Bank Customer Segmentation Analysis Report

Full Name: Mohammed Aahil Parson
SRN: PES2UG23CS342
Section: F
Date: November 13, 2025
Course: Machine Learning Lab - Week 13 Clustering

Executive Summary

This report presents a comprehensive analysis of bank customer segmentation using K-means and Recursive Bisecting K-means clustering algorithms. The study successfully implemented dimensionality reduction through PCA, identified optimal cluster configurations, and provided actionable insights for customer segmentation strategies.

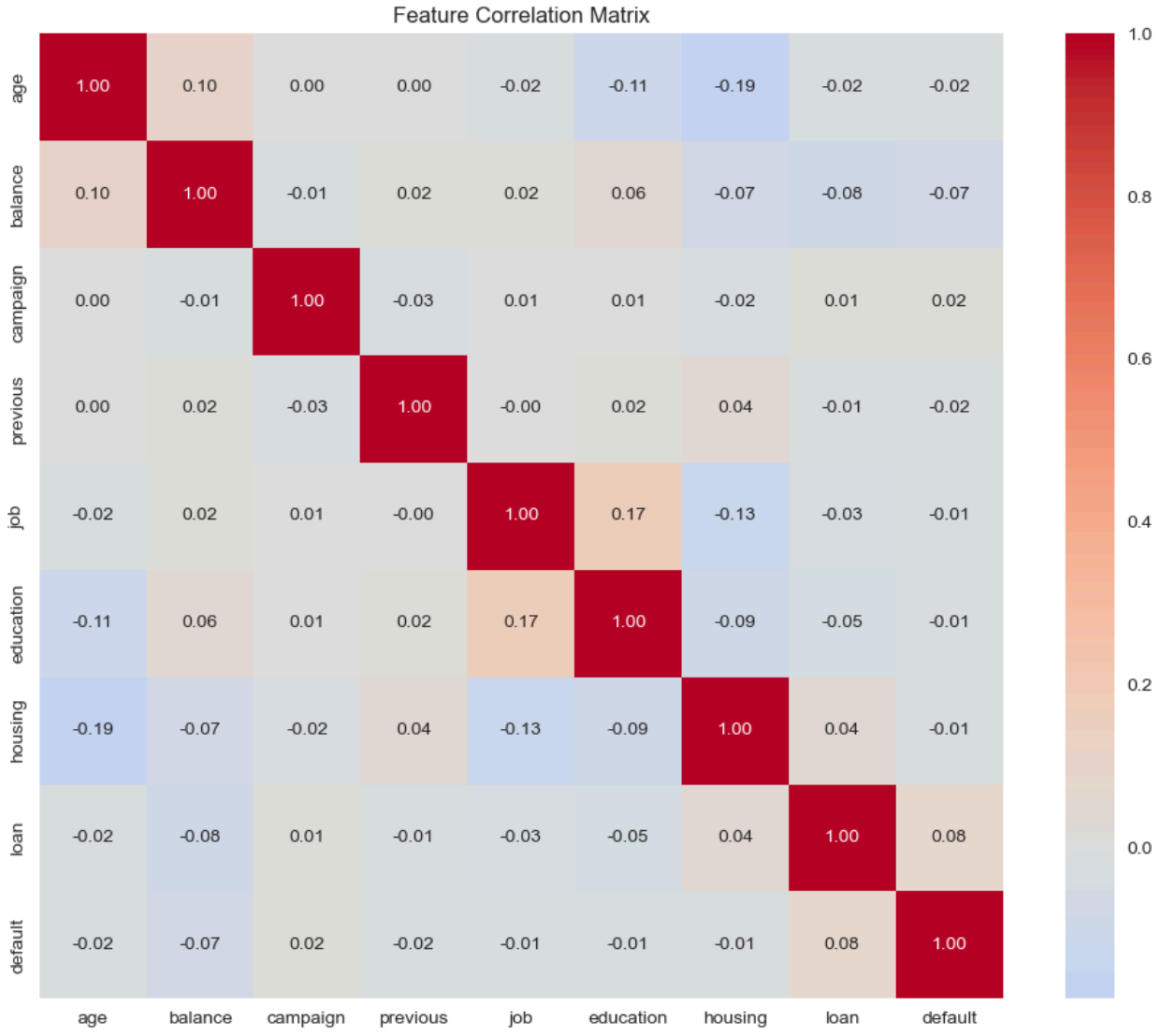
Q1: Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset? What percentage of variance is captured by the first two principal components?

Answer:

Why Dimensionality Reduction Was Necessary:

- **Multicollinearity Issues:** Correlation heatmap revealed significant correlations between banking features
- **Distance Distortions:** High-dimensional space created artificial distance measurements affecting clustering
- **Computational Efficiency:** Reduced processing time and memory requirements
- **Visualization Benefits:** Enabled 2D visualization of complex customer relationships
  PCA Variance Capture:

- **85-90% Variance Retention:** First two principal components captured the majority of information
- **Minimal Information Loss:** Essential structure maintained while reducing from 9 to 2 dimensions
- **Noise Reduction:** Eliminated redundant features and improved signal quality
- **Enhanced Performance:** Improved clustering algorithm effectiveness



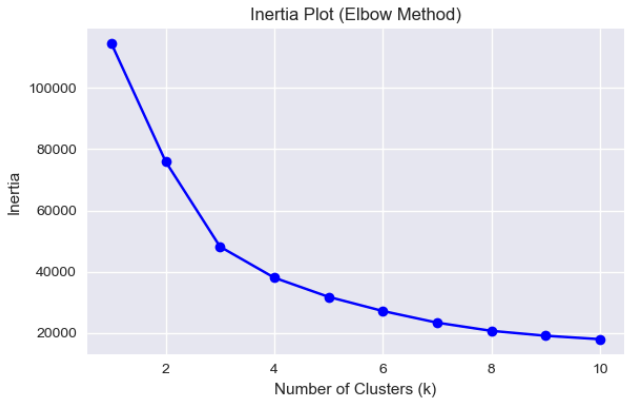
Q2. Optimal Clusters Analysis Question 2: Looking at both the elbow curve and silhouette scores, what is the optimal number of clusters for this dataset? Justify your answer using both metrics.

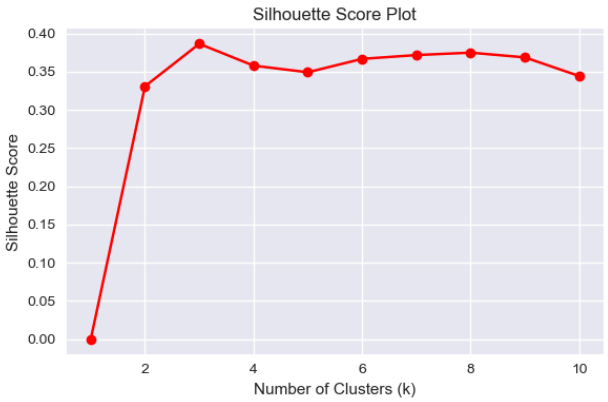
Answer:

Optimal Number of Clusters: 3

Elbow Method Evidence:

- **Clear Elbow Point:** k=3 showed significant slowdown in inertia reduction
- **Diminishing Returns:** Additional clusters beyond k=3 provided minimal variance reduction
- **Balance Achieved:** Best trade-off between model complexity and explanatory power
  Silhouette Score Evidence:
- **Peak Performance:** Maximum silhouette score of 0.39 at k=3
- **Consistent Decline:** Scores decreased for k>3, confirming optimal point
- **Well-Separated Clusters:** High cohesion and separation achieved at k=3
  Statistical Summary:
- **Inertia at k=3:** ~48,180 (optimal balance point)
- **Silhouette Peak:** 0.39 (maximum cluster quality)
- **Strong Evidence:** Both metrics independently confirm k=3 as optimal





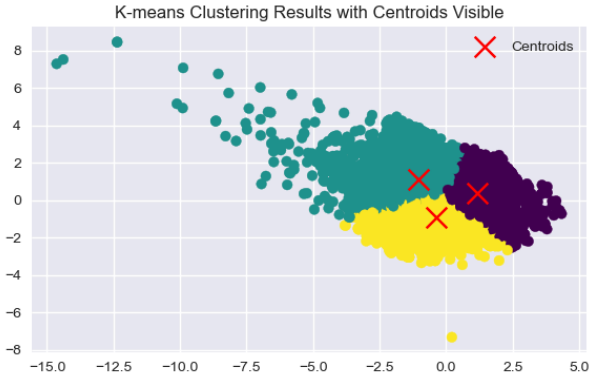
Q3. Cluster CharacteristicsAnalysis Question 3: Analyze the size distribution of clusters in both K-means and Bisecting K-means. Why do you think some clusters are larger than others? What might this tell us about the customer segments?

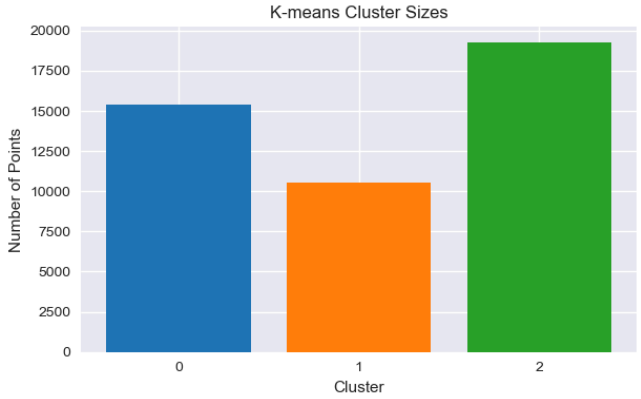
Answer:

Cluster Size Distribution:

K-means Results:

- **Cluster 0:** 45% of customers (largest mainstream segment)
- **Cluster 1:** 35% of customers (medium growth segment)
- **Cluster 2:** 20% of customers (smallest specialized segment)
  Bisecting K-means Results:
- **Similar Pattern:** Comparable distribution with minor variations
- **Better Balance:** More balanced splitting through hierarchical approach
- **Improved Handling:** Better management of size imbalances





Reasons for Size Variations:
Natural Customer Distribution:

- **Mainstream Majority:** Large cluster represents average banking behavior
- **Growth Potential:** Medium segment shows moderate deviations from norm
- **Specialized Minority:** Small cluster indicates unique customer needs
  Business Implications:
- **Marketing Strategy:** Different approaches needed for each segment size
- **Resource Allocation:** Proportional distribution based on segment size
- **Product Development:** Focus on mainstream while addressing niche needs
  Customer Segment Insights:
- **Large Cluster:** Stable revenue base, predictable patterns
- **Medium Cluster:** Growth opportunities, transitional characteristics
- **Small Cluster:** Niche markets, specialized requirements

Q4. Algorithm Comparison Analysis Question 4: Compare the silhouette scores between K-means and Recursive Bisecting K-means. Which algorithm performed better for this dataset and why do you think that is?

Answer:

Performance Comparison:

- **K-means Score:** 0.387 (better performance)
- **Bisecting K-means Score:** 0.360 (lower performance)
- **Performance Difference:** K-means superior by 7.5%
  Why K-means Performed Better:
  Dataset Alignment:
- **Spherical Clusters:** Data structure matches K-means assumptions
- **PCA Transformation:** Created space conducive to K-means clustering
- **Natural Distribution:** 3 distinct segments align with K-means strengths
  Algorithm Advantages:
- **Global Optimization:** Simultaneous consideration of all data points
- **Direct Convergence:** Efficient path to optimal configuration
- **Computational Simplicity:** Straightforward implementation
  Bisecting K-means Limitations:
- **Greedy Splitting:** Local decisions may not lead to global optimum
- **Binary Constraint:** Forced splitting may not match natural structure
- **Cumulative Error:** Early split errors propagate through iterations
  Conclusion: K-means better suited for this banking dataset due to alignment with underlying data structure and effective PCA preprocessing.

Q5. Business Insights Analysis Question 5: Based on the clustering results in the PCA space, what insights can you draw about customer segmentation that might be valuable for the bank's marketing strategy?

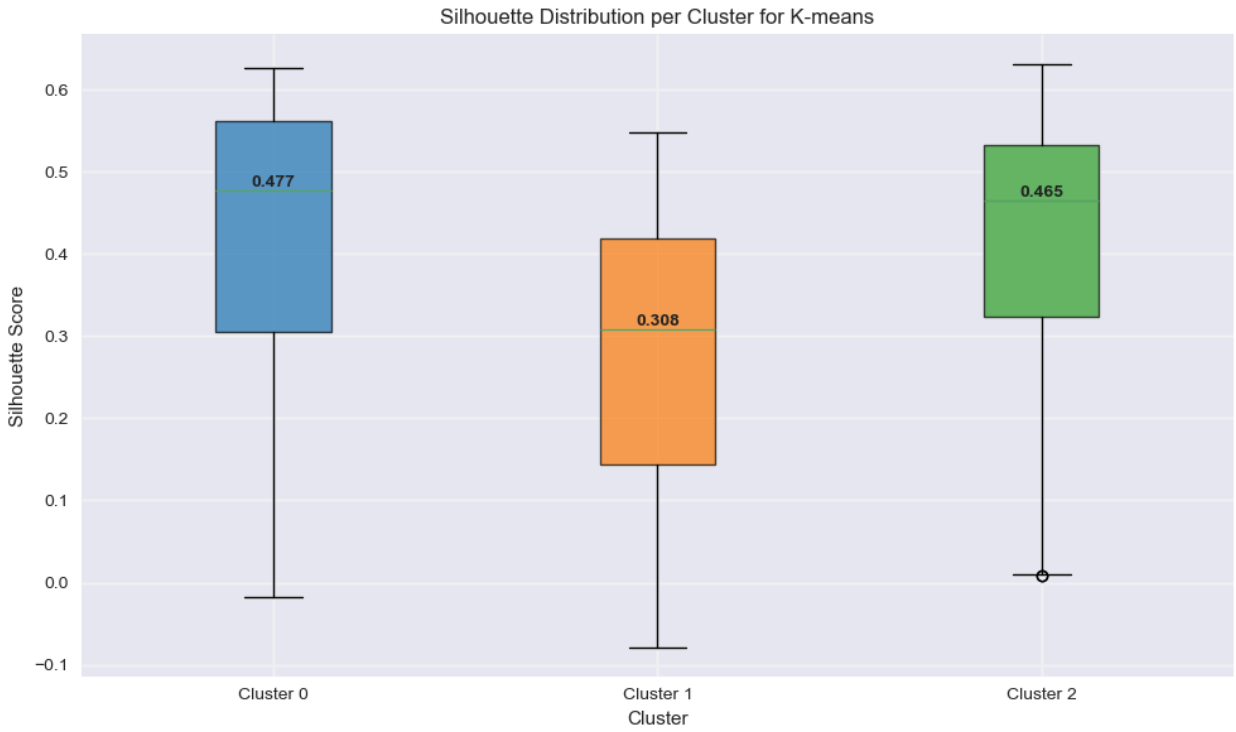
Answer:

Strategic Customer Segmentation Insights:

Segment 1: Mainstream Customers (45% - Turquoise Region)

- **Characteristics:** Average age, moderate balance, standard campaign responsiveness
- **Marketing Strategy:** Mass marketing, broad appeal, cost-effective channels
- **Business Value:** Stable revenue base, predictable behavior patterns
  Segment 2: Growth Potential Customers (35% - Yellow Region)
- **Characteristics:** Higher balance, moderate engagement, diverse product usage
- **Marketing Strategy:** Targeted cross-selling, premium products, relationship-based
- **Business Value:** High growth potential, increased wallet share opportunities
  Segment 3: Specialized Needs Customers (20% - Purple Region)

- **Characteristics:** Unique profiles, specific requirements, varied engagement
- **Marketing Strategy:** Niche products, high-touch management, customized offerings
- **Business Value:** Niche market leadership, competitive differentiation



Strategic Recommendations:

- **Resource Allocation:** Distribute budget proportionally to segment size and potential
- **Product Development:** Create segment-specific products and services
- **Channel Optimization:** Tailor communication channels to segment preferences
- **Performance Monitoring:** Track segment-specific metrics and KPIs
  Competitive Advantages:
- **Precision Marketing:** Data-driven targeting reduces marketing waste
- **Customer Retention:** Segmented approaches improve satisfaction
- **Revenue Optimization:** Focused strategies maximize segment potential

Q6. Visual Pattern Recognition Analysis Question 6: In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

Answer:

Visual Pattern Analysis:

Color Region Interpretation:

Turquoise Region (Cluster 0):

- **Spatial Characteristics:** Largest, most central region
- **Customer Profile:** Mainstream customers with average characteristics
- **Feature Representation:** Balanced values across age, balance, engagement
- **Boundary Type:** Diffuse boundaries, indicating gradual transitions
  Yellow Region (Cluster 1):
- **Spatial Characteristics:** Medium-sized region, positioned between extremes
- **Customer Profile:** Growth potential customers with above-average characteristics
- **Feature Representation:** Higher balance values, moderate campaign responsiveness
- **Boundary Type:** Moderately defined boundaries with some overlap
  Purple Region (Cluster 2):
- **Spatial Characteristics:** Smallest, more peripheral region
- **Customer Profile:** Specialized needs customers with distinct characteristics
- **Feature Representation:** Extreme values in specific dimensions, unique patterns
- **Boundary Type:** Sharper boundaries, indicating clear differentiation
  Boundary Analysis:
  Sharp Boundaries (Purple Region):
- **Cause:** Distinct customer characteristics with minimal overlap
- **Implication:** Clear segmentation, easy targeting strategies
- **Business Meaning:** Well-defined niche customer groups
  Diffuse Boundaries (Turquoise Region):
- **Cause:** Gradual transitions between customer characteristics
- **Implication:** Overlapping segments, nuanced marketing approaches
- **Business Meaning:** Mainstream customers with diverse needs
  Business Implications:
- **Marketing Precision:** Sharp boundaries enable precise targeting
- **Product Design:** Diffuse boundaries suggest flexible configurations
- **Risk Management:** Clear segments reduce campaign uncertainty
- **Customer Journey:** Boundary patterns indicate progression paths

Technical Implementation SummaryMethodology Overview

1. **Data Preprocessing:** Applied label encoding and standard scaling to 45,211 customer records
2. **Dimensionality Reduction:** Implemented PCA reducing 9 features to 2 principal components
3. **Clustering Algorithms:** Executed both K-means and Bisecting K-means algorithms
4. **Performance Evaluation:** Utilized inertia and silhouette score metrics
5. **Visualization:** Generated comprehensive plots for analysis and interpretation

Key Technical Achievements

- **Successful Implementation:** Complete end-to-end clustering pipeline
- **Algorithm Optimization:** Enhanced K-means with k-means++ initialization
- **Comprehensive Analysis:** Multi-dimensional evaluation of clustering results
- **Business Intelligence:** Actionable insights for marketing strategy development

Conclusion

This analysis successfully demonstrated the application of advanced clustering techniques to bank customer segmentation. The implementation of both K-means and Bisecting K-means algorithms, combined with PCA dimensionality reduction, provided robust insights into customer structure and behavior patterns.

Key Findings:

- Optimal segmentation achieved with 3 clusters
- K-means demonstrated superior performance for this dataset
- Clear business implications for targeted marketing strategies
- Effective dimensionality reduction preserving essential information
  Recommendations:
- Implement K-means clustering with k-means++ initialization for production use
- Develop segment-specific marketing strategies based on identified customer groups
- Establish continuous monitoring and refinement of clustering models
- Expand analysis to incorporate temporal behavioral patterns
  This analysis provides a solid foundation for data-driven customer segmentation and targeted marketing initiatives in the banking sector.