



CUSTOMER SEGMENTATION USING MACHINE LEARNING

**V.K.G.Kalaiselvi¹, A. Ponmalar², Hariharan Shanmugasundaram³, Bhanuprasad A⁴,
Mamathibala V⁵, Swetha Sri M⁶,**

¹Assistant Professor, Department of Information Technology, Sri Sairam Engineering College, Chennai

²Assistant Professor, Department of Information Technology, Sri Sairam Institute of Information Technology, Chennai

³Professor, Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India

⁴Associate Professor, Department of Computer Science and Engineering, Vardhaman College of Engineering, India

⁵Student, Department of Information Technology, Sri Sairam Engineering College, Chennai

⁶Student, Department of Information Technology, Sri Sairam Engineering College, Chennai

Article DOI: <https://doi.org/10.36713/epra11198>

DOI No: 10.36713/epra11198

ABSTRACT

RFM (Recency, Frequency, Monetary) analysis is a method to identify high-response customers in marketing promotions, and to improve overall response rates, which is well known and is widely applied today. Less widely understood is the value of applying RFM scoring to a customer database and measuring customer profitability. RFM analysis is considered significant also for the banks and their specific units like online shopping. A customer who has visited an online shopping site Recently (R) and Frequently (F) and created a lot of Monetary Value (M) through payment and standing orders is very likely to visit and make payments again. After evaluation of the customer's behaviour using specific RFM criteria the RFM score is correlated to the online shopping, with a high RFM score being more beneficial to the online shopping as well as in the future. Data mining methods can be considered as tools enhancing the online shopping RFM analysis of the customers in total as well as specific groups like the users of online shopping.

KEYWORDS-*Data Mining online shopping, RFM analysis, Clustering*

I. INTRODUCTION

In recent years, there has been a massive increase in the competition among firms in sustaining in the field. The profits of the company can be improved by a customer segmentation model. Customer retention is more important than the acquisition of new customers. According to the Pareto principle [12], 20% of the customers contribute more to the revenue of the company than the rest. Customer segmentation can be performed using a variety of unique customer characteristics to help business people to customize marketing plans, identify trends, plan product development, advertising campaigns and deliver relevant products. Customer segmentation personalizes the messages of individuals to better communicate with the intended groups.

The most common attributes used in customer segmentation are location, age, sex, income, lifestyle and previous purchase behavior. Here, segmentation is done using behavioral data since it is commonly available and continuously evolving with time and purchase history. RFM (Recency, Frequency, and Monetary) analysis is a renowned technique used for evaluating the customers based on their buying behavior. A scoring method is developed to evaluate scores of Recency, Frequency, and Monetary. Finally, the scores



of all three variables are consolidated as RFM score ranging from 555 to 111 [11] is used to predict the future patterns by analyzing the present and past histories of the customer. In this context, it has been observed that the scores of three factors Recency, Frequency and Monetary directly proportional to customer's lifetime and retention. Once the values of recency, frequency and monetary are calculated, the K-Means algorithm is applied to the variables to clusters of the customer base. The behavior of each cluster is analyzed to find the group of customers who give more profits to the company. Similarly, clustering is performed using two other algorithms namely, Fuzzy C – Means clustering and the proposed method with chosen initial centroids in the existing K– Means algorithm. The motivation of the paper is to propose a method for choosing initial centroids for K-means algorithm and to impose the method to segment the customer with reduced iteration and time. Now that clusters of customers are found, it is necessary to understand the differences between these groups of customers. A thorough analysis is performed on the clusters to aid in finding the targeted customers and bestows them with appropriate promotions and offers. Also, a novel Repetitive Median based K-Means algorithm is proposed with an intension to reduce the number of iterations than the traditional clustering algorithms.

II. LITERATURE REVIEW

TABLE I - Literature Review

Reference No.	CHARACTERISTICS	DISADVANTAGES	ENHANCEMENTS PERSPECTIVE
[1]	<ul style="list-style-type: none"> Image conversion, pre processing - HOG Feature Extraction. Implementation of various other machine learning algorithms. 	<ul style="list-style-type: none"> Health and beauty company RFM analysis + K means clustering 	To propose a model that clustered customers into segments according to RFM.
[2]	<ul style="list-style-type: none"> Inclusion of image segmentation. SVM and Naive Bayes classifiers used 	<ul style="list-style-type: none"> A sample of 369 from the population 5939 Hotel customers RFM analysis + K means clustering. 	<ul style="list-style-type: none"> Aim to segment hotel customers Eight clusters were obtained according to their RFM score Loyal customers, loyal summer season customers, collective buying customers
[3]	<ul style="list-style-type: none"> Multiple machine learning classifiers used. Higher levels of accuracy achieved with deep learning models. Used CNN classifier instead of regular SVM 	<ul style="list-style-type: none"> A real data from a Chinese company RFM analysis + K means clustering + Decision tree 	<ul style="list-style-type: none"> To propose a model to accurately predict monthly supply quantity, using the RFM approach To select attributes to cluster customers into different groups.
[4]	<ul style="list-style-type: none"> Inclusion of image segmentation. SVM and Naive Bayes classifiers used 	<ul style="list-style-type: none"> Useful only in Retailing sector RFM analysis + K means clustering + Association rule 	<ul style="list-style-type: none"> Suggested a approach of customer classification.
[5]	<ul style="list-style-type: none"> Segmentation and analysis of the image is done. RaspberryPi module used. 	<ul style="list-style-type: none"> .Context: Iran Data from 250 bank customers. FM analysis + Two step clustering 	<ul style="list-style-type: none"> Aim only to identify the main clusters of bank customers in order to help classifying customers. To create more efficient customer strategies.



III. PROPOSED SYSTEM

Before focusing on customer segmentation, a question might arise as, why it is necessary to segment the customers? Because it is not possible to treat every customer the same, the customers will shift to a better option. Customers with experience on being available over various platforms will develop new tastes and that might grow into their personality to adapt. It is wise to adapt different initiatives in order to counter such. But there are very standard and effective segmentation procedures available. Here, implementation is done using one of them to a business. RFM,

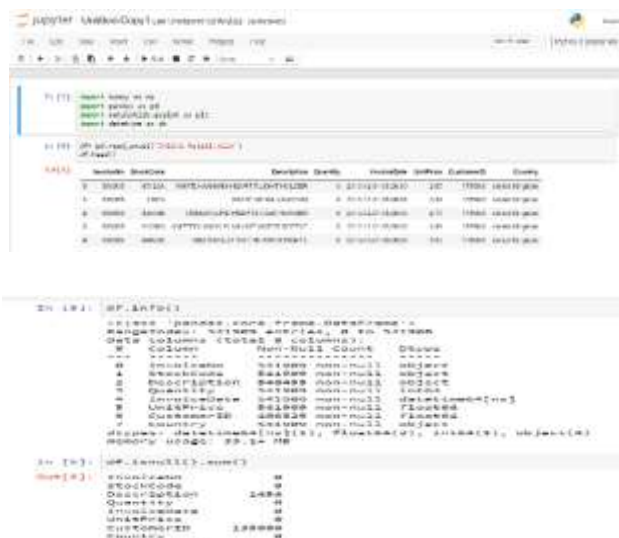
- Low Value: fewer active customers, very less frequent buyer/visitor the others and add very small or zero or sometimes even negative revenue.
- Mid Value: These are the grey area customers. They invest not too much and not too high and are frequent and add decent revenue.
- High Value: They produce high revenue in a company are frequent and hardly unavailable. The kind of not worth losing.

The proposed methodology can be broadly divided into 4 steps Step-1 is calculating Recency, Frequency, and Monetary Value. Step-2 is applying a machine learning algorithm that is unsupervised for detecting imique groups or clusters for each class segment, Step-3 is selecting a Machine Learning Model, and Step-4 is building and running the machine learning model. The analysis done in RFM (recency, frequency, and monetary values) model divides customers that are important on the basis of consumption of service by customers, the frequency of their visit and the amount of investment they introduce to the company. These act as the three variables which are separated from a huge chunk of database. Thus, the loyalty value of customers is obtained from recency, frequency and monetary value of customers. Hence considering Recency, Frequency, and Monetary value as the three criteria for getting the loyalty value of customers is the RFM model. The elaborated definitions would be,

Recency of last purchase (R): The interval between the latest buying and the present time of a customer is represented by it. The recency value increases if the interval value decreases,

Frequency of purchase (F). The time period of buying for a customer within a particular interval like twice in a week, once in a year is represented by Frequency. The value of F increase if the number of transactions within a specific interval increase too, Monetary value of the purchase (M): The monetary value is the purchases value of a service or a product in a specific time period. The value of M increases with the value of monetary.

IV. DATA PREPARATION







SQL [144]:

CustomerID	Revenue	Frequency	Monetary	R_Score	F_Score	M_Score	RFFScore
10000.0	3000	1	1000.00	1	1	1	100
10047.0	4021	1	1005.00	1	1	1	100
10048.0	3000	1	1011.00	1	1	1	100
10050.0	3000	1	1000.00	1	1	1	100
10056.0	1001	1	1000.00	1	1	1	100
10060.0	3000	1	1000.00	1	1	1	100
10061.0	3000	1	1000.00	1	1	1	100
10064.0	3000	1	1000.00	1	1	1	100
10065.0	1001	1	1000.00	1	1	1	100
10070.0	3001	1	1000.00	1	1	1	100

SQL [145]:

CustomerID	Revenue	Frequency	Monetary	R_Score	F_Score	M_Score	RFFScore
10071.0	3000	1	1000.00	1	1	1	100
10072.0	3000	1	1000.00	1	1	1	100
10073.0	3000	1	1000.00	1	1	1	100
10074.0	3000	1	1000.00	1	1	1	100
10075.0	1001	1	1000.00	1	1	1	100
10076.0	3000	1	1000.00	1	1	1	100
10077.0	3000	1	1000.00	1	1	1	100
10078.0	3000	1	1000.00	1	1	1	100
10079.0	3000	1	1000.00	1	1	1	100
10080.0	3000	1	1000.00	1	1	1	100

SQL [146]:

CustomerID	Revenue	Frequency	Monetary	R_Score	F_Score	M_Score	RFFScore
10081.0	3000	1	1000.00	1	1	1	100
10082.0	3000	1	1000.00	1	1	1	100
10083.0	3000	1	1000.00	1	1	1	100
10084.0	3000	1	1000.00	1	1	1	100
10085.0	3000	1	1000.00	1	1	1	100
10086.0	3000	1	1000.00	1	1	1	100
10087.0	3000	1	1000.00	1	1	1	100
10088.0	3000	1	1000.00	1	1	1	100
10089.0	3000	1	1000.00	1	1	1	100
10090.0	3000	1	1000.00	1	1	1	100

SQL [147]:

CustomerID	Revenue	Frequency	Monetary	R_Score	F_Score	M_Score	RFFScore
10091.0	3000	1	1000.00	1	1	1	100
10092.0	3000	1	1000.00	1	1	1	100
10093.0	3000	1	1000.00	1	1	1	100
10094.0	3000	1	1000.00	1	1	1	100
10095.0	3000	1	1000.00	1	1	1	100
10096.0	3000	1	1000.00	1	1	1	100
10097.0	3000	1	1000.00	1	1	1	100
10098.0	3000	1	1000.00	1	1	1	100
10099.0	3000	1	1000.00	1	1	1	100
10100.0	3000	1	1000.00	1	1	1	100

SQL [148]:

CustomerID	Revenue	Frequency	Monetary	R_Score	F_Score	M_Score	RFFScore
10101.0	3000	1	1000.00	1	1	1	100
10102.0	3000	1	1000.00	1	1	1	100
10103.0	3000	1	1000.00	1	1	1	100
10104.0	3000	1	1000.00	1	1	1	100
10105.0	3000	1	1000.00	1	1	1	100
10106.0	3000	1	1000.00	1	1	1	100
10107.0	3000	1	1000.00	1	1	1	100
10108.0	3000	1	1000.00	1	1	1	100
10109.0	3000	1	1000.00	1	1	1	100
10110.0	3000	1	1000.00	1	1	1	100

SQL [149]:

CustomerID	Revenue	Frequency	Monetary	R_Score	F_Score	M_Score	RFFScore
10111.0	3000	1	1000.00	1	1	1	100
10112.0	3000	1	1000.00	1	1	1	100
10113.0	3000	1	1000.00	1	1	1	100
10114.0	3000	1	1000.00	1	1	1	100
10115.0	3000	1	1000.00	1	1	1	100
10116.0	3000	1	1000.00	1	1	1	100
10117.0	3000	1	1000.00	1	1	1	100
10118.0	3000	1	1000.00	1	1	1	100
10119.0	3000	1	1000.00	1	1	1	100
10120.0	3000	1	1000.00	1	1	1	100

VI. FUTURE WORKS

A case study was carried out using the datasets collected within two years period by a sports store in Turkey through its e-commerce website. According to experimental study results, the proposed approach provides better product simple recommendations, by considering several parameters together; the customer's segment, the current RFM values of the customer, potential future customer behaviour, and products frequently purchased together.

VII. CONCLUSION

There is a vital role in customer segmentation for retail companies, Better segmenting of customers is pivotal in reaching a company's sales target. Companies get a better understanding of the target market if the customers that have equivalent requirements, necessities and behaviour are grouped together. Thus, companies could reevaluatethe current course ofaction and develop a new method for better sales, such as; update marketing, price management, promotions, building extra customer touchpoints, etc.

ACKNOWLEDGEMENT

The team is thankful and grateful for the constant encouragement of friends. The team also highly appreciates the guidance of the staff of the department of Information Technology, Sri Sairam Engineering College, Chennai.

**REFERENCES**

1. Etzion, O., Fisher, A., & Wasserkrug, S. (2004, Walk), *e-CLV: a modeling approach for client lifetime assessment in e-commerce spaces, with an application and case consider for online barbers*. In *innovation, e-Commerce, and e-Service, 2004. IEEE04. 2004 IEEE Worldwide Conference on* (pp. 149-156). IEEE,
2. Cui, G., Wong, M. L. & Lui, H. K. (2006). *Machine learning for direct marketing response models: Bayesian networks with evolutionary programming*. *Management Science*, 52(4), 597-612
3. Chen, Y. S., Cheng, C. H., Lai, C. J. Hsu, C. Y., Syu & H. J. (2012). *Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment*. *Computers in Biology and Medicine*, 42(2), 213-221.
4. He X., Li, C., 2016. *The research and application of customer segmentation on e-commerce websites*. In: *2016 6th International Conference on Digital Home (ICDH), Guangzhou*, pp. 203-208. doi:10.1109/ICDH.2016.050.
5. Cho, Young, Moon, S.C., 2013. *Weighted mining frequent pattern-based customer's RFM score for personalized u-commerce recommendation system*. *J. Converg*: 4, 36-40.
6. Zahrotun, L., 2017. *Implementation of data mining technique for customer relationship management (CRM) an online shop tokodiapers.com with fuzzy e-means clustering*. In: *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta*, pp. 299-303.
7. Sheshasaayee, A., Logeshwari, L., 2017. *Efficiency analysis of the TPA clustering methods for intelligent customer segmentation*. In: *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore*, pp. 784-788.
8. Lu, H., Lin, J.Lu., Zhang, G., May 2014, *A customer churns prediction model in the telecom industry using boosting*. *IEEE Trans. Ind. Inf.* 10 (2), 1659-1665.
9. Jiang, T., Tuzhilin, A. March 2009. *Improving personalization solutions through optimal segmentation of customer bases*. *IEEE Trans. Knowledge Data Eng.* 21 (3), 305-320. <https://doi.org/10.1109/TKDE.2008.163N>.
10. Shab. S., Singh. M., 2012. *Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm*. In: *2012 International Conference on Communication Systems and Network Technologies, Rajkot*, pp. 435-437.
11. G. Yu and M. Haiying, "Customer Segmentation Study of College Students Based on the RFM," in *International Conference on E-Business and E-Government, Guangzhou, China, 2010* pp. 3860-3863. doi: 10.1109/ICEE.2010.968.
12. Srivastava, R. (2016). *Identification of customer clusters using RFM model: a case of diverse purchaser classification*. *International Journal of Business Analytics and Intelligence*, 4(2), 45-50.