Application of Unsupervised Learning in Detecting Behavioral Patterns in E-commerce Customers

Divya Udayan J[1], N Moneesh[2], Nehith Sai Vemulapalli[2], Paladugula Pruthvi[2],
Rakshith Sakhamuri [3]
divyaudayanj@am.amrita.edu[1], amenu4aie20150@am.students.amrita.edu[2]
amenu4aie20053@am.students.amrita.edu[2], amenu4aie20054@am.students.amrita.edu[2] 201131@iiitt.ac.in[3]
[1] [2] [3]Department of Computer Science and Engineering, Amrita School of Computing,
Indian Institute of Information Technology, Tiruchirappalli, India

***Abstract*—In the rapidly growing e-commerce world, effectively discerning customer behavior is indispensable for fine-tuning services and bolstering sales. Our research centers around the Online Retail dataset, utilizing unsupervised machine learning to unveil distinct behavioral patterns. Initial data examination was pivotal, ensuring anomalies were addressed, paving the way for reliable results. The novelty of our approach lies in leveraging the Recency, Frequency, and Monetary (RFM) methodology, revealing multi-faceted behavior of customer interactions. This method demystified customer activity into recent engagements, purchasing frequency, and spending magnitude. Utilizing silhouette scores, an optimal clustering number was identified, followed by the application of the KMeans algorithm, which correctly segements customers into discernible behavioral groups. The visualization of these clusters uncovers clear purchase patterns, providing businesses with a lens to refine their marketing endeavors. Conclusively, this method offers businesses an edge, extracting deeper insights into customer behavior and steering optimal growth trajectories.**

***Index Terms*—e-commerce, unsupervised machine learning, RFM methodology, silhouette scores, KMeans clustering.**

I. Introduction

In today's digital era, e-commerce is more than just buying and selling products online. It is about understanding customers and their preferences, predicting future trends, and offering personalized experiences. With the surge of online shopping platforms, every click, purchase, or cart abandonment tells a story about the customer's behavior and preferences. These countless interactions, when analyzed rightly, can offer invaluable insights, helping businesses serve their customers better.

Traditionally, businesses relied on surveys, feedback forms, and direct interactions to understand their customer base. While these methods are valuable, they are often limited in their reach and might not always provide the full picture. Imagine having to ask every customer why they chose to buy a product or why they abandoned their cart. Not only is this method time-consuming, but the answers might also be influenced by biases. Some customers might not remember their reasons, while others might not be willing to share them. Hence unsupervised machine learning, specifically clustering techniques like KMeans[1], gains importance here. One way to categorize such behaviours is by using the KMeans clustering algorithm. Imagine a busy marketplace where each shopper has habits based on their recency, frequency, and monetary spending. KMeans works like an invisible guide, gently grouping together shoppers who exhibit similar shopping patterns. Essentially, it divides the vast sea of diverse shoppers into distinct, manageable groups or 'clusters'. For our study on online retail data, we have employed KMeans to identify these shopper clusters. By doing so, we aim to offer businesses a clearer picture of their clientele, enabling them to tailor strategies for each cluster and enhance customer satisfaction.

Clustering[2], in the context of e-commerce, is like putting customers into different virtual rooms based on how they behave on your platform. So, customers who often buy electronic products might be in one room, while those who frequently purchase books might be in another. By doing this, businesses can tailor their marketing strategies for each 'room' or segment, ensuring a higher likelihood of success.

One of the well-established methods to achieve this segmentation is through the Recency, Frequency, Monetary (RFM) analysis[3]. Let's explore it further: Recency: When was the last time a customer made a purchase? If a customer bought something yesterday, they are more likely to be active and engaged than someone who made a purchase a year ago. Frequency: How often are they buying? A customer who shops five times a month might be a regular shopper, whereas someone who shops once in six months might be an occasional buyer. Monetary: How much are they spending? This helps in identifying high-value customers versus those who spend sparingly. By combining these three metrics, businesses can get a 360-degree view of their customer's shopping behavior. While RFM provides these valuable metrics, deciding how many such 'rooms' or clusters to have isn't straightforward approach. This is akin to a shop owner trying to decide how many sections to have in their shop. Too many, and it becomes chaotic; too few, and it might not serve the purpose.

In the world of online shopping, with vast and diverse customer data, it's essential to group or cluster customers efficiently. To achieve this, we used the KMeans algorithm. However, a pressing question arises: how many clusters or groups should we create? That's where the silhouette score shines. Imagine each customer as a puzzle piece. The silhouette score helps us figure out how well each piece fits within its group compared to neighboring groups. A higher score indicates a better fit, meaning we have grouped the customers more accurately. For our online retail project, the silhouette score guided us in deciding the optimal number of customer clusters, ensuring we didn't merge distinct shopping behaviors or separate similar ones. It's our tool for ensuring our clusters genuinely reflect customer patterns. Recent research has shown that the silhouette score[4] can help in this decision-making process. Think of the silhouette score as a measuring tape. It helps determine how well each customer fits into their 'room' compared to others. A higher score means that the clusters are well-separated and each customer fits well within their assigned cluster.

In our study, we're combining the time-tested RFM analysis with the power of the silhouette score to segment customers of an online retail platform. Our primary dataset comprises diverse online transactions, providing a rich tapestry of customer behaviors. Through this study, our objective is twofold. First, we want to offer businesses a clear lens to view their varied customer base. Second, by understanding these segments, businesses can design strategies that resonate with each group, thereby enhancing their overall shopping experience.

In conclusion, our research stands at a crucial juncture of traditional customer understanding methods and advanced machine learning techniques. E-commerce has transformed the way we shop, and with the integration of data-driven insights, we can make the system even more personalized and customer-centric. The age-old adage, "Customer is King," remains true, and through our research, we aim to provide businesses the tools to treat them like royalty.

II. literature Survey

A K-means clustering algorithm approach[5] by Ching-Hsue Cheng, You-Shyang Chen. This studies the application of the K-means clustering algorithm and its ability to adapt based on changing data. The authors emphasize the flexibility of the method, making it ideal for a variety of datasets. The paper explores the optimization of the algorithm to handle noisy data and its competency in generating clear cluster boundaries, which is particularly useful for intricate tasks such as customer segmentation in the online retail environment.

Recency, Frequency, and Monetary Analysis[6] by David L. Olson, Georg Lauhoff. Focuses on the RFM (Recency, Frequency, Monetary) model, this paper presents an in-depth examination of customer behavior analysis. It emphasizes the significance of understanding customer purchasing habits to optimize marketing strategies. The paper strongly suggests the integration of machine learning techniques with RFM analysis, showcasing improved accuracy in predicting future customer behavior.

Approaches to Clustering in Customer Segmentation[7] by Shreya Tripathi, Aditya Bhardwaj, Poovammal Eswaran Customer segmentation is paramount in business strategies. This paper reviews various clustering methodologies for customer segmentation, arguing that a deeper understanding of different algorithms can lead to more effective segmentations. While K-means is a popular choice, the authors propose that there might be other equally competent, if not better, clustering techniques tailored for specific types of data or business needs.

Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business[8] by Deepali Kamthania, Ashish Pahwa, Srijit S. Madhavan In this study introduces the K-Mode clustering algorithm is used specially designed for categorical data often found in e-commerce. The paper highlights its effectiveness in market segmentation analysis, especially for e-commerce platforms. Visualization strategies are also discussed, helping businesses understand and utilize their segmentation results better.

RFM Analysis of Customer Value Based on Clustering Algorithms: A Case of an Online Retailer in China[9] by Wiharto Wiharto and Esti Suryani. This paper underscores the significance of the RFM model in understanding customer value, especially within the context of China's burgeoning online retail market. Using clustering algorithms, the research provides insights into the buying habits of customers, enabling businesses to tailor their marketing efforts more effectively. The paper suggests that such analyses are crucial for businesses seeking to remain competitive in the dynamic online retail landscape.

III. Dataset

The dataset in focus comes from the UCI Machine Learning Repository, a renowned online resource for machine learning datasets and research. Specifically, the "Online Retail" dataset provides a rich compilation of sales data from an e-commerce business, based primarily in the UK. Each entry in the dataset denotes a specific purchase, detailing essential information like the product description, quantity ordered, invoice date, unit price, customer ID, and the country of the customer. This breadth of information makes the dataset an invaluable asset for discerning patterns in online shopping behaviors. With over 500,000 entries, it provides a comprehensive snapshot of the company's sales and customer interactions. Such a dataset is particularly suitable for our project as it grants a detailed look into the intricacies of e-commerce transactions, offering insights into customer preferences, buying frequency, and expenditure patterns, making it pivotal for understanding and segmenting different customer behaviors.

IV. Methodology

In this section, we provide a detailed explanation of our methodology:

A. Data Acquisition

The first step in the analysis was acquiring data from a reputable source. In this case, the data was sourced from the UCI Machine Learning Repository, specifically the Online Retail dataset. This dataset provides insights into purchase details and customer data. After importing the necessary Python libraries, the dataset was loaded into the environment using the Pandas library. An initial data exploration[9] was conducted to understand the general structure and characteristics of the data. We printed descriptive statistics of the dataset using the describe() method and examined missing values through the isnull().sum() function. This allowed us to understand data distribution and identify potential issues. Data quality is crucial for accurate analysis. Two significant steps were undertaken to ensure data quality: Rows containing missing customer IDs were removed. Any transactions with negative quantities, which could represent returns or faulty entries, were filtered out. Subsequently, a new column named 'TotalPrice' was introduced to compute the total expenditure for each transaction by multiplying the 'Quantity' and 'UnitPrice' columns.

B. *Feature Engineering for RFM Analysis*

One of the cornerstones of customer segmentation is RFM (Recency, Frequency, Monetary) analysis. This method segments customers based on: Recency: How recently a customer has made a purchase. Frequency: How often a customer makes a purchase. Monetary: How much money a customer spends on purchases. To calculate Recency, the most recent purchase date in the dataset was identified, and one day was added to it to set a reference point. Each customer's Recency value was then determined by subtracting their last purchase date from this reference point. Frequency and Monetary values were calculated by counting the number of invoices and summing the 'TotalPrice', respectively, for each customer. Due to varying scales and potential skewness in the RFM values, it was essential to transform and standardize the data. A logarithmic transformation was applied to the RFM data to address skewness and make the data more amenable to clustering. Following this, the StandardScaler from the sklearn library was employed to standardize the values, ensuring they had a mean of zero and a standard deviation of one.

C. *Optimal Cluster Determination*

In any clustering analysis, determining the appropriate number of clusters is of utmost importance, as it can significantly influence the results. To ensure optimal segmentation, we used the silhouette score method, a standard metric in cluster analysis. The silhouette score essentially measures how similar an object (or data point) is to its own cluster when compared to other clusters. Ideally, these scores fall between -1 and 1, with higher scores signifying that the data point is well matched to its own cluster and poorly matched to neighboring clusters.

To implement this, we utilized the KMeans clustering algorithm from the esteemed sklearn library. We conducted iterative fits for potential cluster counts, specifically ranging from 2 to 10. With each iteration, we calculated the silhouette scores and subsequently graphed them. This graphical representation helped in easily pinpointing the cluster count that maximized the silhouette score, thus giving us the optimal number.

D. *Clustering*

With the goal of offering superior customer segmentation based on their RFM (Recency, Frequency, Monetary) metrics, we moved on to the clustering phase. The RFM metrics are crucial as they help gauge a customer's behavior regarding their recent purchases, how often they buy, and their average spend. Harnessing the capabilities of the KMeans clustering algorithm, we segmented customers ensuring homogeneity within clusters. Meaning, customers within a specific cluster exhibited similar purchasing patterns. Such granular segmentation becomes a pivotal tool for businesses, allowing them to design and execute strategies tailored for each segment, ensuring maximum engagement and profitability.

E. *Visualization*

Once the clusters were defined, visual representation became imperative to better understand the group dynamics and variances. To this end, we made use of the Seaborn library, particularly its pairplot function. It facilitated the plotting of RFM metrics against one another, delivering a rich, multi-dimensional perspective of the customer segments. Such a visualization is invaluable. It not only showcases the unique attributes of each cluster but also unravels insights that can guide business decisions, helping to design campaigns and offers that resonate best with each segment.

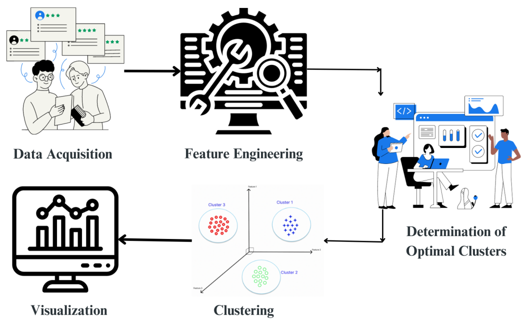


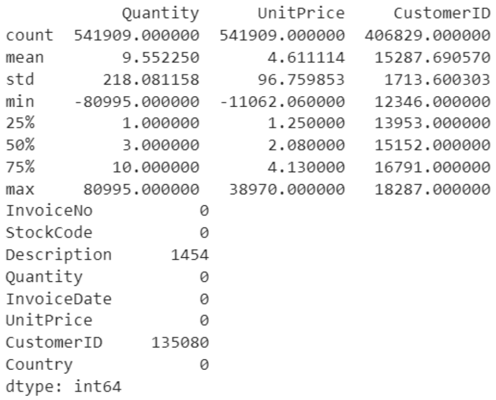
Fig. 1. Basic workflow of the project.

V. Results and Analysis

After using unsupervised learning for e-commerce dataset by using k means, in this process we have used several visualizations in this process, we will explain them in the following sections.

A. *Initial Data Overview*

The dataset provides insights into count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and the maximum value for each numerical column. data.isnull().sum(): This indicates the number of missing values in each column. It is essential to identify and handle missing data to ensure robust analysis.



```
            Quantity      UnitPrice     CustomerID
count  541909.000000  541909.000000  406829.000000
mean        9.552250       4.611114   15287.690570
std       218.081158      96.759853    1713.600303
min    -80995.000000  -11062.060000   12346.000000
25%         1.000000       1.250000   13953.000000
50%         3.000000       2.080000   15152.000000
75%        10.000000       4.130000   16791.000000
max     80995.000000   38970.000000   18287.000000
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

Fig. 2. Statistical summary and overview of the initial Online Retail dataset.

B. Silhouette Score Plot

The silhouette score gauges the consistency within clusters. A high silhouette score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. By plotting silhouette scores against a range of cluster numbers, we aim to determine the most appropriate number of clusters. The plotted graph, titled "Silhouette Score vs Number of Clusters", depicts the silhouette score on the y-axis and the number of clusters (from 2 to 10) on the x-axis. Each point on the graph signifies the silhouette score for that specific number of clusters. The ideal number of clusters corresponds to the highest silhouette score, suggesting the most distinct separation of clusters.

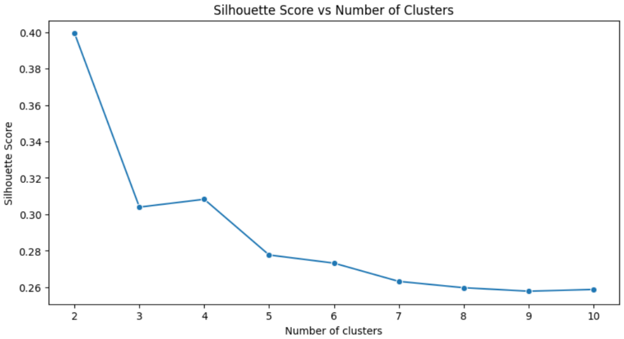
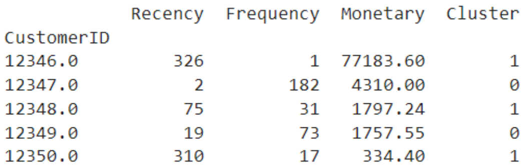


Fig. 3. Silhouette Score Plot illustrating optimal cluster number determination.

C. RFM Clustering Result

After identifying the optimal cluster count using the silhouette score, the data is segmented using the KMeans clustering algorithm. print(rfm.head()): This prints the first few rows of the RFM table, now augmented with a 'Cluster' column. Each row represents a unique customer and their corresponding Recency, Frequency, Monetary values, along with the cluster they belong to. This offers a snapshot of how customers have been grouped based on their purchasing behavior.



```
            Recency  Frequency  Monetary  Cluster
CustomerID
12346.0         326          1  77183.60        1
12347.0           2        182   4310.00        0
12348.0          75         31   1797.24        1
12349.0          19         73   1757.55        0
12350.0         310         17    334.40        1
```

Fig. 4. Tabulated results showcasing RFM-based customer clustering.

D. RFM Visualization with Pairplot

The seaborn's pairplot is an efficient visualization tool for multi-dimensional data. In this context, it plots pairwise relationships across the entire dataset for Recency, Frequency, and Monetary values. With distinct colors for each cluster, it enables visualization of how each cluster is characterized in terms of these three metrics. The plot is titled "Pairplot of RFM values, segmented by Cluster". Each sub-plot visualizes the relationship between two of the RFM metrics, segmented by the cluster color. For instance, one subplot might showcase the relationship between Recency and Frequency, with data points color-coded based on their assigned cluster. This visualization facilitates understanding of patterns and behaviors unique to each cluster.

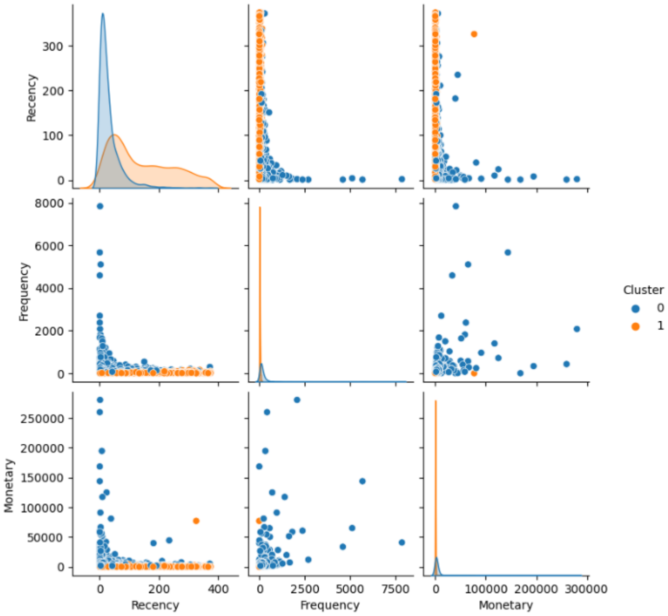


Fig. 5. Pairplot visualization of RFM clusters, highlighting behavioral stratifications.

In summary, through these results and visualizations, the code provides a holistic view of customer segmentation based on their purchasing behavior. The silhouette score plot ensures the accuracy of clusters, the RFM table with cluster assignments details the segment each customer falls into, and the pairplot elucidates the relationships and patterns within the segmented data.

VI. Conclusion

This research, rooted in an exploration of the ”Online Retail” dataset from the UCI Machine Learning Repository, sought to segment customers based on Recency, Frequency, and Monetary (RFM) metrics. The power of KMeans clustering, a machine learning technique, was employed to group customers exhibiting similar buying behaviors. To identify the ideal number of clusters for this exercise, the silhouette score method was applied, which effectively balances both cohesion and separation in the clusters. The results, visually elucidated through a series of figures, facilitated the identification of distinct customer segments, shedding light on various purchasing patterns that can guide targeted marketing strategies. The pairplot further assisted in visualizing multi-dimensional data and emphasized the discernible boundaries among the RFM clusters. The significance of this study lies not only in the insights derived from the current dataset but also in the replicable methodology that can be applied to diverse retail contexts. Future research might delve deeper into personalizing marketing efforts for each segment or explore other clustering algorithms to assess and compare their effectiveness. In essence, this project underscores the value of data-driven decision-making in retail, enhancing both customer experience and business profitability.

References

[1] “K-Means clustering with Mall Customer Segmentation,” Analytics Vidhya, May 25, 2021. https://www.analyticsvidhya.com/blog/2021/05/k-means-clustering-with-mall-customer-segmentation-data-full-detailed-code-and-explanation/.

[2] K. Bindra and A. Mishra, “A detailed study of clustering algorithms,” IEEE Xplore, Sep. 01, 2017. https://ieeexplore.ieee.org/document/8342454.

[3] Wei, Jo-Ting & Lin, Shih-Yen & Wu, Hsin-Hung. (2010). A review of the application of RFM model.

[4] K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette Score,” 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Oct. 2020, doi: https://doi.org/10.1109/dsaa49011.2020.00096.

[5] C.-H. Cheng and Y.-S. Chen, “Classifying the segmentation of customer value via RFM model and RS theory,” Expert Systems with Applications, vol. 36, no. 3, pp. 4176–4184, Apr. 2009, doi: https://doi.org/10.1016/j.eswa.2008.04.003.

[6] Olson, David & Lauhoff, Georg. (2019). Recency Frequency and Monetary Analysis. 10.1007/978-981-13-7181-3-4.

[7] Tripathi, Shreya & Bhardwaj, Aditya & Eswaran, Poovammal. (2018). Approaches to Clustering in Customer Segmentation. International Journal of Engineering & Technology. 7. 802. 10.14419/ijet.v7i3.12.16505.

[8] Kamthania, Deepali & Pahwa, Ashish & Madhavan, Srijit. (2018). Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business. Journal of Computing and Information Technology. 26. 57-68. 10.20532/cit.2018.1003863.

[9] Ashwin, V., Menon, V., Devagopal, A.M., Nived, P.A., Udayan Divya, J. (2023). Detection of Fraudulent Credit Card Transactions in Real Time Using SparkML and Kafka. In Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Lecture Notes in Networks and Systems, vol 540. Springer.

[10] M. Saravanan, Prasad, G., Jagadeesan, M., Raghu Raman, and V Smrithi Rekha, “Group Recommender Model for Boosting and Optimizing Customer Purchases”, in Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, 2012.

[11] Dr. Radhika N. and H., S., “Unsupervised clustering with spiking Neurons by sparse temporal coding and multilayer RBF Networks”, in Conference on Advanced Computer Applications, 2012.

[12] T. J. Devika and Dr. Ravichandran J., “A clustering method combining multiple range tests and K-means”, Communications in Statistics - Theory and Methods, pp. 1-56, 2021.

[13] S. Unnikrishnan, Sreelakshmi, S., and Deepa, G., “Enhancement of accuracy in K-means clustering”, International Journal of Control Theory and Applications, vol. 9, pp. 7619-7626, 2016.

[14] Saraf, E., Pradhan, S., Joshi, S. and Sountharrajan, S., 2022, April. Behavioral Segmentation with Product Estimation using K-Means Clustering and Seasonal ARIMA. In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1641-1648). IEEE.

[15] Don S, “High Dimensional Data Visualization : A Survey”, Journal of Advanced Research in Dynamical and Control Systems, vol. 12, pp. 851-856, 2017.