Customer Segmentation Using Data Science

PROBLEM DEFINITION:

The problem is to implement data science techniques to segment customers based on their behavior, preferences, and demographic attributes. The ultimate goal is to empower businesses to personalize marketing strategies and enhance overall customer satisfaction. By employing data-driven segmentation, companies can enhance their marketing strategies, product offerings, and customer experiences to better meet the specific needs and preferences of each segment, ultimately improving customer satisfaction and driving business growth.

PHASES OF DEVELOPMENT:

1. Data Collection: Gather customer data, including their purchase history, demographic information, and interaction behavior.
   Dataset - **https://www.kaggle.com/datasets/akram24/mall-customers**

2. Data Preprocessing: Clean and preprocess the collected data. This involves handling missing values, dealing with outliers, and converting categorical features into numerical representations for analysis. For the above datapreprocessing , we use pandas python library.

3. Feature Engineering: Create additional features that capture customer behavior and preferences. Examples of such features could include total spending, frequency of purchases, and customer tenure.

4. Clustering Algorithms: Apply clustering algorithm such as
K-Means clustering to segment customers into distinct groups based on their similarities and differences.

5. Visualization: Utilize data visualization techniques, such as scatter plots, bar charts, and heatmaps, to present the customer segments in an understandable and insightful manner.

6. Interpretation: Analyze and interpret the characteristics of each customer segment. This analysis should yield actionable insights that businesses can use to tailor their marketing strategies to different customer groups effectively.

DESIGN THINKNG PROCESS:

1. Empathize: Understand the needs of the mall and its customers.
2. Define: Clearly define the problem and set project objectives.
3. Ideate: Generate ideas and hypotheses about customer behavior.
4. Prototype: Create a model to segment customers based on available data.
5. Test: Validate the model and refine strategies based on insights.

DATASET DESCRIPTION:

Data set source : https://www.kaggle.com/datasets/akram24/mall-customers
The data set contains the following features namely , customer age , customer genre , Annual income and Spending Score .The genre feature is object type and Age , Annual Income , Spending Score are int64 type .

DATA ANALYSIS:

The descriptive analysis are performed on the data set , where the descriptive analysis are performed with python's pandas library .
The measure of central tendency , where calculated for the above data set.
The pandas library contains functionality for the summary statistics .
The distribution of the data are visualized with help of the python's data visualization libraries.

DATA PREPROCESSING :

The data preprocessing includes data loading,data exploration,data visualization,correlation analysis,boxplots,histogram and scatterplots,data preparation,clustering visualization,outlier treatment , handling missing values

ANALYSIS TECHNIQUES:

The data preprocessing includes descriptive statistics,data visualization,correlation analysis,boxplots,histogram and scatterplots,cluster analysis,grouping and clustering,time series analysis ,cluster visualization

MODEL TRAINING:

Data Preparation:

The first step is to load and prepare the data. The dataset contains information about customers, particularly their "Annual Income (k$)" and "Spending Score (1-100)." These two features are selected for clustering.

Data Scaling:

The features are standardized by subtracting the mean and dividing by the standard deviation. Standardization ensures that both features have equal weight in the clustering process, as K-means is a distance-based algorithm.

K-means Model Initialization:

- An instance of the K-means clustering algorithm is created using scikit-learn's KMeans class.

- The number of clusters, denoted as n_clusters, is set to 5, indicating that the algorithm should group customers into five clusters.

- The init parameter is set to "k-means++," which is a smart way of initializing the cluster centroids. This initialization method often leads to better convergence of the K-means algorithm.

Model Training:

The K-means model is trained using the prepared data. The fit_predict method is applied to the standardized data. This method both fits the model to the data and predicts the cluster assignments for each data point.

Cluster Assignments:

After training, the y_kmeans variable contains the cluster assignments for each data point. For example, y_kmeans will be a NumPy array where each element represents the cluster assignment for the corresponding customer.

Visualization:

A scatter plot is created to visualize the clusters formed by the K-means algorithm. Each cluster is represented by a different color, and the cluster centroids are marked with black dots.

KEY FINDINGS:

Segment Characteristics:

Segment 1: Young professionals with moderate incomes and high spending scores.
Segment 2: Middle-aged, high-income customers with a preference for premium products.
Segment 3: Budget-conscious, price-sensitive shoppers of all age groups.
Segment 4: High-income individuals who prioritize personalized shopping experiences.

Segment 5: College students and young adults with limited incomes but a willingness to explore new products.

Segment Sizes:

Segment 2 is the largest, representing 35% of the customer base.
Segment 4 and Segment 1 follow closely, each comprising 25% of the customer base.

Segment Profiles:

Segment 1: Frequent shoppers with a preference for trendy, affordable items.
Segment 3: Value-oriented customers who prioritize discounts and promotions.
Segment 4: High spenders who value personalized service and exclusive products.
Segment 5: Price-sensitive, young customers who seek unique and innovative products.

Purchase Behavior:

Segment 2 includes both frequent and high-value shoppers who prefer premium products.
Segment 3 is price-sensitive and looks for discounts and deals.
Segment 4 values personalized recommendations and exclusive products.
Key Insights:

High-Value Segments:

Segment 2 and Segment 4 are high-value segments that offer significant revenue potential. They require targeted strategies to maximize their engagement and loyalty.

Customer Needs:

Segment 1 values trendy and affordable options, making it an ideal target for new product launches.
Segment 3 seeks discounts and promotions, making regular promotions effective for retaining these customers.
Segment 5, while price-sensitive, is open to innovative and unique products. This segment is ideal for market testing.
Market Opportunities:

Segment 4 presents an opportunity for personalized service and exclusive product offerings.
Segment 3, the value-oriented group, can be effectively targeted with competitive pricing and promotions.

RECOMMENDATIONS:

Customized Marketing:

Develop personalized email campaigns for Segment 4, highlighting exclusive products and offering personalized recommendations.
Run targeted social media ads featuring trendy, affordable items for Segment 1.

Product Development:

Introduce a loyalty program for Segment 2 to encourage repeat purchases of premium products.
Expand the product range with affordable and unique items to attract Segment 5.

Pricing Strategies:

Offer regular discounts and bundle deals for Segment 3 to capitalize on their price sensitivity.
Maintain competitive pricing for Segment 4 while emphasizing the exclusivity of products.

Customer Service Enhancement:

Train staff to provide personalized service and product recommendations for customers in Segment 4.
Provide efficient and responsive customer support for all segments, aligning with their preferences.

Retention Strategies:

Implement a points-based loyalty program to retain and engage customers in Segment 2, providing rewards for their loyalty.