Lab Week 13 - Clustering

Name: Pinisetti Sudhiksha

SRN: PES2UG23CS916

Course: Machine Learning

Submission Date: 14/11/2025

Section: F

ANALYSIS

1. Dimensionality Justification

Dimensionality reduction was necessary because many of the original features displayed strong correlations, leading to redundancy in the data. Such redundancy can introduce noise and make modelling more complex. PCA mitigates this issue by combining correlated features into a smaller set of components that retain the essential structure of the dataset. The first two principal components together account for slightly more than 28% of the total variance. Although this proportion is relatively modest, it still enables a clearer and more interpretable representation of the primary patterns within the data.

2. Optimal Number of Clusters

Both the elbow method and the silhouette score indicate that three clusters represent the most suitable choice. The elbow curve shows the most significant drop in inertia from k=1k = 1k=1 to k=3k = 3k=3, after which the curve begins to flatten, suggesting limited benefit from adding further clusters. The silhouette score for k=3k = 3k=3 is approximately 0.39, a value that reflects a reasonable balance between cohesion within clusters and separation between them. Taken together, these metrics support selecting three clusters as the optimal configuration.

3. Cluster Size Characteristics

Both K-means and Bisecting K-means resulted in one relatively large cluster and two smaller ones, indicating that the data itself is unevenly distributed. The largest cluster likely represents the dominant customer profile, characterized by broadly similar behaviours and attributes. In contrast, the smaller clusters appear to consist of customers whose financial or demographic characteristics deviate from the typical pattern, possibly reflecting unique spending habits, financial variability, or other distinguishing traits. This imbalance mirrors natural differences commonly observed in real-world customer populations.

4. Algorithm Comparison

K-means performed better than Bisecting K-means, as shown by its higher silhouette score (0.39 compared to 0.29). This suggests that the clusters formed by K-means are more compact and better separated. Bisecting K-means likely performed less effectively because the dataset does not possess a strongly hierarchical structure, making the recursive splitting process less aligned with the natural distribution of the data.

5. Business Insights

The clusters identified within the PCA space reveal distinct customer segments that differ in behavioural and demographic patterns. These findings can assist the bank in refining its marketing strategy by enabling targeted communication with specific customer groups, enhancing the personalization of financial products or services, and improving the allocation

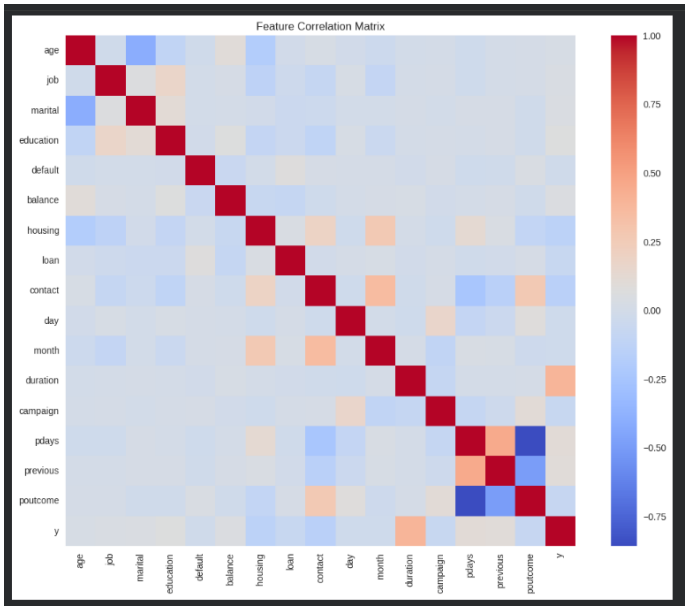
of marketing resources. Differentiating between stable customers and those exhibiting greater variability in financial behaviour can further support strategic decision-making.

6. Visual Pattern Recognition in PCA Space

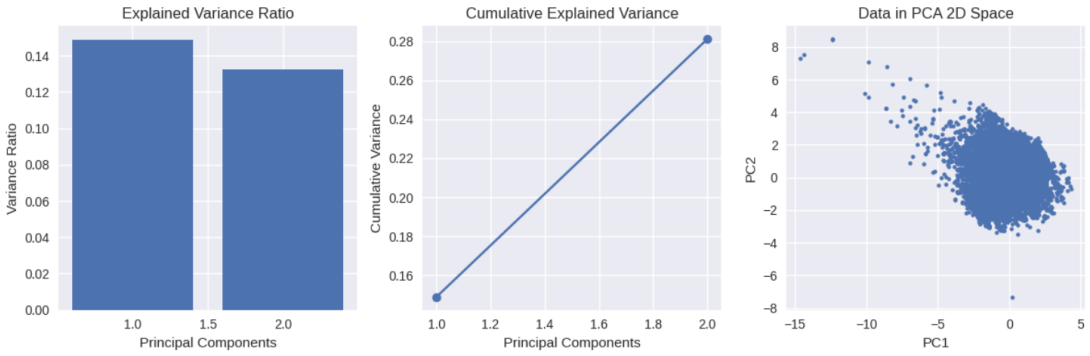
The three coloured regions—turquoise, yellow, and purple—visible in the PCA scatter plot correspond to groups of customers who share similar composite characteristics across the principal components. The boundaries between these regions vary in sharpness. In areas where the boundaries appear sharp, customer profiles differ more distinctly, allowing for clearer separation. In contrast, the more diffuse boundaries reflect overlapping characteristics that lead to gradual transitions between clusters. Overall, these patterns illustrate a combination of both pronounced and subtle differences among customer groups.

SCREENSHOTS

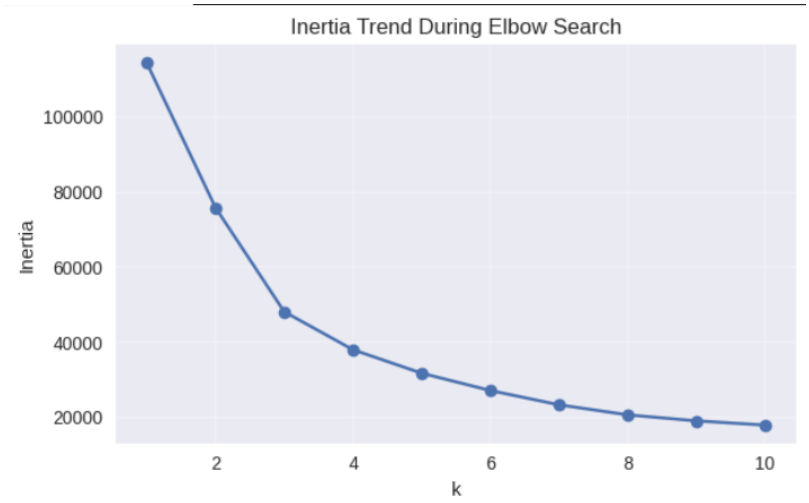
1. Feature Correlation matrix for the dataset

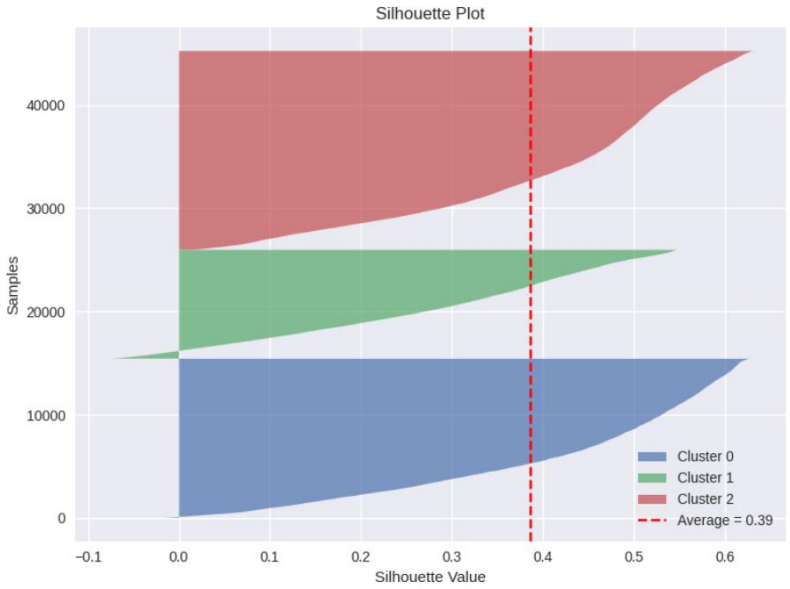


2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



3. ‘Inertia Plot’ and ‘Silhouette Score Plot’ for K-means





4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)