Proceedings of the 5th International Conference on Soft Computing for Security Applications (ICSCSA-2025)
DVD Part Number: CFP256A9-DVD; ISBN: 979-8-3315-9490-9

Enhancing Customer Segmentation: A Novel Approach using Machine Learning Algorithms

1st Ms R.Anitha
Research Scholar,
Department of Computer Science
Vels Institute of Science, Technology
Advanced Studies(VISTAS)
E-mail:anithashivaguru@gmail.com

2nd Dr. Y.Kalpana
Professor,
Department of BCA & IT
Vels Institute of Science, Technology
Advanced Studies(VISTAS)

Abstract— Customer segmentation is needed for every business to survive in the field. Grouping up of customers can be done by finding a similar pattern among customers by analyzing the sales dataset. This research introduces a novel methodology designed to enhance customer segmentation significantly, achieved through the integration of cutting-edge machine learning algorithms, explicitly focusing on clustering and classification techniques, to overcome the inherent limitations and static nature of traditional segmentation methods. The proposed approach uses the well-established RFM model for feature engineering, enabling a comprehensive representation of customer behaviour and value, subsequently employing advanced clustering algorithms to identify distinct customer segments based on these RFM features. To further refine the segmentation, a weighted lifetime value calculation is introduced, serving as a critical classification criterion within each identified cluster, thereby facilitating a more nuanced understanding of customer potential by integrating AI techniques with RFM analysis, the identification of sustainable consumer behaviours is improved, offering more granular insights into consumer behaviour and its impact on sustainability.

Keywords: - Customer Segmentation, Machine Learning, HDBSCAN Algorithm, Clustering Algorithm, Unsupervised Learning, K-Means Algorithm, UMAP, RFM, Customer Lifetime.

I. INTRODUCTION

Customer segmentation is necessary to understand the different consumer groups. Many retail industries depend on machine learning techniques for this process since the traditional method does not work effectively for large datasets. RFM is the foundation model for businesses to gain deeper insights into customer purchase behavior and tailor marketing strategies accordingly. RFM analysis, which stands for Recency, Frequency, and Monetary value, is a widely used method for evaluating customer value based on their purchasing behavior.

Recency is to count the number of days it has been since the last purchase; Frequency is to track how many times purchases are made within a specific period; and Monetary is to assess the overall amount spent by the customer over time. The Weighted Lifetime WL is another feature that can be additionally used with RFM to identify high-value clients, forecast future behavior, and customize marketing initiatives for more engagement and profitability. Machine learning is a crucial tool used for customer segmentation. Unsupervised algorithms are used for clustering the customers. K-Means is the most used clustering algorithm. HDBSCAN is the emerging clustering technique used in our model to enhance the customer segmentation.

II. RELATED WORK

[1][2] Evaluated the unsupervised machine Learning algorithms such as K-Means, DBSCAN, Hierarchical clustering based on RFM feature, and also included a traditional model to analyze the optimization of clustering models. [2] also uses the RFM framework and finds the clusters using K-Means, Mean Shift, Hierarchical, and DBSCAN. [3] Predicts customer purchasing behavior by analyzing sales history and the purchasing behavior of real-time transactional dataset using RFM and K-Means algorithm.

[4] Analyzed Pakistan's largest e-commerce dataset by K-means, Gaussian, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). This study also proposed the RFMT model by introducing Time as an additional feature extending RFM. This study uses the elbow, dendrogram, and Silhouette score, Calinski–Harabasz, Davies–Bouldin, and Dunn index for evaluating the clusters. This study also includes transactional status and season-wise for segmentation. [5] uses RapinMiner 5.2 tools to produce the visual cluster model using K-Means and RFM.

[7] Employs TF-IDF to generate product categories from the descriptions. The Apriori algorithm is used to analyze the purchased products. PCA and T-sne are used to reduce the dimension of data and apply RFM and K-Means for customer segmentation.

[8] This study employs the combination of RFM, Box-Cox transformation, and temporal analysis to predict consumer behavior. Supervised machine learning algorithms like Random Forest, AdaBoost, Extra Trees, LGBM, and XGBoost are also used. In which XGBClassifier and ET have achieved the highest accuracy in forecasting customer lifetime value. [9] also uses RFM and K-Means for segmentation. To enhance the efficiency of segmentation, a decision tree is used to create nested splitting based on Gini index inside each cluster.

[10] To enhance the customer segmentation, Deep Learning and Explainable AI was integrated and proposed a mathematical model to group the demographic data, behavioral patterns, and purchase histories, categorizing customers into distinct clusters united with their preferences and needs. Two dataset Mall customers and e-commerce dataset was used to validate the efficacy of DeepLimeSeg. Lime assigns weights to features based on their contribution to the predictions. Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared are evaluated for validation. The DeepLimeSeg model achieved a low MSE of 0.9412, the average MAE value of 0.9874 and the high R2 value of 0.94152 indicating good predictive accuracy for the spending score.

[11] This study merges the demographic data with RFM as a key feature to analyse the data. Global pizza restaurant chain in Turkey dataset is used in this study. They employed k-means clustering, Apriori association rule mining (ARM) and neural networks. The elapsed time and prediction accuracy were used for evaluation.

[12] UK online retail store dataset was used in this study for customer segmentation using RFM and k-Means algorithm. Elbow method and Silhouette score was used to optimize the segment.

[13] Both behavior-based segmentation and demographic based segmentation was done. Exploratory Data Analysis approach was used to find correlations between features, outliers in the dataset. K-Means, Hierarchical and DBSCAN clustering algorithm were used to segment the real time bank customers using the transactional financial dataset for behavior-based segmentation. K-Modes algorithm was used to segment the same dataset for demographic based segmentation.

[14] The data was extracted from Online Retail Dataset. Q-CUT method is used to obtain RFM Scores. Box–Cox transformation is used to analyze the outliers. This study employs K-Means and DBSCAN.

[15] Taiwan's electronic industry dataset is employed as a case study. Three purposes involved in this study are discretize continuous attributes to enhance the rough sets algorithm; cluster customer value as output (customer loyalty) that is partitioned into 3, 5 and 7 classes based on subjective view, then see which class is the best in accuracy rate; and find out the characteristic of customer in order to strengthen CRM. This study combines RFM value and K-Means algorithm in Rough-Set Theory. Finally, LEM2 algorithm is used to mine classification rules. [16] Multi-layer perceptron (MLP), Support vector machine (SVM) and decision tree classification (DTC) were applied in this study. Multi-class classification is performed through multiple binary classifiers, and the predicted class is voted out. In this research, it is an application tailored neural network having two hidden layers illustrates the structure of MLP used for the prediction of the customers based on the behavioral features.

[17] In this paper, an unsupervised deep learning model called a Self-organizing map (SOP) with an Improved social spider optimization approach has been used for customer segmentation. The Self-Organizing Neural Network (SONN) method is used for clustering and Deep Neural Network (DNN) model is used to classify the customers within the clusters.

[18] The Digikala company, the biggest E-Commerce dataset, was used. This paper proposes an R+FM model that configures the segmentation and applies K-Means algorithm. Interquartile Ranges (IQR) are used to remove the outliers.

III. METHODOLOGY

This study uses Recency, Frequency, and Monetary, along with LifeTime, as features and the K-means algorithm as a clustering technique to segment the Retail Industry's customers. Fig 1 describes the stages carried out in this research.

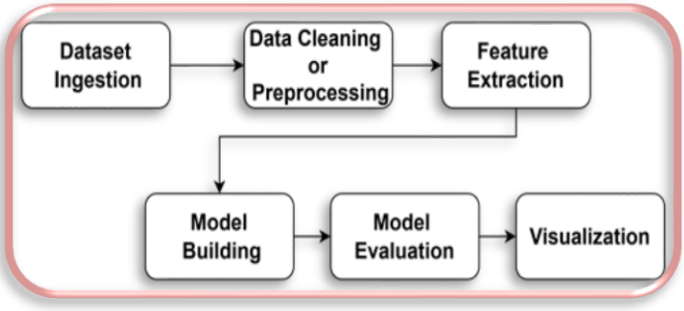

Fig. 1. Schematic View of the Process

The UCI Online Retail Dataset is a publicly available dataset containing transactional data from a UK-based online retail store. It includes information on sales, invoices, customer details, product descriptions, quantities, and prices which includes 541880 records.

A. Data Preprocessing

Data Preprocessing is the process of transforming raw data into a format that is suitable for data analysis.

Dropping the rows which has null value from our Dataframe—Retail using "dropna( )." Then the number of rows was reduced to (401604, 8). Initially, it was (541880, 8).

The Quantity and the UnitPrice fields contain negative values. Change the dataframe by excluding those negative values.

retail = retail[retail['Quantity'] > 0]

retail = retail[retail['UnitPrice'] > 0]

InvoiceNo precedes with the character 'C' it denotes the canceled order. Remove the canceled order from the dataset.

retail = retail[~retail["InvoiceNo"].str.contains("C", na=False)]

B. RFM and Lifetime Value Calculation

Recency value can be calculated using InvoiceDate. Number of day's difference between the last InvoiceDate of each customer and the Specified date. The minimum number of days is a high recency score. Frequency can be calculated using the total count of InvoiceNo. Monetary is calculated using the following formula

$$M_i = \sum_j Total\_Price_j \quad (1)$$

i represents CustomerID
j represents InvoiceNo of respective CustomerID
*Total_Price = UnitPrice*Quantity.*

The Lifetime value is calculated by the formula

$$LifeTime = MAX(InvoiceDate) - MIN(InvoiceDate)$$
$$W = COUNT(InvoiceNo)$$
$$LT = LifeTime * W \quad (2)$$

C. Dataframe Formation

Fig. 2 shows the Dataframe includes Recency, Frequency, Monetary, and Lifetime.

Cust_ID	Frequency	Monetary	Recency	W_Lifetime
12346	1	77184	325	1
12347	182	4310	1	66430
12348	31	1798	74	8742
12349	73	1758	18	73
12350	17	335	309	17

Fig 2. RFML Dataframe

The values in Recency is indirectly proportional in the above table.

Binning the column range from 1 to 5 using Quantile – Cut is shown in the Fig. 3

Cust_ID	F_Score	M_Score	R_Score	L_Score
12346	1	5	1	1
12347	5	5	5	5
12348	3	4	2	4
12349	4	4	4	2
12350	2	2	1	1

Fig 3. R, F, M, and L scores

D. K – MEANS Algorithm

The K-Means Clustering algorithm is unsupervised Machine Learning algorithm used to partition data into distinct clusters based on similarities.

K-Means relies mainly on the Euclidean distance formula to identify the similarity of the data in an iterative way.

$$d = \sum_{k=1}^{K} \sum_{i=1}^{n} (x_i - \mu_k)^2 \quad (3)$$

k represents centers of K cluster, μ_k represents kth center, and x_i represents the i^{th} point in the data set.

The first step involves initializing the centroids, randomly. Next assigning every data to the nearest center group.

$$Z_{ik}^t = \begin{cases} 1 \; if \; k = argmin\|x_i - \mu_z\|, \; ¥z \in \{1,..K\} \\ 0 \; otherwise \end{cases} \quad (4)$$

Next the algorithm recalculates the centroid of each cluster, resulting from the previous step. The algorithm ends if the result of the clustering, in an iteration is the same as the one in the previous iteration.

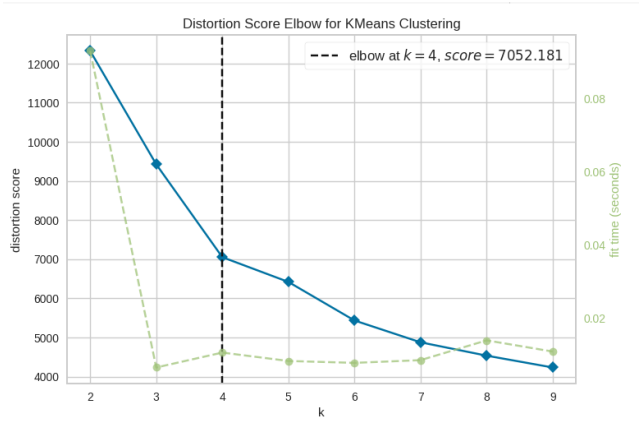

Fig.4 The Elbow Method.

Fig.4 indicates that the elbow point on the plot indicates the optimal number of clusters is 4 with the distortion score 7052.181.

E. Parallel Coordinates Diagram

Fig.5 RFM Parallel Coordinates diagram has three vertical axes, representing Recency (R), Frequency (F), and Monetary (M) scores. Each line in the plot will represent a customer, connecting their scores across these three dimensions. The color of each line typically indicates the cluster the customer belongs to based on the RFM segmentation.

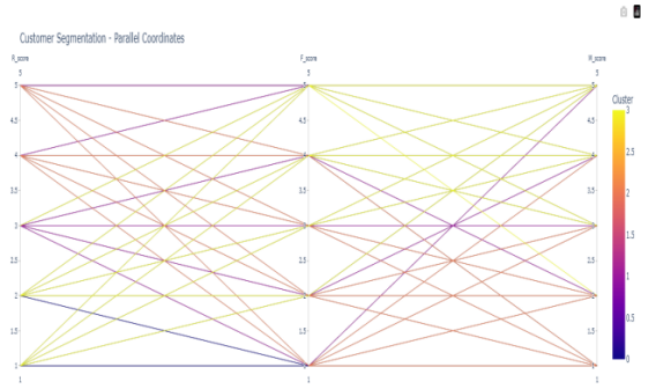

Fig.5 RFM Parallel Coordinates

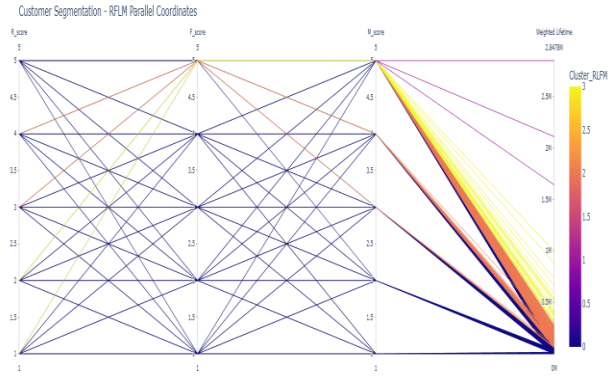

Fig.6 RFML Parallel Coordinates

Fig.6 plot shows how the Weighted Lifetime dimension influences the clustering and how customer groups differ not only in their RFM behavior but also in their calculated lifetime value. The lines extending to a fourth axis, providing a more comprehensive view of customer segments by including the lifetime aspect.

Essentially, the RFML parallel coordinates plot adds an extra dimension Weighted Lifetime to the visualization, allowing for a more understanding of customer segments by incorporating a lifetime value perspective alongside the transactional behavior captured by RFM.

F. HDBSCAN Algorithm

Hierarchical Density-Based Spatial Clustering of Applications with Noise is an unsupervised clustering algorithm it extends DBSCAN by integrating hierarchical clustering techniques. It is designed to identify clusters of varying densities and shapes in data, as well as to handle noise and outliers.

To form a hierarchical clustering structure, it uses core distance and mutual reachability distance.

1) Core Distance

For a point p, the core distance $core_k(p)$ is defined as the distance to its k-th nearest neighbor:

$$core_k(p) = distance\ from\ p\ to\ its\ k\text{-}th\ nearest\ neighbor \quad (5)$$

2) Mutual Reachability Distance

Given two points p and q, the mutual reachability distance $d_{mreach}(p,q)$ is:

$$d_{mreach}(p,q) = MAX(\ core_k(p),\ core_k(q),\ d(p,q)\ ) \quad (6)$$

Where d(p,q) is the original distance between p and q.

This distance effectively expands the original distance based on local density, making it stronger to varying densities.

Fig.7 shows the UMAP embedding HDBSCAN clusters. UMAP- Uniform Manifold Approximation and Projection is a nonlinear dimensionality reduction technique used to embed high-dimensional data into a lower-dimensional space while conserving the data's manifold structure. It builds a fuzzy topological representation of the data, apprehending local and global relationships, and then optimizes a low-dimensional embedding to maintain these relationships.

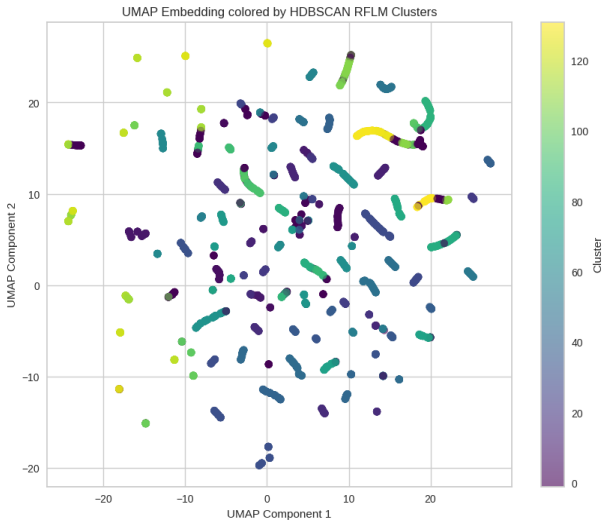

Fig.7 UMAP Embedding HDBSCAN clusters

G. *Evaluation Metric*

In Fig.8, chart specifies the difference of Silhouette score, Calinski-Harabasz score and Davies-Bouldin score between RFM and RFML clusters.

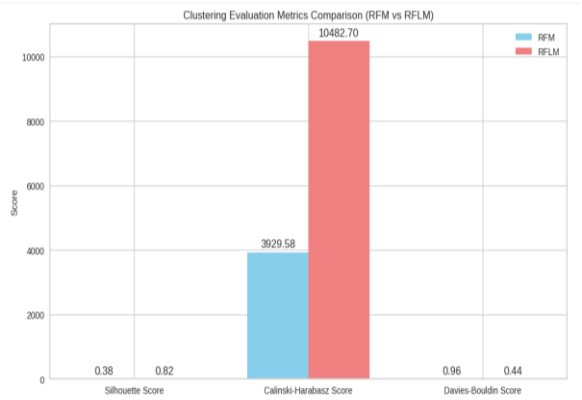

Fig.8 RFM vs RFML

```
RFM Clustering Evaluation:
Silhouette Score: 0.3822617572822178
Calinski-Harabasz Score: 3929.578840377311
Davies-Bouldin Score: 0.9614252914341611
------------------------------
RFMl Clustering Evaluation:
Silhouette Score: 0.8172806515374231
Calinski-Harabasz Score: 10482.699282340536
Davies-Bouldin Score: 0.44344664892149915
```

IV. CONCLUSION

The K-Means algorithm for RFM and RFML resulted in a fixed number of clusters determined by the elbow method. The result of evaluation metrics indicates that the RFML approach yields better-structured clusters based on these metrics in this specific case. The implementation of HDBSCAN identifies multiple clusters and a significant number of noise points. The visualization shows how HDBSCAN identifies dense regions as clusters, and sparse regions as noise. By combining RFM metrics with Lifetime value provides a wide-ranging view of customer value, enabling more effective segmentation, personalized marketing, and improved customer lifetime management.

REFERENCES

[1] Turkmen, Banu. "Customer Segmentation with machine learning for online retail industry." The European Journal of Social & Behavioural Sciences (2022).

[2] Sakaline, Golam, and László Buics. "Advancing Towards Sustainable Retail Supply Chains: AI-Driven Consumer Segmentation in Superstores." Engineering Proceedings 79.1 (2024): 73.

[3] P. Anitha, Malini M. Patil, “RFM model for customer purchase behavior using K-Means algorithm.” Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 5, 2022.

[4] Ullah, Asmat, et al. "Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time." sensors 23.6 (2023): 3180.

[5] Maryani, Ina, et al. "Customer segmentation based on RFM model and clustering techniques with K-means algorithm." 2018 Third International Conference on Informatics and Computing (ICIC). IEEE, 2018.

[6] Doğan, Onur, Ejder Ayçin, and Zeki Bulut. "Customer segmentation by using RFM model and clustering methods: a case study in retail industry." International Journal of Contemporary Economics and Administrative Sciences 8 (2018).

[7] Shen, Boyu. "E-commerce customer segmentation via unsupervised machine learning." The 2nd international conference on computing and data science. 2021.

[8] Arefin, Sydul, et al. "Retail industry analytics: Unraveling consumer behavior through rfm segmentation and machine learning." 2024 IEEE International Conference on Electro Information Technology (eIT). IEEE, 2024.

[9] Chaudhary, Poonam, Vaishali Kalra, and Srishti Sharma. "A hybrid machine learning approach for customer segmentation using rfm analysis." International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020, Volume 1. Singapore: Springer Nature Singapore, 2022.

[10] Talaat, Fatma M., et al. "A mathematical model for customer segmentation leveraging deep learning, explainable AI, and RFM analysis in targeted marketing." Mathematics 11.18 (2023): 3930.

[11] Sarvari, Peiman Alipour, Alp Ustundag, and Hidayet Takci. "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis." Kybernetes 45.7 (2016): 1129-1157.

[12] Shirole, Rahul, Laxmiputra Salokhe, and Saraswati Jadhav. "Customer segmentation using rfm model and k-means clustering." Int. J. Sci. Res. Sci. Technol 8.3 (2021): 591-597.

[13] Aliyev, Musadig, et al. "Segmenting bank customers via RFM model and unsupervised machine learning." arXiv preprint arXiv:2008.08662 (2020).

[14] Lewaaelhamd, Israa. "Customer segmentation using machine learning model: an application of RFM analysis." Journal of Data Science and Intelligent Systems 2.1 (2024): 29-36.

[15] Cheng, Ching-Hsue, and You-Shyang Chen. "Classifying the segmentation of customer value via RFM model and RS theory." Expert systems with applications 36.3 (2009): 4176-4184.

[16] Rahim, Mussadiq Abdul, et al. "RFM-based repurchase behavior for customer classification and segmentation." Journal of Retailing and Consumer Services 61 (2021): 102566.

[17] Wang, Chenguang. "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach." Information Processing & Management 59.6 (2022): 103085.

[18] Tavakoli, Mohammadreza, et al. "Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: a case study." 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE). IEEE, 2018.

[19] Monalisa, S.; Juniarti, Y.; Saputra, E.; Muttakin, F.; Ahsyar, T.K. Customer segmentation with RFM models and demographic variable using DBSCAN algorithm. TELKOMNIKA Telecommunication Computing Electronics Control 2023, 21, 742–749.

[20] Zong, Yi and Enze Pan. “A SOM-Based Customer Stratification Model.” Wireless Communications and Mobile Computing 2022.