Customer Segmentation with RFM analysis using ML models

Manpreet Kaur
Assistant Professor CSE
Chandigarh University
Mohali (140413), India
manpreet.e13653@cumail.in

Astha Sharma
CSE Department
Chandigarh University
Mohali (140413), India
21bcs9826@cuchd.in

Mayank Kishore
CSE Department
Chandigarh University
Mohali (140413), India
21bcs9834@cuchd.in

Singh Surajkant Shubhnath
CSE Department
Chandigarh University
Mohali (140413), India
21bcs9870@cuchd.in

Harsh Verma
CSE Department
Chandigarh University
Mohali (140413), India
21bcs9837@cuchd.in

***Abstract*—The technique of examining the attributes of customers and creating groups based on some specific characteristics is known as customer segmentation. Customer segmentation is a methodology employed to analyze the characteristics of consumers and subsequently categorize them into distinct groups based on specified features. Customer segmentation is a widely employed approach for discerning the identities of the most devoted patrons as well as those that exhibit infrequent engagement with the platform. This study examines a new methodology for consumer segmentation by combining Recency, Frequency, Monetary (RFM) information with sophisticated Machine Learning (ML) models. Through the utilization of transactional data, the calculation of RFM scores is employed, which enables to get valuable insights into client behavior pertaining to the aspects of purchase recency, frequency, and monetary value. The scores play a crucial role in the training of K-means clustering and other related algorithms. The existing machine learning models exhibit potential for development. The study performed has a strong and dependable machine learning model that operates with high efficiency and delivers consistent performance. This is achieved by including the RFM scores of all consumers into the model. The model that has been provided has superior efficiency and robustness compared to all other models, resulting in a notable improvement in silhouette score which comes out to be 0.954319641633133, thus making it 53.8% more efficient then the previously built one. This achievement is particularly noteworthy when considering the conventional machine learning models that have previously been employed.**

Keywords—Customer Segmentation, RFM model, Unsupervised Machine Learning

I. Introduction

The concept of client segmentation was initially presented by Wendell R. Smith in 1956. Nevertheless, despite the passage of several decades, a significant number of organizations, regardless of the size, have failed to implement consumer segmentation strategies, impeding the potential for exponential growth. There are several factors contributing to company failure, and this analysis suggests that a significant cause is the deliberate neglect of customer knowledge acquisition by enterprises. This assertion is supported by the findings of a study on customer segmentation, as shown in the first line of the aforementioned research.

Within the realm of marketing, the significance of customer segmentation cannot be overstated. This strategic undertaking enables businesses to precisely tailor their strategies according to distinct customer categories. Concentrating on particular needs allows companies to enhance their marketing effectiveness, provide personalized experiences, and allocate resources judiciously. Customer segmentation is an invaluable tool for firms across all scales. By comprehending the demands and preferences of the clientele, enterprises may design items and services that are more inclined to fulfill those requirements. Additionally, enterprises possess the ability to generate marketing and sales communications that have a higher probability of connecting with the intended demographic. The use of customer data in customer segmentation has a significant influence on company outcomes. The significance of customer segmentation research is particularly pronounced in the contemporary competitive marketplace, as organizations strive to establish a distinct identity and set themselves apart from the rivals.

By gaining a comprehensive understanding of the wants and preferences of the consumers, organizations may strategically design and offer goods and services that have a higher probability of effectively fulfilling those demands. Additionally, organisations possess the ability to generate marketing and sales communications that are more inclined to resonate with the intended demographic. Customer segmentation is a pivotal strategy that holds significant potential for businesses in the foreseeable future. Through the process of customer segmentation, businesses may gain a comprehensive understanding of the specific investment requirements and strategies for each individual consumer. The implementation of a personalized area for each client will enhance the quality of feedback, thus benefiting the business or corporation.

In the marketing domain, consumer segmentation entails the identification of distinct customer groups based on various characteristics. Demographic traits, such as age and gender, are determined using customer data or surveys. Psychographic traits delve into personality and lifestyle, utilizing surveys and social media analysis. Behavioral traits center on purchasing patterns and loyalty, drawing insights from transaction history and website data. Generational traits classify customers based on birth years, like Baby Boomers or Millennials. Income traits assess economic status, gathered through surveys or transactions. Social class traits infer lifestyle from factors like occupation and education. Usage-based traits analyze product or service usage frequency, benefits sought, and user status.

RFM analysis, an acronym for Recency, Frequency, and Monetary Value, is a highly effective approach for achieving this objective. The RFM research methodology utilizes past consumer transaction data to classify clients into separate categories according to the purchase patterns. These categories provide a fundamental basis for developing customized marketing plans and efforts, eventually enhancing the effectiveness of marketing campaigns and overall client happiness. In the present context, the ways in which machine learning models might enhance and streamline the process of consumer segmentation shall be examined.

The RFM analysis methodology is based on three core measures, namely Recency (R), Frequency (F), and Monetary Value (M), which provide significant and in-depth understanding of consumer behavior. The concept of recency refers to the temporal duration that has passed since a customer's most recent transaction, therefore functioning as an indicator of the customer's current level of involvement with the enterprise. Customers who have made purchases in the recent past tend to exhibit greater levels of involvement and possess a higher likelihood of making subsequent purchases. Frequency is the rate at which a consumer engages in purchasing activities throughout a designated period, commonly spanning one year. Customers that exhibit a high frequency of purchases demonstrate a greater degree of loyalty and value, hence making consistent contributions to the company's overall revenue. Monetary value pertains to the aggregate sum of money that a consumer has expended on purchases within a specified timeframe. This process enables the identification of consumers who possess substantial value and exert a major influence on the company's financial performance.

The dataset under consideration contains 541,910 entries of customer transaction data, comprising a range of parameters including customer ID, nationality, invoice number, and other relevant information.

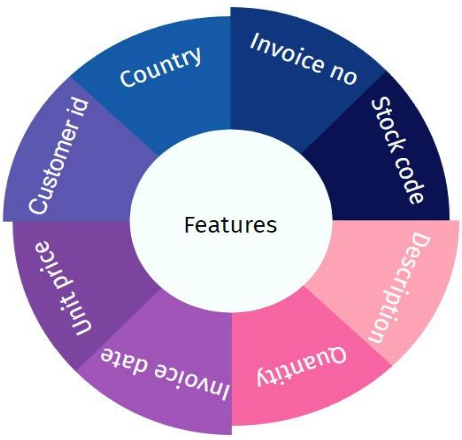

Fig 1: Features in Dataset

Although standard RFM analysis is useful for understanding consumer behavior, the segmentation process may be further improved by using machine learning models, which give increased precision and granularity. Machine learning enhances RFM analysis through its capabilities in data preparation, feature engineering, use of clustering techniques, interpretation of segments, development of customized marketing strategies, and continuous monitoring and adaptation. The use of this dynamic strategy guarantees the preservation of the segmentation's relevance and accuracy in the long term, making a significant contribution to sustained commercial success.

II. Literature review

This part pertains to the existing study conducted by many scholars in the field of Customer Segmentation. The purpose of this section is to highlight and reinforce this information via the use of the provided Literature Review.

In the study, Umakant Mandawkar et al. [1] identified and classified the data into seven distinct segments or clusters. Among these clusters, the data inside three segments had the highest silhouette score, measuring 0.4602. This study highlights the significance of segmentation in achieving company success and explores several methodologies, such as k-means clustering and RFM analysis, for efficiently segmenting clients. Kodinariya et al. [2] conducted a study that explored different methods for identifying the most suitable number of clusters within a dataset. The research examined six distinct techniques for estimating the ideal cluster count in datasets. These methods included the "Rule of Thumb," the "Elbow Method," the "Information Criteria Approach," the "Information Theoretic Approach," the "Silhouette Assisted Choice," and "Cross Validation."

Mohiuddin Ahmed et al. [3] conducted research that explored the k-means clustering technique, its limitations, and the numerous modifications proposed to address these constraints. The article commences with an introductory section elucidating the concept of clustering and its significance. It provides a comprehensive examination of various k-means algorithm variants tailored to tackle initialization issues and manage mixed data types. Moreover, the paper includes a performance assessment employing multiple metrics and an analysis of the associated computational complexity. In the study, Bayer et al. [4]underscored the potential of enhanced segmentation techniques in enabling precise targeting. The dataset was segregated into 25 distinct segments primarily characterized by five levels and five behavioral groups. This division enhances the accuracy of comprehending customer loyalty and the contributing factors to customer value for each individual customer.

In the study, Patel Monil et al. [5] investigated the importance of customer relationship management (CRM) in the context of e-commerce. The authors emphasize the potential of CRM to enhance customer happiness and loyalty through the use of clustering analysis. Additionally, the study provides various clustering methodologies for this purpose. Customer traits are identified, which in turn facilitates the implementation of personalized marketing tactics. Ultimately, these efforts contribute to the company's financial gains and the establishment of enduring customer connections.

The study conducted by Rocio Gonzalez Martinez [6] presented a comparative analysis between RFM Model based on the Fuzzy Linguistic and the conventional based RFM model. The research focused on the business aspect by employing K-means clustering on a 2-tuple model. Furthermore, this manuscript elucidates the use of the RFM model in relation to the retail sector. The study indicates that both models have utility; however, 2-tuples exhibit several advantages in comparison to conventional models.

Subsequent research endeavors may center on the examination of weight vectors through the utilization of test campaigns.

In the study, Christy et al. [7] have presented a unique approach for choosing the initial centroids in the K-means algorithm. This research highlights the comparison between Fuzzy C-means and k-means algorithms, concluding that k-means is a superior technique because of its shorter execution time and fewer iterations required. The study conducted by Mehdi Mohammadzadeh et al.[8] proposes a data mining methodology that utilizes the RFM model to analyze churn behavior. In this methodology, K-means clustering is utilized for customer identification, with the primary objective being the identification of loyal customers to strengthen customer relationships. The secondary aim is to detect customers at risk of churning to enhance the effectiveness of customer retention programs. Mediana Aryuni et. al. [9] conducted a customer segmentation analysis in the domain of Internet marketing harnessing K-means clustering. The focus was on the cluster (AWC) distance using both K-means and K-medoids. The investigation concluded that one of these methods outperformed the other, resulting in a minimal Davies-Bouldin Index (DBI) value of 2. Cuadros et al. [10] introduced an approach for segmentation based on the computation of customer lifetime value (LTV), which encompasses present value, expected value, and customer loyalty. The findings indicate that segments 1 and 4, characterized by higher customer lifetime value (CLV), exhibit the highest level of loyalty. Furthermore, these segments have a favorable ratio of arrears value to income, resulting in a significant value contribution to the organization. On the contrary, segment 3 exhibits unprofitability as seen by its low turnover rate and income that falls below its outstanding value.

Huang et al.[11] in the work has covered analysis of valid customers based on clusters of the enterprise. The work emphasizes the research and analysis of clustering, which is done in five clusters of a customer of the enterprise, leading to the development of the segmentation model, which directly influences the precision of enterprise segmentation. Dhandayudam et al [12] provided a new improved clustering algorithm for segmentation.The study mainly emphasizes the accuracy of different methods like single-link, k-means, complete link, and the improved algorithm, resulting in the fact that sometimes the new algorithm provides better accuracy than all the other algorithms that were based on the number of clusters. In the research, Matthias Carnein et al. [13] have introduced a clustering algorithm based on novel streams designed for the tracking and identification of customer segments over time. The algorithm involved the utilization of 71 micro clusters to summarize data for a dataset containing 500,000 customers.

Tavakoli et al. [14] introduced an R+FM model in the study, which adapts segmentation based on dynamic business fluctuations and employs K-Means clustering to categorize customers. The algorithm was put to the test at Digikala, the largest e-commerce company in the Middle East. The outcomes of the campaign demonstrated that the Segmentation Model led to enhancements in both the purchase count and the average basket value. In the research, Hamerly and Elkan [15] introduced the G-means algorithm as an innovative method for automatically determining the value of k. The researchers applied the G-means and X-means algorithms to cluster two datasets: the NIST dataset and the Pendigits dataset, aiming to conduct a comparative analysis of the performance. The NIST dataset yielded PQ (Partition Quality) scores of 0.177 and 0.024 for G-means and X-means, respectively. Conversely, the Pendigits dataset exhibited PQ scores of 0.196 for G-means and 0.057 for X-means. Hosseini et al. [16] carried out a separate study to address the limitations of the traditional Recency, Frequency, Monetary (RFM) model, which does not consider temporal data changes. To overcome this limitation and enhance the research quality, the scholars incorporated the temporal dimension into the investigation.

In the study conducted by Dawane et al. [17] employed the RFM model to create four distinct clusters, based on purchasing patterns and behaviors of various retail distributors in a Fast-Moving Consumer Goods (FMCG) company operating across the Indian subcontinent. The researchers utilized collected data over a specific timeframe for the analysis. Additionally, Tsai et al. [18] conducted a comprehensive empirical inquiry at an automotive dealership in Taiwan, with a specific focus on customer satisfaction indicators. The RFM Model served as the framework for the analysis. The researchers applied two well-established partitional clustering methods, namely the algorithm of clustering based on k-means and the Expectation Maximization algorithm, to establish four distinct consumer clusters.

A further investigation by Daqing Chen et al. [19] revolved around the implementation of customer-centric marketing and data mining techniques for small-scale online shops. The research incorporated a real-world case study that made use of SAS Enterprise Guide and SAS Enterprise Miner. The investigators employed the RFM Model as the fundamental framework for the research, utilizing the k-means clustering technique and decision tree induction to establish unique and meaningful customer groupings.. In order to enhance the precision of the segmentation, nested segments have been internally generated within the cluster by the use of a decision tree.

M.E. Tsoy et al. [20] conducted an analytical approach involving ABC-analysis instead of the typical customer base division into quintiles. The study involved an extended RFM analysis applied to the context of high-tech product wholesale and retail, which led to the identification of three distinct consumer target groups. In a separate study, Zne-Jung Lee et al. [21] employed two commonly used techniques, clustering and classification, for machine learning. The study introduced clustering and classification methods based on distributed automatic feature engineering (AFE), which utilized artificial bee colonies (ABC) for selecting valuable input data features. In the proposed algorithm, the optimal number of customer segments was determined as k = 4 through AFE, and k-means, Ward's method, and FCM were subsequently used with k to assign labels to all samples.

In the study conducted by V. Mihova et al. [22] utilized K-means clustering to establish three distinct clusters, or segments, of faithful borrowers, which were labeled as "platinum," "gold," and "silver." In each strategy, a thorough analysis is conducted to compare the outcomes of the studied methodologies, using both initial and standardized segmentation variables, as well as the two-step clustering method gained from a previous study conducted by one of the authors. The optimal type of cluster analysis for each method is determined. In a separate investigation conducted by P. Anitha et al. [23], the emphasis is placed on the use of business intelligence for the purpose of identifying novel customers within the Retail Industry. This is achieved through the provision of pertinent and timely data to corporate organizations. The Silhouette Score is employed for the analysis of the Recency, Frequency, and Monetary of the sales, with the aim of identifying an optimal solution.The researchers achieved an ideal Silhouette Score value of 0.362159752 for a clustering analysis involving three clusters.

In the study, that was conducted by Khajvand et al. [24] employed customer lifetime value (CLV) as a means of conducting customer segmentation inside a health and beauty organization. Two approaches were employed. The first way under consideration is the RFM marketing analysis approach. The second strategy being examined is an enhanced version of the RFM analysis method, which incorporates an extra component known as Count Item.The CLV rank was thereafter assigned to each section, taking into consideration its CLV value.

III. METHODOLOGY

In this section of the study, the concentration is on the methodology employed for conducting Customer Segmentation. To initiate the process, a dataset is acquired from the UCI Machine Learning Repository, named the "Online_Retail" dataset. This dataset comprises key attributes including Quantity, InvoiceDate, UnitPrice, CustomerID, InvoiceNo, StockCode, Description and Country. The detailed explanation of these features is as follows:

- InvoiceNo: A unique identifier for each invoice or transaction.
- StockCode: A code or identifier for the specific item or product being sold.
- Description: A textual description of the item or product, offering additional details about the purchase.
- Quantity: It is the quantity of the item or product bought in a particular transaction.
- InvoiceDate: The date and time when the transaction or invoice was issued.
- UnitPrice: The price of a single unit of the item or product.
- CustomerID: A unique identifier for each customer, facilitating tracking of customer transactions.
- Country: The country in which the transaction occurred or the customer's location.

The dataset is frequently employed for the purpose of retail analytics and the examination of customer behavior. The examination of purchase patterns, RFM (Recency, Frequency, Monetary) analysis, and customer segmentation based on buying behaviors holds significant value in comprehending consumer behavior.

During the earliest stages, the dataset undergoes the necessary pre-processing procedures. The starting dataset comprises unprocessed transaction data. The handling of missing values in the 'CustomerID' column involves the removal of rows that include missing 'CustomerID' values. The presence of negative values in the 'Quantity' and 'UnitPrice' columns is confirmed, and entries containing negative 'Quantity' values are thereafter excluded. In order to augment the dataset for analytical purposes, a novel attribute named 'TotalAmount' is created, which denotes the aggregate purchase value associated with each individual transaction. Finally, the column labeled 'InvoiceDate', which was originally in the format of a string representing a date, is transformed into a datetime format. The purpose of these preprocessing stages is to establish data correctness and consistency, creating a foundation for further analysis, such as clustering and RFM (Recency, Frequency, Monetary) analysis.

Then, feature selection is performed on the dataset in order to retain only those features which are crucial for the analysis. The features which proved to be valuable in this research are IncoiceNo, Quantity, InvoiceDate, UnitPrice, CustomerID. All other features are excluded from the dataframe in order to improve the efficiency of the algorithm.

The calculation of Recency, Frequency, Monetary (RFM) values is performed in the second stage.

Recency: The determination of recency involves calculating the duration in days between a customer's most recent purchase and the 'InvoiceDate'. This is accomplished by subtracting the 'InvoiceDate' from the current date. The calculation is performed via the below formula:

Recency(R) = Latest Date - Last Purchase Date

Frequency: The frequency is ascertained by tallying the distinct invoice numbers ('InvoiceNo') for every client, where each unique invoice represents an individual transaction. The mathematical expression representing the concept of frequency is as follows:

Frequency(F) = Count of Unique Invoice Numbers

Monetary: Monetary value is determined by aggregating the 'TotalAmount' for each client, which represents the overall expenditure.The calculation is performed via the prescribed formula:

Monetary(M) = Sum of Total Purchase Amount

RFM Score: The RFM Score is determined by giving values to consumers based on quartiles. Individuals who fall inside the uppermost 25th percentile are awarded the highest score, with subsequent scores decreasing progressively across the quartiles. A numerical score ranging from 1 to 4 is awarded to each value, with a score of 4 being the highest value. The RFM score is determined by aggregating the aforementioned individual ratings. The formula utilized to calculate the RFM score is as follows:

RFM Score = R Score + F Score + M Score

Greater the value of RFM Score, more will be the cruciality of the consumer for the enterprise. Recognizing the distinct attributes of diverse customer segments empowers businesses to customize their products, services, and communication methods. This results in heightened customer satisfaction, as individuals receive offerings aligned with their preferences. The RFM analysis, highlighted in the provided code, underscores this approach by taking into account recency, frequency, and monetary values as pivotal elements in the segmentation process.

The third phase involves the use of K-Means clustering to partition clients into distinct groups based on the transaction behavior, namely RFM values. The K-Means algorithm is employed with the goal of uncovering inherent customer groupings. The algorithm operates iteratively, assigning each customer to the closest cluster centroid and updating the centroids by computing the mean of the customers within each cluster. This iterative process continues until convergence. The visual representation of this process is illustrated in the following figure.

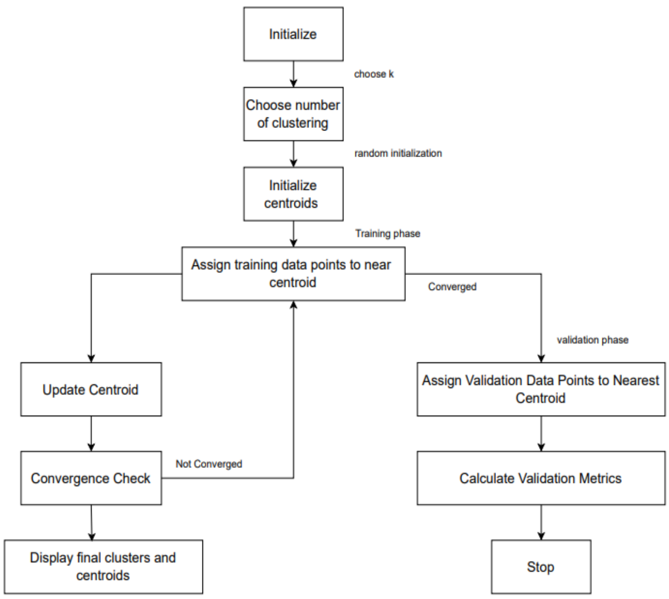

Fig 2: Steps involved in K-means clustering

In this work, four clusters have been created, as determined by the Elbow Method. Creating four clusters has helped in improving the quality of clustering and the efficiency parameters were also improved. These clusters have helped in segmenting the customers into different royalty levels efficiently.

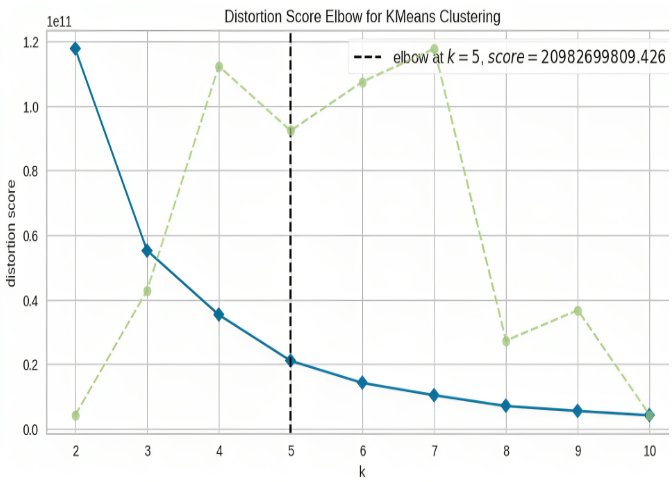

Fig 3: Elbow graph used for determining the K- value

The clusters are assessed based on the centroids, and customer data is grouped accordingly into these segments.

The clustering quality is visualized by a scatter plot, showcasing how well customers are grouped based on Recency and Frequency.

The 'RFM_Loyalty_Level' is then assigned based on this RFM score. For instance, the top 25% may be classified as 'Platinum,' the next 25% as 'Gold,' the 25-50% range as 'Silver,' and the lowest 25% as 'Bronze.' This segmentation proves beneficial in identifying customer loyalty levels, providing a foundation for tailored marketing and retention strategies based on customer engagement and spending patterns.

The following is the graph which shows the clusters formed.

- Platinum is depicted by blue data points
- Gold is depicted by yellow data points
- Silver is depicted by gray data points
- Bronze is depicted by red data points

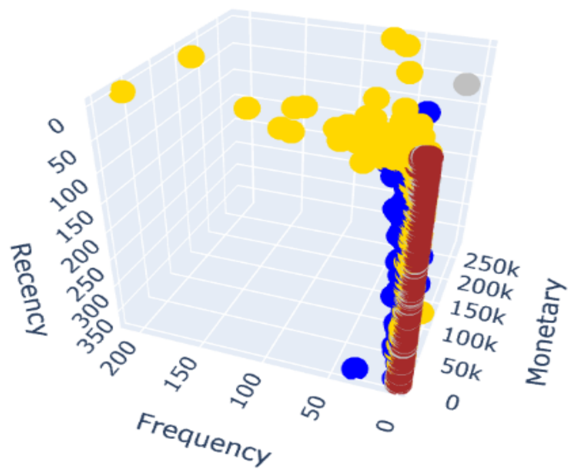

Fig.4: 3D-Scatter Plot for depicting clusters

IV. Results And Discussions

The K-means clustering algorithm employed in the study has exhibited superior performance compared to earlier research in several metrics, including Silhouette Score, davies_bouldin_score, adjusted_rand_score, and fowlkes_mallows_score.

1. **Silhouette Score:** The Silhouette Score is computed for each datapoint by considering the average distance within the same cluster (a) and the average distance to the nearest neighboring cluster (b). The Silhouette Coefficient for a specific sample is then determined.

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)} \quad \text{.......................(1)}$$

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \quad \text{................(2)}$$

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \quad \text{....................(3)}$$

The value of Silhouette Score obtained in this research work is **0.954319641633133**, which is greater than any of the works previously done in this field. The value of Silhouette score closer to 1 indicates a very efficient clustering model.

The previous studies done in this field have a silhouette score of **0.44**. Hence the model proposed in this paper shows a significant increase in its efficiency.

2. **Davies-Bouldin_index:** It is a measure that gauges the average resemblance between each cluster and the cluster that is most similar to it. A lower score is indicative of superior clustering performance. The mathematical expression representing the Davies-Bouldin Score for a given number of clusters, denoted as n, may be expressed as follows:

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i,\ \text{where}$$

$$R_i = \max_{j=1\ldots nc, i\neq j}\left(R_{ij}\right),\ i = 1\ldots n_c \qquad \text{...............(4)}$$

The value of Davies-Bouldin_index obtained in this work is **0.36093416267648226**.

3. **Inertia (Within-Cluster Sum of Squares):** The concept of inertia, quantifies the extent to which the data points inside a cluster deviate from the centroid of that cluster. This metric is sometimes referred to as the within-cluster sum of squares.

$$\sum_{i=1}^{N}(x_i - C_k)^2 \qquad \text{..................................(5)}$$

N is the number of samples.
X is the value of each of those samples.
C is the center of the cluster.

The Inertia Score obtained in the implementation of the clustering model is **35223820551.01195**

4. **Adjusted Rand Index (ARI):** It is a metric used to quantify the degree of resemblance between the real labels and the cluster assignments, while taking into account the possibility of chance. The scale of similarity spans from -1, indicating no resemblance, to 1, indicating perfect similarity.

$$\text{ARI} = \frac{\text{RI}-\text{expectedRI}}{\max \text{RI} - \text{expectedRI}} \qquad \text{..............(6)}$$

In this context, "RI" represents the Rand Index, which assesses the similarity between two sets of clustering results by considering all points that are correctly grouped within the same cluster.

A perfect similarity showing Adjusted Rand Index of **1.0** has been obtained in this work.

5. **Fowlkes-Mallows Index (FMI):** The Fowlkes-Mallows Index (FMI) is a statistic employed to evaluate the geometric mean of accuracy and recall in relation to the correspondence between the true labels and the cluster assignments. The equation representing the Fat Mass Index (FMI) is as follows:

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}} \qquad \text{.........................(7)}$$

The Fowlkes-Mallows Score obtained in the research is 1.0

V. Conclusion And Future Scope

The present work depicts an enhanced K-means clustering model for the purpose of consumer segmentation. The application of K-means clustering was effectively utilized to discern discrete client segments, and the assessment of clustering efficacy was conducted through the utilization of diverse indicators.

The findings from this study indicate that employing K-means clustering is a valuable approach for customer segmentation. This approach offers organizations significant insights that can be utilized to customize the marketing tactics, improve customer retention rates, and enhance the overall efficacy of the marketing efforts. In addition to evaluating the quality of clusters, several metrics such as the silhouette score and Davies-Bouldin Index were utilized, yielding encouraging outcomes.

This research report provides a comprehensive basis for new avenues of investigation and enhancement in the field of consumer segmentation. Researchers have the option to explore sophisticated techniques in feature engineering. In addition to the usual RFM (Recency, Frequency, Monetary) characteristics, it is possible to insert other data such as client demographics, purchase categories, or seasonality considerations. The extension of client segments has the ability to enhance the refinement, equipping organizations with a greater abundance of actionable knowledge.

This study largely centered on the K-means clustering methodology; however, it is important to note that there are other alternative approaches, such as hierarchical clustering, DBSCAN, and spectral clustering. The assessment of these options and the determination of the appropriateness for certain business scenarios may result in enhanced segmentation outcomes. The investigation of personalized marketing methods holds promise for future academic inquiry. The use of machine learning and recommendation systems into segmentation has the potential to enhance the customization of marketing efforts according to the distinct tastes of individual consumer segments, hence potentially augmenting customer engagement and loyalty. This study is expected to provide a foundation for prospective progress in consumer segmentation, facilitating businesses in effectively interacting with the customers through improved accuracy, ethical considerations, and personalized approaches.

References

[1] A. Ullah *et al.*, "Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time," *Sensors (Basel)*, vol. 23, no. 6, p. 3180, Mar. 2023, doi: 10.3390/s23063180.

[2] R. G. Martínez, R. A. Carrasco, J. García-Madariaga, C. P. Gallego, and E. Herrera-Viedma, "A comparison between Fuzzy Linguistic RFM Model and traditional RFM model applied to Campaign Management. Case study of retail business.," *Procedia Computer Science*, vol. 162, pp. 281–289, Jan. 2019, doi: 10.1016/j.procs.2019.11.286.

[3] V. Mihova and V. Pavlov, "A customer segmentation approach in commercial banks," presented at the AIP Conference Proceedings, Oct. 2018, p. 030003. doi: 10.1063/1.5064881.

[4] M. Mohammadzadeh, Z. Z. Hoseini, and H. Derafshi, "A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran," *Procedia Computer Science*, vol. 120, pp. 23–30, Jan. 2017, doi: 10.1016/j.procs.2017.11.206.

[5] I. Bueno, J. Velasco, R. A. Carrasco, and E. Herrera-Viedma, "A geospatial model of RFM analysis: An application to tourism in the Iberian Peninsula," *Procedia Computer Science*, vol. 214, pp. 825–830, Jan. 2022, doi: 10.1016/j.procs.2022.11.247.

[6] B. Cooil, L. Aksoy, and T. Keiningham, "Approaches to Customer Segmentation," *Journal of Relationship Marketing*, vol. 6, pp. 9–39, Jan. 2007, doi: 10.1300/J366v06n03_02.

[7] "Cartesian K-Means | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 06, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/6619232

[8] R. Ait Daoud, A. Amine, B. Bouikhalene, and R. Lbibb, "Combining RFM model and clustering techniques for customer value analysis of a company selling online," in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, Nov. 2015, pp. 1–6. doi: 10.1109/AICCSA.2015.7507238.

[9] J. Bayer, "Customer segmentation in the telecommunications industry," *The Journal of Database Marketing & Customer Strategy Management*, vol. 17, Sep. 2010, doi: 10.1057/dbm.2010.21.

[10] C.-F. Tsai, Y.-H. Hu, and Y.-H. Lu, "Customer segmentation issues and strategies for an automobile dealership with two clustering techniques," *Expert Systems*, vol. 32, Oct. 2013, doi: 10.1111/exsy.12056.

[11] A. J. Cuadros and V. E. Domínguez, "Customer segmentation model based on value generation for marketing strategies formulation," *Estudios Gerenciales*, vol. 30, no. 130, pp. 25–30, Jan. 2014, doi: 10.1016/j.estger.2014.02.005.

[12] J. An, H. Kwak, S. Jung, J. Salminen, and B. J. Jansen, "Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data," *Soc. Netw. Anal. Min.*, vol. 8, no. 1, p. 54, Aug. 2018, doi: 10.1007/s13278-018-0531-0.

[13] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J Database Mark Cust Strategy Manag*, vol. 19, no. 3, pp. 197–208, Sep. 2012, doi: 10.1057/dbm.2012.17.

[14] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," *Procedia Computer Science*, vol. 3, pp. 57–63, Jan. 2011, doi: 10.1016/j.procs.2010.12.011.

[15] G. Hamerly and C. Elkan, "Learning the k in k-means," in *Advances in Neural Information Processing Systems*, MIT Press, 2003. Accessed: Oct. 06, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2003/hash/234833147b97bb6aed53a8f4f1c7a7d8-Abstract.html

[16] M. Hosseini and M. Shabani, "New approach to customer segmentation based on changes in customer value," *Journal of Marketing Analytics*, vol. 3, Sep. 2015, doi: 10.1057/jma.2015.10.

[17] M. Y. Smaili and H. Hachimi, "New RFM-D classification model for improving customer analysis and response prediction," *Ain Shams Engineering Journal*, p. 102254, Apr. 2023, doi: 10.1016/j.asej.2023.102254.

[18] S. Savaresi and D. Boley, "On the performance of bisecting K-means and PDDP," *Proc. of the 1st SIAM Int. Conf. on Data Mining*, Apr. 2001, doi: 10.1137/1.9781611972719.5.

[19] J. Wu and Z. Lin, "Research on customer segmentation model by clustering," presented at the ACM International Conference Proceeding Series, Jan. 2005, pp. 316–318. doi: 10.1145/1089551.1089610.

[20] T. Kodinariya and P. Makwana, "Review on Determining of Cluster in K-means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 90–95, Jan. 2013.

[21] S. Chavhan, R. C. Dharmik, S. Jain, and K. Kamble, "RFM analysis for customer segmentation using machine learning: a survey of a decade of research," *3C TIC*, vol. 11, no. 2, pp. 166–173, Dec. 2022, doi: 10.17993/3ctic.2022.112.166-173.

[22] V. Dawane, P. Waghodekar, and J. Pagare, "RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention." Rochester, NY, May 25, 2021. doi: 10.2139/ssrn.3852887.

[23] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1785–1792, May 2022, doi: 10.1016/j.jksuci.2019.12.011.

[24] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 10, pp. 1251–1257, Dec. 2021, doi: 10.1016/j.jksuci.2018.09.004.

[25] M. Tsoi and V. Shchekoldin, "RFM-analysis as a tool for segmentation of high-tech products' consumers," Oct. 2016, pp. 290–293. doi: 10.1109/APEIE.2016.7807074.