A Data-driven Approach for Customer Segmentation Using RFM Analysis and K-means Clustering

Dr. Arpana Chaturvedi[1], Prof. Praveen Malik[2], Mr. Dipaker Jugran[3]
[1]*Department of AI/ML/IT, New Delhi Institute Of Management*
[2]*Department of IT and DA, New Delhi Institute Of Management*
[3]*PGDM, New Delhi Institute Of Management*

***Abstract*—The process of customer segmentation involves the division of a customer base into smaller groups or segments, based on shared characteristics such as age, gender, interests, and spending habits. The use of machine learning algorithms offers an automated and improved means of customer segmentation by training on vast datasets of customer data and using the resulting models to predict the appropriate segment for a new customer. This technology enables businesses to customize their marketing and sales strategies to target specific segments, resulting in more focused and effective communication with customers and potentially higher conversion rates. In this paper, a model based on unsupervised clustering algorithm, specifically the k-means method, in conjunction with Recency, Frequency, and Monetary (RFM) was employed to segment customers into distinct clusters based on their specific traits. Banking transactions data was used in this study, resulting in the identification of four clusters with unique characteristics that can be utilized for effective marketing purposes.**

Further, this study successfully demonstrated a real-world application of machine learning in the realm of customer segmentation, with tremendous potential for implementation in the banking industry. The deployment of machine learning algorithms to automate the process of customer segmentation holds the promise of improving the accuracy and efficiency of this critical function, leading to better targeting of specific customer segments and increased effectiveness of marketing and sales efforts. The significance of this study lies in its ability to showcase the transformative power of machine learning in driving innovation and optimization in the banking sector, thereby enhancing the overall customer experience and driving sustainable business growth.

***Index Terms*—Customer Segmentation, Machine Learning, Unsupervised Clustering, k-means Algorithm, RFM Analysis, Banking Industry, Targeted Marketing**

I. INTRODUCTION

Customer segmentation is the practice of categorizing a customer base into smaller groups, or segments, based on common characteristics. This approach enables businesses to effectively concentrate their marketing efforts and customize their products or services to the unique requirements and preferences of each segment. Segmentation can be carried out using various criteria, including demographics (such as age, gender, income, and education level), geographic location, and behaviour (such as purchasing habits or brand loyalty), and other factors. Some businesses combine these techniques to create more detailed and accurate customer segments. The goal of customer segmentation is to understand the needs and wants of different groups of customers and to create more personalized and relevant experiences for them and once customer segments have been identified, businesses can develop tailored marketing messages and strategies for each group. This might involve creating different product offerings, customizing marketing campaigns to speak directly to each segment, or providing personalized customer service and support.

Customer segmentation enables companies to refine their marketing efforts, improve customer experience, reduce marketing costs, and gain a competitive advantage. Firstly, customer segmentation is a valuable tool for targeted marketing. By grouping customers with similar characteristics, companies can create personalized marketing campaigns that cater to the unique needs and preferences of each group. As a result, this increases the likelihood of converting potential customers into loyal customers. Tailored marketing campaigns also allow for more precise measurement of return on investment (ROI) and, as a

result, enable more efficient allocation of marketing resources.

Secondly, customer segmentation can enhance the overall customer experience. By understanding the characteristics of each group of customers, companies can tailor their offerings to meet the specific needs of each segment. This can lead to higher customer satisfaction and retention rates. Additionally, personalized services and offerings can result in positive word-of-mouth recommendations, further increasing customer loyalty. Thirdly, customer segmentation can help companies reduce marketing costs. By understanding the characteristics of each group of customers, companies can create marketing campaigns that are more likely to appeal to each segment. This leads to more efficient use of marketing resources, as companies can avoid spending money on marketing campaigns that are unlikely to yield results.

Finally, customer segmentation can provide companies with a competitive advantage. By understanding their customers better than their competitors, companies can differentiate themselves from their competition by offering unique products and services that meet the needs of each segment. This can help companies attract new customers, increase market share, and ultimately increase profitability.

There are several techniques that can be used to perform customer segmentation:

- Association rules: Association rules are a statistical method used in customer segmentation that can help companies understand patterns and relationships between different customer behaviours and characteristics. Association rules are used to identify correlations between different variables and can provide insights into customer behaviour and preferences. Association rules analysis involves identifying frequent co-occurrence patterns in customer transactions. This analysis can reveal relationships between different items that customers purchase, such as products or services. For example, if customers who buy product A also tend to buy product B, this pattern can be identified through association rule analysis. Once identified, this pattern can be used to create targeted marketing campaigns that promote both products together, or to offer a discount or promotion for customers who purchase both products. Association rules can also help identify customer segments with similar buying behaviours or preferences. By analysing customer transactions and identifying patterns in their behaviour, companies can group customers into different segments based on their shared characteristics. This information can then be used to create targeted marketing campaigns that appeal to each segment's unique preferences.
- Collaborative filtering: In collaborative filtering, customer data is used to identify patterns and similarities between customers. This can involve analysing customer transactions, purchasing history, search behaviour, and other data points. Once these patterns are identified, they can be used to group customers into different segments based on their shared characteristics and preferences. By analysing a customer's purchase history and search behaviour, collaborative filtering can identify products or services that are likely to appeal to that customer. This information can then be used to create personalized marketing campaigns or product recommendations that are more likely to result in a conversion.
- Machine Learning: Machine learning is a powerful tool used in customer segmentation that enables companies to analyse large amounts of customer data and identify patterns and relationships between different variables. By leveraging machine learning algorithms, companies can identify customer segments based on their unique characteristics and preferences, enabling more effective targeting and personalized marketing. One of the most common machine learning techniques used in customer segmentation is clustering. Clustering algorithms group customers into different segments based on their shared characteristics and preferences. This can involve analysing a wide range of customer data, such as demographic information, purchasing history, search behaviour, and more. Once customer segments have been identified through clustering, companies can create targeted marketing campaigns that are tailored to each segment's unique preferences. This can include personalized product recommendations, targeted advertisements, and promotions or discounts that are specifically designed to appeal to each segment. Using machine learning algorithm on consumer data to divide them into clusters will be the focus of our study.

II. ROLE OF MACHINE LEARNING IN CUTOMER SEGMENTATION

Customer segmentation is a vital aspect of marketing strategy that involves dividing customers into groups with similar characteristics or behaviours. In recent years, machine learning has emerged as a powerful tool for customer segmentation, thanks to its ability to the process of analysing large volumes of data to uncover hidden patterns and relationships that may not be immediately evident to human analysts is commonly referred to as data analysis. This approach involves the use of specialized software and statistical algorithms to examine and interpret vast amounts of data. The aim is to gain insights and make informed decisions based on the findings. By employing data analysis techniques, organizations can identify trends and patterns that may not be apparent to human analysts and use this information to improve their operations, enhance customer experiences, and drive business growth.

One key advantage of using machine learning for customer segmentation is the ability to identify highly nuanced segments based on a wide range of variables, such as demographics, behaviours, and interests. Machine learning algorithms can analyse these variables and create highly accurate and precise customer segments, allowing businesses to develop more targeted and effective marketing campaigns. Another advantage of using machine learning for customer segmentation is scalability. Machine learning algorithms can process large datasets quickly and efficiently, allowing businesses to analyse and segment their customer base at scale. This is particularly useful for businesses with large and complex datasets that would be difficult or impossible to analyse manually.

Machine learning also offers greater customization in customer segmentation. Machine learning algorithms can create highly personalized customer segments that take into account individual customer preferences and behaviours. This level of customization can lead to more personalized and effective marketing campaigns that better resonate with individual customers. Finally, machine learning algorithms can continue to learn and adapt over time, refining segmentation models as new data becomes available. This means that customer segments can be continuously updated and improved, allowing businesses to stay up-to-date with changing customer needs and preferences.

In the context of machine learning, customer segmentation can be performed using a variety of algorithms, such as clustering algorithms, decision trees, and neural networks (Grossi, 2008).

III. K MEAN CLUSTERING

K-means clustering (S. Na, 2010) is a widely used unsupervised machine learning technique that involves dividing a dataset into distinct clusters. The technique, as described by S. Na in 2010, aims to group similar data points together and separate dissimilar data points into different clusters based on their distance from one another. The main goal of K-means clustering is to create clusters that are as different from each other as possible while keeping data points within each cluster as similar as possible. This approach is used to uncover patterns in data and identify relationships between variables.

In K-means clustering, the algorithm starts by randomly selecting K initial centroids from the dataset. During the clustering process, each data point is initially assigned to the nearest centroid, and new centroids are computed based on the mean of all the points assigned to them. This iterative assignment and centroid updating process is repeated until either the centroids no longer change significantly, or a pre-specified maximum number of iterations is reached. This helps to ensure that the clusters are formed in a way that is optimal for the given dataset, and that the resulting clusters are as distinct from one another as possible. By following this approach, K-means clustering enables users to effectively analyse complex datasets and identify important patterns and relationships that may not be apparent through other analytical methods. The quality of the resulting clusters depends on the choice of initial centroids and the number of clusters (K). The algorithm can be run multiple times with different initial centroids to improve the chances of finding a good clustering solution.

In marketing and customer analytics, K-means clustering is a popular technique. The goal of customer segmentation is to group customers based on similar characteristics or behaviours, so that targeted marketing strategies can be developed for each segment. Once the clusters have been formed, they can be analysed to gain insights into customer behaviour and preferences. For example, marketers can examine

the characteristics of each cluster to determine which products or services are most popular among each segment, and use this information to develop targeted marketing campaigns.

Overall, K-means clustering is a powerful and widely used technique in data analysis and machine learning, particularly in applications in customer segmentation.

IV. RFM ANALYSIS

RFM (A. Joy Christy, 2021) stands for Recency, Frequency, and Monetary Value. It is a customer segmentation model used in marketing and customer relationship management to identify groups of customers based on their purchase behaviour.

Recency, in the context of customer behaviour analysis, refers to the time elapsed since a customer's most recent purchase. Customers who have made a purchase more recently are deemed more valuable than those who have not made a purchase in a longer time frame. Frequency is another aspect of customer behaviour analysis and refers to how often a customer makes purchases. Customers who make frequent purchases are deemed more valuable than those who make fewer purchases. Finally, monetary value represents how much a customer spends on their purchases. Customers who spend more money are considered more valuable than those who spend less. These three measures of customer behaviour analysis, recency, frequency, and monetary value, are commonly referred to as RFM analysis.

Using these three factors, RFM analysis helps businesses identify their most valuable customers and develop targeted marketing strategies to retain and upsell to these customers. By segmenting customers based on their RFM scores, businesses can tailor their marketing and communication efforts to each group, resulting in more effective marketing campaigns and increased customer loyalty.

RFM analysis can be enhanced with machine learning algorithms to improve its accuracy and effectiveness in customer segmentation. Machine learning algorithms can analyse large amounts of customer data and identify patterns and trends that may not be immediately apparent using traditional RFM analysis.

V. LITERATURE REVIEW

Machine learning-based customer segmentation is a method for grouping people that have similar traits from a customer base. This is typically done using a variety of statistical and machine learning techniques, such as clustering, decision trees, and neural networks. In literature review, there are many studies have been proposed and developed various methods for customer segmentation using machine learning. Clustering algorithms, such as k-means and hierarchical clustering, have been popularly used for customer segmentation. Decision tree and Random Forest have also been used in customer segmentation by finding the relationship between customer features and their purchase behaviour. Neural networks, specifically Self-Organizing Maps (SOM) and Artificial Neural Networks (ANN) have also been applied in customer segmentation by learning the underlying structure of the customer data.

Additionally, there are many studies which have used various feature selection methods such as Principal component analysis (PCA), Factor analysis (FA) and independent component analysis (ICA) for customer segmentation. Some studies have also used ensemble methods such as Bagging and Boosting for better performance.

Overall, customer segmentation using machine learning is an active area of research, with many different techniques and approaches being developed and tested. The choice of technique will depend on the specific characteristics of the data and the goals of the analysis.

(Badea, 2013) discusses the performance of neural networks and support vector machines, two of the most well-liked machine learning algorithms, in a segmentation process. To segment the customer and use the appropriate marketing methods, various clustering approaches have been presented in (Monil, 2020) . It also explores whether a hybrid clustering method could do better than a single model. (H. S. Al-Amin) proposes a customer segmentation approach based on decision trees. The study aims to help retailers understand their customers better by dividing them into segments based on their purchasing behaviour and demographic characteristics. The study uses a dataset of customers from a retail store in Bangladesh and applies decision tree algorithms to segment the customers. The authors use several decision tree

algorithms, including ID3, C4.5, and CART, to identify the best algorithm for customer segmentation.

(Hossain, 2017) proposes a customer segmentation approach based on clustering algorithms. It uses a dataset of 3000 customers and applies two clustering algorithms, K-Means (a centroid-based clustering algorithm) and DBSCAN (a density-based clustering algorithm), to segment the customers. The authors evaluate the effectiveness of these algorithms based on various clustering performance measures such as the Silhouette coefficient, Dunn index, and Calinski-Harabasz index. (Wang, 2022) proposes an approach for customer segmentation in digital marketing using deep learning with swarm intelligence. For effective consumer segmentation, a self-organizing map with an improved social spider optimization strategy has been deployed as an unsupervised deep learning model. Modified social spider optimization, a swarm intelligence model, is used in a feature engineering process to analyse the customer data and pick the behavioural attributes of the consumer. Following that, the clients are grouped utilising a self-organizing neural network (SONN). Customers are categorised using the Deep Neural Network (DNN) model based on the clusters.

In another study (Alkhayrat, 2020), authors compare two methods for dimensionality reduction in telecom customer segmentation: deep learning and principal component analysis (PCA). The study found that both methods were effective in reducing the dimensionality of the data, but deep learning was more effective in preserving the information content of the original data than PCA. This suggests that deep learning may be a better method for telecom customer segmentation, as it can better capture the underlying patterns in the data. In (N. Gankidi, 2022), the task was completed using a Python with Machine learning approach. After the data was loaded, it was visualized with bar plots based on different variables. Important variables were identified, and the method called the Elbow method was used to obtain the number of clusters or groups. These variables were then fit into the K-means model in a passive voice. The details and mentalities of customers would be provided by this plot, thereby helping companies to improve their products and techniques to increase their sales.

In another study (Papetti, 2019), consumer information pertaining to the consumption of cannabis products, including flowers and concentrates, was leveraged in conjunction with the Recency-Frequency-Monetary (RFM) framework to effectively segment the consumer base. Machine learning algorithms, such as k-means and agglomerative clustering, were employed to attain the final outcomes. The resultant analysis revealed approximately five to six distinct clusters, each with distinct purchasing tendencies and behaviours.

VI. METHODOLOGY USED

For our study on customer segmentation using machine learning, A publicly available dataset of Bank Customer Segmentation Dataset (BCSD) is used. The Bank Customer Segmentation Dataset (BCSD) is a collection of more than 1 million transactions made by over 800,000 customers of a bank in India. The dataset can be accessed through the following link: https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation. It includes a variety of information, such as customer age (date of birth), location, gender, account balance at the time of the transaction, transaction details, and transaction amount. This data can be used to analyze customer behavior, identify patterns, and develop effective marketing strategies to target specific customer segments.

Table 1. Description of the dataset

Column Name	Unique Values
TransactionID	10485647
CustomerID	884265
CustomerDOB	Date Value
CustGender	2
CustLocation	9355
CustAccountBalance	Continuous Value
TransactionDate	Date Value
TransactionTime	Time Value
TransactionAmount (INR)	Continuous Value

1) Exploratory Data Analysis:

The dataset analyzed in this study consisted of a vast amount of transactional data, encompassing 10485647 transactions from a total of 884265 customers. The dataset contained 9 distinct features, as detailed in table 1, that were used to extract insights and identify patterns within the data. An examination of the transactional data revealed that Mumbai had the highest number of transactions, followed closely by New Delhi. This observation is depicted in figure 4,

which highlights the variation in transaction volume across different geographic locations.

	TransactionID	CustomerID	CustomerDOB	CustGender	CustLocation	CustAccountBalance	TransactionDate	TransactionTime	TransactionAmount (INR)
0	T1	C5841053	10/1/94	F	JAMSHEDPUR	17819.05	2/8/16	143207	25.0
1	T2	C2142763	4/4/57	M	JHAJJAR	2270.69	2/8/16	141858	27999.0
2	T3	C4417068	26/11/96	F	MUMBAI	17874.44	2/8/16	142712	459.0
3	T4	C5342380	14/9/73	F	MUMBAI	866503.21	2/8/16	142714	2060.0
4	T5	C9031234	24/3/88	F	NAVI MUMBAI	6714.43	2/8/16	181156	1762.5

Fig. 1 Overview of the Dataset

Furthermore, the variation over time in the total customer account balance and transaction amount was also explored. Figure 2 illustrated that the total customer account balance reached its peak in September 2016, while the total transaction amount peaked in August 2016. This finding provides valuable insights into the transactional behavior of customers over time.

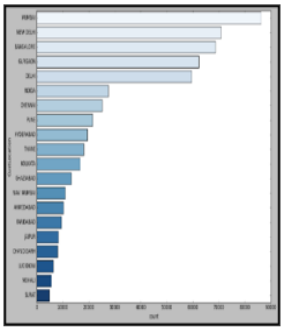

Fig. 2(a) Count of Transaction According to its location.

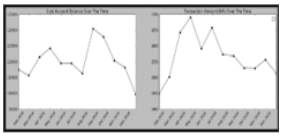

Fig. 2(b) Variation in 'Account Balance' (left) and 'Total

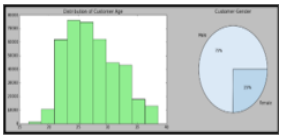

Fig. 2(c) Distribution of 'Age' (left) and 'Gender' (right)

Further EDA revealed that a mere 25% of all transactions were conducted by female customers, while the remaining 75% were performed by male customers. The ages of the customers at the time of transaction were calculated using their birth dates and the transaction date, resulting in a range of ages from 15 to 60 years. It was found that the age distribution of the customers followed a normal distribution, indicating a balanced spread of ages across the dataset as illustrated in figure 3.

A. Feature Engineering:

Feature engineering is an important step in the machine learning process, especially for customer segmentation. It involves creating new features (i.e., variables) from the existing data that can help improve the accuracy of a model. In our study, following feature engineering has been done:

1) Demographic information:

- age_at_purchase: Age at the time of transaction of each customer is created by subtracting customer's date of birth from the Date of transaction.

$$Age_{at_{purchase}} = Cust_{DOB} - Transaction_{date}$$

- income_range: Customers are divided into 3 income groups according to their Account balance- Low Income, Medium Income, and High Income.

2) RFM Features:

- Monetary: A scoring system was devised to evaluate customers based on the monetary value of their transactions. Customers were assigned a score from 1 to 5, with a score of 5 indicating high spenders and a score of 1 indicating low spenders. The scoring bins were determined based on the distribution of transaction amounts, with the aim of ensuring a fair and representative evaluation of customer spending behavior.
- Recency: In this feature engineering process, a numerical scale ranging from 1 to 5 was employed to quantify the recency of a customer's last transaction. Customers who have conducted multiple transactions recently are assigned the highest score of 5, while those who have not made any purchases for an extended period are given a score of 1. This scoring approach enables a more granular representation of the recency of transactions.
- Frequency: To assess the frequency of customer transactions in a more standardized and comparable manner, a feature was created in this study that assigned a score on a scale of 1 to 5, based on the number of transactions made by each customer. Customers with a high volume of transactions were given the highest score of 5, while those with a lower number of transactions were given a score of 1. This approach allowed for a more objective and standardized assessment of customer transaction frequency

3) Aggregated statistics: We will calculate summary statistics such as the total number of transactions, the total amount spent, and the average transaction amount for each customer. These features can provide valuable insights into the customer's spending habits.

By incorporating these engineered features into our model, we created a more accurate representation of the customer segments and their behaviour.

B. Machine Learning Model:

K-means is an unsupervised machine learning algorithm used for clustering data points. It aims to partition a given set of n data points into k clusters. The objective of the algorithm is to minimize the sum of squared distances between each data point and its corresponding cluster's centroid. The algorithm iteratively updates the centroid of each cluster and assigns each data point to the nearest centroid until the centroid positions converge or a maximum number of iterations is reached. The outcome of the algorithm is a partition of the data points into k clusters, where each data point is assigned to the closest centroid based on its distance. K-means clustering is a commonly used technique in data mining and pattern recognition for grouping similar data points together and separating dissimilar ones into different clusters.

1) Clustering Steps:

a) Initialization: The algorithm starts by randomly selecting k centroids, which will serve as the initial cluster centres.

b) Assignment step: The algorithm determines the Euclidean distance between each data point and each of the k centroids for each data point. Next, the data point is assigned to the cluster with the nearest centroid.

c) Recalculation step: The centroids of each cluster are then recalculated as the mean of the data points in the cluster.

d) Repeat steps b and c until convergence: This process is repeated until the centroids no longer move or change, indicating that the clusters have stabilized.

e) Result: The result of the k-means algorithm is k clusters, with each cluster containing a set of data points.

2) Elbow Method

A popular method for determining the ideal number of clusters in a given dataset is the elbow method. It involves fitting the data to a clustering model for varying values of k, the number of clusters, and plotting the explained variance (i.e., the within-cluster sum of squares) as a function of k. The resulting plot resembles an arm with an elbow, and the elbow point, where the rate of reduction in within-cluster sum of squares slows significantly, represents the optimal number of clusters. This method enables data analysts to identify a suitable number of clusters that best captures the variance in the data, and thus obtain meaningful insights into the underlying patterns and structure of the dataset. In this study, the elbow method was utilized to determine the optimal number of clusters for the given dataset.

3) Evaluation Metrics:

- Silhouette Coefficient: The Silhouette Coefficient is a metric that measures the similarity of a data point to its own cluster compared to other clusters. It is a numerical value that falls between -1 and 1, where a score of 1 indicates a strong match between the data point and its cluster, and a score close to -1 indicates that the data point is incorrectly assigned to a cluster. A score of 0 suggests that the data point is located near the decision boundary between two clusters, making it challenging to determine its appropriate cluster assignment. The Silhouette Coefficient is frequently used in clustering algorithms to assess the quality of the resulting clusters and optimize the model's performance. The Silhouette Coefficient is a useful tool for determining the optimal number of clusters for a given dataset and for evaluating the quality of a clustering algorithm.
- Inertia: Inertia is a measure used in K-means clustering to evaluate the effectiveness of the algorithm in grouping similar data points together within a cluster. In the context of clustering analysis, the term "inertia" refers to a metric that measures the sum of the squared distances between each data point within a cluster and the centroid of that cluster. In other words, it calculates how far the data points are from the centre of the cluster. This metric is used to evaluate the quality of a clustering solution, with lower values of inertia indicating that the clusters are more tightly packed around their centroids. The goal of clustering algorithms, such as K-means, is to minimize the inertia to create more accurate and meaningful clusters. The goal of K-means clustering is to minimize the inertia, which is achieved by optimizing the location of the cluster centroids. The inertia value is used as a criterion to determine the optimal number of clusters, by comparing the inertia values across different values of k (the number of clusters). Typically, as the number of clusters increases, the inertia value decreases, since the data points are more closely grouped together.

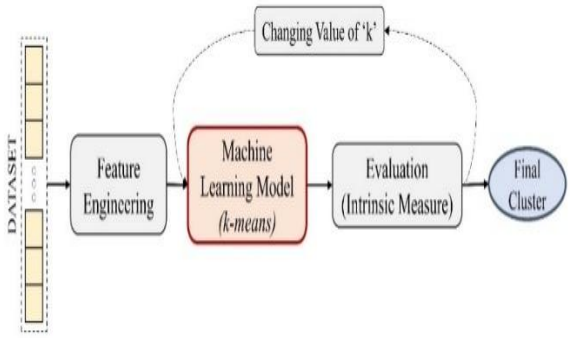

Fig 2. Overview of the workflow

VII RESULTS AND DISCUSSION

To segment customers into distinct groups, a K-means clustering model was trained on pre-processed data that had undergone feature engineering. In each iteration of the model, the value of 'k', which determines the number of clusters, was adjusted, and graphs were plotted to visualize the performance of the model. Two metrics, Inertia and Silhouette Coefficient, were used to evaluate the effectiveness of the model at different values of 'k'. The Inertia is a measure that represents the sum of the squared distances of each data point to its nearest cluster center in cluster analysis. The lower the Inertia value, the more compact and well-separated the clusters are. On the other hand, the Silhouette Coefficient is a metric that measures the compactness and separation of the clusters. It provides a score that ranges from -1 to 1, where a score closer to 1 indicates that the data point is well-matched to its own cluster and poorly matched to neighbouring clusters, while a score closer to -1 indicates the opposite. The Silhouette Coefficient is used to evaluate the quality of clustering results and determine the optimal number of clusters to use in a given dataset.

Figure 6 depicts a graphical representation that illustrates the correlation between Inertia and the number of clusters for the dataset in question. Inertia is a metric that quantifies the total sum of squared distances of each data point from its nearest cluster center. This graph visually shows how Inertia changes as the number of clusters increases, allowing for a determination of

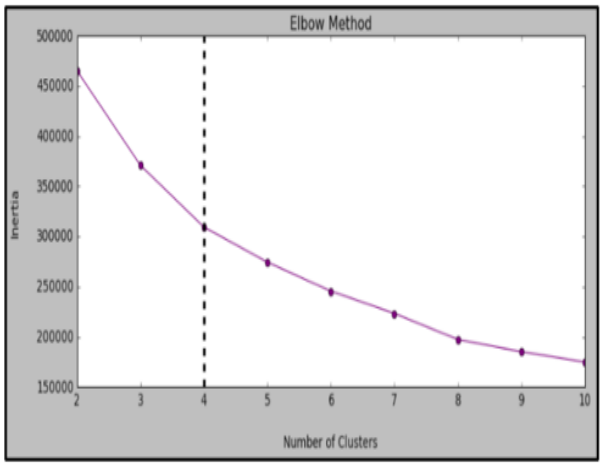

Fig. 4 Inertia vs Number of Clusters

the optimal number of clusters required for effective customer segmentation.

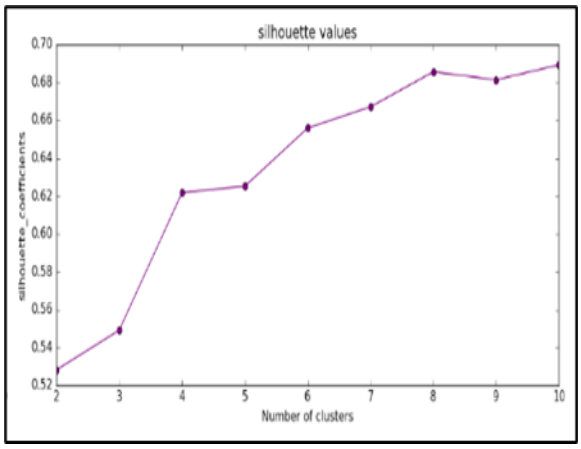

Fig.5. Silhouette Coefficient vs Number of Cluster

Inertia decreases as the number of clusters increases. However, it also shows an elbow-shaped curve at k = 4, indicating that adding more clusters beyond this point does not result in a significant decrease in Inertia. Thus, the ideal number of clusters for this dataset is determined to be 4, and this number is used for the final cluster prediction. This optimal number of clusters ensures that the data is effectively partitioned into homogeneous groups, allowing businesses to gain meaningful insights into their customer segments and tailor their marketing strategies accordingly.

Once the final prediction was made using the machine learning algorithm, a dendrogram was generated as shown in Figure 8. The dendrogram serves as a visual representation of how the clusters were computed,

providing insights into the underlying patterns and relationships among the data points. In the dendrogram, the four predicted clusters are represented by different colors, allowing for easy identification and interpretation of the clusters. The structure of the dendrogram also provides information on the similarity and dissimilarity of the data points, as well as the distances between them.

By examining the dendrogram, businesses can gain a better understanding of the clustering results and use this information to develop targeted marketing strategies for each cluster. For instance, they can focus on retaining customers in the high-value clusters while identifying opportunities for upselling and cross-selling to customers in the low-value clusters.

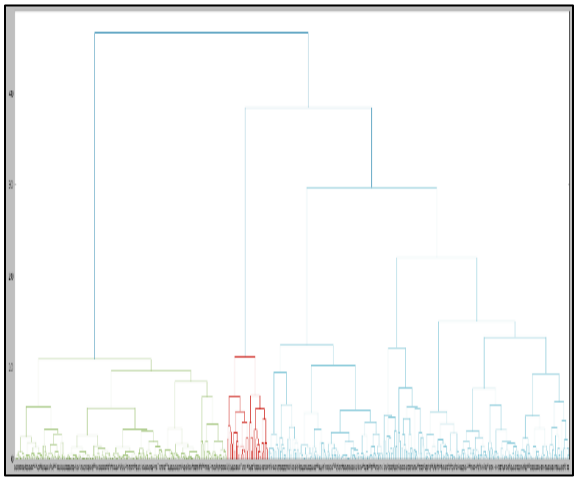

Fig.6. Dendrogram

After conducting the prediction, the dataset was categorized into four distinct clusters, each labelled as 'Cluster 0', 'Cluster 1', 'Cluster 2', and 'Cluster 3'. Figure 9 provides a visualization of the number of data points assigned to each cluster by the K-means algorithm. It can be observed from the graph that Cluster 0 has the highest number of data points assigned to it, indicating that it represents the largest customer segment. Conversely, Cluster 1 has the lowest number of data points, signifying that it represents the smallest customer segment among the four clusters. By understanding the size of each cluster, businesses can better comprehend the distribution of their customer base and the relative importance of each segment.

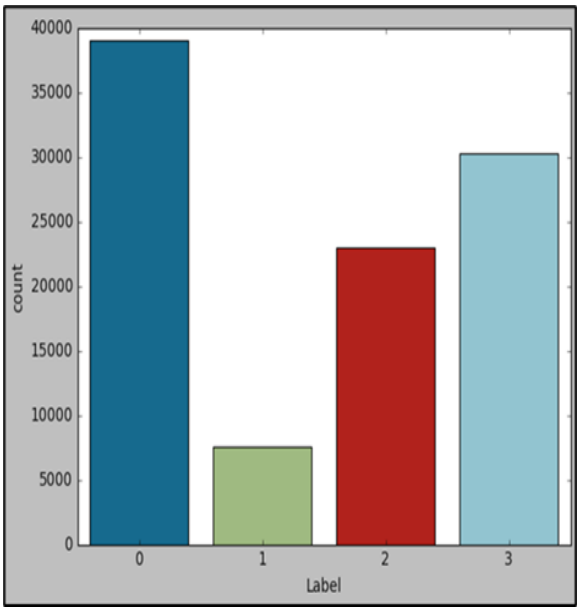

Fig.7. Count of datapoints in each cluster

In Figure 10, the radar chart displays the weight distribution of each feature in each cluster. It shows that in Cluster 0, gender is the most influential feature, while the other features have negligible weight in this cluster. On the other hand, Cluster 3 comprises the most valuable customers as they demonstrate high frequency and recency of transactions, coupled with high monetary value. This indicates that these customers have made recent purchases and tend to make purchases frequently, spending a significant amount of money on each transaction. Therefore, it is crucial for businesses to prioritize their efforts in retaining and cultivating a strong relationship with these customers, as they are likely to have a significant impact on the overall revenue and success of the business.

Further, it was observed that Cluster 1 comprises customers who spend a considerable amount of money on their purchases, but they do not make frequent transactions. In contrast, Cluster 2 includes customers who make frequent transactions but spend relatively less money on each transaction. Therefore, it is imperative for businesses to adopt distinct marketing strategies for each of these customer segments. For customers in Cluster 1, the focus should be on increasing the frequency of their transactions while maintaining their high monetary value. In contrast, for customers in Cluster 2, the emphasis should be on increasing the monetary value of their transactions

while ensuring they continue to make frequent purchases. By understanding the unique characteristics of each customer segment, businesses can tailor their marketing efforts to optimize their revenue and customer satisfaction.

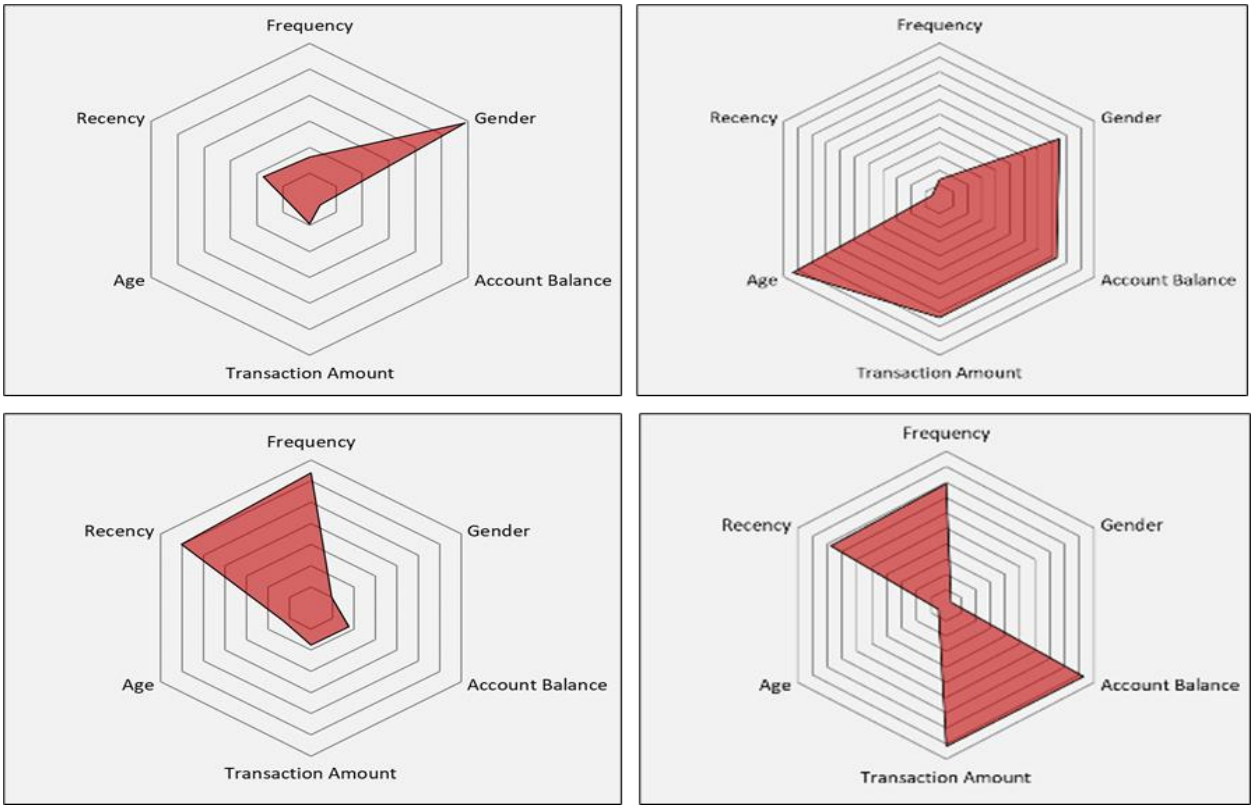

Fig.8. (From left) Cluster 0; Cluster 1; Cluster 2; Cluster 3

VIII CONCLUSION

This study aimed to utilize machine learning techniques to segment customers into distinct groups, thereby enabling businesses to tailor their marketing strategies and improve customer experience.

The dataset was subjected to feature engineering to extract meaningful insights, and a K-means clustering model was trained to categorize customers into clusters based on their behaviour patterns. The elbow method and Silhouette Coefficient were employed to identify the optimal number of clusters, and four distinct clusters were ultimately identified and labelled.

The resulting clusters were analysed to obtain a deeper understanding of their characteristics, and a radar chart was used to visualize the weights of each feature within each cluster. This provided insights into the preferences and behaviour patterns of different customer segments, allowing businesses to develop targeted marketing strategies and enhance customer satisfaction.

In conclusion, this study demonstrates the value of sophisticated machine learning algorithms in customer segmentation, and highlights how businesses can utilize such techniques to gain valuable insights into their customers and enhance their marketing strategies.

IX. ACKNOWLEDGMENT

This research was supported by various research work done by researchers and published papers. I thank all of them as their research work has provided the insight and expertise to me and their work has greatly assisted my research.

REFERENCES

[1] A. Joy Christy, A. U. (2021). RFM ranking – An effective approach to customer segmentation,

Journal of King Saud University - Computer and Information Sciences, 1251-1257.

[2] Alkhayrat, M. A. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. Journal of Big Data, 7.

[3] Badea, I. S. (2013). Customer segmentation in private banking sector using machine learning techniques. Journal of Business Economics and Management, 14, 923-939.

[4] Grossi, E. a. (2008). Introduction to artificial neural networks. European journal of gastroenterology & hepatology, 19, 1046-54.

[5] H. S. Al-Amin, M. M. (n.d.). Customer Segmentation Using Decision Trees: A Study on the Retail Industry. 2020.

[6] Hossain, A. S. (2017). Customer segmentation using centroid based and density-based clustering algorithms. 3rd International Conference on Electrical Information and Communication Technology, 1-6.

[7] Mishra, S. a. (2017). Principal Component Analysis. International Journal of Livestock Research, 1.

[8] Monil, P. a. (2020). Customer Segmentation Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 2104--2108.

[9] N. Gankidi, S. G. (2022). Customer Segmentation Using Machine Learning. 2022 2nd International Conference on Intelligent Technologies (CONIT), 1-5.

[10] Papetti, R. H. (2019). Customer Segmentation Analysis Of Cannabis Retail Data: A Machine Learning Approach. The University of Arizona.

[11] S. Na, L. X. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 63-67.

[12] Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. Information Processing & Management, 103085.