Customer Segmentation using RFM with K-Means clustering

Pawan Rama Mali
Race Academy of Corporate Excellence
REVA University
Bengaluru, India.
pawanramamali.ai01@race.reva.edu.in

Sabyasachi Sengupta
Race Academy of Corporate Excellence
REVA University
Bengaluru, India.
sabyasachisengupta.ai01@race.reva.edu.in

Jay.B.Simha
Race Academy of Corporate Excellence
REVA University
Bengaluru, India.
jb.simha@reva.edu.in

***Abstract*— Customer segmentation is the process of grouping customers by specific likeness (demographics, interests, behavior, etc.) Creating customer segmentation enables a business to target specific groups of customers and personalize marketing for each group. In order to execute and apply the scientific approach using K-Means algorithm, the real time transactional and retail dataset are analyzed. Spread over a specific duration of business transactions, the dataset values and parameters provide an organized understanding of the customer buying patterns and behavior across various regions. This study is based on the RFM (Recency, Frequency and Monetary) model and deploys dataset segmentation principles using K-Means Algorithm. The results thus obtained about sales transactions are compared with various parameters like Sales Recency, Sales Frequency and Sales Volume.**

Keywords—customer-segmentation, RFM model, k-means clustering

I. INTRODUCTION

An e-commerce business wants to keep customers active in purchasing. The business would like to understand customer purchasing behavior to prioritize marketing by customer groups and send relevant promotions to existing customers.
Segmenting customers will allow the e-commerce business to create personalized marketing for each individual group and marketing outreach can be prioritized for each group of customers based on the recency of their last purchase. Customer promotions can be personalized based on how often customers purchases and the average amount spend.

II. LITERATURE REVIEW

Customer segmentation is the first step in improving customer's journey and achieving customer engagement [1]. It is a process of dividing the customer base into distinct and homogeneous groups in order to develop differentiated marketing strategies based on their characteristics [2]. Customer segmentation intends to support and develop different business tasks or activities regarding marketing goals and can be analyzed by appropriate analytical techniques or tools.
Recency, Frequency and Monetary (RFM) analysis is a marketing technique in analyzing customer behavior such as how recently a customer has purchased, how often the customer purchases, and how much the customer spends. It could improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions [3]. The advantage is that the customers behavior can be captured by using a relatively small number of features, which improves the transparency of the target selection models that are developed [d]. The RFM variables are appropriate for capturing the specifics of the customers purchase behavior [d].

RFM was model was proposed by Hughes (1994), and it is a model that segments the customer depending on if the customer is important customer or not. It can be defined as

1) Recency of the last purchase (R). R represents recency, which refers to the interval between the time that the latest customer interaction and present. The shorter the interval is, the bigger will be the R value.

2) Frequency of the purchases (F). F represents frequency, which refers to the number of transactions in a particular period, for example, two times of one year, or two times of one quarter or two times in one month. The higher the frequency is, the bigger will be the F value.

3) Monetary value of the purchases (M). M represents monetary, which refers to total amount spent by customer in a particular period. The higher the monetary is, the bigger will be the M value. [6]

III. METHODS AND MODELS

This section talks about the proposed design flow, algorithm which has been used and the framework for this below experiment.

A. Selecting clustering Algorithm

K-means is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycenter's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position. [8][7][9]. We are using K-means algorithm to segment the customers buying pattern.

B. Methodology

Process of taking the online dataset to understand how customers are segmented, and then evaluate those customer segments as describe in Figure 1.

Step 1: Online customer data is chosen for our current experiment. On importing the data, we do exploratory data analysis to understand the data composition and spread, to come up with our RFM analysis. In this analysis we try to find unique customers, if there are any missing values or unexpected values which needs to be treated before our RFM segmentation.
Step 2: Data cleaning process will have removal of missing values, duplicate values and negative quantity values were also observed removed.

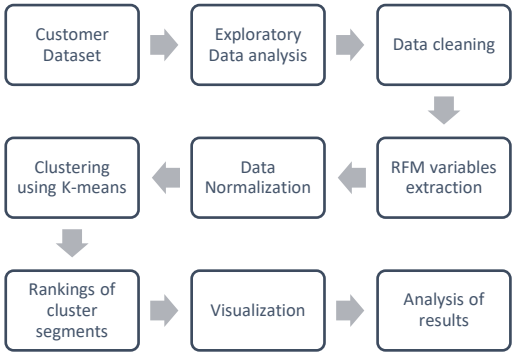

Figure 1 Proposed flow for customer segmentation analysis.

Step 3: RFM value extraction, is done after cleaning the data, to remove the data anomalies.

Recency- In order to check the recency of a customer, need to pick the max date from all the transactions present in our data set to calculate recency for each transaction for the customer by checking the most recent transactions.
Frequency- In order to calculate the frequency, calculate how frequently the customer has been purchasing. Analysis can be further added to add for individual stock codes as well.
Monetary- It refers to amount the spent by the customer. Total sum of the purchased products gives an indication how is the customer buying.
Step 4: On performing the RFM data analysis, it was noticed that data for RFM parameters have wide range due to base units are not same. Recency is in number of days, Frequency is in number of times customer purchased and monetary value is the currency amount. Normalization is performed on the dataset for RFM values to be scaled within 0 to 1. Min max feature scaling is used as per **Equation 1** to normalize the RFM values.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where x is an original value, x' is the normalized value.

Step 5: K-means clustering is the process of grouping a set of physical or abstract objects into groups of similar objects. A cluster is a collection of objects that are similar to one another within the same cluster (having similar features) and are dissimilar to the objects in other clusters [10]. K-means is one of the well-known algorithms for clustering, and it has been used extensively in various fields including data mining, statistical data analysis and other business applications. The K-means algorithm for partitioning is base on the mean value of the objects in the cluster. MacQueen sug- gested the term K-means for describing an algorithm that assigns each item to the cluster with the nearest centroid (mean)[l]. Based on the concept above, the computing process for K-means is presented as follows:
1: Partition the items into K initial clusters. Firstly, partition the items (m objects) into K initial clusters.
2: Proceed through the list of items. Assign an item to the cluster whose centroid is nearest (distance is computed by using Euclidean distance with either standardized or un-standardized observations) and re-calculate the centroid for the cluster receiving the new item or for the cluster losing the item.
3: Repeat Step 2 until no more reassigning. Rather than starting with a partition of all items into K pre- liminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2. The final assign- ment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points. Experience suggests that most major changes in assignment occur with the first reallocation step.

K-Means algorithm calculates its centers iteratively[11]. Let D={ x /i=1,…,n} be a data set
having K-clusters, C={ ic /i=1,…,K} be a set of K centers and F ={x/x is member of cluster k} be the set of j
samples that belong to the k-th cluster. K-Means algorithm minimizes the following function which is defined as a cost function:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - (\overline{x_\iota})|^2 \quad (2)$$

where x is the point of the data set, x is the centroid of I the i^{th} cluster. The K-means algorithm calculates cluster centers iteratively as shown in **Eq 2**.
Algorithm 1: Classic K-means algorithm
(1) Initialize K centers locations (1c ,…, kc) using random sampling.
(2) Assign each ix to its nearest cluster center k. c
(3) Calculate new kc centers as shown in **Eq. 3**.: =

$$C_k = \frac{\sum_{x_i \in F_k} x_i}{|F_k|} \quad (3)$$

(4) Repeat steps 2 and 3 till the cluster centers remain the same.

Step 5: Once the cluster segments have been formed, the cluster labels needs to be mapped to the original data set for all customers. RFM data columns are 3 hence ranking operation is performed on an m rows and 3 column matrix. Ranking of cluster groups is performed by applying the below **Eq 4.**

$$\sum_{n=0}^{2} x_{n+1} * 2^n \quad (4)$$

Where n is total number RFM columns =3 and x_{n+1} is the value for 1,2,3 column, respectively.

IV. Experiments and Results

The above methodology is applied on an online customer data set containing 541909 observations

Table 1 Dataset Description

Column Label	Description
InvoiceNo	Invoice number Nominal, a 6-digit integral number uniquely assigned to each transaction
StockCode	Item code Nominal, a 5-digit integral number uniquely assigned to each distinct product
Description	Item name Nominal
Quantity	The quantities of each Item per transaction Numeric
InvoiceDate	Invoice Date and time Numeric, the day and time when each transaction was generated
UnitPrice	Unit price Numeric
CustomerID	Customer number Nominal, a 5-digit integral number uniquely assigned to each customer
Country	Country name Nominal, the name of the country where each customer resides

Table 2

Index	CustomerID	Stock Code	Recency	Frequency	Monetary
0	12346	23166	325	1	77183.6
1	12347	16008	246	1	6
2	12347	17021	182	1	10.8
3	12347	20665	246	1	17.7
4	12347	20719	1	4	34

Table 1., shows the features and their description in the data set. ***Table 2***, shows the first five records of the customer and their product buying spread for Recency, Frequency and Monetary value

The data post cleaning process is transformed with log function (log (x)) so that recency, frequency and monetary data is normally distributed as shown in *Fig. 2.*

Data is then scaled using min max scaling function so that the values of recency, frequency and monetary lie between 0 and 1 as shown in *Fig. 3.*

One this data is scaled between 0 and 1, we try to segment the data by calling the k-means algorithm. Test is performed for 3,5,9 clusters respectively and ranking is on the cluster segments is applied by grouping the clusters and taking the mean of recency, frequency and monetary values.

For cluster value k=3

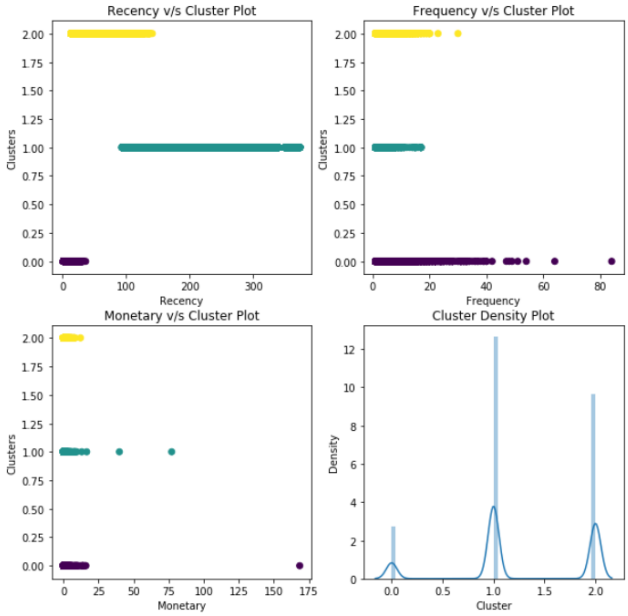

Figure 4

Table 3

index	cls	Recency	Frequency	Monetary	Customer Count	rank_ poswt
0	0	7.463203	2.734072	76.19875	29350	6
2	2	51.011849	1.511902	33.208884	102713	2.216 027
1	1	229.360044	1.205197	24.224616	134739	1

For cluster value k=5

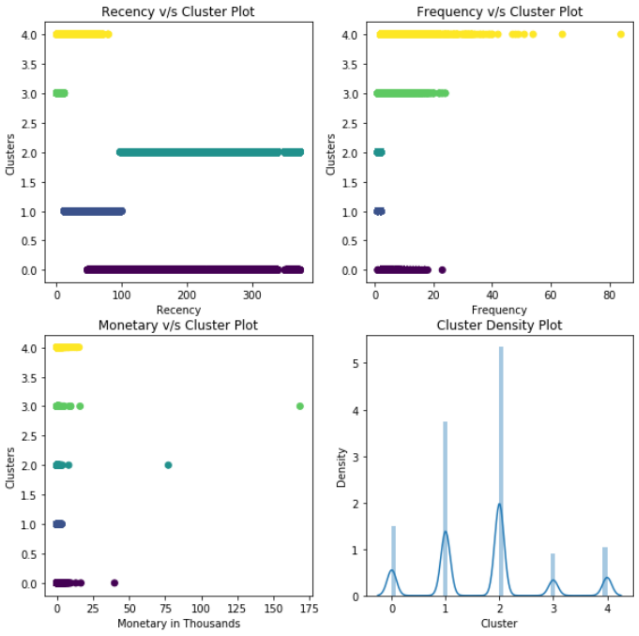

Figure 5

Table 4

index	cls	Recency	Frequency	Monetary	Customer Count	rank_poswt
4	4	23.157638	3.774637	107.043401	22355	6.402874
0	0	148.731605	2.559498	74.044394	31707	5.516657
3	3	4.816205	1.810403	45.972889	19130	3.107244
2	2	237.184612	1.010518	17.693472	113996	1.210526
1	1	50.225023	1.037004	16.005862	79614	0.641199

For cluster value k=9

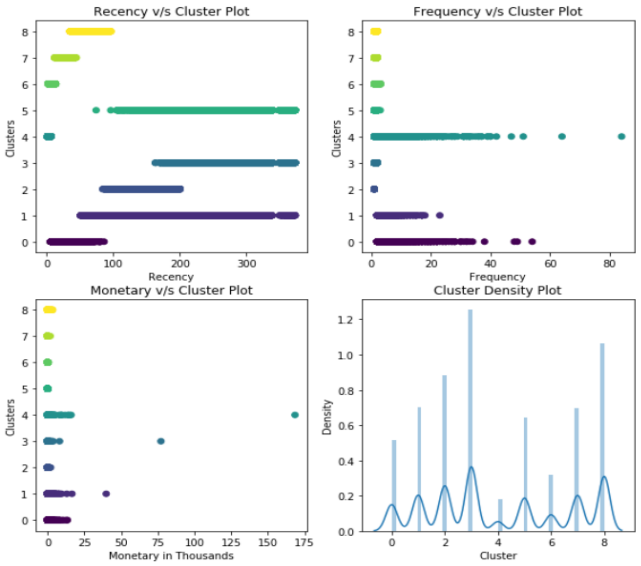

Figure 6

Table 5

index	cls	Recency	Frequency	Monetary	CustomerCount	rank_poswt
0	0	26.361737	3.549911	102.608541	21969	6.222753
4	4	1.916667	3.738279	126.255559	7764	6
1	1	146.940562	2.556091	73.425671	29880	5.70593
3	3	280.942164	1.017784	25.553387	53475	3.299406
2	2	138.140817	1	17.686718	37673	2.731582
8	8	63.098102	1.037635	17.728599	45463	2.63284
6	6	6.995326	1.197938	13.94327	13479	2.150371
7	7	25.461554	1.029947	12.490012	29652	2.061468
5	5	270.787194	1.041134	3.13091	27447	1.053233

Fig.4., shows the result for cluster k=3 plotting cluster spreads for recency, frequency and monetary values (in thousands). Also, it shows the overall cluster spread for the data. Cluster=1 having the density in terms of customers segmented to it. Ranking of segments is show in for *Table 3.* Its observed that cluster 0 is having highest rank among the other two clusters.

Fig.5., shows the result for cluster k=5 cluster spreads for recency, frequency and monetary values (in thousands). Cluster=5 having the density in terms of customers segmented to it. Ranking of segments is show in for *Table 4.* Its observed that cluster 4 is having highest rank among the other four clusters.

Fig. 6., shows the result for cluster k=9 cluster spreads for recency, frequency and monetary values (in thousands). Cluster=3 having the density in terms of customers segmented to it. Ranking of segments is show in for *Table 5.* Its observed that cluster 0 is having highest rank among the other eight clusters.

V. Conclusion

As per the analysis there are 3 major customer segments defined as per *Fig. 4.* for cluster 3. It is also observed that as cluster value increases for k=5 and k=9, cluster segments with higher recency gets higher rank, in comparison to clusters with lower recency by higher frequency and monetary value. Opportunity areas are present to work on a negative weight for recency so as define a correct ranking for a segment.

With these results, a marketing strategy can be proposed for the above experiment [5] to map the cluster segments. Customers with high monetary, high frequency and low recency values can be considered as 'Platinum' customers since they can improve marketing the product by word of mouth. For k=3, cluster 0 can be the platinum segment of customers. For k=5, cluster 4 can be the platinum segment of customers. For k=9, cluster 0 and 4 can be the platinum segment of customers.

Customers with low monetary, low frequency and high recency values are areas where marketing team needs to spread brand awareness for higher sales as the customers have not being buying the products too often as well as for a smaller price. For k=3, cluster 1 can be the target segment of such customers. For k=5, cluster 1 and 2 can be the target segment of such customers. For k=9, cluster 5 and 7 can be the segment of such customers.

With these we can get into deeper dive into future studies to understand the buying behavior of customer, thereby offering the customers the products as per their liking.

Table 6

RFM Pattern	Marketing Strategies	
	Short Term	Long Term
R↓F↑M↑ Platinum	• Improve word-of-mouth marketing • Product review • Promote new product and bundling program	• Give Rewards and customer loyalty program • Online and offline marketing treatment activities: book review, workshop, intimate meeting • Special customer services for complaint and live chat
R↑F↑M↑ Gold	• Cross selling and bundling program • Loyalty program offering	• Improve conversion to loyalty program • Live chat • Conversion to platinum customers
R↑F↓M↓ Iron	• Improve brand activation (awareness), creative marketing campaign • Flash sale program of discount product • Cross selling	• Improve word of mouth marketing to new customer • Improve cross selling product • Customer services using chat bot • Conversion to gold customers

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R.

B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] Fotaki G, Gkerpini N & Triantou A I 2012 Online customer engagement management (Netherland: Utrecht University)

[2] Tsiptsis K & Chorianopoulos A 2009 Data mining techniques in CRM: Inside customer segmentation

[3] Birant D 2011 Data mining using RFM analysis Knowledge-Oriented Applications in Data Mining InTech

[4] Kaymak U 2001 Fuzzy target selection using RFM variables Proc. Joint 9th IFSA World Congress and 20th NAFIPS Int. Conf. (Cat. No. 01TH8569) vol 2(C) pp 1038–43

[5] Customer Segmentation on Online Retail using RFM Analysis: Big Data Case of Bukku.id Mohamad Mohamad Abdul Kadir1, Adrian Achyar ICEASD 2019, April 01-02, Indonesia

[6] Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Applications, 36(3 PART 1), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

[7] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent Classification & Clustering Of Lung & Oral Cancer through Decision Tree & Genetic Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering,2015

[8] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques, International Journal of Computer Science & Mobile Computing,2015

[9] Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018, 135–139. https://doi.org/10.1109/CTEMS.2018.8769171

[10] [Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco: Morgan Kaufmann Publishers.

[11] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of5th berkeley symposium on mathematical statistics and probability (pp. 281–297). Berkeley: University of California Press.

[12] Qin, X., Zheng, S., Huang, Y., & Deng, G. (2010). Improved K-Means algorithm and application in customer segmentation. APWCS 2010 - 2010 Asia-Pacific Conference on Wearable Computing Systems, 224–227. https://doi.org/10.1109/APWCS.2010.63