Grenze International Journal of Engineering and Technology, June Issue

New RFM+WL Model for Effective Customer Segmentation using Machine Learning

R. Anitha[1] and Y. Kalpana[2]
[1]Research Scholar, Department of Computer Science, Vels Institute of Science, Technology Advanced Studies (VISTAS), India
Email: anithashivaguru@gmail.com
[2]Professor, Department of BCA & IT, Vels Institute of Science, Technology Advanced Studies (VISTAS), India

***Abstract*—Customer segmentation is a powerful tool for exploring customer insights to tailor their marketing strategies and improve customer retention. Traditional RFM (Recency, Frequency, Monetary) analysis provides a strong basis for segmenting customers based on purchase behavior. However, by incorporating customer "Lifetime" i.e., recurring value as an additional feature and machine learning techniques, businesses can gain deeper insights and create more precise customer segments. This study explores a novel approach that combines RFM +WL and machine learning to optimize customer segmentation. By applying clustering algorithms, we can identify distinct customer segments based on their RFM+WL attributes. This refined segmentation enables businesses to prioritize high-value customers, personalize marketing campaigns, and implement target retention strategies. The proposed approach offers a powerful tool for companies to maximize customer lifetime and drive sustainable growth.**

***Index Terms*— Customer Segmentation, Machine Learning, Clustering Algorithm, Unsupervised Learning, K-Means Algorithm, RFM, Customer Lifetime.**

I. INTRODUCTION

Customer segmentation plays a vital role in marketing analytics and CRM. It involves effective customer management for greater enrichment, targeting, and retention initiatives.

Common methodologies for customer segmentation exemplify demographic, geographic, psychographic, and behavioral classifications. Such classes include age, gender, income, occupation, and education. The customers could be divided into segments based on demographic characteristics like population density and regional features. This is based on factors like social status, personality, and lifestyle.

Elements of behavioral segmentation include purchase behavior, usage rate, brand loyalty, occasion, and customer journey. The main driving factor of this study is to realize the implementation of the RFM value in customer base segmentation concerning their purchasing habits.

RFM provides a simple and actionable framework for assessing and segmenting customers based on three core dimensions of their behavior.

Recency (R): How recently a customer purchased or interacted with the business.

Frequency (F): How often a customer purchases or engages with the business within a given period.

Monetary (M): How much money a customer spends during a specific time frame.

This study enhances the RFM analysis by incorporating the 'Weighted LifeTime' value as an additional feature.

Grenze ID: 01.GIJET.11.2.392

It also explores the application of machine learning techniques to segment the customers that can identify and address evolving customer insights. The main objective of this research is to evaluate the K-Means clustering algorithm in segmenting customers.

II. Related Work

A.J. Christy et al 2021. In this literature, Segmentation is done using RFM analysis and algorithms like K-Means, Fuzzy c-Means, and RM K-Means algorithm. The RM K- Means algorithm took less iteration to complete the clustering process.
M.Y. Smaili et al. 2023 found that adding diversity D as a parameter along with RFM significantly improves the RFM model by analyzing the strong correlation between D and the old model parameters. Segmentation with the RFM-D model generated different quality clusters than the RFM. The main objective is to mention the importance of analyzing all the variables and their correlations to choose the right parameters to consider in the marketing model in addition to the classical variables R, F, and M.
Thanh Ho et al. 2023, used cohort analysis to analyze customer retention rates and recommend marketing strategies for each segment. Developed a new model RFM-D, Demography data was combined with traditional RFM and applied with K-Means and K-Prototype algorithms.
The study of Asmat Ullah et al. 2023, proposed a novel approach model RFMT (RFM + Time) that analyzed Pakistan's largest e-commerce dataset by introducing K- Means, DBSCAN, and Gaussian. This study includes transaction status and season-wise segmentation.
Israa Lewaa 2023, computed the RFM score for the e-commerce dataset and performed the segmentation using machine learning clustering algorithms like K-Means and DBSCAN.

III. Methodology

A. RFM Analysis

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.

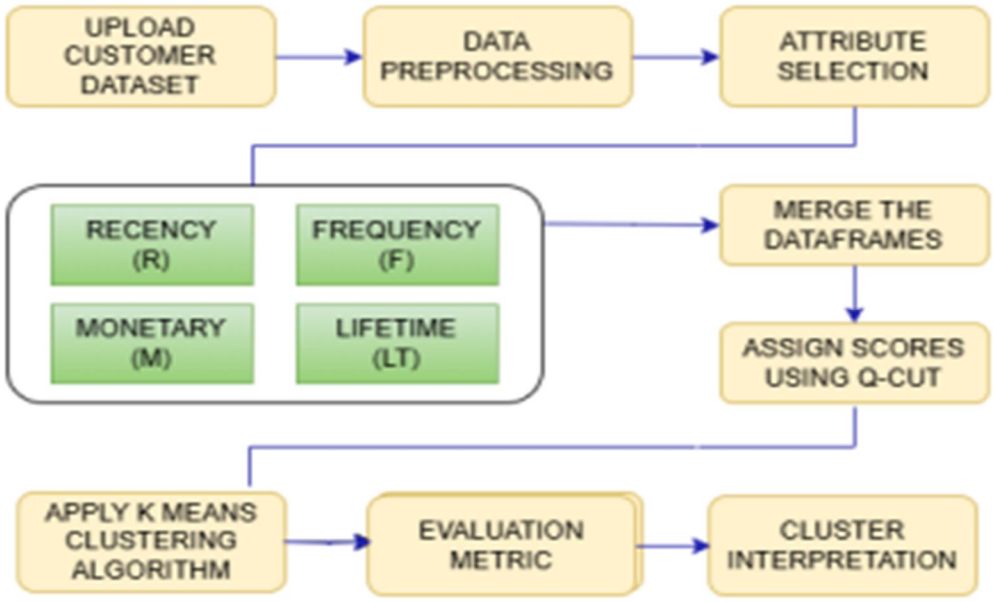

Figure 1. Schematic View of The Process

B. Dataset

Acquired the Customer Purchase Dataset from the UCI Repository. It was UK's E-Commerce Dataset based on customer's purchasing behavior i.e., Online Retail Dataset which consists of eight fields and 541880 records. Table.1 describes the attributes of the dataset. Fig.2 shows the sample data.

TABLE I. CUSTOMER DATASET

S. No	Field	Description
1.	InvoiceNo	Six-digit Integer. InvoiceNo with 'C' denotes canceled order.
2.	StockCode	Five-digit Integer assigned to unique product.
3.	Description	Name of the Product
4.	Quantity	The number denotes the number of each product per transaction.
5.	InvoiceDate	Date/Time of each transaction.
6.	UnitPrice	Numeric Value that represents the price of a product in sterling.
7.	CustomerID	The Five-digit Integer is assigned to each customer.
8.	Country	Customer's Location in String.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

Figure. 2 Snapshot of Customer Dataset

C. Data Preprocessing

Data Preprocessing is the process of transforming raw data into a format that is suitable for data analysis.
Dropping Null Values: In particular, the unique ID of the customer, CustomerID, contains a Null value in many rows. We are dropping those rows from our Dataframe—Retail using "dropna( )." Then the number of rows was reduced to (401604, 8). Initially, it was (536641, 8).
Data Cleaning: The values in the Quantity and the UnitPrice fields contain negative values. Removing those negative values using the following statement
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)]

D. Attribute Selection

For RFM Analysis, we need to calculate Recency(R), Frequency(F), and Monetary (M).
Calculating Recency(R): The InvoiceDate is the main attribute used to calculate recency. Find the difference in the days between the last InvoiceDate of each customer and the Specified date. The minimum number of days is a high recency score.
Calculating Frequency(F): The attribute used to calculate frequency is InvoiceNo. When the customer purchases the product, InvoiceNo. is generated automatically. Counting the occurrence of InvoiceNo. for each customer group by CustomerID will give a value for frequency for each customer.
Calculating Monetary(M): To find the value of monetary(M), we have to calculate the Total price the customer spends on each purchase using the attributes UnitPrice and Quantity.
Total_Price = UnitPrice*Quantity.
Sum Up the Total_Price for each customer group by CustomerID will give a value for monetary for each customer.

$$M_i = \sum_j \text{Total\_Price}_j$$

i represents CustomerID
j represents InvoiceNo of respective CustomerID
Calculating LifeTime (LT): The novelty of our research work is adding LifeTime value along with RFM. To measure the LifeTime value, find the difference between the first and the recent purchase date using the attribute InvoiceDate. To add weightage LifeTime is multiplied by the number of purchases the customer made in that period.
LifeTime = MAX(InvoiceDate) - MIN(InvoiceDate)
W = COUNT(InvoiceNo) [for each unique customerID]
WL = LifeTime * W

E. Merge The Dataframes:

Combine the dataframes Recency, Frequency, Monetary, and Lifetime into one dataframe. Fig.3 depicts the result of dataframes merging of each customer. Unlike other columns, values in Recency is indirectly proportional.

F. Assign Scores Using Q-Cut:

Binning the column range from 1 to 5 using Quantile – Cut. Fig.4 shows the binning results.

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

Figure. 3 Snapshot of RFM_df Dataframe

Cust_ID	F_Score	M_Score	R_Score	WL_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

Figure. 4 Snapshot of R, F, M, and L scores

G. K – Means Algorithm

(M.Y. Smaili and H. Hachimi)
The K-Means Clustering algorithm is unsupervised Machine Learning algorithm used to partition data into distinct clusters based on similarities. Below are the steps in Kmeans algorithm.

- Choose the number k of clusters and initialize randomly k centroids.
- Assign each data point to the nearest centroid, form k clusters.
- Recalculate the centroids as the mean of all data points assigned to each cluster.
- Repeat the steps 2 and 3 until the centroids no longer change significantly or a specified number of iterations is reached.

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

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

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. For each i =

$$Z_{ik}^{t} = \begin{cases} 1 \; if \; k = argmin\|x_i - \mu_z\|, \; \forall z \in \{1, \ldots K\} \\ 0 \; otherwise \end{cases} \quad \ldots\ldots \quad (2)$$

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.
The input given for this algorithm is the 'rfml_df' dataset and the features are R_Score, F_Score, M_Score, and WL_Score. The number of clusters can be found using Elbow method and Silhouette score.

H. Evaluation Metric

The Elbow Method is used to identify the optimal number of clusters (K) for K-Means clustering. This includes creating a graph of the distortion score, or Within-Cluster Sum of Squares (WCSS), as a function of the number

of clusters and identifying the point where the decrease rate changes significantly, known as the 'elbow point'. The elbow point on the plot in Fig.5 indicates the optimal number of clusters is 4 with the distortion score of 226534.638. This correlation matrix in the Fig.6 shows how the features correlated. To calculate the weighted lifetime, frequency i.e., number of customer purchases is used. So, F_Score and WL_Score are highly correlated with the value 0.84.

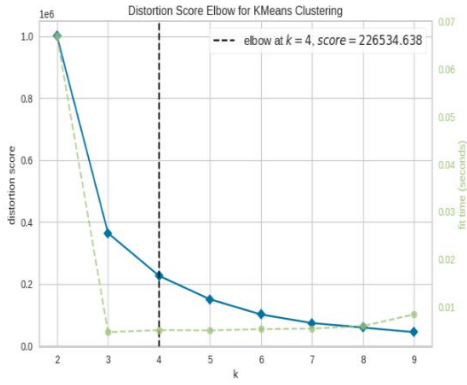

Figure 5. The Elbow Method

Figure 6. Correlation Matrix

The average silhouette score is: 0.6027304922952303
The Calinski-Harabasz Index is: 3200.396838437067
The Davies-Bouldin Index is: 0.6361563183726398

Average silhouette score: 0.6027 indicates a moderate to high level of cohesion in clusters and relatively strong separation between clusters. It also states that data points are effectively placed in corresponding clusters. There is a measurement of 3200.40 for the Calinski-Harabasz Index, which defines high cluster well-compacted with high separation by the ratio measure between cluster scattering and within-scattering. It is also essential to highlight that the calculated Davies-Bouldin Index showed 0.6361, implying that each cluster has insignificant overlap and is pretty different from its neighboring cluster.

I. Cluster Interpretation

Matplotlib is used to visualize the customer purchase behavior dataset. Figures 7 - 10 show the number of customers distributed in Monetary, Frequency, Recency, and Lifetime values.
Clustering using the K-Means algorithm is done. Four clusters are formed based on RFM+WL scores.

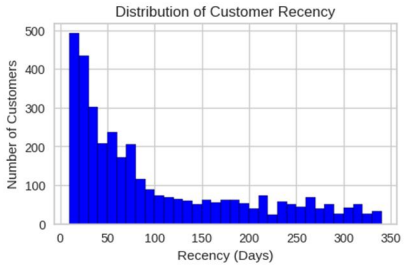

Figure 7. Recency

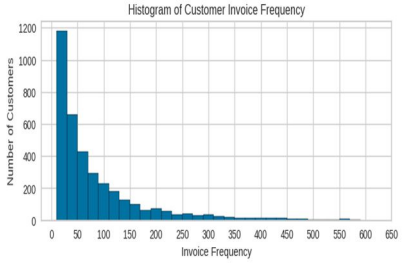

Figure 8. Frequency

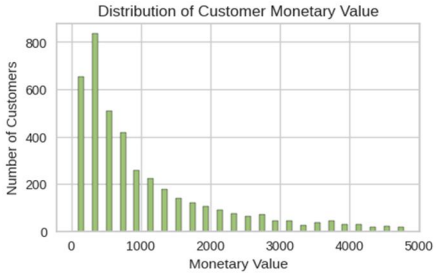

Figure 9. Monetary

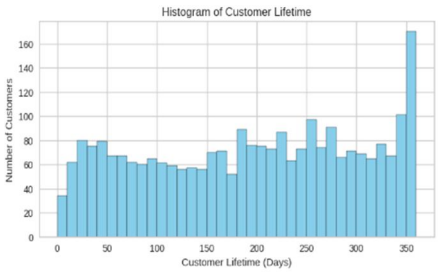

Figure 10. Lifetime

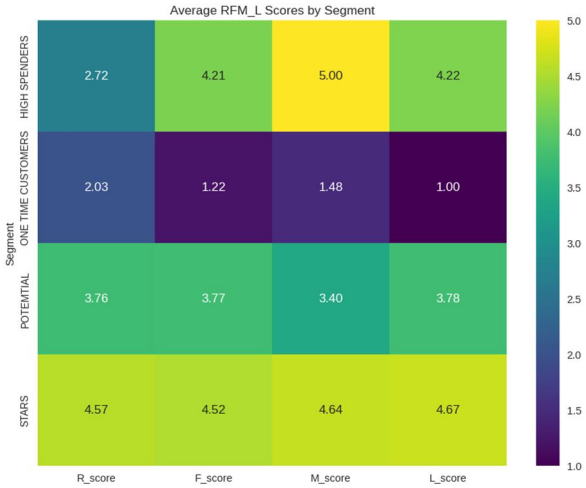

Figure 11. Heat Map of Clusters

Fig 11. The heat Map of four clusters shows the average scores of RFM+WL. The average value of M_score is 5.0 for the "HIGH SPENDERS" cluster. In "ONE-TIME CUSTOMERS" the average value of WL_score is 1.0. All the scores range from 4.0 to 5.0 in the cluster "STARS" and 3.0 to 4.0 in the cluster "POTENTIAL".

IV. Conclusion

By understanding the category of customers, the company can make many effective decisions. This study uses the RFM+WL approach to analyze customers based on their purchase behavior. The weighted lifetime is the new approach to classify the customer incorporated in this study. Lifetime value only calculates the days between the first and last purchases. Weighted lifetime improves the lifetime value by multiplying it by the number of purchases made in that period.

The machine learning algorithm created four distinct clusters. The results of evaluation metrics clearly state that the clustering model had strongly regrouped the data into clearly defined and distinguishable clusters.

The following result shows the number of customers in each cluster.

STARS	942
ONE-TIME CUSTOMERS	1400
POTEMTIAL	1566
HIGH SPENDERS	238

32.3% of the total were one-time customers indicating a need for improved retention and re-engagement strategies to reduce churn and improve lifetime. 36.1% were potential customers. This group can be converted into high-value customers by providing personalized offers. The small yet impactful segment is High spenders with 5.5% contributing significantly to revenue. To retain this group personalized services are essential. The Stars with 21.7% are loyal, high-frequency customers who consistently contribute to revenue. Strategies like exclusive rewards, personalized communication, and loyalty programs can strengthen their engagement.

References

[1] 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, 33 (10) (2021), pp. 1251-1257

[2] Moulay Youssef Smaili and Hanaa Hachimi. "New RFM-D classification model for improving customer analysis and response prediction". Ain Shams Engineering Journal 14(2023). 102254.

[3] Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D-S, Le, and T-G. (2023). "An extended RFM model for customer behaviour and demographic analysis in retail industry". Business Systems Research, 14(1), 26-53.

[4] Ullah, A et al. "Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time". Sensors 2023, 23, x.

[5] Lewaaelhamd, I. (2024). "Customer segmentation using machine learning model: an application of RFM analysis". Journal of Data Science and intelligent Systems2(1), 165–172.

[6] Fotaki, G.; Spruit, M.; Brinkkemper, S.; and Meijer, D. "Exploring big data opportunities for online customer segmentation". International Journal of Business Innovation and Research. (IJBIR) 2014, 5, 58–75.

[7] B. Kaur et P. K. Sharma, “Implementation of customer segmentation using integrated approach”, International Journal of Innovative Technology and vol. 8, no 6, p. 3, 2019.
[8] Y. Qiu, P. Chen, Z. Lin, Y. Yang, L. Zeng, et Y. Fan, “Clustering analysis for silent telecom customers based on K-means++”, in 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, vol. 1, p. 1023 1027.
[9] K. R. Kashwan et C. Velu, “Customer segmentation using clustering and data mining techniques”, Int. J. Comput. Theory Eng., p. 856 861, 2013.
[10] Mathew, A.; Scholar, P.G.; Jobin, T.J. “Role of big data analysis and machine learning in ecommerce customer segmentation”. In Proceedings of the National Conference on Emerging Computer Applications (NCECA), Online, 17 June 2021; p. 189.
[11] Erlich, Z., Gelbard, R., and Spiegler, I. (2016). “Evaluating a positive attribute clustering model for data mining”. Journal of Computer Information Systems, 43(3), 100-108.
[12] Chen, D., Sain, S. L., and Guo, K. (2012). “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining”. Journal of Database Marketing and Customer Strategy Management, 19(3), 197-208.
[13] Turkmen, B. “Customer segmentation with Machine learning for online retail industry”. Eur. J. Soc. Behav. Sci. 2022, 31, 111–136.
[14] T. C. Chen et al., “Application of data mining methods in grouping agricultural product customers,” Math. Probl. Eng., vol. 2022, 2022.
[15] D. Deepa, A. Sivasangari, R. Vignesh, N. Priyanka, J. Cruz Antony, and V. GowriManohari, “Segmentation of shopping mall customers using clustering,” pp. 619–629, 2023.
[16] Supangat S and Mulyani Y: “Customer loyalty analysis using recency, frequency, monetary (RFM) And K-Means cluster for Labuan Bajo Souvenirs in online store”. Journal of Information Systems and Informatics (2023).
[17] Asha, V., Binju Saju, Singh Navnit Dhirendra, Yuvraj Kaswan, Prajwal G C and S. P. Sreeja: “Machine Learning based prototype for customer segmentation using RFM”. 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), IEEE (2023).
[18] A. Agrawal, P. Kaur, and M. Singh: “Customer segmentation model using K-means clustering on E-commerce”. 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), IEEE, Erode, India (2023).
[19] Alzami, Farrikh et al., “Implementation of RFM method and K-Means algorithm for customer segmentation in E-Commerce with streamlit”. ILKOM Jurnal Ilmiah (2023).
[20] Sheshasaayee, A and Logeshwari, L., 2017. “An efficiency analysis on the TPA clustering methods for intelligent customer segmentation”. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, pp. 784–788.
[21] Tong, L., Wang, Y., Wen, F. and Li, X., Nov. 2017. “The research of customer loyalty improvement in telecom industry based on NPS data mining”. China Commun. 14 (11), 260–268.
[22] Hartoyo, H., Manalu, E., Sumarwan, U., and Nurhayati, P. (2023). “Driving success: A segmentation of customer admiration in automotive industry”. Journal of Open Innovation: Technology, Market, and Complexity, 9(2), 100031.
[23] Gustriansyah, R., Suhandi, N., and Antony, F. (2020). “Clustering optimization in RFM analysis based on k-means”. Indonesian Journal of Electrical Engineering and Computer Science, 18(1), 470–477.
[24] Shirole, R.; Salokhe, L and Jadhav, S. “Customer segmentation using RFM model and k-means clustering”. Int. J. Sci. Res. Sci. Technol. 2021, 8, 591–597.
[25] Vinit Dawane, Prajakta Waghodekar and Dr. Jayshri Pagare. (2021). “RFM analysis using k-means clustering to improve revenue and customer retention”. International Conference on Smart Data Intelligence. 3852887.
[26] D. Birant, “Data mining using RFM analysis,” in Knowledge-Oriented Applications in Data Mining, no. iii, K. Funatsu, Ed. In Tech, 2011, pp. 91–108.