

Customer Segmentation with Unsupervised Learning

Leveraging machine learning to unlock actionable customer insights for financial institutions

Agenda

01	02
Problem Statement	Dataset Overview
Define business challenges and segmentation objectives	Explore data characteristics and feature composition
03	04
Approach & Methodology	Results Analysis
Detail preprocessing and algorithmic strategies	Compare K-Means, DBSCAN, and Hierarchical methods
05	06
Business Insights	Key Takeaways
Translate technical findings into strategic value	Summarize findings and outline future directions

Problem Statement

Financial institutions face increasing pressure to understand their customer base with precision and depth. Traditional demographic segmentation falls short in today's complex financial landscape.



Targeted Marketing

Optimize campaign effectiveness by identifying customer preferences, spending behaviors, and product affinity patterns for personalized messaging



Customer Retention

Proactively identify at-risk segments through behavioral analysis and implement tailored retention strategies before churn occurs



Risk Management

Enhance fraud detection capabilities and assess credit risk by understanding typical vs. anomalous customer behavior patterns

Unsupervised learning provides the analytical foundation to discover these hidden customer patterns without predetermined assumptions.

Dataset Overview

Data Characteristics

- **Source:** UCI/Kaggle Credit Card Dataset
- **Sample Size:** ~9,000 customer records
- **Task Type:** Unsupervised learning (no target labels)
- **Data Quality:** Real-world financial transaction data

Key Features

- Customer income levels and credit limits
- Account balance and payment history
- Purchase patterns and transaction frequency
- Cash advance behavior and installment usage

This dataset represents authentic customer behavior patterns, making it ideal for developing production-ready segmentation models that financial institutions can implement immediately.



Understanding Outliers in Financial Data

What Are Outliers?

Data points that deviate significantly from normal customer behavior patterns, statistically defined as values beyond $1.5 \times \text{IQR}$ from quartile boundaries.



Fraud Detection

Unusual spending spikes, irregular transaction timing, or atypical merchant categories often signal fraudulent activity requiring immediate investigation



Risk Assessment

Extreme credit utilization, erratic payment patterns, or sudden behavioral changes help identify high-risk customers for proactive intervention



Data Quality

Outliers reveal data collection errors, system glitches, or processing anomalies that could compromise model accuracy and business decisions



VIP Opportunities

High-value customers with exceptional spending or unique behaviors represent premium segments worth specialized attention and tailored services

📄 **Our Strategy:** We retained outliers during initial analysis, then used DBSCAN to systematically identify and separate 386 outlying customers for dedicated analysis alongside main segments.



Approach & Methodology

Our comprehensive analytical framework combines rigorous data preparation with multiple clustering approaches to ensure robust customer segmentation results.



Data Cleaning & Preprocessing

Systematic removal of duplicates, handling missing values, and feature standardization to ensure algorithm compatibility



Dimensionality Reduction (PCA)

Principal Component Analysis to reduce noise, improve visualization, and optimize computational efficiency while retaining critical variance



Multi-Algorithm Approach

Comparative analysis using K-Means, DBSCAN, and Hierarchical Clustering to identify the most effective segmentation strategy

This methodical approach ensures our segmentation results are both statistically sound and practically applicable for financial decision-making.

Data Cleaning & Preprocessing Pipeline

1

Duplicate Removal

Identified and eliminated redundant customer records to prevent clustering bias and ensure unique customer representation

2

Irrelevant Data Filtering

Removed non-predictive features and customer identifiers that don't contribute to behavioral segmentation patterns

3

Data Type Optimization

Converted categorical variables and ensured proper numerical formatting for machine learning algorithm compatibility

4

Missing Value Treatment

Applied domain-appropriate imputation strategies, considering financial context and feature relationships

5

Feature Standardization

Normalized scale differences between income, balance, and transaction features to prevent algorithmic bias

Each preprocessing step was carefully validated to maintain data integrity while optimizing clustering performance.

Principal Component Analysis (PCA)

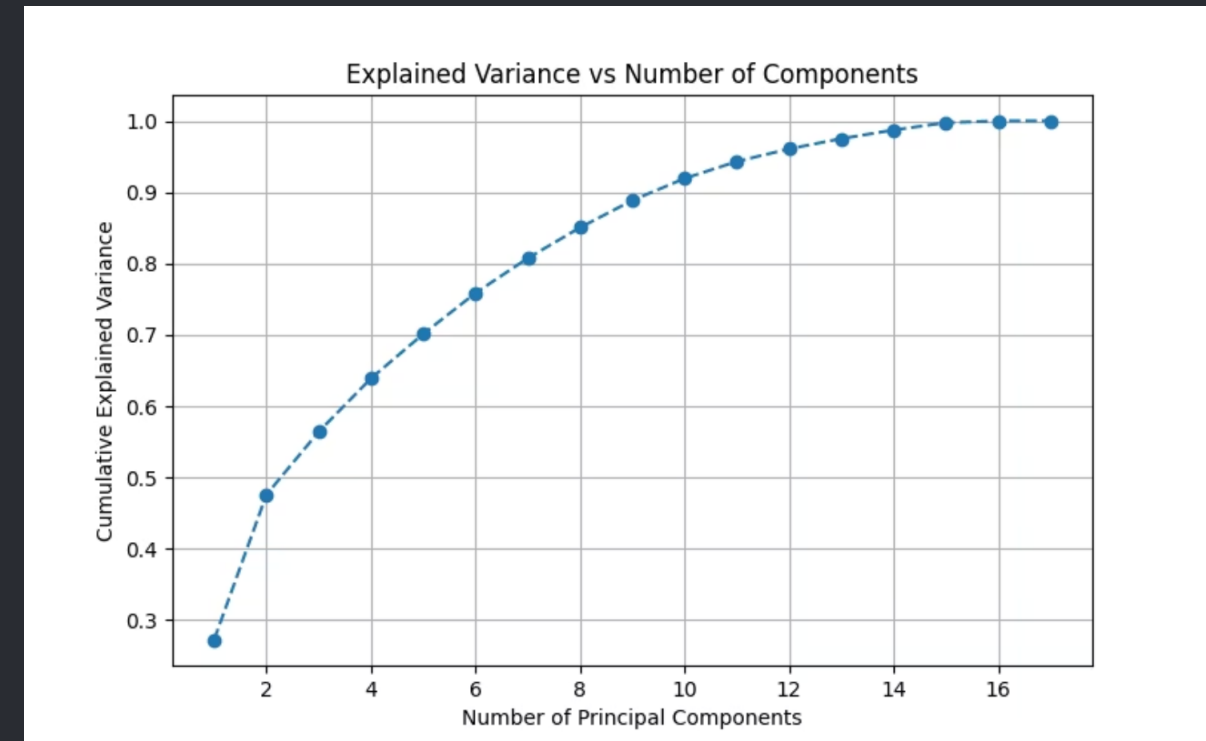
Dimensionality Reduction Strategy

PCA transforms our high-dimensional customer feature space into a more manageable representation while preserving critical behavioral patterns.

Key Benefits:

- **Noise Reduction:** Eliminates irrelevant feature variations
- **Enhanced Visualization:** Enables meaningful 2D/3D cluster plotting
- **Computational Efficiency:** Reduces time and memory complexity
- **Optimal Representation:** 8 components capture 80%+ variance

This dimensionality reduction maintains the essence of customer behavior while creating a foundation for more effective clustering algorithms.



Evaluation Metrics Framework

Comprehensive model assessment requires multiple complementary metrics to ensure clustering quality and business relevance.

1

Silhouette Score

Range: 0 to 1 **higher is better**

Measures intra-cluster cohesion versus inter-cluster separation. Values above 0.5 indicate strong, well-defined segments.

2

Dunn Index

Interpretation: **higher is better**

Ratio of minimum inter-cluster distance to maximum intra-cluster distance. Higher values indicate well-separated, compact clusters.

3

Davies-Bouldin Index

Interpretation: **lower is better**

Average similarity between clusters and their most similar neighbors. Lower values indicate better cluster separation.

4

Calinski-Harabasz Index

Interpretation: **higher is better**

Ratio of between-cluster to within-cluster variance. Higher values indicate denser, more distinct customer segments.



K-Means Clustering Results

Performance Assessment

0.146

Silhouette Score

Below acceptable threshold

1.649

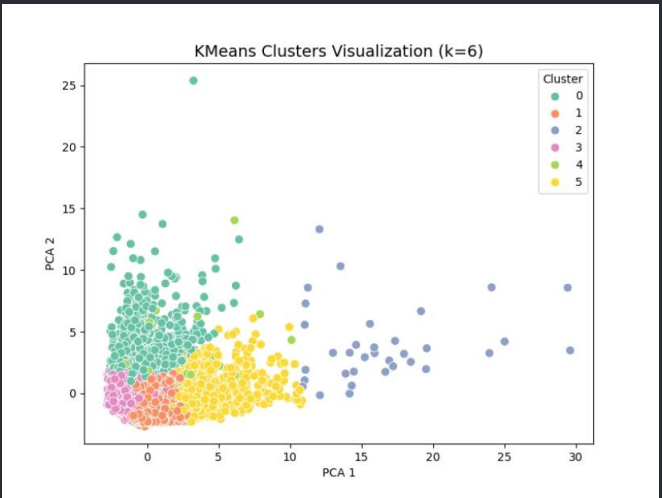
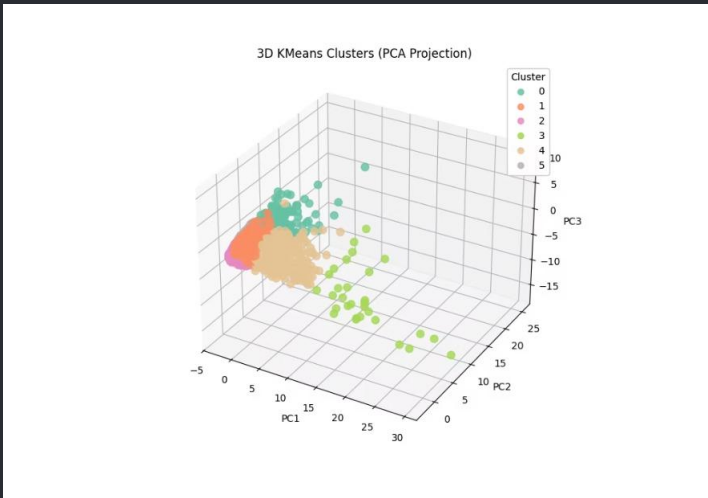
Davies-Bouldin

High inter-cluster similarity

0.002

Dunn Index

Poor cluster separation



Key Finding: K-Means struggled with the complex, non-spherical nature of financial customer behaviors, producing overlapping segments with limited business interpretability.

K-Means Clustering Results

Business Impact: The poor separation metrics indicate these clusters would provide limited value for targeted marketing campaigns or risk assessment strategies. Alternative approaches are necessary for actionable customer insights.


DBSCAN Analysis Results

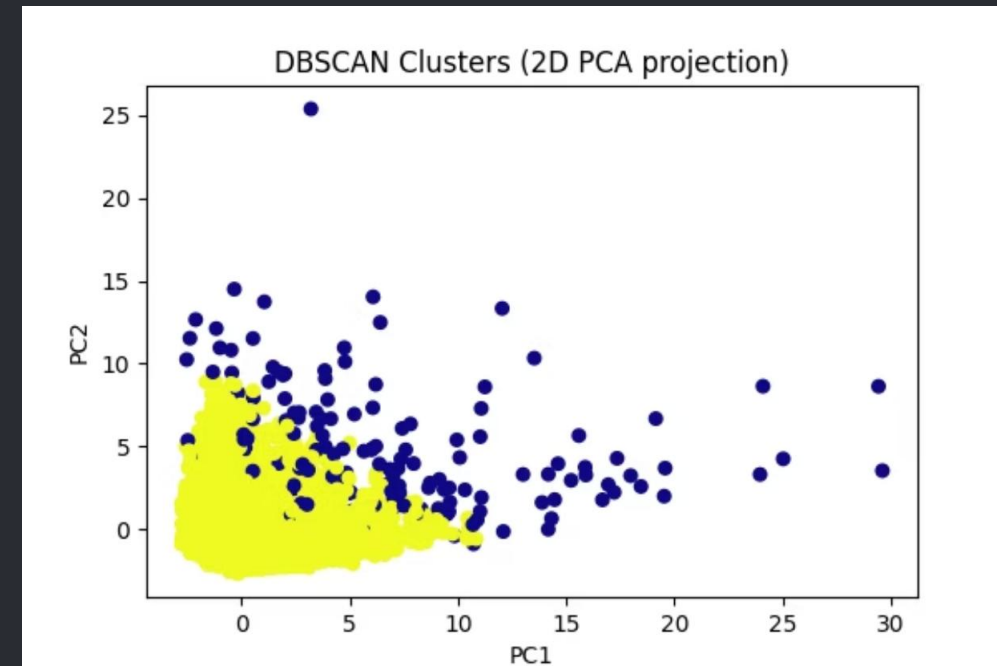
Algorithm Performance

DBSCAN's density-based approach revealed significant challenges with our financial dataset's characteristics.

Key Findings:

- **Outlier Detection:** Successfully identified numerous outlying customers
- **Parameter Sensitivity:** Results highly dependent on epsilon and min-points selection
- **Clustering Collapse:** Majority of data treated as noise or single large cluster
- **Metric Failure:** Silhouette and DBI scores undefined due to poor cluster formation

 **Conclusion:** While DBSCAN excels at outlier detection, its assumption of uniform density doesn't align with the varied behavioral patterns in financial customer data.



Strategic Recommendation: DBSCAN's strength in outlier identification makes it valuable as a preprocessing step for fraud detection, but alternative clustering methods are needed for meaningful customer segmentation.

Hierarchical Clustering Results

0.156

Silhouette Score

Below acceptable threshold

1.668

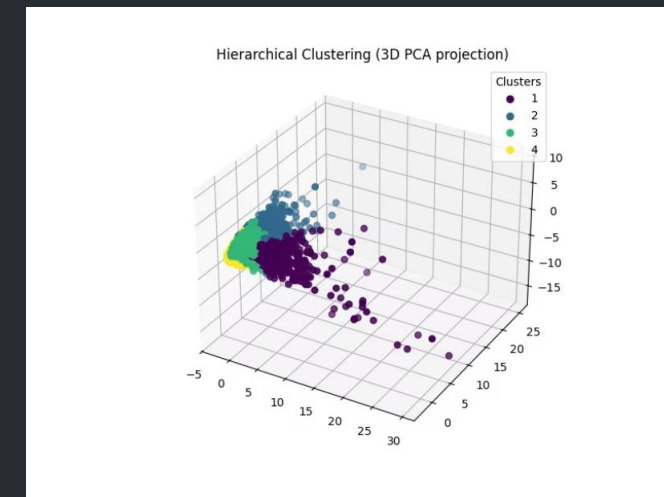
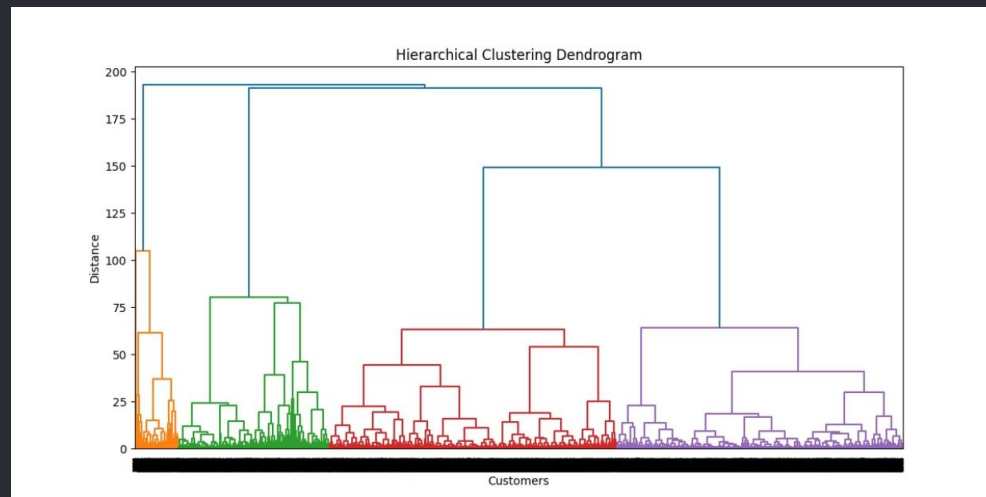
Davies-Bouldin

High inter-cluster similarity

1450.842

CHI

Poor cluster separation



📌 **Conclusion:** While Hierarchical clustering results, might be better than k-means, its still not ideal, but its a good way to visualize the formation of clusters. Limited Business use.

Hybrid Approach (DBSCAN + K-Means)

0.231

Silhouette Score

Best among others

1.4706

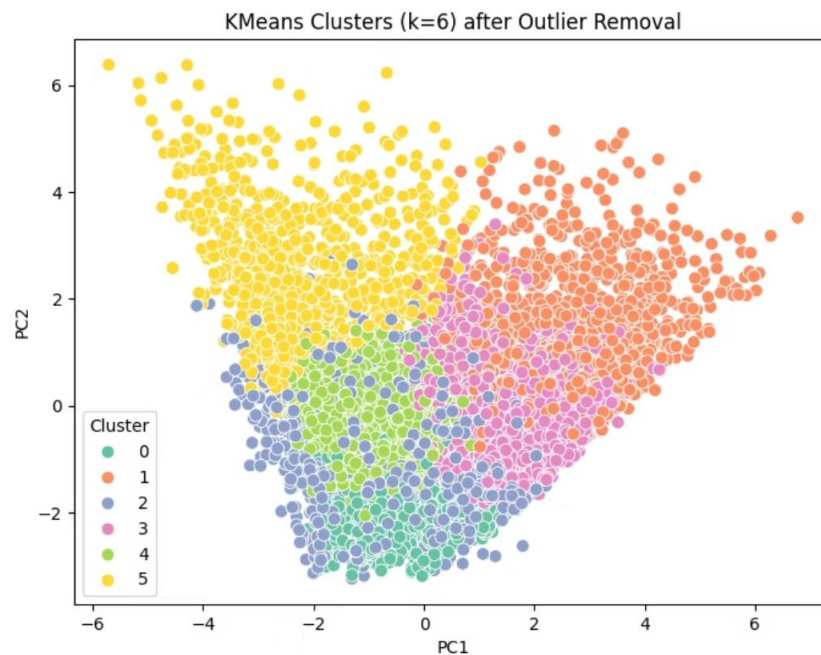
Davies-Bouldin

High inter-cluster similarity

386

Outliers removed

Data points removed



Conclusion: Best approach, I could find, by removing outliers using DBSCAN and then making clusters using k-means.

6 clusters formed

Advantages: Has better silhouette score than others and better clusters formation.

Disadvantages: Might have remove important data points as outliers.

Solution: To run algorithm on only outliers set of data points.

Business Insights

Cluster 0 – Low activity customers → Upsell campaigns

Cluster 1 – Premium high spenders → Loyalty & rewards

Cluster 2 – Disciplined payers → Cross-sell products

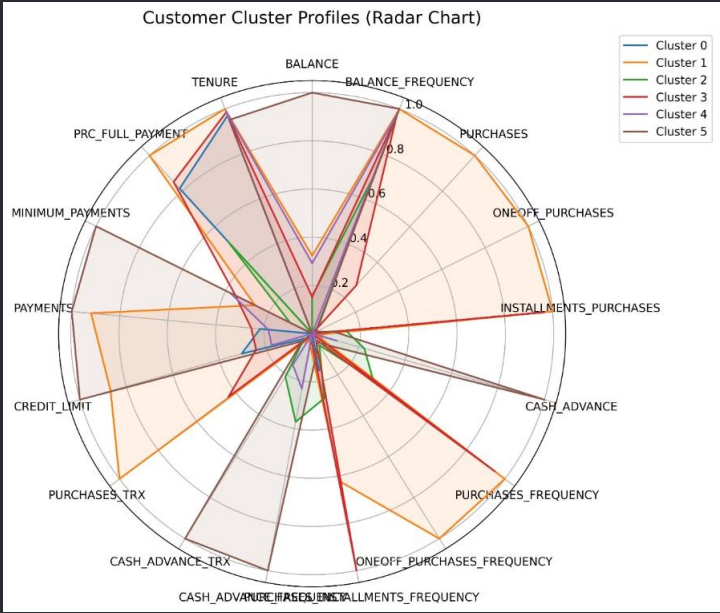
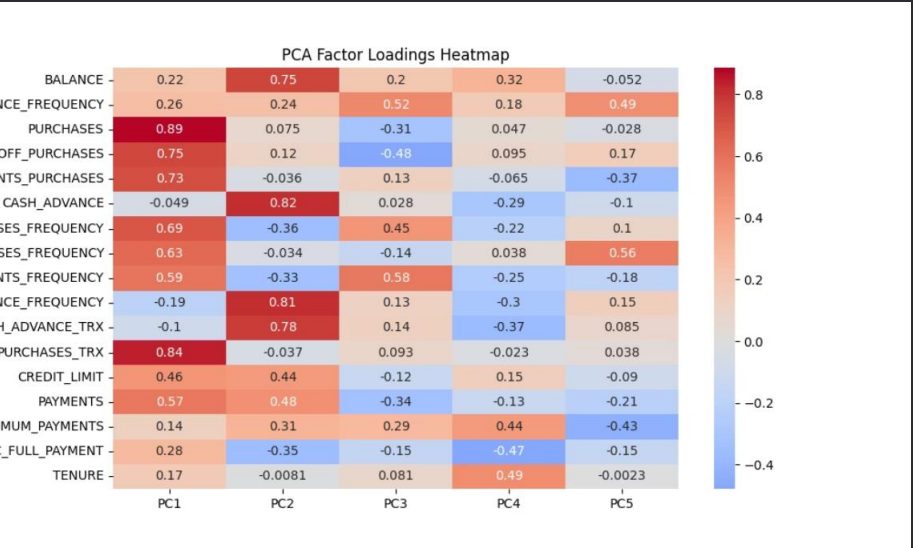
Cluster 3/4 – Balanced spenders → EMI/instalment offers

Cluster 5 – Frequent small transactions → Cashback programs

6 meaningful customer groups identified

Outliers provide insights for fraud/risk monitoring

Segmentation enables targeted marketing strategies



Future work

Build interactive Streamlit dashboard

Where users can add their own dataset, optimal unsupervised algorithm will be automatically selected and output will be provided.

Technologies Used

Python

Scikit-learn

Pandas and NumPy

Matplotlib and Seaborn

Thank you