Q2 - Customer Segmentation

Team Neura

Intellihack 5.0

1. Introduction

In today's competitive e-commerce landscape, understanding customer behavior is key to targeted marketing and personalized services. This task focuses on segmenting customers into three distinct groups, Bargain Hunters, High Spenders, and Window Shoppers, by analyzing their interactions on the platform. The segmentation is achieved using the k-means clustering algorithm, which partitions the dataset into clusters based on similarity in behavior.

2. Data Overview

The dataset contains six key features:

- **customer_id:** Unique identifier for each customer.
- **total_purchases:** The total number of purchases made.
- **avg_cart_value**: The average monetary value of items in the customer's cart.
- **total_time_spent:** The total time (in minutes) spent on the platform.
- **product_click:** The number of products viewed.
- **discount_count:** The number of times a discount code was used.

These features help in capturing different aspects of customer behavior. The hidden clusters correspond to:

- **Bargain Hunters**: Frequent buyers with lower cart values and high discount usage.
- **High Spenders**: Customers making fewer but higher-value purchases.
- **Window Shoppers**: Users who spend more time browsing with low purchase frequency.

3. Methodology

3.1 Data Preprocessing

Missing Value Analysis:
The dataset was initially examined for missing values. Appropriate imputation techniques (such as mean or median imputation) were applied to ensure no missing data would skew the clustering results.

Feature Scaling:
Given that k-means clustering relies on distance calculations, feature scaling is essential. The StandardScaler was applied to normalize the dataset so that each feature contributes equally. Standard scaling transforms the data to have a zero mean and unit variance, ensuring that features with larger ranges do not dominate the distance metrics.

3.2 Exploratory Data Analysis (EDA)

A detailed EDA was conducted to understand the distribution and relationships between features:

- **Histograms:** To observe the distribution of each feature.
- **Correlation Matrix:** To identify potential linear relationships among variables.
- **Pair Plots & Scatter Plots:** To visually assess the relationships between different pairs of features.
- **Box Plots**: To detect outliers and understand the spread of data.

These visualizations guided the data preprocessing steps and provided preliminary insights into which features might influence the clustering outcomes.

3.3 Clustering Process with k-Means

Algorithm Choice:
K-means clustering was selected for its simplicity and efficiency in handling large datasets. It iteratively assigns data points to clusters based on the nearest centroid and updates centroids until convergence.

Optimal Cluster Determination:
Two key metrics were used:

- **Inertia:** Measures the sum of squared distances between data points and their assigned cluster centroids.
- **Silhouette Score:** Evaluates how similar an object is to its own cluster compared to other clusters.

By plotting the inertia against various values of k (elbow method) and examining the silhouette scores, it was determined that k = 3 was optimal for this dataset.

3.4 Model Training and Evaluation

The model was trained on the scaled dataset using k-means with k=3. Post-training, the clustering performance was evaluated based on the following accuracies:

- Cluster 0: 60%
- Cluster 1: 80%
- Cluster 2: 60%

These accuracies reflect the model's ability to correctly identify and separate the customer segments, as determined by domain-specific validation and the alignment with the expected behavior of Bargain Hunters, High Spenders, and Window Shoppers.

4. Insights Gained

Cluster Characteristics:

1. **Bargain Hunters:** Cluster members exhibited high purchase frequency coupled with lower average cart values and high discount usage.
2. **High Spenders**: This cluster had fewer transactions but higher average cart values, indicating premium purchasing behavior.
3. **Window Shoppers**: Characterized by high browsing times and product views with low purchase activity.

Visualization Benefits:
EDA techniques such as pair plots and box plots helped in visualizing the separation between clusters, thereby reinforcing the decision of choosing k=3.

Performance Metrics:
Inertia and silhouette scores provided quantitative backing to the clustering decisions, ensuring that the chosen model configuration aligned with the inherent structure of the data.

5. Challenges Faced

Handling Missing Values:
Missing data required careful imputation to avoid bias. Selecting the right strategy (mean vs. median) was essential to maintain the integrity of the dataset.

Scaling Sensitivity:
Since k-means is sensitive to the scale of features, improper scaling could have led to misinterpretation of distances, thereby affecting cluster assignments.

Choosing the Right Number of Clusters:
Although the elbow method and silhouette scores indicated k=3, verifying the clusters against known business segments (via domain knowledge) was challenging and required iterative tuning.

Outlier Impact:
Outliers can significantly distort cluster centroids. Identifying and handling outliers was necessary to ensure robust clustering.

6. Suggestions for Improvement

Outlier Treatment:
Implement systematic outlier detection and removal techniques before clustering to prevent undue influence on cluster centroids.

Algorithm Variations:
Experiment with alternative clustering algorithms such as hierarchical clustering, DBSCAN, or Gaussian Mixture Models. These methods can sometimes reveal structure that k-means might miss, especially if clusters are not spherical.

Parameter Tuning:
Fine-tuning the initialization of centroids and the number of iterations in the k-means algorithm may lead to more stable clustering outcomes.

7. Conclusion

The customer segmentation project successfully employed k-means clustering to partition an e-commerce customer base into three meaningful segments: Bargain Hunters, High Spenders, and Window Shoppers. The report discussed each step, from data cleaning and EDA to feature scaling, model training, and evaluation. Despite challenges such as handling missing values and scaling issues, the process led to actionable insights that can inform targeted marketing strategies. Future improvements can be made by exploring alternative clustering methods, enhancing preprocessing techniques, and applying dimensionality reduction to further refine the segmentation.

7. References

[1] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[2] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.