spending	25.73	79.36
High Income, Low Spending	88.20	17.11
Moderate Income, Low Spending	26.30	20.91

Each cluster represented a group of customers with similar purchasing behavior. For instance, one group had high income but low spending, indicating potential targets for special promotions.

V. Discussion

Segmentation revealed meaningful patterns in customer behavior. Businesses can now apply tailored strategies for each segment, such as:

- Offering discounts to low-spending high-income groups
- Rewarding loyal high-spending customers
- Creating budget-friendly deals for low-income segments

VI. Conclusion

This research demonstrates how customer segmentation using K-Means clustering can uncover valuable insights from e-commerce customer data. The clustering revealed meaningful patterns in customer behavior that can directly inform marketing strategies and resource allocation. By identifying groups such as high spenders and low-engagement customers, businesses can tailor their approaches to different customer types more effectively.

Acknowledgment

The author would like to thank the open-source community and dataset providers for making the data accessible.

References

[1] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters.

[2] Vijay Choudhary, "Customer Segmentation Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

[3] Kotler, P., and Keller, K.L., "Marketing Management," 15th ed., Pearson Education, 2016.