Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 260 (2025) 616–625

Procedia Computer Science

www.elsevier.com/locate/procedia

Seventh International Conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R-2024)

A Framework for Customer Segmentation to Improve Marketing Strategies Using Machine Learning

[a]Aya Ashraf *, [b] Christina Albert Rayed, [a] Nancy Awadallah Awad, [b] Heba M.Sabry

[a] *Faculty of Management Sciences, Sadat Academy, Cairo, Egypt*
[b] *Faculty of Computers and Information, Sadat Academy, Cairo, Egypt*

Abstract

It is hard for the marketing team to set a strategy without dividing the customers into groups. Clustering is a well-known machine-learning technique that can be used to implement customer segmentation. It is an unsupervised learning method that creates clusters by dividing a dataset into many valuable subclasses. In online retail datasets, algorithms such as K-means, Mini Batch K-means, Spectral Clustering, and Fuzzy K-means are employed to categorize customers according to their Recency, Frequency, and Monetary (RFM) features. After analyzing the Silhouette Score, the K-means achieved a higher score, 0.432619, which implies that this algorithm achieved comparable cluster cohesion and separation levels. This paper aims to develop a framework for customer segmentation using machine learning to improve marketing strategies.

Keywords: Customer Segmentation; RFM (Recency, Frequency, Monetary); Machine Learning; Clustering

1. Introduction

Machine learning is a domain of research dedicated to extracting knowledge from data. It lies at the intersection of statistics, artificial intelligence, and computer science. Unsupervised learning and supervised learning are the two modes of learning that fall under the umbrella of machine learning. Unsupervised learning is models that are trained

* Corresponding author.
E-mail address: aya.ashraf@sadatacademy.edu.eg

1877-0509
10.1016/j.procs.2025.03.240

on unlabelled data, enabling them to automatically extract and learn features and patterns from the data. One of the categories within unsupervised learning is clustering. Clustering algorithms partition the data into distinct groups based on their similarities. These algorithms group similar items, allowing for the identification of meaningful clusters within the data. One practical application of clustering is customer segmentation [1] [2].

Customer Segmentation involves grouping customers based on shared traits or behaviors instead of using a uniform approach for everyone. It is achieved by analyzing geographic conditions, economic factors, demographics, and behavioral patterns. This process relies on data encompassing geographic conditions, economic factors, demographic characteristics, and behavioral tendencies [3]. Customer segmentation is essential in marketing as it enables the customization of market plans for each customer segment, supports business decision-making, identifies products associated with each segment, manages demand and supply, identifies potential client bases, and focuses on those bases [4]. Categorizing customers according to shared characteristics or behaviors allows businesses to comprehend their customers better and customize offerings to match specific requirements, resulting in increased revenue. Customer segmentation also brings advantages such as personalized, targeted marketing strategies, improved product placement, creation of new products based on segment preferences, and overall higher revenue generation [5].

Recency, Frequency, and Monetary (RFM) are variables that segment customers and analyze their purchasing behavior. Recency denotes the duration since the customer's latest order. It is measured in terms of the number of days. Frequency answers how many orders or purchases the customer made within a given timeframe. It represents the count of all orders or purchases made by the customer. Monetary value reflects the money the customer spent during a specific period. It is calculated by summing all the order amounts within the given timeframe. Higher values for Frequency and Monetary are preferable, while lower values for Recency are considered better [6] [7].

This paper focuses on customer segmentation based on their purchase behavior attributes that are RFM by using machine learning using the "Online Retail" dataset through various clustering algorithms. It also makes a comparative analysis of these algorithms using Silhouette Score to choose the most appropriate. This paper aims to provide marketing teams with insightful data that can guide their decision-making processes and aid them in creating more effective and targeted marketing strategies. By understanding customer behavior and categorizing them into distinct segments, the approach can be tailored to reach better and connect with the target audience. This paper is structured into five sections: introduction, literature review, methodology, applying customer segmentation framework, and future work and conclusion.

2. Literature Review

In previous research on customer segmentation using machine learning algorithms, it was found that the K-means clustering algorithm and the Elbow Method are the most popular choices for determining the optimal number of clusters. This algorithm has been utilized by researchers in order to divide customers into groups according to a variety of segmentation variables. Several studies utilized the K-means algorithm to segment customers based on spending score and annual income like [3] [8] [9]. Additionally, [5] expanded on this approach by incorporating the K-means and Agglomerative algorithms. Other studies like [7] [10] [11] applied the K-means algorithm to segment customers based on RFM variables.

Despite the extensive use of the K-means algorithm, these studies face several limitations that can obstruct their effectiveness and impact. The most common issues include using small datasets, leading to less reliable results as smaller samples may not accurately represent the larger population. Additionally, many studies depend on a single algorithm, typically K-means, without exploring the potential benefits of other clustering algorithms that may offer better insights. Algorithms such as DBSCAN, hierarchical clustering, Mini Batch K-means, Spectral Clustering, or Fuzzy K-means might provide alternative perspectives that enhance the understanding of the data. Furthermore, there is often a lack of actionable recommendations for each identified customer cluster. Without clear, practical strategies tailored to each segment, the findings remain challenging to implement in real-world business scenarios.

3. Methodology

The methodology outlined in this paper provides a detailed framework for customer segmentation using unsupervised machine-learning algorithms. It also evaluates several different machine-learning algorithms in order to choose the one that is most appropriate.

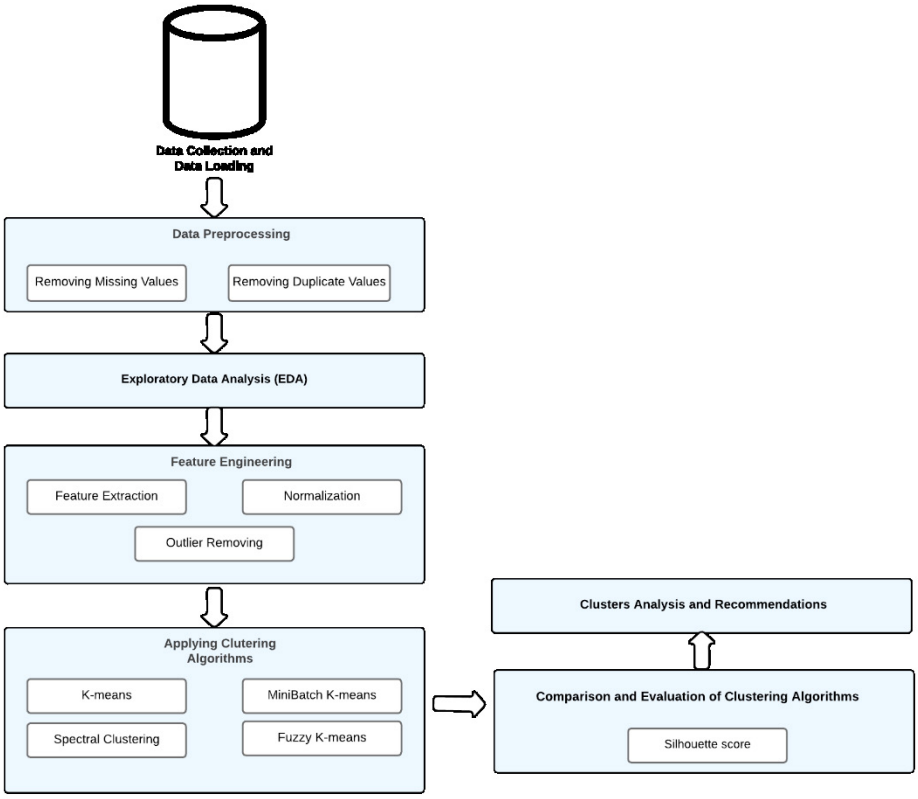

Fig. 1. Customer Segmentation Framework

As shown in **Figure 1**, the framework stages include data collection and data loading, preprocessing, Exploratory Data Analysis (EDA), feature engineering, application of clustering algorithms, comparison and evaluation of clustering algorithms, and cluster analysis.

3.1. Data Collection and Data Loading

Data is the facts expressed as facts that are analyzed for various computations and are given to answer research questions [13]. Data can be categorized at the highest level as categorical or numerical— the questions "What type" and "Which category" are answered with categorical data. There are two types of categorical data: ordinal and nominal. Data lacking a natural order and a sequence between the different categories is called nominal data. On the other hand, the data that can be categorized in ascending and descending order known as data ordinal. Numerical data is information that may be calculated to provide an answer to the "how many" or "how much" questions. Discrete and continuous data are categories for numerical data. Data that can only be calculated in whole numbers is called discrete data. Decimals cannot be used to represent this kind of data. Data that can be counted in decimals and has no theoretical gaps between data points is called continuous data [14].

The procedure of acquiring information to gain a deeper understanding of the research topic is referred to as data collection. In the initial research phase, data collection can enhance outcome quality by minimizing the probability of errors during the project. Data collection methods are categorized into two main types: Primary Data Collection Methods and Secondary Data Collection Methods. Primary data denotes information that is unpublished and acquired directly, remaining unaltered by any person. Primary data can be obtained through diverse methods, such as interviews, questionnaires, and mechanical instruments. On the other hand, secondary data refers to information acquired from published sources, indicating that it has been previously collected by another party for a distinct purpose and may be utilized for alternative research initiatives. Secondary data sources comprise books and scholarly articles [13] [15].

The process of reading and making data accessible is known as data loading. It can be divided into a number of primary categories, such as reading text files and more effective disk formats, loading data from databases, and communicating with network sources like web APIs [16].

3.2. Data Preprocessing

Data preprocessing is acrucial stage in data analysis, involving transforming raw data into a valuable dataset. Preprocessing encompasses several essential steps, including handling missing, categorical, and duplicate values. These steps are vital for cleaning and transforming the datasets and preparing them for effective data analysis. Through this process, the datasets are refined and modified, ultimately facilitating the creation of a valuable data analysis [17].

3.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a method that involves examining and understanding data through techniques such as visualization and summary statistics. In the initial stages of data exploration, obtaining an analytical summary of the data is crucial. Creating meaningful visual representations is essential as it helps determine necessary data transformations and generate model ideas. While data analysis is often associated with tasks like "number crunching," mathematical formulas, and algorithms, the significance of graphical analysis for visualization is frequently overlooked [18].

3.4. Feature Engineering

Feature engineering involves extracting relevant characteristics from raw data and converting them into suitable formats that a machine-learning model can effectively utilize. It plays a vital role in the overall machine learning pipeline as it can significantly simplify the modeling process by providing appropriate features and enhancing the quality of output results. Various phases involve feature engineering, scaling, and outlier removal.

Outlier refers to an observation that deviates significantly from the other values within the dataset. The outliers should be removed to avoid model inefficiency. The input scale can affect models that are smooth functions of the input. Feature scaling aims to alter the scale of a feature. Min-Max Scaling and Standardization are two typical methods of feature scaling. In Min-Max Scaling, each feature value (i.e., the value of a feature in a specific data point) is adjusted within the range of [0, 1]. This is achieved by compressing or stretching all feature values based on the minimum and maximum values of the feature across the entire dataset. On the other hand, Standardization involves subtracting the mean of the feature across all data points and dividing it by the variance. This process results in a scaled feature with a mean of 0 and a variance of 1 [19].

3.5. Application of Clustering Algorithms

Clustering is a fundamental concept in unsupervised learning that involves identifying patterns or structures within unclassified data. Clustering is considered the most important problem in unsupervised learning as it aims to organize objects into groups based on their similarities. These algorithms can be adjusted to figure out the number of clusters to identify, providing flexibility in defining the granularity of these groups. A cluster refers to a set of objects that are similar to each other but dissimilar to objects in other clusters. The output of clustering can be visually represented,

and dimensionality reduction techniques such as the Principal Component Analysis (PCA) algorithm can be utilized to reduce the number of dimensions. The formation of groups in clustering is based on the similarities among the data points.

There are various clustering methods, including partitioning, hierarchy, and overlapping. Partitioning clustering ensures that each data point belongs to only one cluster, also known as exclusive clustering. Examples of partitioning clustering are the K-means and Mini Batch K-means algorithms. In hierarchical clustering, each data point is initially assigned to a separate cluster. Subsequently, the closest clusters are merged in an iterative manner, which ultimately results in a reduction in the total number of clusters. To assign data points to multiple clusters with varying degrees of membership, overlapping clustering makes use of fuzzy sets. An example of an overlapping clustering algorithm is the fuzzy k-means neural network [1] [2].

3.6. Comparison and evaluation of clustering algorithms

Due to the absence of labelled information in unsupervised learning algorithms, determining the efficacy of the learned output poses a significant challenge. Evaluating the usefulness of such algorithms is difficult since there is no definitive correct output against which to compare. When it comes to evaluating the effectiveness of clustering, many distinct metrics of evaluation are utilized, such as the Cohesion, Separation, Silhouette Coefficient, SSE (Sum of Squared Errors), and Dunn Index. Cohesion: Measures the compactness within clusters and is an intra-cluster metric. Separation: Quantifies the dissimilarity between clusters and is an inter-cluster metric. Silhouette Coefficient: Combines cohesion and separation metrics into a single measure, commonly used for assessing clustering quality. SSE (Sum of Squared Errors): Represents the sum of squared errors within clusters. Dunn Index: Evaluates the ratio of the smallest distance between data points from different clusters to the largest distance between clusters. These evaluation metrics offer various insights into the efficacy of clustering and dimensionality reduction models [1] [2].

3.7. Clusters Analysis and Recommendations

In the final stage of the methodology, cluster analysis is conducted to identify groups within the dataset based on similarities or patterns. This analysis helps uncover natural groupings, providing valuable insights into customer segmentation. Once the clusters are identified, recommendations are developed for each cluster's specific behaviors. These recommendations provide the distinct needs and preferences of each segment. For example, clusters with low engagement may benefit from reactivation campaigns, while high-value clusters might require personalized attention and exclusive offers to maintain their loyalty. Moderate engagers could be targeted with initiatives designed to boost their spending and engagement levels, aiming to convert them into high-value customers.

4. Applying Customer Segmentation Framework

The Applying the customer segmentation framework section presents the outcomes of the customer segmentation framework using machine learning. The results focus on the framework's effectiveness in identifying customer segments based on RFM and comparing multiple algorithms. They also show how marketing performance can be achieved through recommendations for each segment. The results confirm machine learning-based customer segmentation potential as a valuable tool for marketers aiming to optimize strategy and boost customer satisfaction.

The initial phase of the framework is data collection, which is an essential stage. This paper employs the "Online Retail" dataset obtained from the University of California, Irvine (UCI), a public research univrsity located in Irvine, California, United States. This Online Retail dataset is comprehensive and includes all transactions executed by a UK-based retailer from December 1, 2009, to December 9, 2011. The company primarily offers distinctive all-occasion giftware, setting its products apart in the marketplace. The dataset indicates that a substantial segment of the company's customer comprises wholesalers, signifying that it operates as a business-to-business (B2B) entity. This aspect enhances the analysis by offering insights into bulk purchasing behaviors and wholesale market dynamics.

Table 1. Features Description.

Column Name	Description	Type
InvoiceNo	Every transaction is allocated a distinctive 6-digit number, with a 'c' at the beginning indicating a cancellation.	Categorical
StockCode	Each individual product is assigned a unique 5-digit number.	Categorical
Description	The product's name.	Text
Quantity	The amount of each product in every transaction.	Numeric
InvoiceDate	The date and time at which a transaction is initiated.	Date & Time
UnitPrice	The product's unit price, in British pounds (£).	Numeric
CustomerID	A unique 5-digit number is designated to each customer.	Numeric
Country	The country in which the customer resides.	Categorical

The dataset contains 1,044,848 records and eight features: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. These features comprise categorical, textual, and numerical data. Detailed information about these features is presented in **Table 1**.

Data preprocessing is the second step. The dataset experiences multiple cleaning phases, encompassing the handling of absent values and duplicates. This cleaning process guarantees data accuracy, which is essential for clustering algorithms. Missing values are eliminated, inconsistent data is eliminated, and duplicate entries are deleted.

The third step involves conducting Exploratory Data Analysis (EDA). During the analysis, the dataset had 5,863 unique customers and 4,626 unique stocks involved in the transactions. The overall quantity sold from 1/12/2009 to 9/12/2011 was 10,514,002 units. Also, the total sales generated reached £17,086,246.198.

Additional exploration focused on analyzing individual years within this time range. In 2011, transactions occurred between January 4 and December 9. The quantity sold during this year was 4,845,163 units, resulting in sales totaling £8,178,314.084. Similarly, in 2010, transactions from January 4 to December 23, with the quantity sold reaching 5,270,295 units, generated sales of £8,229,651.044. This analysis offers significant insights into the scale of the transactions, emphasizing the growth and financial success attained over the reviewed years.

After analyzing the top 10 purchased products in 2010 and 2011, the "White Hanging Heart" product stood out as one of the most popular items across both years; this finding underlines a consistent consumer preference for this product over the two years, highlighting its enduring appeal and market relevance.

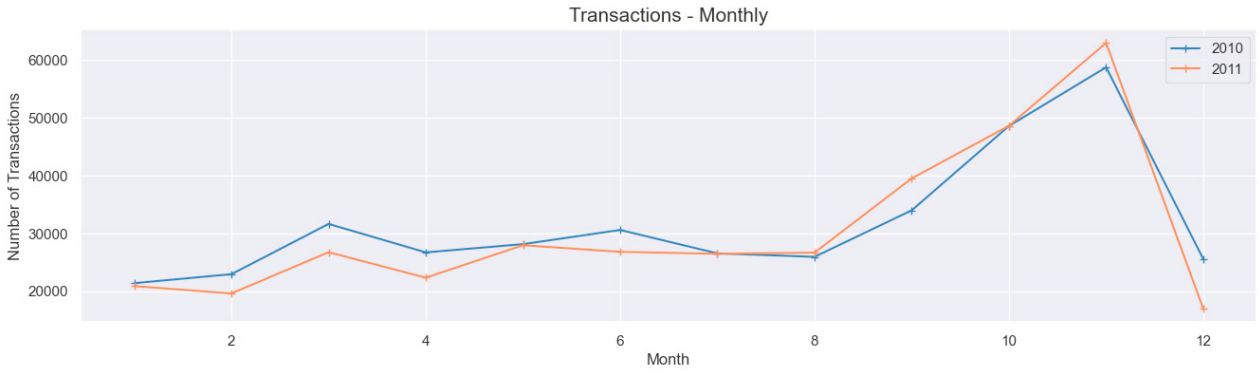

Fig. 2. Monthly Transactions

After totaling the monthly transactions for 2010 and 2011, November was the month with the highest number of transactions. This significant finding is clearly illustrated in **Figure 2**, which provides a detailed visual representation of the transaction data across the different months, highlighting November's peak activity. Also, the United Kingdom accounted for 89% of the total transactions in 2010 and 88% in 2011.

In the feature engineering step, the aggregation technique was employed to extract RFM features, as shown in **Table 2**. The RFM features were subjected to outlier detection and removal, reducing the dataset to 4,875 remaining customers. To standardize the data and facilitate further analysis, the MinMaxScaler scaling technique was applied.

Table 2. RFM Features.

	CustomerID	Recency	Frequency	Monetary
0	12346	0.439620	0.222222	0.188168
1	12347	0.001357	0.777778	0.612820
..	..	..	..	..
5862	18287	0.055631	0.666667	0.382648

4.1. Applying Machine Learning Algorithms

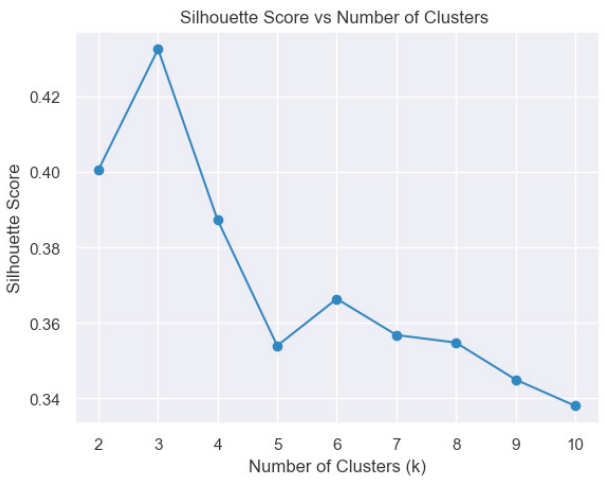

Fig. 3. Optimal Number of Cluster in K-means

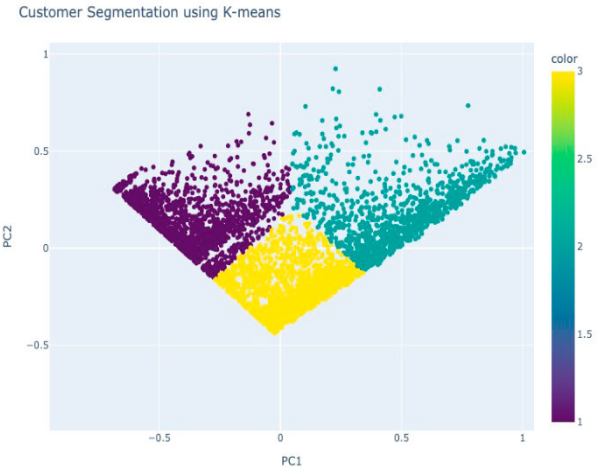

Fig. 4. K-means Clustering

The experiment's objectives are to segment customers by finding the optimal number of clusters and comparing the quality of clusters of the machine learning algorithms using the Silhouette score. This evaluation helps decide which machine learning algorithm best suits selected data. In all experimental results, the dataset is analyzed using different clustering algorithms.

K-means clustering was initially applied to divide customers categorized according to RFM features, and the Silhouette Score analysis demonstrated that the optimal quantity of clusters was 3, as it achieved the highest score, as shown in **Figure 3**. This result indicates that partitioning customers into three distinct groups has the most meaningful and well-separated clusters. The cluster assignments obtained from k-means were then visualized using a scatter plot. Principal Component Analysis (PCA) was used to reduce the dimensionality of the RFM data from three to two dimensions for clear representation, as shown in **Figure 4**. Additionally, the Mini Batch k-means, Spectral Clustering, and Fuzzy k-means algorithms were also used to segment customers based on RFM features, and the Silhouette Score analysis indicated that the optimal number of clusters is 3 for all these algorithms, similar to the findings with the K-means algorithm.

4.2. Comparing and Evaluating Clustering Algorithms

The measured parameters, Silhouette score, are employed to evaluate and compare the performance of different clustering algorithms. This matrix assesses the quality of the clusters generated by the algorithms. By examining these

cluster quality measures, the optimal amount of clusters for each algorithm can be determined, allowing for the identification of the most suitable option for the chosen data type. This analysis offers significant insights into the efficacy and suitability of each algorithm for customer segmentation.

Table 3. Clustering Algorithms Comparison.

	K-means	Mini Batch K-means	Spectral Clustering	Fuzzy K-means
Silhouette Score	0.432619	0.430609	0.427383	0.432366

The clustering evaluation results provide valuable insights into the performance of different clustering algorithms across various metrics. Upon analyzing the Silhouette Score, the K-means achieve a higher score, as shown in **Table 3**. It implies that this algorithm achieved comparable levels of cluster cohesion and separation.

4.3. Clusters Analysis and Recommendations

The K-means algorithm achieved a high score, which was selected as the basis for implementing the clustering analysis.

Table 4. Clusters Analysis.

Cluster	Number of Customers	Monetary	Frequency	Recency
1	1696	0.109468	0.081040	0.657136
2	1155	0.461346	619529	0.126503
3	2024	0.134076	0.149978	0.129056

In **Table 4**, The customers are categorized into three clusters based on their number, monetary value, frequency of engagement, and Recency of interactions. Cluster 1 consists of 1,696 customers who exhibit a low economic value (0.109468), infrequent engagement (0.081040), and high Recency of interaction (0.657136), indicating they are relatively inactive. Cluster 2 consists of 1,155 customers with a high monetary value (0.461346), frequent engagement (0.619529), and low Recency (0.126503), representing the most valuable and engaged segment. Cluster 3 includes 2,024 customers who have moderate monetary value (0.134076), moderate engagement frequency (0.149978), and low Recency (0.129056), indicating a balanced but moderately active segment. This classification allows targeted strategies to enhance customer value and engagement tailored to each cluster's specific characteristics.

Table 5. Description and Recommendation for each Cluster.

	Segment's Name	Description	Recommendations
Cluster 1	Inactive Customers	This cluster represents customers who have low spending levels, low engagement frequency, and haven't interacted with the store recently.	It is recommended to implement reactivation strategies to re-engage these customers and stimulate their participation.
Cluster 2	High-Value Customers	This cluster represents customers who have high spending levels, high engagement frequency, and have recently interacted with the store.	They are considered the most valuable segment and require personalized attention to maintain their loyalty and satisfaction.
Cluster 3	Moderate Engagers Customers	This cluster represents customers with moderate spending levels, moderate engagement frequency, and moderate recency of interactions.	As this segment represents a high portion of customers, targeted efforts should be made to increase their engagement and spending, with the aim of converting them into high-value customer segments.

Table 5 shows that customers are categorized into three segments: Inactive Customers, High-Value Customers, and Moderate Engagers Customers. Inactive Customers exhibit low spending, low engagement, and recent inactivity and need to be reactivated to re-engage them. High-value customers who display high spending and frequent

engagement require personalized attention to maintain their loyalty and satisfaction due to their significant value to the business. Moderate Engagers Customers have moderate spending, engagement, and interaction recency. Since they represent a substantial portion of the customers, targeted efforts should focus on increasing their engagement and spending to convert them into high-value customers.

Table 6. RFM Table with Cluster and Category Columns.

	CustomerID	Recency	Frequency	Monetary	Cluster	Cluster's Name
0	12346	0.439620	0.222222	0.188168	1	Inactive Customer
1	12347	0.001357	0.777778	0.612820	2	High-Value Customer
..	..	..	..	..	..	
5862	18287	0.055631	0.666667	0.382648	2	High-Value Customer

After analyzing the three clusters, two new columns were added to the RFM table: "Cluster" and "Cluster's Name" as shown in **Table 6**. The "Cluster" column assigns each customer a number based on their recency, frequency, and monetary values, while the "Cluster's Name" column provides a meaningful classification, such as "Inactive Customers," "High-Value Customers," or "Moderate Engagers Customers." These labels make understanding each customer's behavior easier and allow for a more personalized approach to segmenting and targeting customers based on their unique activity levels and spending habits.

5. Conclusion and Future Work

In conclusion, a machine-learning framework for customer segmentation to improve marketing strategies is presented in this paper. The results showed its effectiveness in identifying customer segments based on RFM attributes and generating valuable insights. Through data collection, preprocessing, analysis, feature engineering, and clustering, the paper segmented customers into three clusters based on the k-means algorithm. Using the Silhouette score determines the optimal number of clusters and the best algorithm. The analysis identified distinct segments with specific strategy recommendations, including inactive customers, high-value customers, and moderate engagers. Implementing these strategies can improve engagement, loyalty, and marketing performance, offering a valuable approach to leveraging machine learning for effective marketing and customer satisfaction.

In future work, a model to predict customer monetary value for the upcoming year using machine learning algorithms to capture subtle differences in behavior and spending patterns will developed. This model will inform targeted marketing strategies, focusing on valuable customer segments and tailoring engagement efforts to maximize customer lifetime value. Additionally, it will help identify and nurture potential high-value customers, ensuring efficient marketing resource allocation. This data-driven approach will enhance marketing effectiveness, customer relationship management, and business outcomes by targeting customers with the highest projected ROI.

References

[1] S. Dridi, "Unsupervised Learning - A Systematic Literature Review," 2021.

[2] A. Müller and S. Guido, Introduction to Machine Learning with Python, US: O'Reilly Media, 2017.

[3] Y. Kushwaha and D. Prajapati, "Customer Segmentation using K-Means Algorithm," International Journal Of Creative Research Thoughts (IJCRT), 2020.

[4] T. Mani, K. P. Kumar and S. G. Teja, "Customer Segmentation in Shopping Mall Using Clustering In Machine Learning," International Research Journal of Engineering and Technology (IRJET), vol. 9, no. 3, 2022.

[5] P. Warale and H. Lone, "Cluster Analysis: Application of K-Means and Agglomerative Clustering for Customer

Segmentation," Journal of Positive School Psychology (JPSP), vol. 6, no. 5, 2022.

[6] J. M. John, O. Shobayo and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," Multidisciplinary Digital Publishing Institute (MDPI), 2023.

[7] V. Dawane, P. Waghodekar and J. Pagare, "RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention," International Conference on Smart Data Intelligence (ICSMDI), 2021.

[8] M. Seshashayee and S. Dileep, "Customer Segmentation Using Machine Learning," International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), vol. 4, no. 5, 2022.

[9] M. Sathyanarayana, S. Dhanish, P. S. Kumar and A. N. Reddy, "Mall Customer Segmentation Using Clustering Algorithm," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 11, no. 1, 2023.

[10] D. Devarapalli, A. S. Virajitha, G. S. Veera, A. S. Sravya, B. T. Keerthi and A. P. Devi, "Analysis of RFM Customer Segmentation Using Clustering Algorithms," International Journal of Mechanical Engineering, vol. 7, no. 1, 2022.

[11] R. Shirole, L. Salokhe and S. Jadhav, "Customer Segmentation using RFM Model and K-Means Clustering," International Journal of Scientific Research in Science and Technology (IJSRST), vol. 8, no. 3, 2021.

[12] A. Joshi and H. Tiwari, "An Overview of Python Libraries for Data Science," Journal of Engineering Technology and Applied Physics (JETAP), vol. 5, no. 2, 2023.

[13] H. Taherdoost, "Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects," International Journal of Academic Research in Management (IJARM), vol. 10, no. 1, 2021.

[14] P. Ranganathan and G. Nithya, "An Introduction to Statistics – Data Types, Distributions and Summarizing Data," Indian Journal of Critical Care Medicine (IJCCM), 2019.

[15] S. A. Mazhar, R. Anjum and A. I. Anwa, "Methods of Data Collection: A Fundamental Tool of Research," Journal of Integrated Community Health (JICH), vol. 10, no. 1, 2021.

[16] W. McKinney, Python for Data Analysis, Sebastopol: O'Reilly Media, 2022.

[17] N. Pandey, P. K. Patnaik and S. Gupta, "Data Pre-Processing for Machine Learning Models using Python Libraries," International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 4, 2020.

[18] R. Snee, "Using Exploratory Data Analysis," 2020.

[19] A. Zheng and A. Casari, Feature Engineering for Machine Learning, Sebastopol: O'Reilly Media, 2018.

[20] S. Dridi, "Unsupervised Learning - A Systematic Literature Review," 2021.

[21] T. Mani, K. P. Kumar and S. G. Teja, "Customer Segmentation in Shopping Mall Using Clustering In Machine Learning," International Research Journal of Engineering and Technology (IRJET), vol. 9, no. 3, 2022.

[22] M. Sathyanarayana, S. Dhanish, P. S. Kumar and A. N. Reddy, "Mall Customer Segmentation Using Clustering Algorithm," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 11, no. 1, 2023.