Customer Segmentation Report

1. Introduction

Customer segmentation is a crucial process in understanding customer behavior and optimizing marketing strategies. By categorizing customers into distinct groups based on their purchasing patterns, browsing habits, and discount usage, businesses can tailor their marketing efforts, personalize customer experiences, and improve overall engagement.

This report presents an in-depth analysis of customer behavior on an e-commerce platform using clustering techniques to identify three distinct customer groups: Bargain Hunters, High Spenders, and Window Shoppers. These segments exhibit unique shopping behaviors:

- Bargain Hunters are deal-seeking customers who frequently purchase lower-value items and heavily rely on discounts.
- High Spenders make fewer but high-value purchases and are less influenced by discounts.
- Window Shoppers browse extensively, view a large number of products, but make very few purchases.

To uncover these hidden clusters, we performed Exploratory Data Analysis (EDA) to understand the distribution and relationships among features. We then applied feature scaling to normalize numerical attributes before implementing clustering algorithms. Through model selection and evaluation, we identified the optimal clustering technique that best represents the underlying patterns in the dataset.

This report details the methodology used, including data preprocessing, model selection, and evaluation. The findings are visualized through graphs and charts, demonstrating the separation of customer segments. Finally, we discuss potential business insights and recommendations based on the identified clusters to enhance marketing strategies and customer engagement.

2. Exploratory Data Analysis (EDA)

2.1 Dataset Overview

The dataset contains information about customer interactions, purchases, and browsing behaviors on an e-commerce platform. It consists of **999 records** and includes the following six features:

1. **customer_id** – A unique identifier for each customer.
2. **total_purchases** – The total number of purchases made by the customer.
3. **avg_cart_value** – The average value of items in the customer's cart.
4. **total_time_spent** – The total time (in minutes) the customer has spent on the platform.
5. **product_click** – The number of products viewed by the customer.
6. **discount_counts** – The number of times the customer has used a discount code.

Statistical Summary

The dataset exhibits a diverse range of customer behaviors, as seen from the descriptive statistics:

- **Total Purchases:** Customers have made between **0 and 32** purchases, with a mean of **11.57**.
- **Average Cart Value:** The average cart value ranges from **$10.26 to $199.77**, with a mean of **$75.46**.
- **Total Time Spent:** Customers have spent between **5.12 and 119.82** minutes on the platform, with an average of **49.35** minutes.
- **Product Clicks:** The number of products viewed varies from **4 to 73**, with an average of **28.24** clicks.
- **Discount Usage:** Customers have used between **0 and 21** discount codes, with an average of **4.31**.

This dataset provides valuable insights into different shopping behaviors, enabling us to identify distinct customer segments.

2.2 Handling Missing Values

Missing values were handled using the median imputation method to maintain data integrity without introducing bias.

Initially, the dataset had missing values in the following columns:

- **total_purchases**: 20 missing values
- **avg_cart_value**: 20 missing values
- **product_click**: 20 missing values

After handling missing values, all columns now contain complete data, ensuring consistency for analysis.

2.3 Data Standardization

Feature Scaling and Standardization

Since clustering algorithms rely on distance-based metrics, it is essential to ensure that all numerical features contribute equally to the analysis. Given that our dataset includes features with different units and ranges—such as total_purchases (ranging from **0 to 32**) and avg_cart_value (ranging from **$10.26 to $199.77**)—we applied **standardization** to bring all features to a common scale.

To achieve this, we used **StandardScaler** from sklearn.preprocessing, which transforms the numerical features by centering them around zero and scaling them to have unit variance. The transformation follows the formula:

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

where:

- X is the original feature value

- μ is the mean of the feature
- σ is the standard deviation of the feature

Implementation

The following steps were performed to standardize the dataset:

1. **Loading the Dataset:** The original dataset was loaded into a Pandas DataFrame.
2. **Selecting Numerical Features:** The customer_id column was excluded, leaving five numerical features (total_purchases, avg_cart_value, total_time_spent, product_click, and discount_counts).
3. **Applying StandardScaler:** The selected features were transformed using StandardScaler(), ensuring that they had a mean of 0 and a standard deviation of 1.
4. **Validation:** We verified the transformation by checking that the mean of the scaled features was approximately **0**, and the standard deviation was approximately **1**.

3. Model Selection

3.1 Clustering Algorithm

We used K-Means clustering to segment the customers into three predefined groups based on various characteristics. To determine the optimal number of clusters, we employed the Elbow Method, which analyzes the Within-Cluster Sum of Squares (WCSS) by plotting the WCSS values against the number of clusters. The point where the rate of decrease in WCSS slows down (the "elbow") indicates the most appropriate cluster count, ensuring the best balance between model simplicity and accuracy.

3.2 Cluster Interpretation

After applying the clustering algorithm, customers were categorized into the following segments:

- **Bargain Hunters**: Frequent buyers of low-cost items, heavily relying on discounts.

- **High Spenders**: Customers who make fewer but high-value purchases and rarely use discounts.
- **Window Shoppers**: Customers who browse extensively but rarely make purchases.

4. Model Evaluation

4.1 Cluster Validation

- **Silhouette Score Analysis**: A high silhouette score indicates well-separated clusters.
- **Centroid Analysis**: Examining cluster centers revealed expected behavioral trends.
- **Distribution Analysis**: Visualizing distributions of each feature within clusters validated our assumptions.

5. Cluster Visualization

We used the following visualizations to interpret cluster distributions:

Plot 1: K-Means Clustering in 2D Space

In this plot, the K-Means clustering result is visualized by plotting the total_purchases on the x-axis and avg_cart_value on the y-axis. Each data point represents a customer, colored according to the cluster it belongs to. The color bar shows the cluster index.

- **Key Steps:**
    1. K-Means is applied to the scaled customer data (df_scaled).
    2. The clusters are assigned to a new column cluster in the dataframe.
    3. The results are visualized with a scatter plot using the total_purchases and avg_cart_value features.

Plot 2: PCA Visualization of the Clusters

This plot reduces the dimensionality of the data to two principal components using PCA (Principal Component Analysis). The goal is to project the customer data into a 2D space to visualize how the clusters are distributed in a reduced space.

- **Key Steps:**
    1. PCA is applied to the numerical_features in the dataset.
    2. The PCA results (two principal components) are stored in a new dataframe (df_pca).
    3. The clustering result is added to the PCA dataframe (df_pca['cluster']).
    4. A scatter plot is used to visualize the clusters in the 2D PCA space.

6. Findings and Business Insights

6.1 Marketing Strategies for Each Segment

- **Bargain Hunters:**

  Target price-sensitive customers with **personalized discount recommendations** based on their browsing and purchasing behavior. Offer **exclusive deals** and **flash sales** to create urgency, and implement loyalty programs that reward frequent bargain hunters with discounts for future purchases.
- **High Spenders:**

  Focus on **tiered loyalty programs** offering exclusive benefits like **early access to products**, **premium memberships**, and **personalized shopping experiences**. Providing VIP treatment, such as private shopping events, will enhance their experience and drive customer loyalty.
- **Window Shoppers:**

  Engage with **targeted content** and **personalized product recommendations** to move them toward conversion. Highlight **limited-time offers** and use **urgency tactics** like countdown timers. **Product reviews** and **user-generated content** can help build trust and increase the likelihood of conversion.

6.2 Future Improvements

- **Exploring Additional Features:**
  Incorporate **browsing patterns** (e.g., time spent on product pages) and **session duration** to better understand customer behavior. Integrating data from **social media activity** and **customer interactions** could provide deeper insights into customer preferences and enhance segmentation accuracy.
- **Alternative Segmentation Approaches:**
  Experiment with **hierarchical clustering** to reveal multi-level customer segments, or use **DBSCAN** to identify clusters with irregular shapes and outliers. These methods may offer more nuanced insights into customer behavior and improve targeting strategies.

7. Conclusion

This analysis successfully identified three distinct customer segments, each with unique behaviors and preferences, providing valuable insights for targeted marketing efforts. By understanding the characteristics of each segment, businesses can tailor their promotional strategies to meet the specific needs and motivations of each group. This approach not only enhances customer engagement but also improves the overall customer experience, fostering long-term loyalty. Furthermore, the insights gained from this segmentation can be leveraged to optimize resource allocation, personalize communications, and refine product offerings, ultimately driving increased conversions and maximizing revenue. These findings lay the groundwork for more efficient marketing campaigns and better customer retention strategies.