MALL CUSTOMER SEGMENTATION USING MACHINE LEARNING ALGORITHMS

Ayomide Otinwa

Otinwaayomide1@gmail.com

South East Technological University

Abstract- **This paper focuses on applying customer segmentation to a "SHOPPING MALL CUSTOMER DATASET", which helps identify customers to improve targeted marketing strategies and customer relations management. The dataset contains four features that are standardized for consistency among variables. Five different clustering algorithms: K-means, K-means++, Agglomerative Hierarchical, Gaussian Mixture, and Fuzzy C-Means clustering were built to find the best algorithm. The implementation showed that K-means achieved a silhouette score of 0.41 with 4 clusters and K-means++ achieved the same metrics, while Agglomerative Hierarchical Clustering had a silhouette score of 0.36 with 5 clusters. In addition, the Gaussian Mixture clustering achieved a silhouette score of 0.41 with 4 components, and Fuzzy C-Means had a silhouette score of 0.41 with 4 clusters. These results show that Fuzzy C-Means clustering performed better at clustering the customers over the other methods for the dataset and provided meaningful insights for segmentation. This study contributes to the importance of clustering techniques in understanding consumer behaviors and improving marketing outcomes.**

Keywords- Customer Segmentation, K-means, Agglomerative Hierarchical, Gaussian Mixture, Fuzzy C-Means, Silhouette score, Clustering, Shopping mall customer information dataset, consumer behaviours.

I. INTRODUCTION

The business industry continues to rapidly change and to improve the marketing techniques and concurrently increase sales, segmenting of customers is very important for marketing products to customers in specific divisions of the business. Customer segmentation is the process of splitting clients into groups according to similar attributes. The algorithms used in machine learning provide an alternative to normal segmentation approaches, which fall short in allowing businesses to study large data and find inconsistencies [1]. The objective of this research is to identify business information that can maximize customer interactions, improve marketing strategies and raise overall business performance by applying clustering techniques such as K-means clustering, Agglomerative Hierarchical clustering, Gaussian Mixture clustering and Fuzzy-C means clustering on a shopping mall customer data. In a very competitive business environment, companies may increase customer satisfaction and meet sales targets by adjusting their advertising tactics to suit each customer segment's needs using advanced algorithms which increases customer satisfaction and prolonged loyalty [2].

The sections below will provide a view of the related works, methodology used, evaluation techniques applied, results, conclusion and considerations for future works.

II. BACKGROUND AND RELATED WORKS

A. Background

Increased customer satisfaction, targeted marketing campaigns, and business expansion all depend on an understanding of consumer behavior. Thus, customer segmentation is the business strategy of dividing up customers based on common features, which allows for the focusing on of very particular needs of customers. Advanced techniques can be used to help businesses identify hidden patterns in customer behavior and categorize customers. Allocating resources effectively, improved decision-making and targeting makes this one of the most crucial ways to maximize overall business performance.

B. Related Works

Developing business strategies and improving customers' experiences within businesses require effective customer segmentation. To achieve this, we outline a thorough process for client segmentation inside e-commerce frameworks in order to further the research on machine learning algorithms[1]. Preprocessing, feature engineering, and model selection are the next steps in their process, which entails obtaining data from various sources, including transaction records and demographic profiles. They emphasize the need of training and evaluating models to fine-tune accuracy, highlighting clustering techniques like K-means, hierarchical clustering, and DBSCAN for unsupervised segmentation. On the other hand, in the context of mall customer segmentation, a different study explores the possibilities of hierarchical clustering over K-means. In order to provide more detailed insights into consumer behavior, this study used univariate clustering based on income and bivariate clustering including spending scores to identify five distinct clusters. The study also emphasizes the role of bottom-up dendrograms in illustrating the univariate and bivariate hierarchical structure of the customer data, and enhancing interpretability[2].

In contrast to these methods, a different paper highlights the value of meticulous feature selection and preprocessing after data collection from transaction records to increase model accuracy. With the quality of the data playing a critical role in determining segmentation outcomes, this approach addresses the need for robust predictive modeling[3]. However, a different study goes deeper into K-means clustering and further explains the algorithm's iterative structure. Large mall datasets can benefit greatly from it as it assigns data points to clusters based on proximity metrics such as frequency, and monetary (RFM) value. The elbow method is also used to optimize the number of clusters[4]. As additional backing for these methods, a different research project broadens on earlier research by classifying customers according to their spending patterns using a variety of clustering approaches, such as grid-based models. This study emphasizes the importance of good data for successful segmentation by examining the role of principal component analysis (PCA) in reducing the dimensions of the data and identifying the underlying patterns in the data[5]. The reviewed works of literature provide a range of approaches, emphasizing on the significance of picking the right clustering technique, and implementing comprehensive preprocessing to improve the segmentation accuracy.

These studies used different methodologies such as k-means, hierarchical clustering, and DBSCAN to segment customers which have contributed to our knowledge of consumer behavior, limitations remain in handling overlapped clusters. This work will enhance the segmentation accuracy and give more insight tot eh segments by applying more algorithms in conjunction with the previously applied techniques. This should result in more valuable insights and proper handling of customers overlapping behavior.

III. METHODOLOGY

A. Clustering

The machine learning technique used in this work is the clustering technique. Clustering is an unsupervised machine-learning technique in which data points are assigned to different groups based on their similarities. It ensures data points closer to each other should be similar to those far from each other. There are several clustering techniques, but five clustering techniques were implemented for this work.

B. Data Collection

The first milestone in the project was the collection of data from known and open-source data sources. The data ("Shopping Mall Customer Segmentation Data") was sourced from Kaggle's inventory of datasets which was published publicly by Zubair Mustafa. The data contains 5 features and 15,079 samples of customer data.

C. Data Features

The dataset consists of the following features:

Customer ID: Customer Identification Number of each customer in the dataset.

Age: The age of each customer in the dataset.

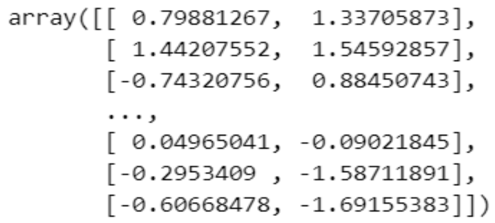
Gender: The gender of each customer in the dataset.

Annual Income: The annual yearly income of each customer.

Spending score: Spending habit score of each customer assigned by mall

D. Data Preparation

The collected data was pre-processed to remove unnecessary features (“Customer ID” & “Age”) from the dataset which did not have any observable correlation with the data. And was further scaled using the standard scaler to have uniformity across all features in the dataset to remove redundancies in the data and result in better data quality.



```
array([[ 0.79881267,  1.33705873],
       [ 1.44207552,  1.54592857],
       [-0.74320756,  0.88450743],
       ...,
       [ 0.04965041, -0.09021845],
       [-0.2953409 , -1.58711891],
       [-0.60668478, -1.69155383]])
```

Fig. 1. Standardization of dataset

E. Algorithms used

1) K-means Clustering

K-means clustering is an unsupervised machine learning algorithm used to cluster data into k clusters based on the similarity in features. It clusters data points based on the number of clusters k assigned.

Step-1: Determine the k number of centroids

Step-2: Assign clusters to the closest centroid using the Euclidean distance formula

$$dist(x,y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$$

Fig. 2. Euclidean distance

Step-3: Update centroids using the mean of each cluster points

Step-4: Repeat steps ii and iii until its stabilized.

2) Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering is a bottom-up clustering method where a cluster of data points is hierarchically formed based on similarities among them. Every data point is treated as its own cluster and then iteratively joins the closest clusters to it until all points have been merged into one cluster or the desired number of clusters chosen.

Step-1: Initialize each data point as its own cluster.

Step-2: Calculate the distance between all the cluster pairs

Step-3: Merge the clusters with less distance or high similarity. This is also based on the linkage criterion. i.e. Single, complete, or ward linkage.

Step-4: Repeat until a number of clusters preferred is formed.

3) Gaussian Mixture Clustering:

GMM is the clustering methodology based on the idea that the data observed stems from a mixture of several Gaussian distributions.

4) Fuzzy C-Means Clustering:

Fuzzy C-Means is a clustering method whereby data points can belong to multiple clusters with different degrees of membership. It assigns a probability to each data point for all clusters.

Step-1: Find the initial value for c, m & v

Step-2: Calculate each data point in reference to the preferred cluster with the equation

$$\mathrm{u_{ij}} = \frac{1}{\sum_{k=1}^{c} \frac{\|xj - vi\|}{\|xj - vk\|}^{2/(m-1)}}$$

Fig 3

Step-3: Calculate the new number of centers with the equation

$$\mathrm{V_i} = \frac{\sum_{j=1}^{N} xj * u_{ij}^{m}}{\sum_{j=1}^{N} u_{ij}^{m}}$$

Fig 4

Step-4: Repeat ii and iii until **u** is different from the value in the previous stage and less than the threshold.

C = number of clusters

M = fuzzy value

V = initial centres

IV. EVALUATION METHODS

A. Elbow Method:

The elbow method is a technique used predominantly for K-means clustering and works by plotting the within-cluster sum of squares often referred to as inertia against the number of clusters. The point at which the curve breaks forms the shape of an elbow [3].

Step-1: Choose a range of values for k, fit the model and compute the within-cluster sum of squares

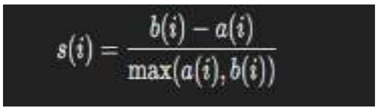
Step-2: Plot the wcss against the number of clusters k.

Step-3: Determine the elbow from the curve in the graph

Step-4: Select the optimum value for k in the graph

B. Silhouette Score:

Silhouette Score is an evaluation technique that gives the similarity of each point in its cluster to other clusters. It ranges from -1 to 1; where a score of -1 means that the data points are poorly clustered and there is no covalent structure in the data and a silhouette score of 1 indicates that there is a strong structure and well clustered data points. The silhouette score s(i) is calculated as:



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

Fig. 5. Silhouette score formula

Where:

a(i) is the average distance from i to other points in the cluster.

b(i) is the average distance from i to other clusters.

V. IMPLEMENTATION AND RESULTS

In this work, five clustering algorithms were applied: K-means clustering, K-means++ clustering, Agglomerative Hierarchical Clustering, Gaussian Mixture clustering, and Fuzzy C-Means clustering to segment the customers in a shopping mall dataset. The idea was to find interesting customer clusters using attributes to make appropriate business marketing strategies. The Silhouette Score and elbow method were used to evaluate each algorithm to know the optimum number of clusters needed and to know how best it fits the model contrary to other clusters.

A. Evaluation

For all the clustering algorithms, we calculated the silhouette score which ranged from -1 to 1, with values closer to 1 meaning there was a strong clustering of points. The results of each results can be seen below:

1) K-means and K-means++ Clustering:

Two evaluation methods were applied to the K-means and K-means++ algorithm to evaluate the number of clusters. The elbow method which uses the within-cluster sum of squares against the number of clusters to find the optimum cluster k showed that the optimum number of clusters required in the dataset is three as seen by the bend in the graph below. The graph shows a bend(elbow) at four being the optimum number of clusters k.

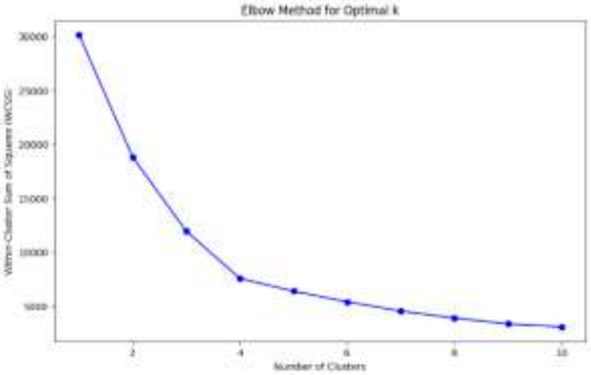


Fig. 6. K-means and K-means++ Elbow Graph

Meanwhile, evaluating the number of clusters with the silhouette plot gave an optimum score of 0.41 with the number of clusters k being 4. The silhouette score suggests that the points were moderately separated and distinguishable from each cluster point.

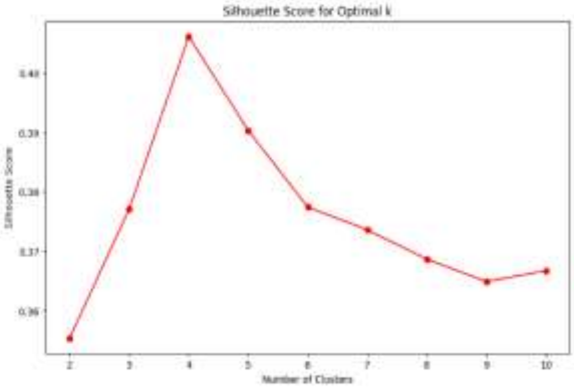


Fig. 7. K-means and K-means++ Silhouette plots

The silhouette plot was used to visualize the distribution of all the clusters made, to determine how well each data fits into its cluster compared to others.

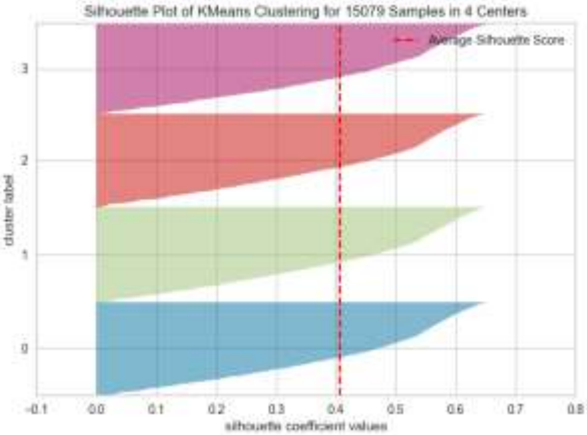


Fig. 8. K-means Silhouette Coefficient plot

2) Agglomerative Hierarchical Clustering:

The Hierarchical clustering algorithm had a silhouette score of 0.36 which produced 5 clusters. The 5 clusters show a simpler segmentation among the customers based on the features. And these 5 clusters can be used to perform marketing strategies based on it. The average linkage was used to derive the distance between the pairs of points in the two clusters.

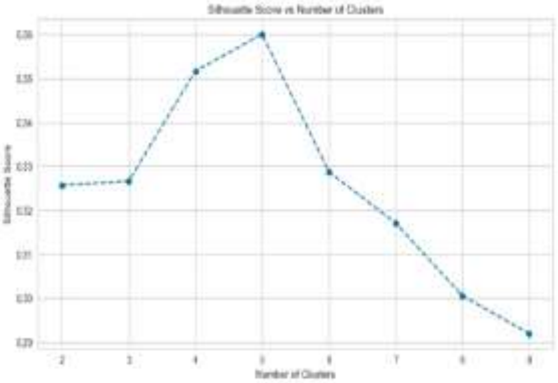


Fig. 9. Hierarchical Silhouette plot

For a more descriptive understanding of the clustering, a dendrogram was used to display the relationship between the points and visualize the hierarchy of the clusters using a bottom-up approach. The dendrogram is shown below starting from each data point and merging to form a single cluster.

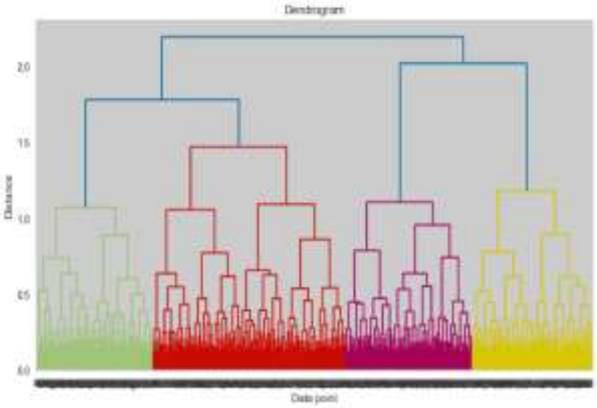


Fig. 10. Hierarchical Dendrogram

3) Gaussian Mixture Clustering:

The Gaussian Mixture had 4 clusters with a silhouette score of 0.41. This indicates a strong model based on how much meaningful pattern in the data the model has unraveled and has been segmenting the data into a very dissimilar cluster that considerably aligns well with the ground structure. The silhouette score has good separation between the clusters.

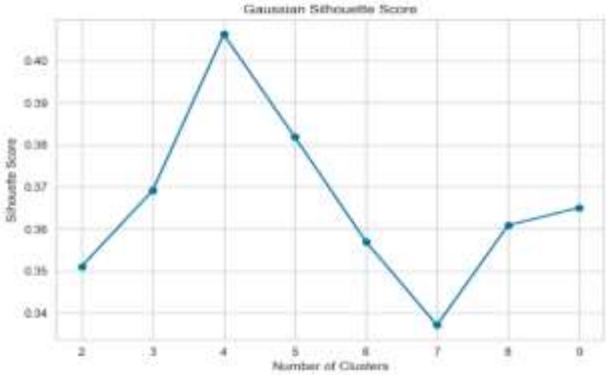


Fig. 11. Gaussian Mixture silhouette plot

4) Fuzzy C-Means Clustering:

This algorithm had a silhouette score of 0.41 with an optimum of 4 clusters. The score was similar to that of the K-means clustering. But Fuzzy C-means allows a data point to belong to more than one cluster with different membership degrees. The result of this algorithm shows that the data formed well-separated clusters even while using fuzzy means. The diagram below represents the silhouette scores for each cluster.

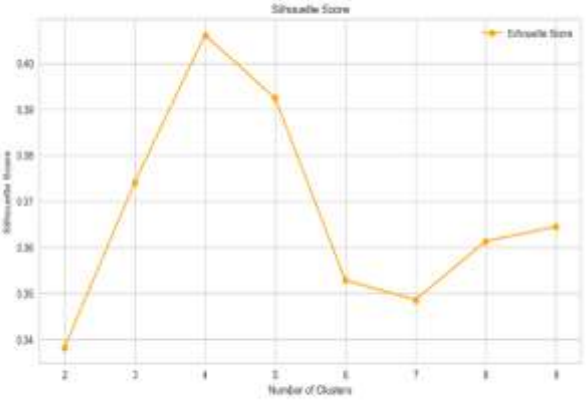


Fig. 11. Fuzzy C-Means Silhouette plot

B. Visualization of results

Below, scores of the Silhouette Coefficient for each algorithm are shown for evaluating the effectiveness of each clustering methods.

ALGORITHM	SILHOUETTE COEFFICIENT
K-means	0.4062145400231557
K-means++	0.4062145400231557
Agglomerative	0.3601076101336627
Gaussian Mixture	0.40626587748790594
Fuzzy C-Means	0.4061128624298844

Table 1. Silhouette Coefficient table

The different clusters were plotted in 3D. The results show different clustering obtained from K-Means and K-Means++. K-Means++ better initializes centroids for the stability of the solution, while Agglomerative Clustering emphasizes the hierarchical relationship, hence yielding slightly different segmentations that capture subtler similarities. GMM finds clusters with probabilistic boundaries and is thus flexible for use with overlapping clusters. Fuzzy C-Means allows for soft overlap in classes by letting partial membership across multiple classes reflect gradual transitions between groups.

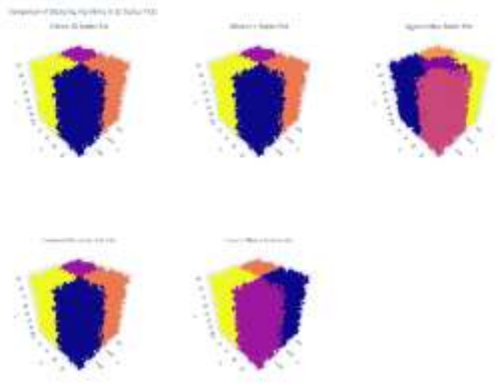


Fig. 12. Visualization of cluster Results

These visualizations illustrate that each algorithm interprets the structure in the data differently and gives insight into how to choose a proper methodology with regard to one's data and purposes of clustering.

C. Error Analysis

The performance of the algorithms using the silhouette score was moderate for K-means, K-means++, Gaussian Mixture, and Fuzzy C-Means with a score of approximately 0.41, which shows a good quality of clustering. However, Agglomerative Hierarchical Clustering performed slightly worse with a score of 0.36 which shows less clear clusters. Most of the low silhouette scores were located at the boundaries where the customers overlapped. Gaussian Mixture and Fuzzy C-Means reduced some boundary effects but left doubt in overlapping segments.

Comparing the models to the baseline, K-means, Gaussian Mixture, and Fuzzy C-Means showed more advantages. K-means had a silhouette score of 0.41 but was not effective at handling overlapping clusters. The Gaussian Mixture improved the segmentation by using probabilistic assignments, thus handling overlapping customer behaviors in a better way. Fuzzy C-Means went further by allowing partial memberships, offering insights into transitional customer behaviors. These models had better performance on the dataset by addressing K-means' limitations.

Misclassification at the cluster boundaries was common for all the algorithms. Data Points around the boundaries shared characteristics with more than one cluster which caused the ambiguity. K-means often misclassifies these types of points especially when the clusters are not well separated. The Gaussian Mixture model overcame these errors through probabilistic assignments, but its performance degraded when the probabilities across the clusters were alike. And Further, Fuzzy C-Means showed better handling of boundary points by allowing partial memberships but sometimes lost clarity in segmentation for points with weak memberships.

VI. CONCLUSION

These results can be applied to customer segmentation in business environments. By applying any of the clustering algorithms, businesses could use different marketing strategies based on the groups and businesses. If an in-depth approach is needed, businesses could apply Fuzzy C-Means clustering which allows for customers to be clustered into various segments.

In conclusion, businesses should choose an algorithm that suits their businesses, especially in cases where some customers could be deemed outliers even though they are not. The performance of these algorithms can be useful for customer segmentation but will vary based on their dataset.

VII. FUTURE WORKS

This project acts as a starting point for how clustering methods may be applied to customer segmentation. However, more work can be done for the models to become more effective and applicable. Including more customer features, like demographics, purchase history, or behavioral metrics, would strengthen the model. Also, the inclusion of ensemble methods will be promising in combining strengths from several clustering models making it more robust, especially for large and complex datasets. Deep learning models, including autoencoder or neural network-based clustering methods, can be used to get more informative insights about hidden features and their relationships in customer data. These could enhance the accuracy and applicability of the customer segmentation model.

REFERENCES

[1] R. M, C. Vijai, K. Srivastava, N. Kalyan, B. Pravallika, and A. Dutt, “Application of Machine Learning Algorithms for Customer Segmentation in E-Commerce Management,” in 2024 International Conference on Science Technology Engineering and Management (ICSTEM), Coimbatore, India: IEEE, Apr. 2024, pp. 1–5. doi: 10.1109/ICSTEM61137.2024.10560944.

[2] A. Afzal et al., “Customer Segmentation Using Hierarchical Clustering,” in 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India: IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/I2CT61223.2024.10543349.

[3] Varad R Thalkar, “Customer Segmentation Using Machine Learning,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., pp. 207–211, Dec. 2021, doi: 10.32628/CSEIT217654.

[4] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, “Customer Segmentation using K-means Clustering,” in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India: IEEE, Dec. 2018, pp. 135–139. doi: 10.1109/CTEMS.2018.8769171.

[5] S. R. Regmi, J. Meena, U. Kanojia, and V. Kant, “Customer Market Segmentation using Machine Learning Algorithm,” in 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India: IEEE, Apr. 2022, pp. 1348–1354. doi: 10.1109/ICOEI53556.2022.9777146.

[6] S. Na, L. Xumin, and G. Yong, “Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm,” in 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China: IEEE, Apr. 2010, pp. 63–67. doi: 10.1109/IITSI.2010.74.

[7] D. N. Rajalingam and K. Ranjini, “Hierarchical Clustering Algorithm - A Comparative Study,” Int. J. Comput. Appl., vol. 19.