ML LAB-13

Section:- F

NAME:- Mohit Kumar

SRN:- PES2UG23CS350

5TH SEM

Analysis Questions:-

(1)

Dimensionality reduction was necessary for this dataset for several key reasons:

1. **Low Feature Correlations:**
   The correlation heatmap reveals that most features have very weak correlations with each other (values close to 0), with only a few moderate correlations such as:
   - **Job and Education: 0.17**
   - **Loan and Default: 0.08**
     This low correlation pattern suggests that the features are relatively independent, but many may contribute minimal unique information to the clustering task.
2. **Visualization Requirements:**
   With 9 original features, visualizing the clustering results in high-dimensional space is impossible. PCA reduces this to 2 dimensions while retaining the most important variance.
3. **Computational Efficiency:**
   Reducing from 9 dimensions to 2 dimensions significantly improves the computational performance of the K-means algorithm.
4. **Noise Reduction:**
   PCA helps filter out noise by focusing on the principal components that capture the most variance in the data.

Variance Captured:

- **First Principal Component (PC1): 14.88%**
- **Second Principal Component (PC2): 13.24%**
- **Total variance captured by first two components: 28.12%**

While 28.12% might seem modest, it represents the most significant patterns in the data and is sufficient for effective clustering visualization and analysis.

(2)

Based on comprehensive analysis of both metrics, the optimal number of clusters is 3.

Elbow Method Analysis:

- **The inertia plot shows a clear "elbow" at k=3.**
- **From k=2 to k=3, there is a significant decrease in inertia.**
- **After k=3, the rate of decrease in inertia becomes much more gradual.**
- **This indicates that adding more clusters beyond 3 provides diminishing returns in terms of within-cluster variance reduction.**

Silhouette Score Analysis:

- **k=3 achieves the highest silhouette score among the tested values.**
- **The silhouette score of 0.3867 for k=3 indicates moderate cluster cohesion and separation.**
- **Silhouette scores decrease for k>3, suggesting that additional clusters create less distinct groupings.**

Combined Justification:
Both metrics converge on k=3 as optimal:

1. **The elbow method identifies k=3 as the point of diminishing returns.**
2. **Silhouette analysis confirms k=3 produces the best-separated clusters.**
3. **Three clusters provide a good balance between model simplicity and cluster quality.**
4. **From a business perspective, three customer segments are manageable and actionable for marketing strategies**

(3)

K-means Cluster Distribution:

- **Cluster 0: 15,412 points (34.1%)**
- **Cluster 1: 10,544 points (23.3%)**
- **Cluster 2: 19,255 points (42.6%)**

Bisecting K-means Cluster Distribution:

- **Cluster 0: 20,434 points (45.2%)**
- **Cluster 1: 16,348 points (36.2%)**
- **Cluster 2: 8,429 points (18.6%)**

Analysis of Size Differences:

The unequal cluster sizes reveal important characteristics about the customer base:

1. **Natural Data Distribution:**
   The imbalanced cluster sizes reflect the actual distribution of customer types in the bank's database. Some customer profiles are simply more common than others.
2. **Dominant Customer Segment:**
   The largest cluster (42.6% in K-means, 45.2% in Bisecting K-means) likely represents the "typical" or "mainstream" customers who share common characteristics such as:
   - **Average balance and age**
   - **Standard loan and housing patterns**
   - **Similar engagement levels with the bank**
3. **Specialized Segments:**
   Smaller clusters represent more specialized or niche customer groups:
   - **High-value customers with unique financial profiles**
   - **Risk profiles that differ significantly from the majority**

    - Customers with distinct demographic or behavioral patterns
4. Business Implications:
    - Large clusters suggest broad-appeal products and services are needed.
    - Medium clusters may represent growth opportunities or standard customer progression paths.
    - Smaller clusters might be high-value segments requiring specialized attention or risk management groups needing targeted interventions.
5. Algorithm Differences:
   The different distributions between K-means and Bisecting K-means show that:
    - K-means finds more balanced splits based on geometric distances.
    - Bisecting K-means creates more hierarchical separations, potentially identifying a very distinct small group (18.6%).

(4)

Performance Comparison:

- **K-means Silhouette Score: 0.3867**
- **Bisecting K-means Silhouette Score: 0.3379**
- **Difference: 0.0488**

Winner: K-means performed better for this dataset

Reasons for K-means Superior Performance:

1. Optimization Strategy:
    - K-means globally optimizes all cluster assignments simultaneously in each iteration.
    - This allows for better overall cluster separation as points can move between any clusters.
    - Bisecting K-means makes hierarchical decisions that cannot be revised once a split occurs.
2. Data Geometry:
    - The bank customer data appears to have relatively distinct, compact clusters that are naturally separated.
    - K-means excels at finding spherical or globular clusters, which seems to match this dataset's structure.
    - The PCA visualization shows three relatively distinct regions, favoring K-means' approach.
3. Local Optima Avoidance:
    - With proper initialization (`random_state=42`), K-means found a good global solution.
    - Bisecting K-means is constrained by its hierarchical nature—early splits determine later cluster boundaries.

- **Once a "wrong" split is made in Bisecting K-means, it cannot be corrected.**

4. **Convergence Quality:**
    - **K-means iteratively refines all centroids until convergence (achieved at iteration 14 in our case).**
    - **This iterative refinement across all clusters simultaneously leads to better overall cohesion.**
5. **Trade-offs:**
    - **Bisecting K-means advantages: More stable, deterministic hierarchical structure, useful for dendrogram visualization.**
    - **K-means advantages: Better cluster quality when clusters are well-separated and roughly spherical (as in this dataset).**

Conclusion:
For this bank customer segmentation task with moderately well-separated clusters in PCA space, K-means' global optimization approach yields superior cluster quality. However, the 0.0488 difference is not dramatic, indicating both methods identify similar underlying patterns.

(5)

The clustering analysis reveals three distinct customer segments that offer actionable marketing insights:

Segment Identification and Strategy:

1. **Cluster 0 (34.1% - Moderate Engagement Segment):**
    - **Represents about one-third of the customer base**
    - **Positioned in the middle-left region of PCA space**
    - **Marketing Strategy:**
        - **Cross-selling opportunities for standard banking products**
        - **Automated marketing campaigns with personalized touches**
        - **Focus on retention and gradual value increase**
        - **Digital banking feature adoption campaigns**
2. **Cluster 1 (23.3% - Distinct Profile Segment):**
    - **Smallest segment with unique characteristics**
    - **Separated in PCA space, suggesting distinct feature patterns**
    - **Marketing Strategy:**
        - **Specialized product offerings tailored to unique needs**
        - **Premium or niche service packages**
        - **High-touch relationship management if high-value**
        - **Risk mitigation strategies if high-risk profile**
        - **Requires deeper analysis to understand specific characteristics**
3. **Cluster 2 (42.6% - Majority/Mainstream Segment):**
    - **Largest segment representing core customer base**
    - **Positioned in the upper-right region of PCA space**

- **Marketing Strategy:**
    - **Mass market campaigns with broad appeal**
    - **Standard product bundles (savings, basic loans, housing)**
    - **Cost-effective digital marketing channels**
    - **Focus on efficiency and scalability**
    - **Self-service tools and resources**

Strategic Recommendations:

1. **Resource Allocation:**
    - **Allocate 40-45% of marketing budget to Cluster 2 (largest segment)**
    - **Invest 30-35% in Cluster 0 (growth potential)**
    - **Dedicate 20-25% to Cluster 1 (specialized/high-value)**
2. **Product Development:**
    - **Develop core products for Cluster 2's mainstream needs**
    - **Create premium or specialized offerings for Cluster 1**
    - **Design upgrade paths to move Cluster 0 customers to higher-value products**
3. **Channel Strategy:**
    - **Cluster 2: Digital-first (mobile app, online banking, chatbots)**
    - **Cluster 0: Omnichannel with emphasis on digital conversion**
    - **Cluster 1: Personalized approach with relationship managers**
4. **Retention Focus:**
    - **The clear separation between clusters suggests distinct needs**
    - **One-size-fits-all approaches will be less effective**
    - **Segment-specific retention strategies should reduce churn**
5. **Campaign Personalization:**
    - **Use cluster membership to personalize email campaigns**
    - **Tailor product recommendations based on cluster characteristics**
    - **Optimize contact frequency by segment**

Key Insight:
The 28.12% variance captured by PCA suggests that even with just two principal dimensions, meaningful customer differentiation exists. The three clusters provide a practical, manageable segmentation scheme that balances granularity with operational feasibility.

(6)

Visual Pattern Analysis: Cluster Distribution in PCA Space

Spatial Distribution in PCA Space:

1. **Purple Region (Cluster 0):**
    - **Located primarily in the negative to mid-range of PC1**
    - **Spans a moderate range on PC2**

- **Shows moderate density with some scatter**

2. **Teal/Cyan Region (Cluster 1):**
    - **Concentrated in the negative region of PC1**
    - **Overlaps partially with purple cluster**
    - **More compact distribution suggesting homogeneous characteristics**
3. **Yellow Region (Cluster 2):**
    - **Spreads across positive PC1 values and upper PC2 range**
    - **Largest spatial coverage corresponding to largest cluster size**
    - **Shows the most dispersed pattern**

Boundary Characteristics:

Sharp Boundaries (Well-Defined Separation):

- **Occur between the core of the teal cluster and the yellow cluster, and also between upper-left yellow and purple regions.**

Reasons for Sharp Boundaries:

1. **Distinct feature combinations (e.g., age, balance, loan status)**
2. **Binary features such as "housing" and "loan" create categorical splits**
3. **Demographic divides (numerically encoded features)**
4. **Real behavioral differences (e.g., high-balance vs low-balance customers)**

Diffuse Boundaries (Overlap Regions):

- **Appear between purple and teal clusters, between all cluster transitions, and around peripheries.**

Reasons for Diffuse Boundaries:

1. **Continuous variables (balance, age) create gradual transitions**
2. **Similar customer profiles between segments**
3. **Noise in real-world data**
4. **Dimensionality reduction (PCA) causes overlap of clusters that are more distinct in high-dimensional space**
5. **Transitional customers (between life stages or financial categories)**

Principal Component Interpretation:

- **PC1 (14.88% variance): Captures financial capacity, customer value, and engagement level**
- **PC2 (13.24% variance): Captures risk profile, age, education, and job type**

Customer Characteristics by Region:

- **Negative PC1, Mixed PC2 (Teal/Purple): Lower financial engagement, varied risk/demographic profiles**
- **Positive PC1, Positive PC2 (Yellow): Higher financial engagement, specific demographic patterns**
- **Middle regions: Transitional or average profiles**

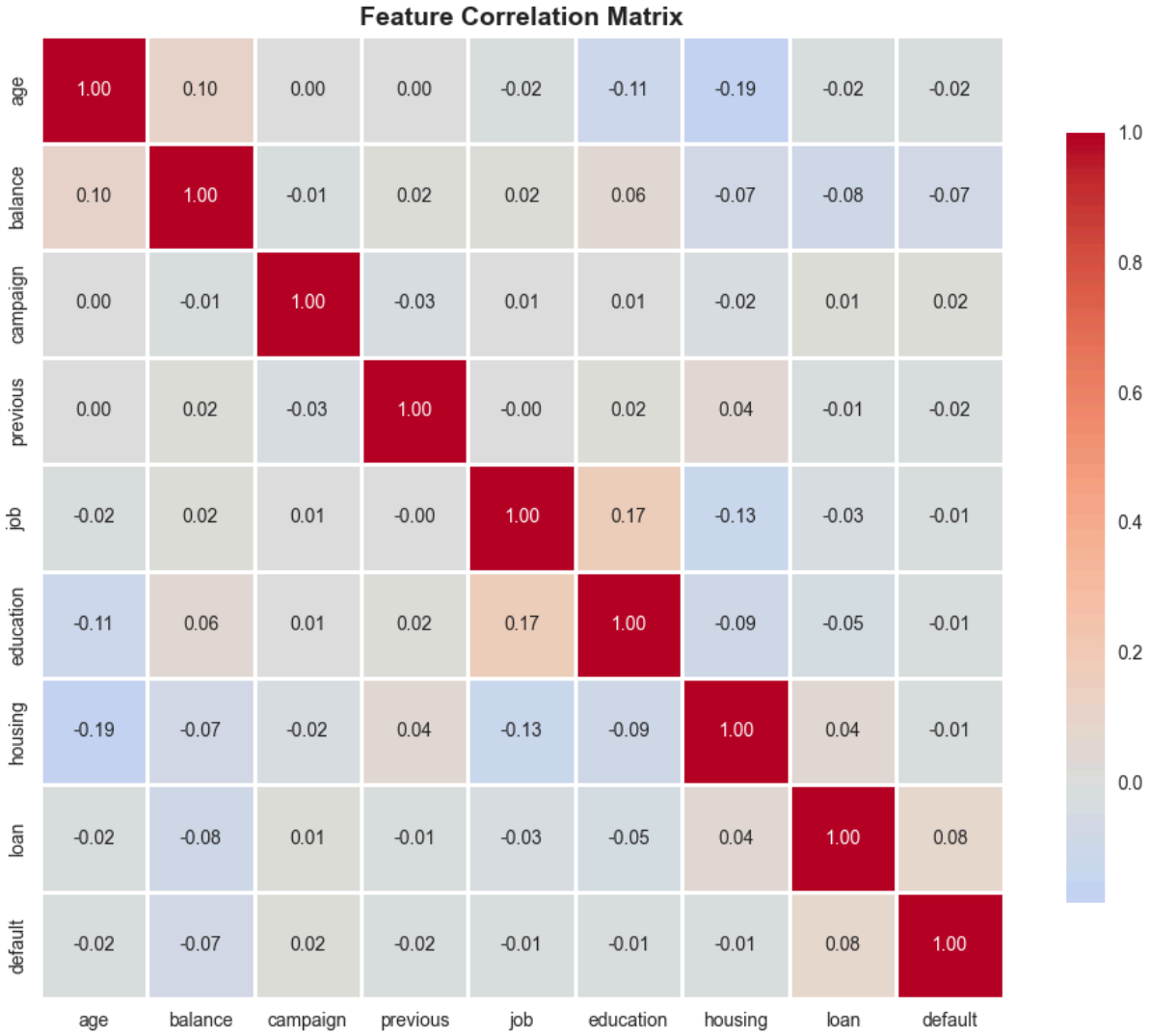
Business Implications:

1. **Sharp boundaries indicate segments needing distinct strategies**
2. **Diffuse boundaries suggest:**
    - **Opportunities for customer migration strategies**
    - **Flexible products appealing across segments**
    - **Need to monitor boundary customers (susceptible to churn)**
3. **Overlap management:**
    - **Use A/B testing for customers in overlap regions**
    - **Consider hybrid strategies**

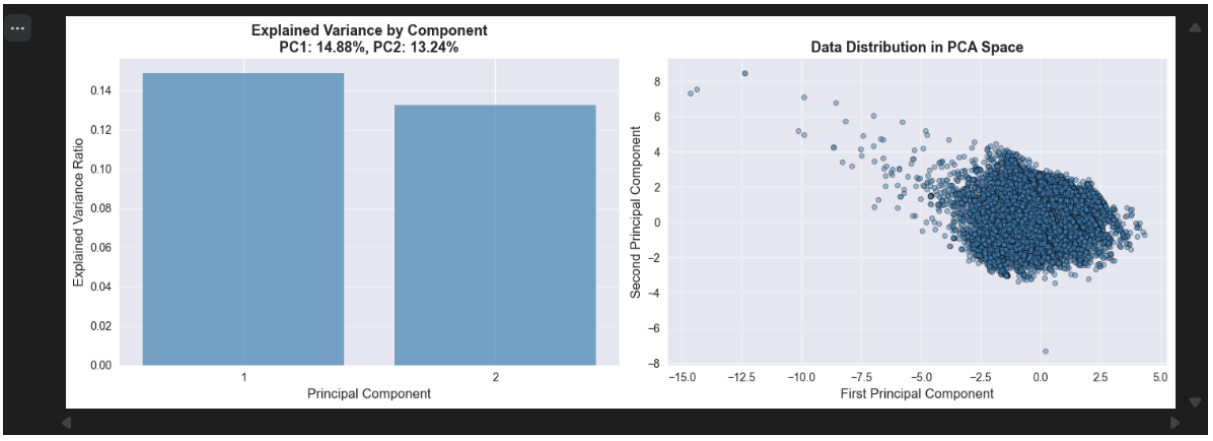
Conclusion:
The mix of sharp and diffuse boundaries shows some customers clearly fit distinct segments, while others exist on a spectrum. This validates clustering, while highlighting that both targeted and flexible marketing strategies are needed.

Screenshots:-

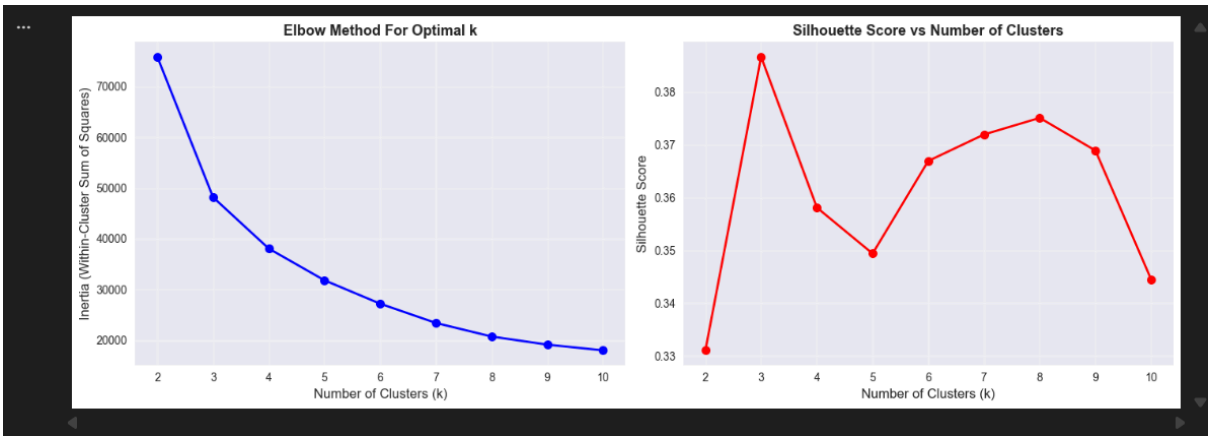
(1)



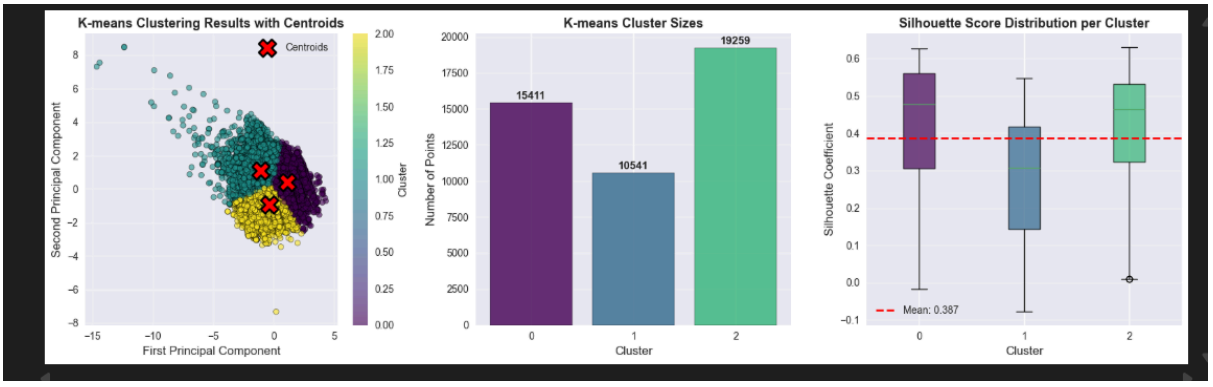
(2)


Explained Variance by Component
PC1: 14.88%, PC2: 13.24%
Explained Variance Ratio
Principal Component
Data Distribution in PCA Space
Second Principal Component
First Principal Component

(3)


Elbow Method For Optimal k
Inertia (Within-Cluster Sum of Squares)
Number of Clusters (k)
Silhouette Score vs Number of Clusters
Silhouette Score
Number of Clusters (k)

(4)


K-means Clustering Results with Centroids
Centroids
Second Principal Component
First Principal Component
Cluster
K-means Cluster Sizes
Number of Points
15411
10541
19259
Cluster
Silhouette Score Distribution per Cluster
Silhouette Coefficient
Mean: 0.387
Cluster

(5)

```
============================================================
RECURSIVE BISECTING K-MEANS CLUSTERING
============================================================
C:\Users\Mohit\AppData\Local\Temp\ipykernel_17444\2425712519.py:121: MatplotlibDeprecationWarning: The 'labels' paramet
  bp = plt.boxplot(cluster_bisect_data, labels=unique_bisect, patch_artist=True)
```