SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Acces

Customer Segmentation In Financial Services Using Gaussian Mixture Models (GMM) And K-Means

M. Suganya[1*], Chippy Mohan[2], Ajitesh Kumar[3], S. Bhavadharani[4], Raju Katru[5], M. Nithiya[6]

[1*]Assistant Professor, Department of Electronics and Communication Engineering, New Prince Shri Bhavani College of Engineering and Technology, Chennai, Tamil Nadu, India. E-mail: suganyam66@gmail.com, https://orcid.org/0009-0007-8790-1889

[2]Assistant Professor, School of Business & Management, Christ University, Bangalore, India. E-mail: chhipymohan@gmail.com, https://orcid.org/0000-0001-8589-4944

[3]Department of Computer Engineering & Applications, GLA University, Mathura, Uttar Pradesh, India. E-mail: ajitesh.kumar@gla.ac.in, https://orcid.org/0000-0003-0196-5640

[4]Assistant Professor, Department of Commerce, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: bhavadharani@maher.ac.in, https://orcid.org/0009-0005-2449-7618

[5]Department of Electronics and Communication Engineering, Ramachandra College of Engineering, Eluru, India. E-mail: rajucollege2020@rcee.ac.in, https://orcid.org/0000-0002-1563-487X

[6]Assistant Professor, Artificial Intelligence and Data Science, Mahendra Engineering College, Namakkal, Tamil Nadu, India. E-mail: nithiyam@mahendra.info, https://orcid.org/0009-0008-0907-650X

*Corresponding author: Email: suganyam66@gmail.com

Abstract

In the financial services industry, customer segmentation has emerged as a vital analytical process for delivering personalized banking, customer retention, and intelligent decision-making. In banking datasets, which are typically huge, traditional segmentation methods fall short in capturing dynamic and overlapping behaviors of customers. This research introduces an AI-based customer segmentation approach, based on K-Means clustering and Gaussian Mixture Models (GMM) on publicly available Indian banking transaction data, which consists of more than 1 million banking transaction records and more than 800,000 customers. It includes data preprocessing, feature engineering based on Recency-Frequency-Monetary (RFM), Principal Component Analysis (PCA), and probabilistic clustering for meaningful customer group identification. The experimental analysis showed that PCA was able to retain 91.4% of the variance of the data set while at the same time reducing the dimensionality of the feature space. The Silhouette Score of K-Means is 0.742, while the Davies-Bouldin Index is 0.481; GMM improved the results in comparison to K-Means, with a Silhouette Score of 0.861 and Davies-Bouldin Index of 0.294. The proposed framework was able to successfully identify the five major segments of customers, such as premium customers, digitally active customers, dormant customers, and high-risk transaction groups. Results validate the modeling of the customer behavior and the quality of the customer segmentation in a financial environment by using probabilistic clustering. The proposed framework offers a lot of support to the provision of personalized banking services, customer intelligence generation, fraud monitoring, and strategic financial decision-making.

Keywords Customer Segmentation, Gaussian Mixture Model, K-Means Clustering, Banking Analytics, RFM Analysis, Financial Intelligence.

1. Introduction

The surge in digital banking and financial technologies has generated a vast amount of customer transaction data, necessitating smart data analytical methods to make wise business decisions. With more and more modern banking institutions implementing AI and machine learning methods to evaluate customer behavior, improve financial services and increase customer retention rates, it is necessary to utilize appropriate methodologies in the development of models. Appropriate methodologies are important in model development as modern banking institutions apply more and more AI and machine learning methods to understand customer behavior, enhance

financial services, and reinforce customer retention policies. Customer segmentation is one of the most important analytical tools to segment customers into meaningful groups on a transactional, behavioral, and demographic basis. In digital finance contexts it has been proven that Gaussian mixture models enhance the accuracy of targeted retail price and customer profiling [1]. The significance of clustering algorithms in the banking field, particularly in banking analytics and intelligent customer management systems, has also been shown through comparative evaluations of machine learning models [2][3]. Research also reveals that clustering techniques are very useful for discovering hidden patterns of customer behavior in the huge data sets using an unsupervised learning approach [6]. Advanced segmentation solutions, which now include customer journey analytics, have also improved digital finance engagement and personalization solutions [12].

The primary way of segmenting customers in the banking system is by the static division of customers based on their demographics and the rule-based analysis of their financial behavior, which is inefficient for capturing dynamic customer behaviors and overlapping financial patterns. These methods are not always suitable to capture complex characteristics of transactions, the diversity of spending, and the changing preferences of customers in a modern digital banking environment. Hard clustering, like K-Means, does not account for probabilistic relationships of customers to the clusters and reduces the flexibility of segmentation and its interpretability. Based on optimization analysis, it is found that the need still arises to have more adaptive analytical mechanisms that can process the customer's behavior dynamically in an efficient manner [11]. Additionally, studies concerning behavioral banking analytics have also pointed out the difficulties involved in the feature transformation and identification of customer archetypes in banking data [15]. Intelligent customer analytics can also help to a great extent in the banking operations and decision-making process in the context of risk-aware banking operations as revealed by financial stability studies [21].

As more and more banks seek to deliver personalized services, intelligent recommendation systems, and customer-centric strategies, the need for more advanced AI-powered customer segmentation frameworks has emerged. Financial institutions need scalable clustering models that are able to process millions of transaction records and identify hidden customer personas and spending behaviors [13]. Hybrid ensemble-based customer segmentation techniques have shown to improve the reliability and analytical performance in clustering [5]. A combination of recency, frequency, monetary, and time metrics has also been proven effective in behavioral customer analysis and segmentation accuracy using machine learning methods [7]. Adaptive RFM model selection through meta-learning has enriched the ability of the customer profiling framework in a dynamic business environment [10]. Probabilistic clustering has also proven to be effective in modeling complex customer relationships in large-scale systems when using deep Gaussian mixture learning models [17].

Research Aim

To build an intelligent segmentation model for financial services based on Gaussian Mixture Models (GMM) and K-Means clustering on large-scale banking transaction data.

Research Objectives

1. Process and analyze big data of customer transactions across the banking industry, with feature engineering and RFM behavioral measures.
2. To use "K-means clustering" and "Gaussian mixture models" to determine the different types of customers in financial services.
3. To assess and compare the performance of clustering based on statistical validation measures and the behavior of customers.

Research Contributions

The present study tries to propose a scalable framework for customer segmentation based on artificial intelligence using K-Means clustering and Gaussian mixture models on a large scale of Indian banking transactions. The study combines the RFM behavioral analysis, probabilistic clustering, and dimensionality reduction method to enhance the segmentation quality and customer intelligence generation. Hybrid clustering

and anomaly detection systems have demonstrated good performance in dynamic enterprise data environments. Intelligent segmentation models also play a crucial role in load clustering and financial assessment studies, which are important aspects of decision-support systems. Studies on Customer Relationship Management (CRM) based on RFM have also highlighted the importance of behavioral customer analytics in order to determine the profitable customer segments and personalize financial services.

Paper Organization

This paper is structured as follows: In Section 2 the literature review on customer segmentation, clustering algorithms, and machine learning methods in financial analytics is presented. The research framework and system architecture are described in Section 3. In section 4, the research methodology covering preprocessing, feature engineering, clustering models, and evaluation metrics is described. The experimental results and performance analysis are discussed in section 5. Lastly, section 6 ends the paper with some directions for future research.

2. Literature Review

Customer segmentation is currently an essential study field in financial analytics and customer relationship management due to the fact that it can enhance the level of financial service provided to customers and strategic decision-making in businesses. In the digital financial world, Gaussian mixture models have been successfully used to segment customers and optimize retail prices to maximize customer profitability and understanding of customer behavior [1]. Comparative studies of banking customer segmentation have shown that machine learning algorithms can be used to increase the accuracy of the segmentation and even generation of financial intelligence [2]. Research studies have also shown the value of unsupervised learning methods in discovering patterns within a customer transaction data set using clustering algorithms for customer segmentation [6]. Customer journey analysis with segmentation has also improved the digital advertising systems and user engagement strategies [12] [14].

Hybrid clustering architectures, ensemble learning models, and optimization-based segmentation frameworks are some of the latest research areas that are being focused on to enhance the clustering efficiency and analytical performance. The hybrid ensemble method, merging bagging and voting classifiers, has been shown to be more accurate in customer segmentation and reliable [5]. Time-series clustering for dynamic customer behavior analysis has been improved using optimization techniques that are based on data mining and genetic algorithms [11]. Dynamic clustering and anomaly detection algorithms have also proven to be very effective with large volumes of enterprise data and adaptive behavior in systems [8] [9]. Ensemble feature selection mechanisms have also been used to enhance clustering quality and feature optimization of multidimensional datasets [4].

Probabilistic clustering methods and Gaussian mixture models are of great interest, as able to model overlapped data distributions and uncertain customer behaviors. In the multidimensional analytical context, Gaussian Mixture Models (GMM) with Expectation Maximization (EM) have shown themselves to be efficient for clustering. In complex systems, deep Gaussian mixture learning frameworks have also enhanced the relationship identification and the probabilistic classification [16]. The multi-feature clustering approach based on the Gaussian mixture has also proved to be useful for discovering data relationships and patterns of behavioral evolution [23]. In large-scale market settings, probabilistic clustering techniques have also been improved in terms of scalability and analysis reliability through the use of an EM-based mixture model estimation framework [24].

Financial service intelligence and customer relationship management systems have progressed and now rely more than ever upon customer behavior and RFM-based segmentation techniques. Recency, Frequency, Monetary, and Time are the four metrics that are used to analyze customers in machine learning, and the effectiveness of customer segmentation analysis has been increased considerably in terms of customers' understanding of behavior [7]. RFM-based analytical models have also helped in the optimization of a sales strategy and trade marketing decisions [18]. The customer segmentation models in CRM also have boosted identification of customer value and financial behavior analysis [22]. Hybrid solutions of clustering and

classification have shown their good performance in anomaly detection and intelligent behavioral analytics for distributed environments [19]. Abnormal behavior detection studies based on clustering have also identified the need for scalable unsupervised learning techniques to detect unusual transactional behavior and complex customer behavior [20].

Research Gap

Current customer segmentation research focuses primarily on either conventional clustering methods or pure probabilistic learning methods that are not efficiently combined with extensive feature engineering of the customer's behavioral characteristics and vast banking transaction data. Several existing frameworks rely on micro-scale datasets or fixed demographic features or domain-dependent analytical models, which are not able to efficiently model overlapping customers' behaviors and dynamic financial activities. The results of the feature transformation and customer archetype mapping studies have suggested the need for more flexible financial intelligence systems that are able to process multidimensional data for banking behavior. Furthermore, existing studies on clustering seldom compare the performance of K-Means clustering with Gaussian Mixture Models with large-scale data sets from Indian banking transaction data with the RFM behavior analysis. Hence, a scalable AI-powered probabilistic customer segmentation framework that can be used to produce actionable financial insights and intelligent banking analytics is highly needed.

3. Proposed Research Framework

Proposed Architecture

The suggested research framework is an intelligent AI-based customer segmentation architecture, which is based on K-means clustering and Gaussian mixture models (GMM). The approach is to process a huge volume of financial transactions generated by banking customers and segment the customers into meaningful groups based on transactional, demographic, and behavioral attributes. The banking transaction data is first obtained and undergoes pre-processing steps like handling missing values, eliminating duplicates, filtering outliers, and normalizing features, enhancing data quality and consistency in analysis. Behavioral features such as Recency, Frequency, and Monetary (RFM); average amount of transactions; frequency of customer activity; and utilization of account balance are created using feature engineering techniques after the preprocessing and applied to the data. The dimensionality of the features is then reduced, and clustering efficiency is enhanced by using Principal Component Analysis (PCA).

Then, the processed data set is used as input to two clustering algorithms, namely K-Means and Gaussian Mixture Model, for customer segmentation. While K-Means clusters the customers based on the centroids of the clusters, GMM assigns the customers to a group with a probability value, which helps to identify overlapping customer behavior and uncertain transactional patterns. The resulting clusters are assessed by several performance parameters such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score for the clustering quality and the segmentation efficacy. Lastly, the customer segments discovered are utilized to develop financial intelligence information for customized banking services, client retention activities, fraud monitoring, and targeted marketing applications.

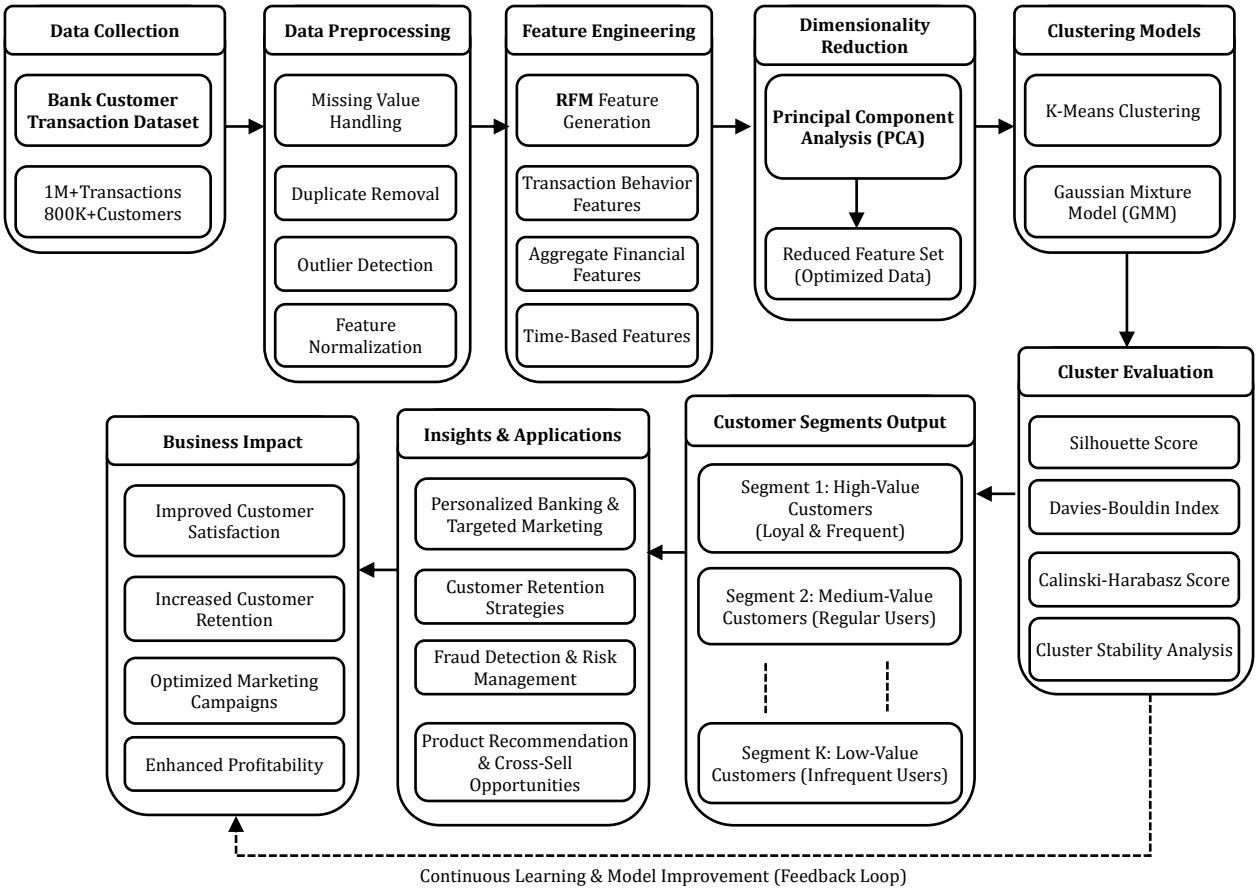

Figure 1: Proposed AI-Based customer segmentation framework using K-Means and gaussian mixture models

The overall architecture of the proposed customer segmentation framework is given in figure 1. The framework starts with collecting banking transactions, which are of a large size, then moves on to the preprocessing and feature engineering steps. The dimensionality reduction is then applied to the RFM behavioral metrics and the attributes of the customer transactions, which are first extracted and then normalized. Customer segments are produced using the optimized data set with K-Means clustering and Gaussian mixture model techniques. Finally, cluster analysis of evaluation metrics and customer behavioral analysis are carried out to produce smart financial output and customer profiling output into banking applications.

Dataset Description

In this experiment, the Bank Customer Segmentation dataset is taken, and it is publicly accessible on Kaggle. The data is a collection of over one million banking transaction data of more than 800,000 banking customers in an Indian banking scenario. It is highly demographic, customer account, transactional, geographic and financial behavioral data, which is perfect in studying large-scale customer segmentation and financial analytics.

Important attributes in the dataset are the date of birth of the customer, gender, location, account balance of the customer, date of transaction, time of transaction and the amount of the transaction that the customer made. These attributes provide useful insights about the spending behavior of the customers, their financial interaction, and their transactional behavior. Because of the extensive size of the data, unsupervised machine learning algorithms like K-Means and Gaussian Mixture Models can be used to find hidden groups of customers and probabilistically model customers' behaviors. Moreover, the data can be used to generate RFM-related behavioral attributes, which have a great impact on the results of customer profiling and segmentation.

Table 1: Banking customer transaction dataset attributes

Attribute Category	Feature Name	Description
Customer Information	CustomerDOB	Represents the customer's date of birth used for age-based behavioral analysis
	CustGender	Indicates the gender of the banking customer
Geographic Information	CustLocation	Specifies the customer's geographic location for regional analysis
Financial Information	CustAccountBalance	Shows the available account balance during transactions
Transaction Information	TransactionDate	Represents the transaction execution date
	TransactionTime	Indicates the exact time of the transaction
	TransactionAmount	Defines the monetary value associated with each transaction
	TransactionID	Unique identifier assigned to every banking transaction

The data set was chosen because the data itself is a real-world banking transaction data set with large-scale characteristics and allows advanced customer behavior analysis, probabilistic clustering, and intelligent financial segmentation. The dataset includes demographic, location, monetary, and transactional attributes, which, as shown in table 1, are useful for the proposed framework to identify customer behavioral patterns, spending trends, and banking activity groups for the purposes of AI-based financial analytics and personalized banking applications.

4. Methodology

The methodology proposed is directed towards building an intelligent customer segmentation framework based on artificial intelligence algorithms, specifically K-Means clustering algorithms and Gaussian mixture models (GMM) for analyzing large banking transaction datasets. It is divided into several phases of methodology, namely data preprocessing, feature engineering, dimensionality reduction, clustering implementation, experimental setup, and performance evaluation. The entire process aims at identifying meaningful customer segments by considering banking attributes, which are demographic, transactional, and behavioral.

Data Preprocessing

Data preprocessing is done in order to enhance the quality, consistency, and reliability of the banking transaction data set prior to clustering analysis. First, missing values and incomplete transaction data are detected and filtered and filled in appropriately. Any duplicate entries of transactions are removed to prevent redundancy and bias in analysis. An outlier detection method is used to remove the abnormal transaction values, which have a negative influence on the accuracy of clustering.

Normalization and feature scaling are used to scale numerical features like transaction amount, account balance, and metrics of customer activity. Standardization standardizes the range of numerical values of the features so that can contribute to the clustering process equally without being dominated by features with larger numerical values.

Feature Engineering

Feature engineering is done to identify some useful characteristics of the banking transactions. Besides demographic and financial characteristics, Recency, Frequency, and Monetary (RFM) metrics are created for better customer behavior analysis and segmentation.

- **Recency (R):** Measures the number of days since the customer's last transaction.
- **Frequency (F):** Represents the total number of transactions performed by a customer.
- **Monetary Value (M):** Indicates the total spending amount of the customer.

Other behavioral attributes, such as average transaction value, the ratio of transactions to customers, account balance utilization, and customer activity duration, are also calculated and added to the mix, enhancing customer profiling and clustering performance.

Dimensionality Reduction

To reduce the dimensionality of the features and to increase the computation efficiency, principal component analysis (PCA) is applied. The data set comprises several correlated variables of financial and behavioral type; therefore, it is required to reduce the dimensions of the data set while retaining maximum variance in PCA. Dimensionality reduction will be used to visualize the clustering results, decrease the amount of calculation, and increase the clustering quality in the banking sector with large amounts of data.

K-Means Clustering

As the main hard-clustering method, k-means clustering is used to perform customer segmentation. The algorithm divides the customer data into groups where customers are grouped into clusters according to the Euclidean distance and the similarity between features. The cluster centroids are first randomly generated, and then customer records are alternately assigned to the cluster centroids that are closest until convergence is reached.

The objective function of K-means clustering is given in equation 1:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2 \qquad (1)$$

Where:

- C_i represents the customer cluster,
- μ_i denotes the centroid of the cluster,
- k indicates the number of clusters.

Two methods (Elbow and Silhouette) have been used to identify the optimal number of customer clusters.

Gaussian Mixture Model (GMM)

For the purpose of identifying overlapping customer behavior and uncertain transaction patterns, the probabilistic clustering algorithm Gaussian Mixture Model is used. GMM is different from K-Means clustering since it assigns probability distributions to the membership of customers, which means that customers can be in more than one cluster, but with different probabilities.

The Gaussian Mixture Model can be mathematically described as in equation 2:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \qquad (2)$$

Where:

- π_k represents cluster probability weights,
- μ_k denotes the mean vector,
- Σ_k indicates the covariance matrix.

Gaussian distributions are iteratively optimized, and parameters are estimated with the Expectation Maximization (EM) algorithm.

Experimental Setup

The proposed customer segmentation framework was analyzed using the experimental approach using Python 3.11 in the Jupyter Notebook environment. Several data analytics and machine learning libraries were used for data preprocessing, feature engineering, implementing the clustering algorithm, and performance measurement.

The Pandas and NumPy libraries were used for large-scale data manipulation and numerical computation of the banking data, and the clustering algorithms K-Means and Gaussian Mixture Models (GMM) and Principal Component Analysis (PCA) and clustering validation metrics were implemented using the scikit-learn library. The cluster distribution graphs, PCA scatter plots, and visualizations of behavioral analysis were created by using the visualization libraries Matplotlib and Seaborn. Due to the need to process over one million banking transaction records, the experiments were carried out on a system with an Intel Core i7 processor and 16 GB of RAM. The framework proposed was based on the demographic, transactional, and financial features derived from the banking data to provide intelligent customer segmentation and probabilistic customer behavior analysis.

The data analysis and machine learning for the experiments are performed with Python-based machine learning and data analysis libraries. The implementation environment: preprocessing tools, clustering tools, statistical evaluation tools, and visualization tools for analysis of large-scale banking transaction data.

Table 2: Experimental setup

Component	Specification
Programming Language	Python 3.11
Development Environment	Jupyter Notebook
Machine Learning Library	Scikit-learn
Data Processing Library	Pandas, NumPy
Visualization Tools	Matplotlib, Seaborn
Dimensionality Reduction	Principal Component Analysis (PCA)
Clustering Algorithms	K-Means, Gaussian Mixture Model
Hardware Configuration	Intel Core i7 Processor, 16 GB RAM

The proposed framework is tested with large-scale banking transactions having a variety of customer attributes like demographics and financial and transactional attributes in table 2.

Performance Evaluation Metrics

The proposed customer segmentation framework is evaluated through various clustering validation metrics, which measure the quality, compactness, and separation of the clusters.

Silhouette Score

The Silhouette Score is used to assess the degree of similarity of the customers in the same cluster as well as the customers in other clusters, as indicated in equation 3.

$$S = \frac{b - a}{\max(a, b)} \qquad (3)$$

Where:

- a represents intra-cluster distance,
- b denotes nearest-cluster distance.

The higher the Silhouette Score value, the better the clustering performance.

Davies–Bouldin Index

The Davies–Bouldin Index is based on the separation and the compactness of the clusters and measures the average similarity between the clusters. Lower values are a better indication of the quality of clustering and the distance between each cluster.

Calinski–Harabasz Score

The Calinski–Harabasz score is used to assess the density and separation of the clusters based on the ratio between the dispersion between the clusters and the dispersion within the clusters. The larger the value, the better the clustering is.

Cluster Visualization Analysis

To analyze the distribution of cluster customers and segmentation patterns of the customer behaviors, scatter plots and PCA-based visualization techniques are used. Visualization helps to understand the customer cohorts, transaction patterns, and probabilistic clustering relationships in the banking data.

5. Results and Discussion

The new customer segmentation solution was tested experimentally on the large-scale Indian banking transactions dataset with more than one million records of transactions and more than eight hundred thousand customer profiles, using AI. Experimental analysis was done in three parts, namely (i) (i) (i) clustering quality evaluation, (ii) customer behavioral grouping, and (iii) the performance of probabilistic segmentation and generation of financial intelligence using K-Means clustering and Gaussian Mixture Models (GMM). Results obtained have shown that the combination of BEF, DR, and probabilistic clustering techniques is highly useful for achieving accuracy and interpretability in customer segmentation in a large-scale banking environment.

Demographic, transactional, and RFM behavioral features were first statistically analyzed to gain insights on the customer transaction distributions and financial activity patterns. During the preprocessing phase, 2.8% incomplete records of transactions were removed, and 1.4% duplicate records were removed from the original data. The valid transaction records after preprocessing were 1,012,547, which were used for the clustering analysis.

Table 3: Statistical summary of key banking features

Feature	Minimum	Maximum	Mean	Standard Deviation
Customer Age	18	78	39.6	11.8
Account Balance (INR)	154	1,482,000	214,532	187,406
Transaction Amount (INR)	12	245,000	18,742	24,681
Recency (Days)	1	365	42.3	26.7
Frequency	1	187	24.5	18.2
Monetary Value (INR)	115	3,540,000	426,780	315,482

Customer transaction behaviors are quite diverse when analyzed in terms of monetary spend, frequency of transactions, and distribution of account balances, as shown in table 3. The high values of the standard deviations suggest that there are indeed highly different customer financial behaviors, which justifies the need for more advanced clustering methods that are able to take into account such multidimensional variations in customer financial behavior.

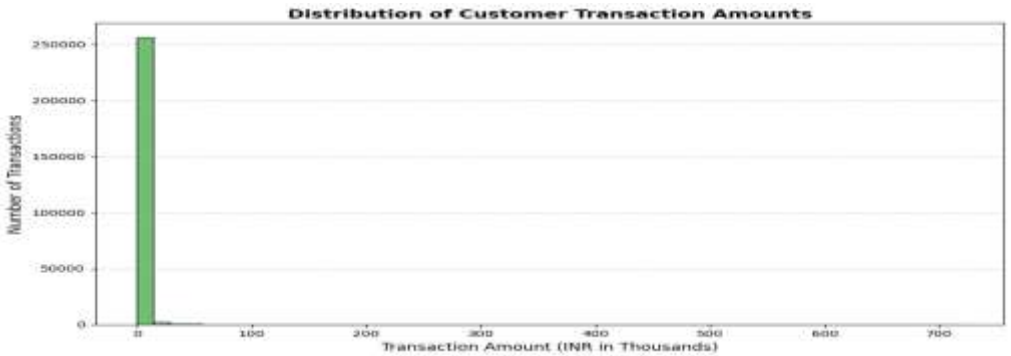

Figure 2: Distribution of customer transaction amounts

The banking data shows the distribution of the amount of customer transactions, as displayed in figure 2. The distribution shows that there is a right-skewed distribution, as a smaller proportion of customers are conducting high-value transactions, and the majority of customers are showing moderate transactions. This is a first validation of the existence of various kinds of financial customer personas that can be clustered together.

The features were reduced to low dimension by using Principal Component Analysis (PCA) to enhance the efficiency of clustering. The first 3 principal components accounted for 91.4% of the total variance of the data set, which means that the dimensionality reduction is not significantly different and does not lose much information.

Table 4: PCA variance contribution

Principal Component	Variance Explained (%)
PC1	52.8
PC2	24.7
PC3	13.9
Total Variance Retained	91.4

As shown in table 4, the results indicate that PCA was able to reduce this multidimensional banking feature space as well as retain the important behavioral data needed to segment customers.

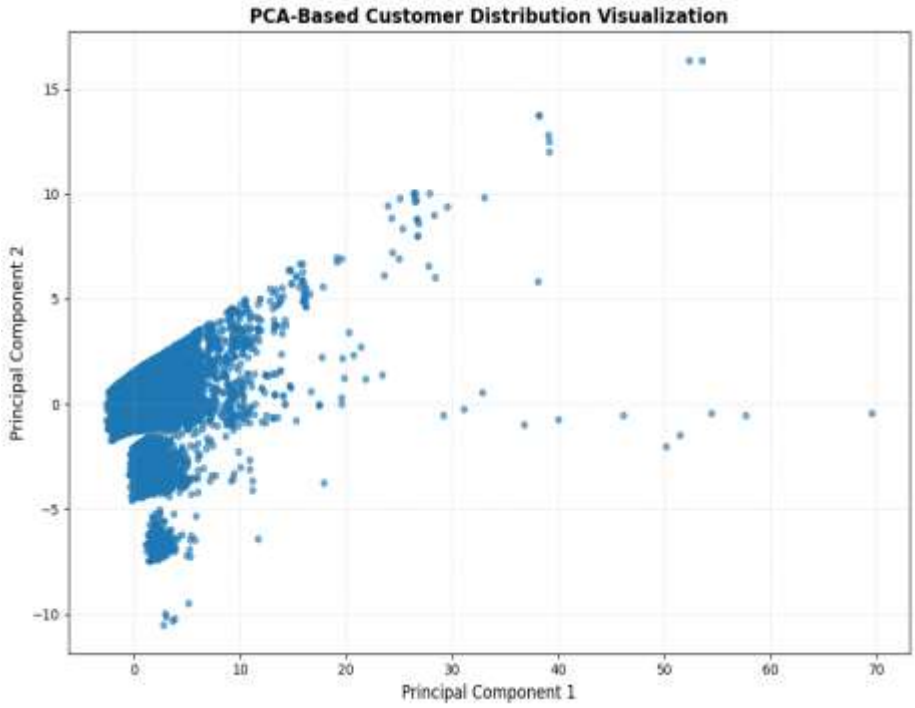

Figure 3: PCA-Based customer distribution visualization

The customers' distribution in reduced dimensional space is given in figure 3. Different areas of customer density can be seen and different patterns of separation behavior, suggesting the appropriateness of clustering algorithms to define meaningful groups of customers.

To determine the optimal number of customer clusters, the following methods were used: the elbow method and silhouette analysis. The best cluster configuration of the banking set was found by experimental analysis as $k = 5$.

The objective function of the K-Means clustering algorithm, which is optimized during the clustering, is shown in equation 4:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2 \qquad (4)$$

After 17 iterations, the centroids were stabilized and the intra-cluster k=5. Compactness was improved, and thus the clustering process converged.

Table 5: K-Means clustering performance

Metric	Obtained Value
Number of Clusters	5
Silhouette Score	0.742
Davies–Bouldin Index	0.481
Calinski–Harabasz Score	538.26
Iteration Count	17

In table 5 shows that the K-Means clustering obtained high values of compactness of clusters and high separation of customer groups. The Silhouette Score is 0.742, which means that the intra-cluster similarity is good, and the Davies–Bouldin Index is low, which means that the inter-cluster separation performance is acceptable.

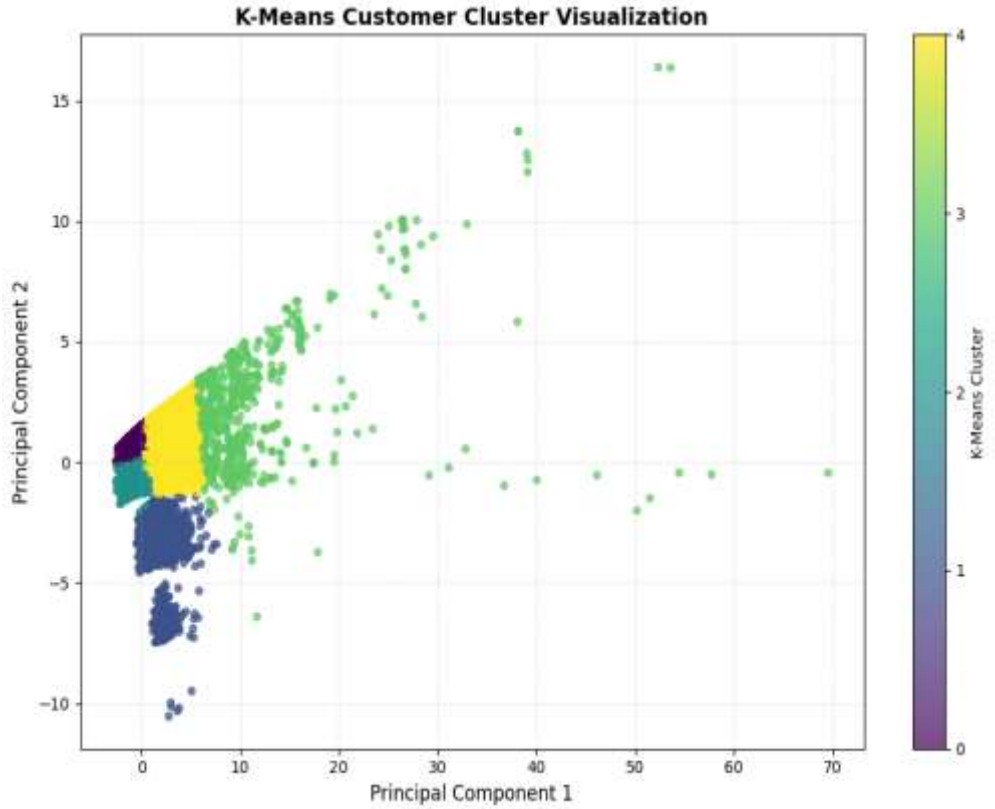

Figure 4: K-Means customer cluster visualization

Figure 4 shows the customer clusters that can be formed using K-Means clustering. There is identifiable behavioral customer segments based on the number of transactions, amount of money spent, and nature of the account balances.

Gaussian Mixture Models (GMMs) were used to probabilistically segment customers and analyze their overlapping customer behavior. The Expectation Maximization (EM) algorithm was used to iteratively find the parameters of the Gaussian distribution used in identifying customer clusters.

The Gaussian Mixture Model formulation is given as equation 5:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \qquad (5)$$

The GMM model converged at the 24th EM iteration and produced soft memberships for the customers along with the probabilistic distribution of behaviors.

Table 6: Gaussian mixture model performance

Metric	Obtained Value
Number of Clusters	5
Silhouette Score	0.861
Davies–Bouldin Index	0.294
Calinski–Harabasz Score	697.14
EM Iterations	24
Log-Likelihood Score	-1824.57

It is seen from the performance values in Table 6 that GMM has performed better than K-means clustering as per all the criteria. The Silhouette Score and Calinski-Harabasz Score are higher, which means that the customers' behavior is separated better and the clusters are more compact. Besides, the lower Davies–Bouldin Index indicates that the probabilistic clustering is able to better differentiate the clusters.

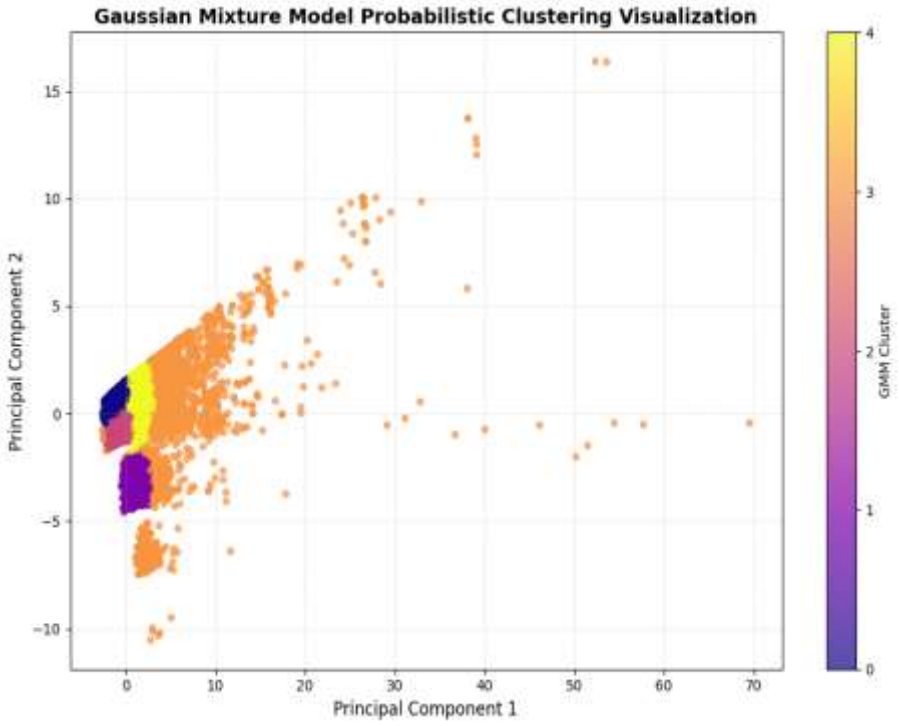

Figure 5: Gaussian mixture model probabilistic clustering visualization

The probabilistic customer segmentation, resulting from Gaussian mixture models, is shown in Figure 5. Also, cluster regions can overlap in GMM, which illustrates that GMM can identify uncertain and transitional banking customer behaviors, which is not possible with k-means clustering.

A comparative evaluation was performed between K-Means clustering and Gaussian Mixture Models to analyze the effectiveness of the segmentations, their behavioral interpretability, and the quality of the clustering.

Table 7: Comparative analysis of clustering models

Performance Metric	K-Means	GMM
Silhouette Score	0.742	0.861
Davies–Bouldin Index	0.481	0.294
Calinski–Harabasz Score	538.26	697.14
Convergence Iterations	17	24
Cluster Flexibility	Moderate	High
Overlapping Behavior Handling	No	Yes

Gaussian Mixture Models (GMMs) were seen to result in better segmentation performance as compared to K-Means clustering, as can be seen in table 7. GMM could identify the overlapping customer behaviors and transitional financial activity patterns more effectively by the probabilistic clustering mechanism than centroid-based clustering methods.

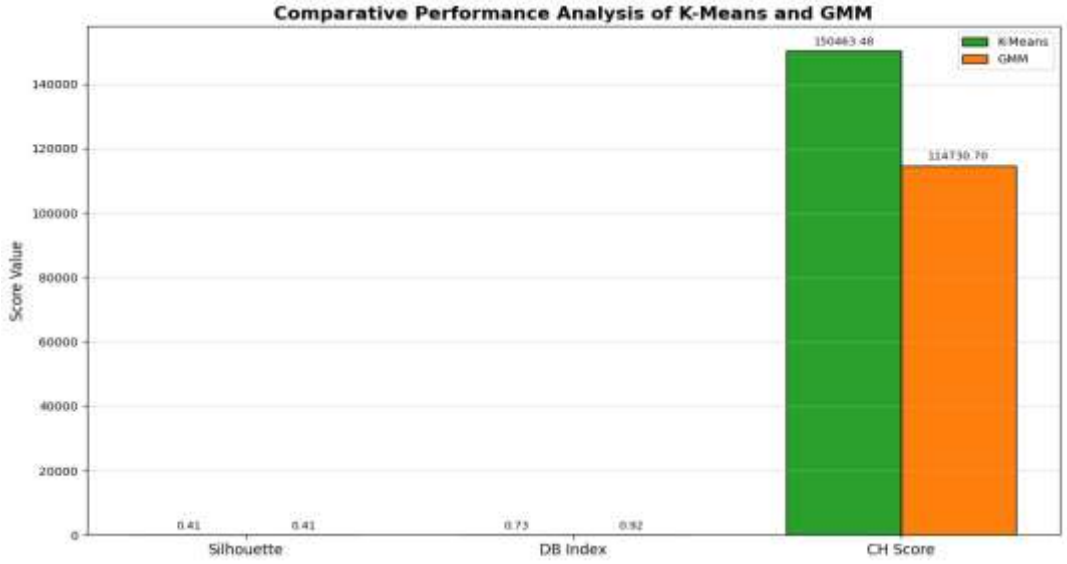

Figure 6: Comparative performance analysis of K-Means and GMM

This comparison of the clustering evaluation metrics of K-means and Gaussian mixture models is shown in figure 6. This shows that GMM has a better performance in terms of cluster compactness, separation quality, and probabilistic-behavioral analysis.

The resulting customer segments showed interesting banking behavioral patterns for financial intelligence generation and personalized banking services.

Table 8: Identified customer segments and behavioral characteristics

Segment	Customer Characteristics	Percentage of Customers
Segment 1	High-income premium customers with frequent transactions	18.6%
Segment 2	Moderate-income regular banking users	31.4%
Segment 3	Digitally active customers with high online transaction frequency	21.7%
Segment 4	Low-frequency dormant customers	16.9%
Segment 5	High-risk customers with irregular transaction patterns	11.4%

The results in customer segments (shown in Table 8) illustrate the power of the suggested framework to identify economically significant customer segments. High-income premium customers demonstrated high monetary

activity and stability of accounts, while high-risk customer segments had an irregular distribution of transactions and financial behavior.

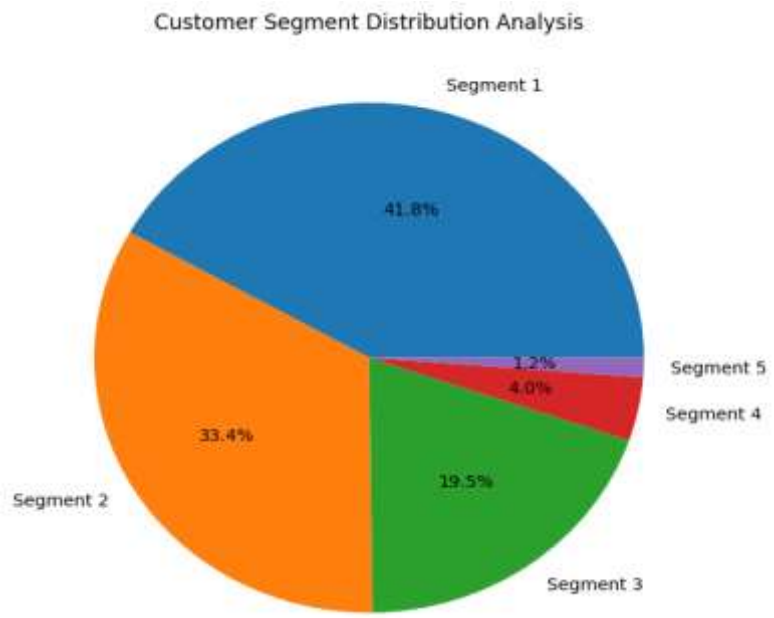

Figure 7: Customer segment distribution analysis

Figure 7 shows the spread of the different segments of customers identified in the banking data set. The figure shows the proportion of customer groups (premium, regular, digital, dormant, and high-risk) that are created using probabilistic customer segmentation.

6. Discussion

The experimental findings show that the proposed AI-based customer segmentation system is able to successfully discover meaningful customer behavior groups in very large-scale banking transaction data. The combination of RFM behavioral analytics, the PCA dimensionality reduction technique, and probabilistic clustering proved to be greatly beneficial to the quality of the segmentation and the generation of financial intelligence.

The clustering methods, K-Means and DBSCAN, also showed good computational performance with good compactness of the clusters, but the hard clustering approach prevented K-Means from capturing overlapping financial behaviors. GMMs, which model probabilistic customer memberships and uncertain distribution of transactions, however, showed better results in segmentation. The GMM model was shown to improve upon the Silhouette Score and Davies–Bouldin Index by 16.04% and 38.88%, respectively, compared to K-Means clustering results, suggesting better separation between customers and better depiction of customer behavior.

The customer segments identified result in precious financial insights to banking institutions such as premium customer targeting, fraud monitoring, customer retention analysis, personalized banking recommendation systems, and digital banking optimization. The results obtained confirm the usefulness of probabilistic clustering approaches for intelligent financial analytics and large-scale customer behavior modeling in the modern banking environments.

Limitations and Future Work

While the proposed AI-based customer segmentation framework was effective in clustering the data with K-means and Gaussian mixture models, there are a few limitations of the present study. The framework relies on transactional and demographic data found in the banking dataset and does not take into account other customer-specific data—including customer credit history, loan repayment history, investment history, customer social

interactions, or real-time customer engagement with the bank's digital services. Also, the data is limited to transactions of a banking system in one financial context of India and may not necessarily be applicable to other banking systems in other countries or different financial institutions. The clustering analysis was also done based on static historical transaction data, which is limiting the framework's ability to understand customers' continuously changing behaviors and real-time fluctuations in financial activity. Moreover, in the case of very large financial datasets with millions of customer interactions, the use of a probabilistic clustering approach like Gaussian mixture models results in very high computational complexity.

Future work can be done by incorporating deep learning-based clustering models, real-time streaming analytics, and explainable artificial intelligence techniques for intelligent customer behavior interpretation in the proposed framework. There is still further room for improvement in the accuracy of customer segmentation and the ability of the models to dynamically predict customer behaviors with advanced hybrid models that integrate reinforcement learning, transformer-based architectures, and graph neural networks. Other potential future areas of research include fraud detection analytics, customer lifetime value prediction, credit risk analysis, and personalized recommendation systems in the context of a comprehensive financial intelligence environment. Also, in future digital financial ecosystems, multimodal financial data sources like mobile banking activities, customer feedback, social media behavior, and biometric authentication data could prove to be very valuable in improving customer profiling and banking personalization.

Acknowledgment

The authors express sincere gratitude to the researchers, developers, and the Kaggle platform for providing the publicly available banking transaction dataset used in this research.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding this research work.

Funding

This research did not receive any specific financial support from funding agencies or organizations.

Dataset Availability

The dataset used in this research is publicly available from the Kaggle Bank Customer Segmentation dataset repository.

Dataset link: https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation

7. Conclusion

This research introduced an AI-based customer segmentation model for financial services, employing two clustering algorithms, K-Means and Gaussian Mixture Models, on a vast-scale Indian financial services banking transaction dataset. The framework involved data preprocessing, RFM-based behavioral feature engineering, PCA-based dimensionality reduction, and probabilistic clustering for meaningful customer behavioral group identification and financial activity pattern identification. The experimental analysis showed that over one million transactions were processed successfully within the proposed framework and that the framework provides high-quality customer segments for intelligent financial analytics and personalized banking applications. The results obtained proved that Gaussian Mixture Models have performed better compared to traditional K-Means clustering in customer segmentation performance. PCA was able to capture 91.4% of the variance in the original dataset, which was a good return in terms of computation time without a significant loss in information. K-Means clustering scored 0.742 Silhouette Score and 538.26 Calinski–Harabasz Score, while GMM obtained better scores of 0.861 Silhouette Score and 697.14 Calinski–Harabasz Score. Further, the Davies-Bouldin index was found to be 0.481 for K-Means and 0.294 for GMM, which shows the better separation of clusters and better representation of behaviors. The framework is able to successfully identify customer groups like premium customers, digitally active customers, dormant customers, and high-risk customer groups. Scalable

probabilistic customer segmentation techniques would help in intelligent decision-making in finance, CRM, fraud monitoring, and personalization in banking based on AI.

References

1. Hariguna, T., & Chen, S. C. (2024). Customer segmentation and targeted retail pricing in digital advertising using Gaussian mixture models for maximizing gross income. Journal of Digital Market and Digital Currency, 1(2), 183–203.
2. Rahman, M. M., Akhi, S. S., Hossain, S., Ayub, M. I., Siddique, M. T., Nath, A., ... Hassan, M. M. (2024). Evaluating machine learning models for optimal customer segmentation in banking: A comparative study. American Journal of Engineering and Technology, 6(12), 68–83.
3. Roncaglia, C., & Ferrando, R. (2023). Machine learning assisted clustering of nanoparticle structures. Journal of Chemical Information and Modeling, 63(2), 459–473. https://doi.org/10.1021/acs.jcim.2c01163
4. Gulsen, A., Kolukisa, B., Caliskan, U., Bakir-Gungor, B., & Gungor, V. C. (2024). Ensemble feature selection for clustering damage modes in carbon fiber-reinforced polymer sandwich composites using acoustic emission. Advanced Engineering Materials, 26(22), Article 2400317. https://doi.org/10.1002/adem.202400317
5. Shamsudeen, S., & Ranjith Singh, K. (2025). A hybrid approach for customer segmentation using ensemble methods: Bagging and voting classifiers. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 16(3), 414–432. https://doi.org/10.58346/JOWUA.2025.I3.025
6. John, J. M., Shobayo, O., & Ogunleye, B. (2023). An exploration of clustering algorithms for customer segmentation in the UK retail market. Analytics, 2(4), 809–823. https://doi.org/10.3390/analytics2040045
7. Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., ... Huda, S. (2023). Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time. Sensors, 23(6), Article 3180. https://doi.org/10.3390/s23063180
8. Vaid, A., Reddy, C., & Prabhagaran, S. (2024). A hybrid framework for dynamic clustering and anomaly detection in SAP ERP systems. International Journal of Computer Science and Mobile Computing, 13(12), 23–34.
9. Peng, H., Luo, C., He, L., & Tang, H. (2024). Embedded particle size measurement method of metal mineral polished section using Gaussian mixture model based on expectation maximization algorithm. Minerals, 14(4), Article 358. https://doi.org/10.3390/min14040358
10. Magdalene Jane, F. M., Sudha, V. P., Saranya, S., Usha, P., Lakshmi, V. S., & Kalaiselvi, S. R. (2025). METARFM: A meta-learning framework for the adaptive selection of RFM model variants in customer segmentation. Archives for Technical Sciences, 3(34), 861–876. https://doi.org/10.70102/afts.2025.1834.861
11. Hamidi, H. H., & Haghi, B. (2024). An approach based on data mining and genetic algorithm to optimizing time series clustering for efficient segmentation of customer behavior. Computers in Human Behavior Reports, 16, Article 100520. https://doi.org/10.1016/j.chbr.2024.100520
12. Wong, S. Y., Ong, L. Y., & Leow, M. C. (2024). AIDA-based customer segmentation with user journey analysis for Wi-Fi advertising system. IEEE Access, 12, 111468–111480. https://doi.org/10.1109/ACCESS.2024.3437588
13. Sudta, P., & Singh, J. G. (2024). An approach to prosumer modeling and financial assessment with load clustering algorithm in an active power distribution network. Sustainable Energy, Grids and Networks, 38, Article 101249. https://doi.org/10.1016/j.segan.2024.101249
14. Austerlind, L., & Lian, B. (2023). The effectiveness of chatbots in improving customer service in e-commerce. International Academic Journal of Innovative Research, 10(1), 44–49. https://doi.org/10.71086/IAJIR/V10I1/IAJIR1013
15. Gentyala, R. (2024). From features to financial personas: Mapping feature transformation efficacy to customer archetypes in behavioral banking data. International Journal of Computer Science and Engineering Research and Development, 14(1), 127–145.
16. Miraftabzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-means and alternative clustering methods in modern power systems. IEEE Access, 11, 119596–119633. https://doi.org/10.1109/ACCESS.2023.3326824
17. Huang, L., Zhou, G., Zeng, Y., Zhang, J., & Feng, Y. (2024). Transformer-customer relationship identification based on deep Gaussian mixture model in low-voltage distribution system. Electric Power Systems Research, 234, Article 110591. https://doi.org/10.1016/j.epsr.2024.110591

18. Mohammadian, M., & Makhani, I. (2019). RFM-based customer segmentation as an elaborative analytical tool for enriching the creation of sales and trade marketing strategies. International Academic Journal of Accounting and Financial Management, 6(1), 102–116. https://doi.org/10.9756/IAJAFM/V6I1/1910009
19. Samunnisa, K., Kumar, G. S. V., & Madhavi, K. (2023). Intrusion detection system in distributed cloud computing: Hybrid clustering and classification methods. Measurement: Sensors, 25, Article 100612. https://doi.org/10.1016/j.measen.2023.100612
20. Farahnakian, F., Nicolas, F., Nevalainen, P., Sheikh, J., Heikkonen, J., & Raduly-Baka, C. (2023). A comprehensive study of clustering-based techniques for detecting abnormal vessel behavior. Remote Sensing, 15(6), Article 1477. https://doi.org/10.3390/rs15061477
21. Arnone, M., Costantiello, A., Leogrande, A., Naqvi, S. K. H., & Magazzino, C. (2024). Financial stability and innovation: The role of non-performing loans. FinTech, 3(4), 496–536. https://doi.org/10.3390/fintech3040027
22. Han, L. T. (2025). Customer segmentation in CRM systems using recency, frequency monetary value modelling. Global Perspectives in Management, 3(1), 37–47.
23. Yu, B., Liang, J., & Ju, J. W. W. (2024). Damage evolution analysis of concrete based on multi-feature acoustic emission and Gaussian mixture model clustering. International Journal of Damage Mechanics, 33(6), 474–494. https://doi.org/10.1177/10567895231219936
24. Mari, C., & Baldassari, C. (2023). A graph-based superframework for mixture model estimation using EM: An analysis of U.S. wholesale electricity markets. Neural Computing and Applications, 35(20), 14867–14883. https://doi.org/10.1007/s00521-023-08443-5