Joy Esangbedo

CS 549- H01 Project Report (Completed)

921530499

TOPIC: PROBLEM DEFINITION

This project seeks to apply unsupervised machine learning techniques, specifically the K-Means Clustering Algorithm, to predict the customer segment population in a supermarket mall. It will also assist in understanding and identifying segments of customers who are likely to converge as "Target Customers." Convergence in this context implies a group of customers who exhibit similar behavior or characteristics that make them particularly interesting or valuable to the business.

Generally, the problem is outlined by applying unsupervised machine learning (i.e. K means) to segment customers by iterating the clustering process to arrive at the optimal cluster outcome based on their structure, with a focus on grouping target customers. The chosen language is python on Google Colab because it is easy to use and accessible. The outcome will provide actionable insights for the marketing team to optimize their strategies and enhance engagement with specific customer segments.

PROJECT OBJECTIVE

The primary goal is to provide valuable insights to the marketing team using K-means clustering to identify object that are similar to one another and unrelated to objects in other groups. This will help to understand the characteristics of these target customers, the marketing team can tailor strategies and campaigns more effectively. This involves creating a sense of customer preferences, behaviors, and needs, ultimately aiding the planning and implementation of targeted marketing strategies.

DATA SET

The dataset is related to customer segmentation concepts, which consist of 200 rows and 5 columns with no missing values. The available data for the 5 columns includes customerID, the unique id assigned to the customer, age refers to the customer's age, gender refers to either 'female' or 'male' customers, annual income(k$) refers to the earnings of the customer in a year, and a spending score refers to the score assigned by the mall based on defined parameters reflecting customer behavior and purchasing data. This dataset can effectively be used for establishing machine learning models to create customer segment groups based on their demographics and behavioral patterns.

Link to dataset: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python/download?datasetVersionNumber=1

SYSTEM DESCRIPTION

The project simulation adopts python language on Google Colab with its focus on k-means clustering algorithm to identify customer segment groups that converge as 'target customer' for specific products in a mall. Hence, the model to be used is as follows:

EXPLORATORY DATA ANALYSIS

The EDA analyzes the pie chart, data types, distributions, cardinality of features with findings showing the gender disparity in the customer groups. By visualizing the data graphs like scatter plots and boxplots, the findings revealed some outliers. Customers were segmented by attributes like age, gender, annual income, and spending score. The split-apply-combine rule was used to understand patterns within specific subgroups of the data. These features are further evaluated to identify correlations, and cluster the customers into segments by age, income and spending.

K-MEANS CLUSTERING

K-means clustering algorithm was used for segmenting the customer population into distinct groups for target marketing. K-Means is an unsupervised machine learning algorithm used for clustering or grouping similar data points together. In the context of customer segmentation, the goal is to partition and predict customers into distinct groups based on their demographic details, and customer behavior to provide insights and targeted market strategy. This data is also depicted using data preprocessing and modelling techniques, which encompasses categorical data, scaling features, and 'one-hot code' to compute the model's performance.

Similarly, the elbow method helps to determine the optimal number of clusters in segment customer groups based on customer behavior and demographics. This involves running the K-Means algorithm with different values of K and plotting the sum of squared distances within clusters (inertia) against K. The "elbow" point is where adding more clusters provides diminishing returns.

LIBRARIES USED

- numPy: This was used to support numerical computing.
- seaborn: This was used to make it easier to create informative visualizations.
- scikit-learn: A tool for clustering and model selection.
- yellowbrick: This was used for finding the optimal number of clusters in KMeans clustering.
- matplotlib: This was used for creating static, interactive, and animated plots.
- pandas: This was used for data cleaning, exploration, and preprocessing tasks.
- warnings: This was used to ignore warning messages.
- sklearn.preprocessing: This was used for data preprocessing tasks like scaling.

RESULT AND CONCLUSIONS

The EDA evaluated the data aspects, distributions and derived insights that shaped further analysis. Thus, the pie chart of the gender distribution provides a quick view that majority (around 55%) of customers are females. This overall insight can help drive marketing decisions targeted towards women. Boxplots of annual income and spending scores by gender indicate that male customers generally earn higher incomes, but the spending range is similar across genders. This means income level driven communication may need to be more gender focused. An annual income vs spending score scatter plot visualization indicates most customers lie in the mid income range of $15k-$60k with spending scores between 30-80. Interesting outlier segments are very high-income earners (>$120k) but surprisingly low spenders (<$40). As the customer age range is very wide (18-70), it needs to be scaled to a range of 0-1 before applying any clustering algorithms. Annual income can also be log transformed to handle skewness. This scaling provided a more meaningful customer groups during segmentation modelling. Therefore, EDA was executed first, followed by scaling and one-hot encoding, and finally model building.

In K Means Clustering, the Elbow Method visualizer fits a KMeans model on the dataset for a range of cluster numbers from 1 to 10 for k = 3, and 3 to 9 for k = 5 showing the randomness in selecting a cluster. It plots the Kmeans inertia for each model with an optimal k value of 5 to eliminate underfitting or overfitting data, which proved difficult to predict. The customer details calculated were assigned to the closest center applied to the following x and y data points as represented by their results. Based on the outcome of the two choices (Choice A and Choice B), it is evident that Choice A (Age vs Annual Income) predicted cluster of 4 improves the clustering process than choice B (Spending Score vs Annual Income) with the predicted cluster outcome which varies between 3 and 4.