CUSTOMER SEGREGATION IN DATA SCIENCE
PHASE - 2

Introduction

The advancements in Data Science and Machine Learning have enabled us to solve several complex regression and classification problems. However, the performance of all these ML models depends on the data fed to them. Thus, it is imperative that we provide our ML models with an optimal dataset. Now, one might think that the more data we provide to our model, the better it becomes – however, it is not the case. Suppose we feed our model with a huge dataset (with a large no. of features/columns). In that case, it gives rise to the problem of overfitting, wherein the model starts getting influenced by outlier values and noise. This is called the Curse of Dimensionality.

Principal Component Analysis (PCA)

As stated earlier, Principal Component Analysis is a technique of feature extraction that maps a higher-dimensional feature space to a lower-dimensional feature space. While reducing the number of dimensions, PCA ensures that maximum information of the original dataset is retained in the dataset with the reduced no. of dimensions and the co-relation between the newly obtained Principal Components is minimal. The new features obtained after applying PCA are called Principal Components and are denoted as PCi (i=1,2,3…n). Here, (Principal Component-1) PC1 captures the maximum information of the original dataset, followed by PC2, then PC3, and so on.

Data Preprocessing:

Before applying PCA, it's essential to prepare the data. This includes handling missing values, encoding categorical variables, and scaling the data to make all features comparable.

Apply PCA:

PCA aims to reduce the dimensionality of the dataset while retaining as much variance as possible. The steps involved are:

a. Standardization: Standardize the data by subtracting the mean and dividing by the standard deviation for each feature. This step is crucial for PCA as it assumes that the data is centered.

b. Eigendecomposition: Compute the covariance matrix of the standardized data and then perform eigendecomposition to obtain the eigenvalues and eigenvectors.

c. Select Principal Components: Sort the eigenvalues in descending order and select the top 'k' eigenvectors, where 'k' is the desired reduced dimensionality. These eigenvectors represent the principal components.

d. Transform Data: Project the original data onto the 'k' selected principal components to create a reduced-dimension dataset.

Interpretation:

PCA doesn't just reduce dimensionality; it also provides insights into the most important features or dimensions in the data. You can interpret the principal components to understand which variables or combinations of variables are contributing the most to the variance in the data.

Customer Segmentation:

With the reduced-dimension dataset, you can apply clustering techniques like K-means, hierarchical clustering, or DBSCAN to segment customers based on their similarities. The reduced dataset should capture the most important information, making clustering more effective.

Visualize Results:

Visualize the clusters or customer segments in a reduced-dimensional space using scatter plots, heatmaps, or other visualization techniques to gain a better understanding of the relationships between customer groups.

Evaluate and Refine:

After segmentation, it's crucial to evaluate the quality of your clusters using metrics like silhouette score, Davies-Bouldin index, or domain-specific evaluation criteria. You may need to refine the number of clusters or consider adding back some dimensions if necessary.

Application:

Use the customer segments for targeted marketing, product recommendations, or other business strategies tailored to each group's specific needs and preferences.

PCA can be a powerful tool for reducing the dimensionality of customer data while preserving essential information, making it easier to perform effective customer segmentation and extract meaningful insights for business decisions.

Results:

The application of PCA for customer segmentation often leads to several key outcomes:

a. Identified Customer Segments: PCA enables the identification of distinct customer segments based on their similarities and shared characteristics.

b. Reduced Dimensionality: By reducing the dimensionality of the data, PCA simplifies the customer segmentation process, making it more manageable and interpretable.

c. Improved Interpretation: PCA provides insights into the most critical features or dimensions in the data. You can understand which variables contribute the most to the variance in the dataset, aiding in the interpretation of customer segments.

d. Enhanced Clustering: Clustering algorithms applied to the reduced-dimension dataset often yield more accurate and meaningful customer segments. This can lead to improved targeting, personalization, and marketing strategies.

e. Visualizations: Visual representations of customer segments in a reduced-dimensional space help stakeholders gain a better understanding of the relationships between different customer groups.

Conclusion:

PCA is a valuable technique for customer segmentation that simplifies data while retaining its essential information. Through PCA, businesses can identify meaningful customer segments and create more targeted marketing strategies. The results often include improved clustering accuracy, enhanced customer insights, and the ability to tailor business strategies to specific customer needs and preferences. By utilizing PCA, businesses can make data-driven decisions and enhance customer experiences.