Bank Customer Segmentation Using Clustering Algorithms

Gideon Mpungu
Department of Computer Science
Makerere University
mpungu.gideon@students.mak.ac.ug

December 29, 2023

Abstract

This paper investigates the effectiveness of three clustering algorithms – K-Means, Affinity Propagation, and Hierarchical Clustering – in segmenting a bank's customer base based on credit data. The analysis reveals valuable insights into customer heterogeneity, with each algorithm contributing unique perspectives. K-Means provides a straightforward segmentation into three clusters, readily interpretable based on age and credit preferences. Affinity Propagation automatically determines the optimal number of clusters, identifying a distinctive fourth segment characterized by younger customers with high credit needs. Hierarchical Clustering visually depicts the data's hierarchical structure and confirms the three-cluster solution from K-Means. The findings hold significant implications for the banking industry, particularly in targeted marketing, personalized product offerings, and risk management. Future research directions include ensemble clustering, incorporating additional data, and predictive modeling. By leveraging the power of diverse clustering algorithms, banks can unlock valuable insights into their customer base, paving the way for a more competitive and customer-centric banking experience.

1 Introduction

Exploring Idea 2, our research delves into clustering algorithms, a realm beyond the extensively explored classification algorithms. This study focuses on implementing and analyzing the application of KMeans, Affinity Propagation, and Hierarchical Clustering, particularly in the context of customer segmentation within the banking industry. The landscape of customer understanding is constantly evolving, demanding innovative approaches to extract hidden patterns and segment diverse populations. By applying these powerful techniques to banking data, we aim to uncover unique insights into customer behaviors and unlock the potential of these algorithms in real-world scenarios. Conventional classification, while valuable, often relies on pre-defined labels or categories, potentially overlooking nuanced relationships and underlying structures within data. Clustering algorithms, on the other hand, operate without such constraints, grouping data points based on inherent similarities and dissimilarities. This allows us to discover natural clusters within the customer base, revealing unexpected patterns and hidden segments that might be invisible to traditional methods.

Keywords: Clustering Algorithms, EDA, K-Means, Affinity Propagation, Hierarchical Clustering, Silhouette Score, Inertia

2 Literature Review

The ever-evolving landscape of customer behavior demands innovative approaches to decode hidden patterns and segment diverse populations. Clustering algorithms, operating without pre-

defined labels, emerge as powerful tools for customer segmentation, grouping individuals based on inherent similarities and dissimilarities. This literature review explores existing applications of clustering algorithms in banking, identifies research gaps, and highlights the potential of a comparative analysis utilizing KMeans, Affinity Propagation, and Hierarchical Clustering within this specific domain.

2.1 KMeans: Efficiency with Limitations

KMeans, a partitioning-based algorithm, has been extensively applied in customer segmentation within the banking sector. Its simplicity and efficiency make it ideal for handling large datasets. Studies like [1] and [2] have demonstrated its application. However, its sensitivity to outliers and reliance on pre-defined centroids limit its flexibility.

2.2 Affinity Propagation: Unveiling Hidden Relationships

Affinity Propagation offers a message-passing algorithm for cluster identification based on mutual preferences. [3] successfully applied it to segment insurance company data, showcasing its effectiveness in complex customer relationships. However, its computational complexity can be a drawback in banking.

2.3 Hierarchical Clustering: Deciphering Complex Structures

Hierarchical clustering builds a hierarchy of nested clusters, providing insights into customer relationships at different granularity levels. [4] segmented bank customers based on transaction data using Hierarchical Clustering, revealing insights into financial behavior and risk profiles.

2.4 Research Gaps and Uniqueness

Despite existing studies, research gaps remain in the application of clustering algorithms for banking customer segmentation:

- **Comparative Analysis:** Existing studies often focus on individual algorithms, limiting insights into their comparative performance and suitability for specific segmentation tasks within the banking domain.
- **Domain-Specific Challenges:** Banking data presents unique challenges for clustering algorithms, such as mixed data types with varying scales, high dimensionality, and imbalanced clusters with highly heterogeneous spending patterns.

Our research addresses these gaps by:

- **Comparative Analysis:** Applying KMeans, Affinity Propagation, and Hierarchical Clustering to the same banking dataset allows for a comprehensive comparison of their performance.
- **Domain-Specific Adaptation:** Tailoring preprocessing techniques and algorithm parameters to address the challenges of banking data.

3 Methodology

Our research employs a comparative analysis of three clustering algorithms - KMeans, Affinity Propagation, and Hierarchical Clustering - to segment banking customers based on financial data. This section outlines the key steps taken, including dataset selection, preparation, and model selection/optimization.

3.1 Dataset Description

We selected the publicly available dataset from `https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data` for its comprehensive coverage of factors crucial to customer segmentation in the banking sector. This decision was influenced by several considerations:

3.1.1 Factors Considered for Dataset Selection

- **Relevance to Customer Segmentation:** The dataset contains diverse features directly relevant to understanding customer behaviors in the banking domain.
- **Data Composition:** The dataset includes a mix of categorical and numerical features, crucial for diverse analysis techniques in clustering.
- **Usability and Documentation:** The dataset is well-documented and user-friendly, facilitating efficient analysis and reproducibility.
- **Dataset Size:** Its substantial size provides a robust base for meaningful customer segmentation.
- **Public Availability and Ethical Use:** Ensures reproducibility of results and ethical compliance in data usage.

3.1.2 Key Features of the Dataset

- **Demographic Information:** Age, Sex, Job, Housing - to understand customer profiles.
- **Account Details:** Checking account, savings account, credit amount - to analyze financial behaviors.
- **Credit History:** Duration, Purpose - to assess creditworthiness and spending patterns.

These features are pivotal in developing machine learning models that can accurately segment customers, addressing real-life challenges in the banking sector.

3.2 Data Preparation and Exploratory Data Analysis (EDA)

3.2.1 Data Cleaning

1. Handling Unnecessary Column: The first column was identified as an index and set as the DataFrame's index.
2. Analysis of Missing Values and Data Types: Assessed and imputed missing values in "Saving accounts" and "Checking account" with "None" to indicate the absence of an account.

3.2.2 Data Exploration

The correlation plot see Figure 1, shows a linear correlation with Pearson value of 0.62 and very small p-value which makes sense because usually, people take bigger credits for longer periods. When stratifying by sex, see Figure 2, we observe that women tend to be younger than men, however, there is no clear difference between men and women in terms of amount and duration of the credit. Further visualization and analysis revealed insights such as positive correlation between credit duration and amount, and preferences in credit behavior based on age and job category.

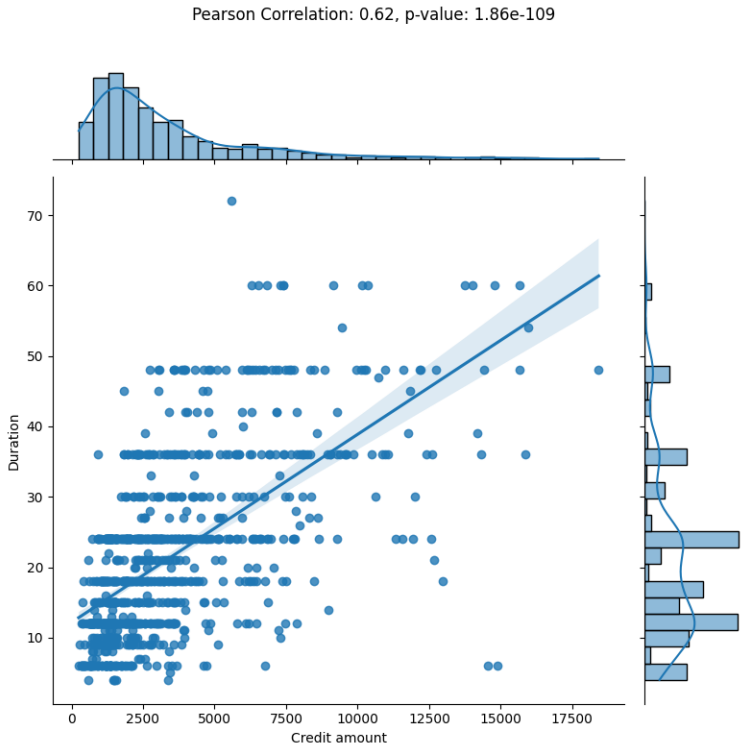

Figure 1: Correlation plot between credit amount and duration

3.2.3 Data Transformation

- KMeans Clustering: Applied logarithmic transformation to features 'Age', 'Credit amount', and 'Duration' to reduce right-skewness and improve clustering results. Also, standardized features to have a mean of 0 and a standard deviation of 1 to ensure equal feature weighting, as KMeans is sensitive to scale.
- Affinity Propagation: No data transformation applied as the algorithm is relatively robust to skewness and scale differences.
- Hierarchical Clustering: Applied logarithmic transformation to features 'Age', 'Credit amount', and 'Duration' to mitigate the impact of skewness on cluster formation, as hierarchical clustering can be sensitive to non-normal distributions.

3.3 ML Model Selection and Optimization

3.3.1 KMeans Clustering

KMeans is a partitioning method that divides the dataset into K clusters. It assigns each data point to the nearest cluster center and iteratively refines these centers. With our dataset, KMeans aimed to segment customers based on similarities in their credit profiles. The number of clusters was optimized, evaluated using the silhouette score and Calinski-Harabasz score to ensure distinct, well-separated customer groups.

3.3.2 Affinity Propagation

Affinity Propagation does not require the number of clusters to be predefined. It works by sending messages between pairs of samples until clusters emerge. This method was used to identify natural groupings in our dataset based on credit behavior. The preference parameter and damping factor were tuned to balance sensitivity to data points and convergence stability.

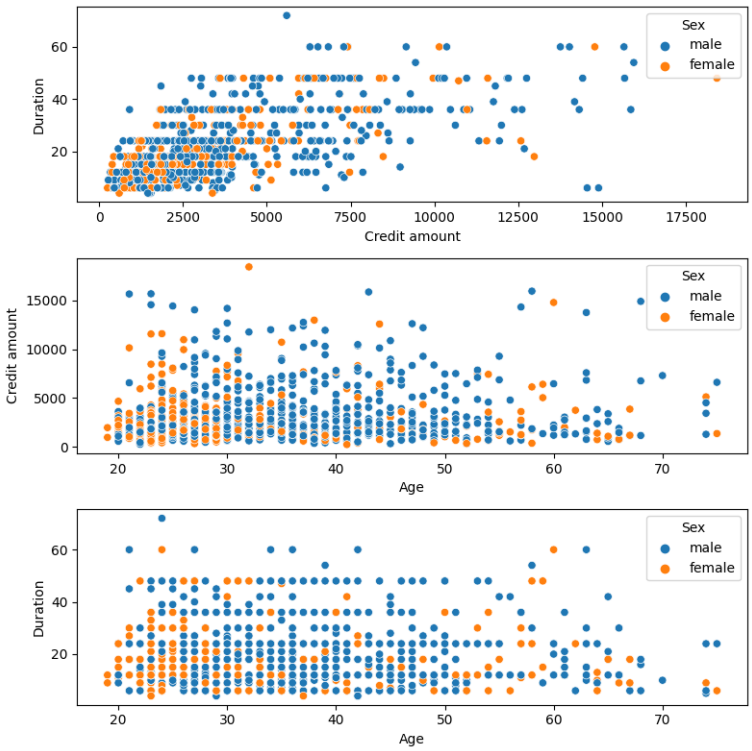

Figure 2: Scatter plots for numerical variables

3.3.3 Hierarchical Clustering

Hierarchical Clustering builds a hierarchy of clusters either through a bottom-up (agglomerative) or top-down approach. In this study, agglomerative helped to uncover hierarchical relationships in customer data, offering insights into different customer segments. We evaluated different linkage criteria and distance metrics to find the most coherent and interpretable clustering structure for our banking dataset.

4 Results and Discussion

4.1 Machine Learning Evaluation Metrics

This section details the key metrics used to evaluate the performance of each clustering algorithm employed for customer segmentation.

4.1.1 K-Means Clustering

- **Silhouette Score:** Measures cohesion within clusters and separation between them. Higher scores indicate better cluster quality. Best scores observed for 3 clusters.
- **Inertia:** Reflects the sum of squared distances within each cluster. Lower inertia indicates tighter clusters. Optimal number of clusters determined as 3.

4.1.2 Affinity Propagation

- **Number of Clusters:** Automatically determined based on data. Segmentation with 4 clusters provided an alternative perspective.
- **Cluster Centers:** Analysis of centers revealed each segment's characteristics.

4.1.3 Hierarchical Clustering (Agglomerative)

- **Dendrogram:** Visualized the hierarchical structure of clusters. Optimal number of clusters determined as 3.
- **Cluster Analysis:** Similar to K-Means, analyzing mean values for each cluster revealed characteristics.

4.2 Comparative Analysis

All three algorithms identified 3 clusters as optimal. However, See Figure 3 comparison of Silhouette Scores across clustering algorithms in our study reveals KMeans as the most effective, achieving a score of 0.30. This indicates relatively good cluster separation and cohesion. Affinity Propagation follows with a score of 0.27, suggesting moderate effectiveness with slightly less distinct clusters. Hierarchical Clustering, with the lowest score of 0.24, indicates its clusters are the least distinct among the three. These results highlight KMeans as the superior method for distinct and cohesive clustering in this dataset, with the other methods providing moderate clustering quality. See Table 1 for a comparison of key aspects of the clustering algorithms.

Table 1: Comparison of clustering algorithms

Feature	**K-Means Clustering**	**Affinity Propagation**	**Hierarchical Clustering**
Number of Clusters	Optimal: 3	Determined by algorithm (4)	Optimal: 3
Silhouette Score	0.3	0.27	0.24
Inertia	Decreases with more clusters	N/A	N/A
Cluster Centers	3 distinct centers	4 unique centers	3 centers aligned with K-Means
Interpretation	Clusters aligned with demographics. See Figure 4	Distinct 4th cluster. See Figure 5	Similar to K-Means. See Figure 6
Strengths	Simple, efficient	Determines cluster count	Visualizes structure
Limitations	Requires cluster count	Sensitive to parameter	Not suited for large datasets

4.2.1 Insights from Clustering Analysis

Each algorithm provided unique insights into customer segmentation. K-Means identified clear and interpretable segments, Affinity Propagation revealed a distinct segmentation that K-Means could not identify, and Hierarchical Clustering offered a different perspective on data structure and customer relationships.

4.2.2 Recommendations

Consider Affinity Propagation for its ability to reveal unique clusters. Use K-Means for its simplicity and interpretability, but carefully evaluate cluster quality. Employ Hierarchical Clustering to visualize hierarchical structure and validate cluster count. Combine insights from multiple algorithms for a comprehensive understanding of customer segmentation.

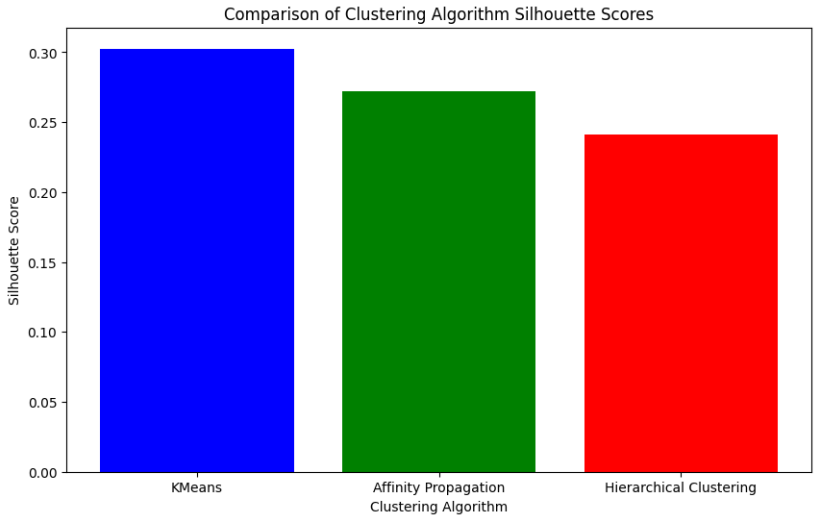

Figure 3: Comparison of clustering algorithms using silhouette scores

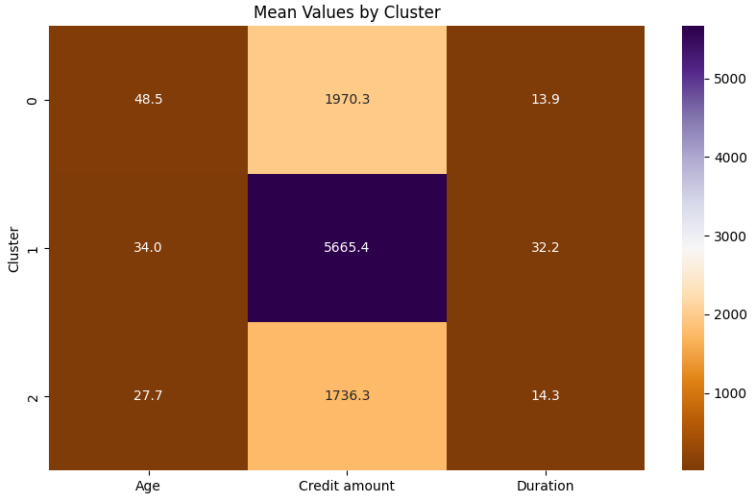

Figure 4: Cluster mean values with kmeans clustering

5 Conclusion

This study's application of K-Means, Affinity Propagation, and Hierarchical Clustering on banking data has uncovered distinct customer behaviors and preferences. K-Means identified clear segments, notably a group characterized by mature customers favoring stability in financial products. Affinity Propagation revealed a unique cluster of younger customers inclined towards high-credit products. Hierarchical Clustering confirmed these findings and added a perspective on customers' varying financial behaviors across different age groups.

These insights are pivotal for banks to tailor their marketing strategies, product offerings, and risk management approaches. For instance, targeting the younger, high-credit-need segment with aggressive investment options, while offering secure, long-term financial planning products to the more conservative segments. The study underscores the value of nuanced customer segmentation in crafting personalized banking experiences and strategic risk mitigation.

Future research should delve deeper into the behaviors within each segment for more refined marketing tactics and explore additional data dimensions to further enrich customer insights. The integration of diverse clustering algorithms can significantly enhance the understanding of customer dynamics, propelling the banking industry towards more informed and customer-

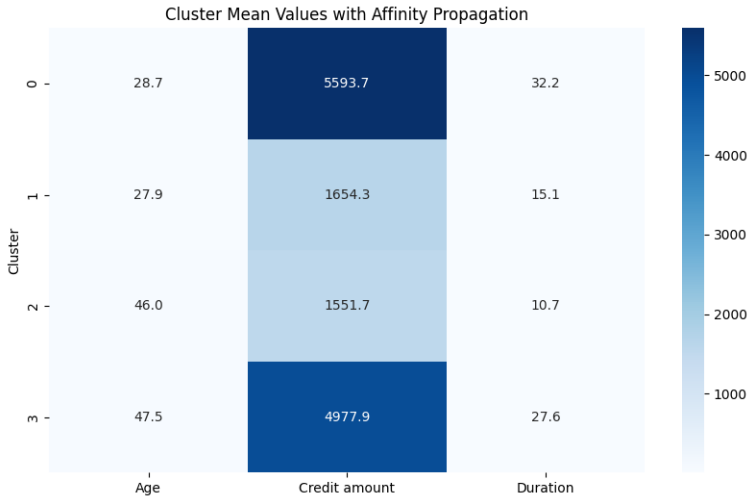

Figure 5: Cluster mean values with affinity clustering

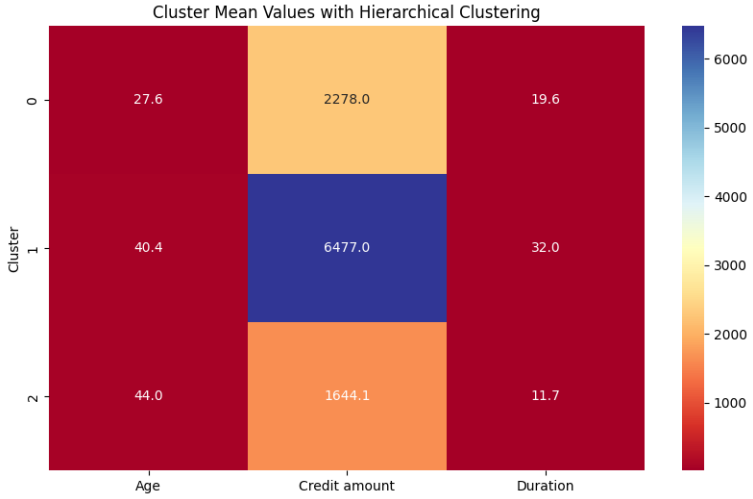

Figure 6: Cluster mean values with hierarchical clustering

centric strategies.

References

[1] M. Aryuni, E. D. Madyatmadja, and E. Miranda, "Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering," in *2018 International Conference on Information Management and Technology (ICIMTech)*, 2018, pp. 1-9.

[2] R. W. S. Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," 2020.

[3] L. Zhang, J. Priestley, J. Demaio, X. Ni, and X. Tian, "Measuring Customer Similarity and Identifying Cross-Selling Products by Community Detection," *Big Data*, 2020.

[4] H. Su-li, "The customer segmentation of commercial banks based on unascertained clustering," in *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, 2010, vol. 1, pp. 297-300.