Market Customer Segmentation Project

Market Segmentation Project Summary

Introduction:

In today's competitive market, understanding customer behaviour is paramount for businesses looking to optimize their marketing strategies and enhance customer engagement. This project aimed to segment the market using a provided dataset, offering a granular view of distinct customer groups and enabling data-driven decision-making.

Methodology:

1. Exploratory Data Analysis (EDA):

Distribution Analysis: Before diving into complex data analysis, it's crucial to understand the nature and distribution of dataset variables. Visual representations, such as histograms, box plots, and density plots, were created to gauge the patterns and trends within the data. This step revealed key insights into variables' spread and potential outliers, setting the stage for subsequent analysis.

Correlation Analysis: Determining the relationship between variables is pivotal to understanding potential dependencies. By assessing correlations, we gleaned insights into which variables moved together, aiding in anticipating multicollinearity issues later in the analysis.

Upon analysing the distribution plots of the variables in the dataset, several key patterns and characteristics were identified:

1. Positively Skewed Distributions: A majority of the variables in the dataset exhibited a positively skewed distribution.

2. Negatively Skewed Distributions: Contrarily, variables such as 'Balance Frequency' and 'Tenure' demonstrated a negatively skewed distribution.

3. Bimodal Pattern: The variables: 'Purchase Frequency' and 'Purchases Install Frequency' showed a bimodal distribution pattern, which is characterized by two distinct peaks. This suggests that there might be two different groups or behaviours within the dataset for these variables. Bimodal distributions often warrant a deeper investigation as they might signify underlying subgroups within the data.

From analysing the correlation, we have the information that some of the variables pair like for example – ‘Purchases’ and ‘One-off Purchases’ or ‘Cash advance Trx’ and ‘Cash advance frequency’ has a very high correlation. That means the data should be reduced through some dimension reduction technique.

2. Data Pre-processing:

Scaling: Considering the varied range and units of different features, it was imperative to bring all variables to a standard scale. Using the Standard Scaler, the dataset was transformed to have zero mean and unit variance, ensuring that no particular feature dominated the subsequent PCA and clustering due to its original scale.

3. Dimensionality Reduction:

Principal Component Analysis (PCA): As datasets grow in features, the curse of dimensionality can become a concern, making computations complex and often less interpretable. PCA was utilized to address this. By transforming the original 17 columns into 9 principal components, we were able to retain over 80% of the data's variance. This reduction ensured that we encapsulated the majority of the data's information in just over half of its original dimensions, streamlining further analysis.

4. Clustering- KMeans Algorithm:

Determining Optimal Cluster Number: One of the critical decisions in clustering is selecting the number of clusters. Using the elbow method, which plots the sum of squared distances for different cluster numbers and identifies the "elbow" where the rate of decrease sharply changes, it was determined that four clusters were optimal for this dataset.

Key Findings:

Cluster 0: The first noteworthy cluster included members who maintained a significantly high balance, which was indicative of their financial behaviour and potentially their purchasing capacity. An even more distinguishing feature of this cluster was their proclivity for cash advances, suggesting a distinct set of financial needs or preferences.

Cluster 3: This cluster was characterized by high-spending members. Their values of purchases were notably higher than members of other clusters. This finding indicates that Cluster 3 members are prime customers, possibly driving a considerable portion of the organization's revenue.

Clusters 2 & 3: Both these clusters housed members who consistently engaged with the organization's products. Their frequent purchases underscored their importance, not just in terms of immediate revenue but also in potential long-term value and loyalty to the organization.

Conclusion:

This project successfully segmented the market into four well-defined groups using advanced data analysis techniques. These segments, each with its unique characteristics and behaviours, present a goldmine of insights for the organization. By understanding and acting upon these segment-specific traits, the organization can not only create more impactful marketing strategies but also foster deeper relationships with its customer base, driving both immediate revenue and long-term loyalty.