Enhanced Customer Segmentation: A Comparative Analysis

Anjana Rao[1], Anushka Nandwani[1], and Arshiya Singh[1]

Electronics and Communication Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India.

anjana005btece23@igdtuw.ac.in, anushka008btece23@igdtuw.ac.in, arshiya010btece23@igdtuw.ac.in

Abstract. Customer segmentation is very important in today's marketing environment to ensure that appropriate resources are utilized as businesses expand. Our project considers the customers' segmentation through K-Means clustering and Gaussian Mixture Models (GMM). We used the mall customer dataset. To determine the optimal number of clusters for K-means we used the elbow method while for GMM, we used the Bayesian Informed Criterion (BIC). By adding eXplainable AI(XAI) features like SHAP and LIME we were able to enhance interpretability of GMM. . The research findings indicate that GMM with XAI offers greater insight that enables business organizations to achieve higher performances.

Keywords: K-Means Clustering, Gaussian Mixture Model (GMM), Explainable Artificial Intelligence (XAI), Customer Segmentation, SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME).

1 Introduction

Customer segmentation is crucial for modern marketing, shifting from basic categorization to sophisticated analysis for better engagement and profitability. Accurate segmentation is vital for tailoring marketing efforts and optimizing resources. This research addresses the challenge of enhancing segmentation accuracy and interpretability through advanced clustering techniques.

1.1 **Research Gap (Problems with Traditional Methods):** While K-Means clustering is popular for customer segmentation it often struggles with complex data and lacks interpretability. Thus, existing approaches fail to provide accurate segmentation and clear insights, leaving a gap in effective marketing strategies.

1.2 **Major Contributions of the Paper:**

- Comprehensive Comparative Study: Provides a detailed evaluation of K Means, GMM without XAI and GMM with XAI focusing on their effectiveness,interpretability and practicality and marketing applications.
- Introduction of Explainable AI (XAI) in Customer Segmentation: Our study combines Gaussian Mixture Model (GMM) with XAI techniques (SHAP and LIME) to enhance interpretability of clustering results.

1.3 **Organization of the Paper:** The paper presents the gaps in research methodology, highlighting the comparison in traditional K-Means, GMM, and GMM with XAI, followed by results and discussion evaluating their accuracy and interpretability, and in the end concludes with insights into how businesses can refine segmentation strategies.

2 Literature Review

Researchers have suggested that the segmentation of customers into clusters provides actionable insights for customer retention, loyalty programs, and engagement strategies. This is particularly useful for digital start-ups looking to enhance customer satisfaction and long-term value [1]. Alves Gomes and Meisen proposed a structured approach to segmentation, emphasizing four stages: data collection, representation, segmentation, and targeted outreach, with a focus on utilizing personalized methods like RFM analysis and advanced machine learning models for better targeting in e-commerce [2]. In fact, findings show that dynamic segmentation enables businesses to continuously update customer clusters in real time based on behavioral changes, ensuring evolving needs are met. This highlights the capability of machine learning to create more responsive customer experiences [3]. Vamsi Katragadda's research suggests that machine learning models, particularly clustering techniques such as K-means and hierarchical clustering, substantially outperform classical approaches for purposes of segmenting customers based on behaviors and preferences in evolution [4]. Several authors also indicate that clustering methods can be applied to segment a customer base into distinct groups based on similar characteristics, allowing businesses to target and understand their customer base [5]. Kotler, the father of marketing himself, believes that any good segmentation must have four criterias: measurability, accessibility, substantiality, and actionability. His research points to this fact that the segments which are differentiated have also got profitability and can be sought using machine learning algorithms over massive data [6].

McCarthy's study revolves around the usefulness of behavioral data, be it in terms of a frequency and recency metric on a purchase basis; how these can predict future value. His research recommends using CLV models other than just demographic data which in fact should update dynamic machine learning reprojecting new values of CLV of users based on their updated behavior [7]. Among the various clustering techniques, GMM is one of the most efficient tools for segmentation. Using a probabilistic model, GMM represents data as a combination of multiple Gaussian distributions, where each of them explains a different cluster. The functionality allows GMMs to model clusters with diverse shapes and sizes, unlike methods that enforce rigid boundaries, such as k-means [8]. Gayam and S. R stress that XAI has been an essential mechanism in amplifying designer confidence in AI-generated forecasts especially on customer segmentation in developing products.By offering feature-based explanations, XAI aids in identifying key variables and fine-tuning the model, ensuring that AI systems remain efficient while maintaining their predictive accuracy [9]. However, Nimmagadda's research underscores that although artificial intelligence presents significant advantages to the realm of e-commerce, it concurrently poses a number of challenges. A primary issue pertains to guaranteeing transparency in the methods by which companies gather, utilize, and disseminate customer information, thus allowing consumers to give informed consent regarding AI applications. Furthermore, the research accentuates the importance of establishing robust security protocols, including sophisticated encryption methods, to protect customer data [10].

3 Methodology

This study evaluates the effectiveness of three clustering techniques: K-Means clustering, Gaussian Mixture Model (GMM) without Explainable AI (XAI), and GMM with XAI. The purpose is to evaluate each method's efficacy, interpretability, and practicality in categorizing mall customers based on behavioral and demographic traits. The study evaluates various methodologies to discover which is most effective for consumer segmentation in marketing applications.

Dataset Description: The analysis utilized the publicly available Mall Customer Datatest comprising 200 entries with attributes such as Customer ID, Gender, Age, Annual income and Spending Score. The spending score reflects how much a customer spends at the mall. Higher spending scores indicate higher levels of client expenditure at the mall. The spending score is graded on a scale from 1 to 100. (https://www.kaggle.com/code/yousefmohamed20/mallcustomer-segmentation-using-kmeans/input).

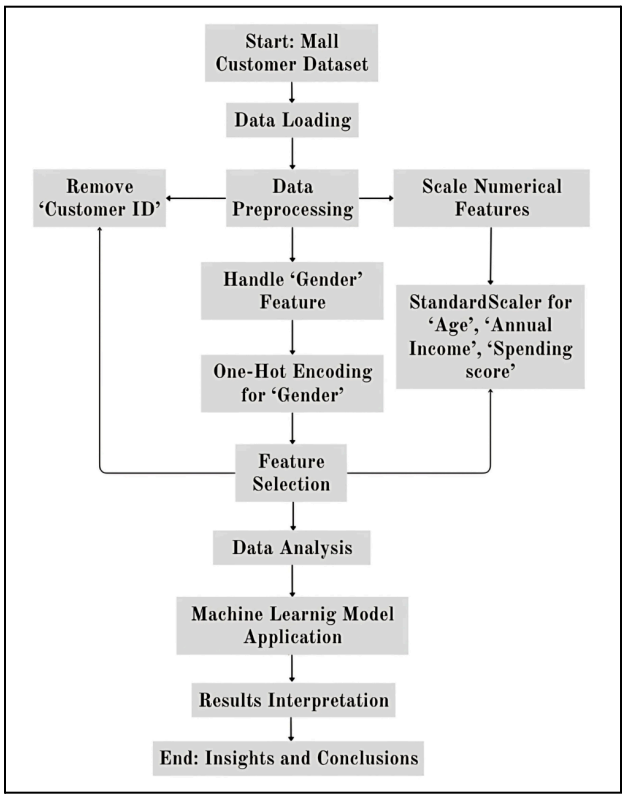


Fig. 1. Methodology flowchart

Pre-Processing steps: 'Customer ID' was excluded from the analysis as it does not provide any useful information regarding customer's behavior, characteristics or preferences and serves only as an identifier.It doesn't provide any information regarding spending patterns, age or income levels that are actually useful for clustering customers based on their behavior. It doesn't contribute any valuable insights or patterns that could enhance the segmentation process. The remaining features were retained as they offer meaningful information for clustering based on demographic and behavioral traits.

To bring all numerical features to the same scale, we applied Standard Scaler from the scikit-learn library. This ensured that variables such as 'Age' and 'Annual Income' which have different units and scales, contribute equally to the clustering process.

One hot-encoding was used to convert the categorical data in the 'Gender' feature to numerical data. This created two binary columns (Gender_Male and Gender_Female) ensuring that the data could be processed by machine learning algorithms.

Model Implementation: The study employed three clustering techniques to identify the customer segments. Each model's implementation is explained below:

K-Means: It is a widely used unsupervised machine learning algorithm. It is preferred for its simplicity and efficiency in partitioning the datasets. It minimizes the variance within each cluster formalized as:

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

Where 'J' is the total variance, 'K' is the number of clusters, 'x' is a data point, 'Ci' is the cluster, and 'μi' is the centroid. We utilized the Elbow method and Silhouette scores to determine the optimal number of clusters. The Elbow method works by plotting WCS (Within-Cluster Sum of Squares) against the number of clusters. WCSS decreases as more clusters are added but at a certain point, the rate of decrease lowers sharply which creates a bend or 'elbow' like structure in e plot that suggests the optimal number of clusters to us.

$$WCSS = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \quad (2)$$

The Silhouette score, calculated as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (3)$$

It assessed clustering quality by measuring how similar data points are within a cluster compared to the nearest cluster. Here 'a' is the average distance within the cluster and 'b' is the distance to the nearest cluster. Higher silhouette score indicates more distinct and accurate clustering. Using this we grouped into clusters depending on their purchasing behavior.

GMM without XAI: It is a soft clustering method. K-means uses a distance-based approach and assigns data points to a single cluster based on distance from centroid but GMM uses a probabilistic approach. It assumes that data points are a mixture of several Gaussian mixtures, each representing a different cluster. Each data point is assigned a probability of belonging to each other which provides a better understanding of the clustering structure. Optimal number of clusters were identified using Bayesian Information Criterion (BIC).

$$BIC = -2.\ln(L) + p.\ln(n) \quad (4)$$

Where ‘L’ is the likelihood of the model, ‘p’ is the number of parameters and ‘n’ is the number of data points. The model with the lowest BIC score was selected for its optimal balance of complexity and fit. Results were visualized using a probabilistic segment radar plot, showing the likelihood of a customer belonging to each cluster.

GMM with XAI: In order to enhance the interpretability of the output obtained using the GMM clustering algorithm, we applied XAI techniques in our analysis. XAI basically provides human-understandable explanations for the decisions generated by AI and machine learning models. The explainable AI techniques used were – SHAP (SHapley Additive exPlanations) and LIME (LocalInterpretable Model-agnostic Explanations) SHAP assesses the contribution of each feature to the cluster assignment and explains how different factors influenced the clustering results. LIME generates local explanations for individual cluster assignments and offers a model's decision-making process for individual cases. Instead of providing global understanding of the dataset it focuses on individual instances of the model. This helped in transforming GMM from a ‘black box’ model into a more transparent system which further helped in understanding the underlying logic of the clustering process.

Table 1. Comparison of K-Means and GMM

Aspect	**K-Means**	**GMM**
Algorithm	Centroid Based	Probabilistic, gaussian mixture
Clustering assignment	Hard (one cluster per point)	Soft (probabilities of belonging to clusters)
Optimal cluster method	Inertia (elbow method)	Bayesian Information Criterion (BIC)
Output	Discrete cluster assignment	Probability distribution over cluster
Interpretability	Direct, but less nuanced	More nuanced, especially with XAI techniques
Cluster shape assignment	Spherical	Ellipsoidal
Flexibility	Less flexible	More flexible

However there is an important point that the chosen methods, XAI techniques and GMM in particular, may be difficult to apply on large scale datasets. The computational cost of them, together with the time complexity, may rise steeply and may limit their applicability in large scale problems.

4 Results and Discussion

Table 2. Result Analysis

Aspect	**K-means**	**Gaussian Mixture Model (GMM)**
Model Performance Metrics	• *Elbow Method:* Optimal number of clusters lies between 4 and 5, based on the elbow point of WCSS. • *Silhouette Score:* The highest score (~0.43) is achieved with 6 clusters, indicating a potential optimal configuration	• *Performance Metrics*: Identifies three customer segments (Segment 0, Segment 1, Segment 2), with assigned probabilities for each.
Optimal Clusters Identification	The Elbow method identifies 4–5 clusters.	Probabilistic approach determines 3 distinct customer segments.
Segmentation Results	Six segments identified based on Age, Income, and Spending Score: 1. Youth, high spending. 2. Young, high earners. 3. Middle-aged, moderate spenders. 4. Developed, high earners, low spending. 5. Older, low-income, low spending. 6. Older, moderate-income, moderate spenders.	Three probabilistic segments: 1. *Segment 0:* Middle-income average consumers (primary). 2. *Segment 1:* Higher income/spenders (secondary). 3. *Segment 2:* Affluent, active buyers (minimal).
Real-time Implementation	• *Model Deployment:* Fit the K-means model with six clusters for ongoing analysis.	• *Segment Assignment:* Use GMM to assign customers to segments with probabilistic accuracy.
Applications	• Supports personalized marketing, tailored product recommendations, customer service improvements, and inventory optimization based on segment needs.	• Facilitates targeted marketing and customized services by leveraging segment membership probabilities.
Visualization	• Use scatter plots or 3D charts to represent customer distribution across the six segments.	• Use radar charts to visualize the probability of customers of the three segments.

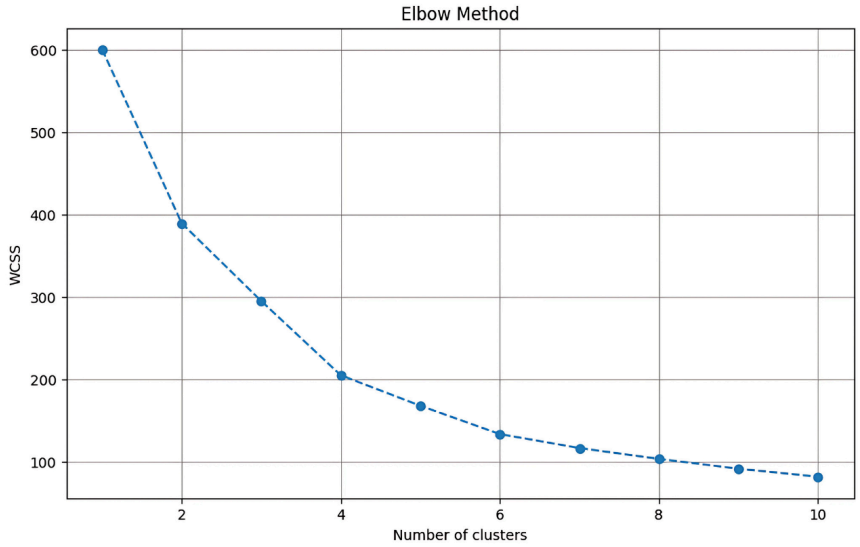


Fig. 2. Elbow Method

Fig. 3. Silhouette Score

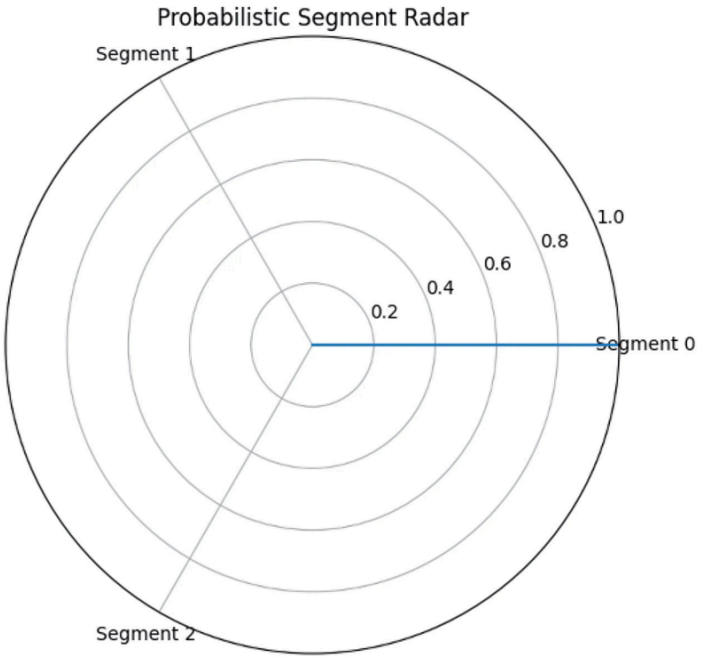


Fig. 4. Performance Metrics radar chart

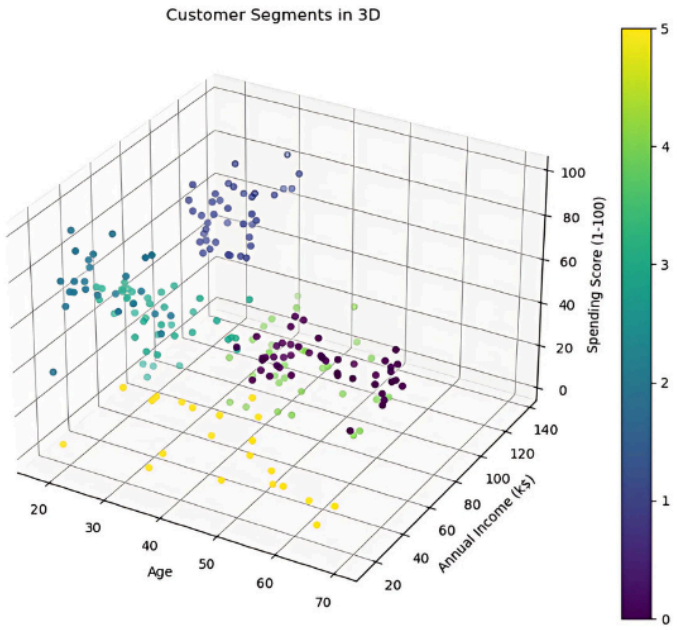


Fig. 5. Customer distribution using scatter plot

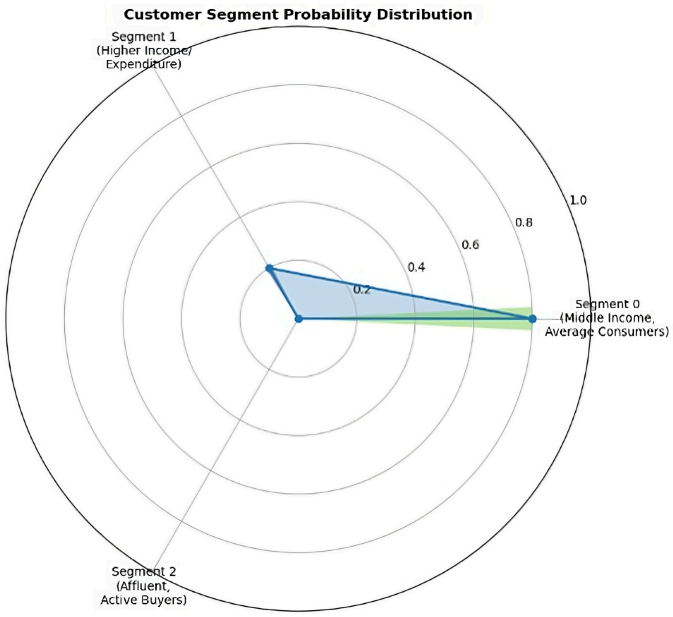


Fig. 6. Radar Chart

Gaussian Mixture Model with XAI (GMM-XAI):

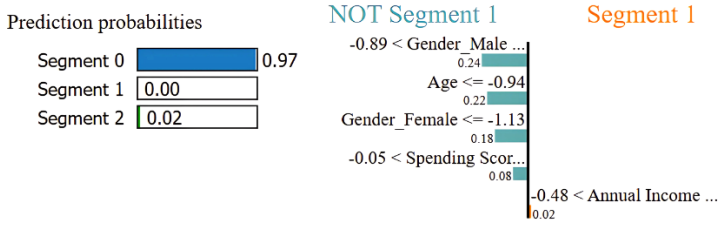
a) Model Performance Metrics: The GMM-XAI model segments the market into three categories. (Segment 0,1 and 2) , with a high degree of separation. For the example customer , the model shows 97% probability of belonging to Segment 0 and near-zero probabilities for Segments 1 and 2.

b) Segmentation Results:

- Segment 0: 97% likelihood for the example customer.
- Segment 1: 0% likelihood
- Segment 2: 2% likelihood

c) Analysis of Customer Segments:

- Segment 0 (High Probability): Potential younger customers, negative SHAP value for age, high spending score,smaller average income, predominantly male.
- Segment 1: Older customers, lower spending score, higher income, predominantly female.
- Segment 2: Intermediate characteristics, blending features from Segments 0 and 1.



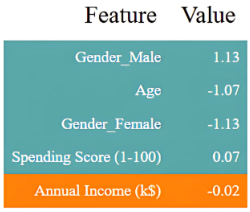


Fig. 7. Segments Analysis

d) Real-Time Implementation:

- Evaluate the GMM model with 3 components fit to the entire feature vectors dataset.
- Store the trained model and the SHAP explainer that are to be used during real-time prediction.

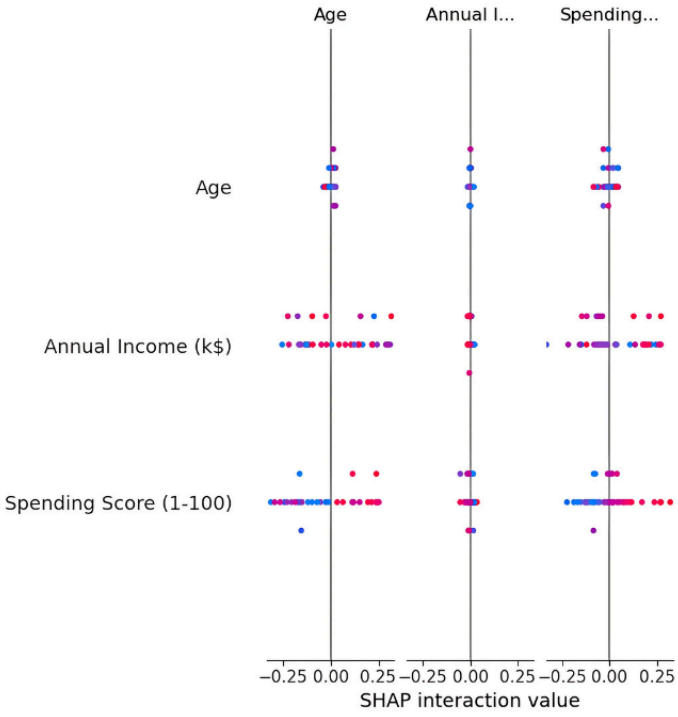


Fig. 8. SHAP interaction

Table 3. Suitability of clustering methods

Parameter	K-Means	GMM with XAI (SHAP, LIME)	GMM without XAI
Computation Speed	Fast, especially for large datasets.	Moderate, increased computational load due to XAI techniques.	Moderate, slower than k-means but faster than GMM with XAI.
Clustering Time	Quick, scales well with increasing data.	Slower due to the added computation of explainability techniques.	Slower, requires more iterations compared to k-means.
Granularity	Lower clusters are defined by distance to centroids.	Higher clusters can capture more complex distributions due to probabilistic nature.	Higher clusters capture complex distributions.
Effect of Data Size	Efficient on large datasets, with highly dimensional data.	Handles larger datasets but may slow down due to XAI overhead.	Handles moderately large datasets, but performance drops with very large data.
Handle Dynamic Data	Static, does not handle dynamic or streaming data well.	Static, but explainability can help interpret changes over time.	Static, not suitable for dynamic data.
Clustering Result Efficiency	Good for well separated clusters.	High, with better interpretability using XAI methods.	High, effective but lacks interpretability.

Implications and Significance of the Research: The combination of machine learning and interpretability technologies is revolutionary in customer segmentation-driven enterprises. Transparency in decision-making processes makes models much more successful by enabling organizations to comprehend not only the judgements made but also the rationale behind them. This promotes trust, improves usability, and complies with moral AI standards.

An advancement in segmentation accuracy and clarity has been made with the use of Gaussian Mixture Models (GMM) enhanced by Explainable AI (XAI) approaches like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). In addition to increasing segmentation accuracy, GMM with XAI identifies the fundamental causes of consumer behavior. This is particularly important for sectors where decision-making and customer experience are directly impacted by comprehending the logic behind AI-driven insights. What's new about this study? It suggests how AI can combine its high level of precision with transparency in decision-making processes so as to make Machine Learning models more accessible and useful to other people. This research not only advances customer segmentation methodologies but also underscores the importance of broader societal impact of Explainable AI promoting ethical AI and transparent data practice.

5 Conclusion

This study compared K-Means, GMM, and GMM enhanced with XAI for customer segmentation, demonstrating the strengths and weaknesses of each technique. K-Means, while fast and simple, proved limited when handling complex customer data, where GMM's probabilistic approach was more effective. GMM with XAI further improved interpretability by using SHAP and LIME to explain individual segmentations, allowing for more actionable insights. Despite the increased computational cost, the integration of XAI offered a clearer understanding of the clustering process, making it highly beneficial for marketing strategies that require transparency and customization.

As is true with many new methodological approaches, future research should attempt to replicate these results using far larger sample sizes and encompassing more diverse populations than has been done here. An attempt should also be made to find out how other clustering algorithms in conjunction with various XAI methods could provide a better balance between the two important factors-the model performance and the model understanding. Nevertheless, the research has contributed to continue by showing the benefits of incorporating XAI in clustering models hence; improving the customer segment models and enhancing customer value. It is in these aspects that further studies in customer segmentation may lead to better definitions for customer segment.

6 References

1. Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, *36*(9), 4995-5005.
2. Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, *21*(3), 527-570.
3. Mozumder, M. A. S., Mahmud, F., Shak, M. S., Sultana, N., Rodrigues, G. N., Al Rafi, M., ... & Bhuiyan, M. S. M. (2024). Optimizing Customer Segmentation in the Banking Sector: A Comparative Analysis of Machine Learning Algorithms. *Journal of Computer Science and Technology Studies*, *6*(4), 01-07.
4. KATRAGADDA, V. (2022). Dynamic Customer Segmentation: Using Machine Learning to Identify and Address Diverse Customer Needs in Real-Time. *IRE Journals*, *5*(10), 278-279.
5. Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018, December). Customer segmentation using K-means clustering. In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 135-139). IEEE.
6. Kotler, P., Keller, K. L., Brady, M., Goodman, M., & Hansen, T.(2016). Marketing Management 3rd edn PDF eBook. Pearson Higher Ed.

7. McCarthy, D. M., & Fader, P. S. (2018). Customer-based corporate valuation for publicly traded non contractual firms. Journal of Marketing Research, 55(5), 617-635.
8. 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.
9. Hu, X., Liu, A., Li, X., Dai, Y., & Nakao, M. (2023). Explainable AI for customer segmentation in product development. *CIRP Annals*, *72*(1), 89-92.
10. Gayam, S. R. (2021). Artificial Intelligence in E-Commerce: Advanced Techniques for Personalized Recommendations, Customer Segmentation, and Dynamic Pricing. *Journal of Bioinformatics and Artificial Intelligence*, *1*(1), 105-150.
11. Nimmagadda, V. S. P. (2022). Artificial Intelligence for Customer Behavior Analysis in Insurance: Advanced Models, Techniques, and Real-World Applications. *Journal of AI in Healthcare and Medicine*, *2*(1), 227-263.
12. Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on customer segmentation technique on ecommerce. Advanced Science Letters, 22(10), 3018-3022.
13. Perumalsamy, J., Krothapalli, B., & Althati, C. (2022). Machine Learning Algorithms for Customer Segmentation and Personalized Marketing in Life Insurance: A Comprehensive Analysis. *Journal of Artificial Intelligence Research*, *2*(2), 83-123.
14. Potla, R. T. (2023). Enhancing Customer Relationship Management (CRM) through AI-Powered Chatbots and Machine Learning. *Distributed Learning and Broad Applications in Scientific Research*, *9*, 364-383.