An Enhanced Customer-Market Segmentation Architecture with Optimized Clustering

[a*]Neelkumar Gandhi, [b]Arjun Krishna SR, [c]Parthasarathy Govindarajan
[a,b] MTech Scholar, [c] Assistant Professor,
School of Computer Science and Engineering,
Vellore Institute of Technology
Vellore – 632014, Tamil Nadu, India
E-mail: neelkumar.g2022@vitstudent.ac.in, arjunkrishna.sr2022@vitstudent.ac.in, parthasarathy.g@vit.ac.in

Abstract

As information technology advances, too much must be generated, stored, and pre-processed to meet the demands of various companies. One marketing tool that might assist a company in promoting and gaining from sales efforts is segmentation. Understanding the idea of predictive modelling and knowing a way to leverage BD (Big Data) for segmentation is crucial for marketing professionals. The boundaries between various fields are becoming less defined, and there is an increasing amount of field interlinking. The goal of the project is to develop market and customer segments based on predictive modelling with BD that is already there within the organisation. A complex prediction model can be used to construct client segments, according to the findings of an empirical investigation. Market segmentation recognises that not all customers have the same interests, buying habits, or preferences. Market segmentation seeks to increase the strategic and targeted nature of a company's marketing initiatives in contrast to general coverage of all potential customers. A company can improve its chances of making sales and better manage its resources by creating specialized strategies for certain commodities in relation to target groups. Customer segmentation, which is the procedure of making cluster of same type thinking people. The clustering technique aids in improving comprehension of customers in terms of both static demographics and dynamic behaviour. The simple yet effective RFM approach may be used to segment the market. RFM analysis is used to examine consumer behaviour, including how lately (recency), frequently (frequency), and financially (spend) the customers make purchases (monetary). In this study, data mining has been used to group products into categories based on recent sales, frequent sales and total amount spent. This study has put out a brand-new k-Means methodology for RFM analysis. For e-commerce platforms, consumer segmentation may successfully save marketing costs while also increasing customer happiness. The output measure is compared with the existing RFM models.

Keywords: Customer segmentation, K-Means clustering, Data pre- processing, RFM, Data-mining

1. INTRODUCTION

An unsupervised kind of machine learning approach said as cluster analysis. Clustering seeks to create groupings or clusters within a data made up of data points. The cluster consists of similar kind of entities based on their attributes.

Selling products generates revenue for retail enterprises. The retail network is made up of a number of subsidiaries that are spread over many geographical areas. The company would not fully understand customer needs and market opportunities at these. The target consumer must be the company's primary focus in order to maximise profit while creating a win-win situation for both parties. One way to improve the outcome of a win-win situation is through customer segmentation.

2. Research Method

2.1 K-Means Clustering

The data is divided into several groups using the unsupervised type of learning technique K-Means Clustering. Take k is the optimal number of clusters. The concept of a centroid based algorithm will be explained in the working explanation of k-means. Each data point is there with the nearest k-center. Data points that are nearer to a specific k-center are used to construct different unique clusters.

2.2 RFM Analysis

Recency, Frequency, and Monetary (RFM) analysis is a behaviour-based procedure for segmenting customers. It organises the customers based on their previous purchase activities. How much, how recently, and how often a consumer purchased. To better serve

its clients, RFM categorises them into different groups. It helps managers find new clients so they may do more profitable business. There is a group of clients who make large purchases, but what if they just made one or how recently? Do they regularly purchase our goods? Additionally, it helps managers execute an effective advertising campaign for tailored service. [1][2]

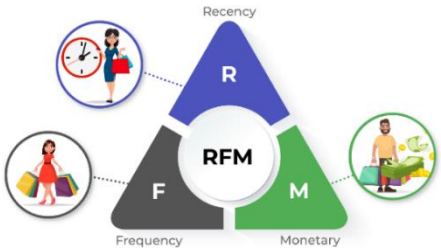

Figure 1. RFM

Image Source: meedium.com/capillary-data-science/58804480d232

Recency: When did the user most recently perform an activity (such as log in or place an order)?

Frequency: How frequently does the user do this action?

Monetary: What is the total monetary value of this user over the course of their lifetime?

2.3 Architecture

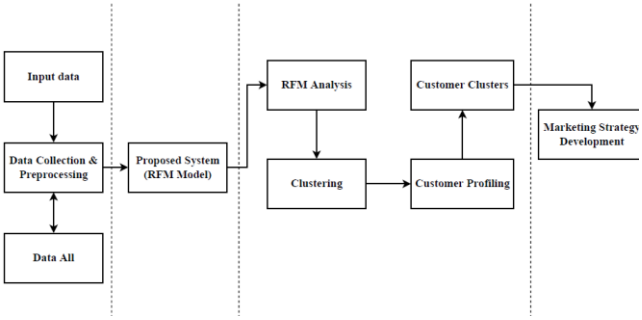

Figure 2. Proposed System

In above, proposed system we can observed that after pre-processing of data we apply RFM analysis and customer profiling we will get optimized clusters.

3. Data Analysis

3.1 About the Data

Dataset is made up of transactional data from December 2010 to December 2011. The dataset contains 540k transactions and 4372 Unique Customers.

Table 1. Online Retail Dataset

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Sales
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010	2.55	17850	United Kingdom	15.3
536365	71053	WHITE METAL LANTERN	6	01-12-2010	3.39	17850	United Kingdom	20.34
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010	2.75	17850	United Kingdom	22
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010	3.39	17850	United Kingdom	20.34
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010	3.39	17850	United Kingdom	20.34

Table 2. Data cleaning techniques

Sl. No.	Technique	Dataset	Description
1.	Data cleaning	Online Retail	Ignore the tuples, Fill the Missing values, dropna
2.	Data cleaning	Online Retail	Remove negative values
3.	Clustering	Online Retail	Noisy data (eliminate Outliers)

3.3 Data Preparation

3.3.1 RFM Calculation

Table 3. RFM Values for each customer

index	Customer ID	Recency	Frequency	Monetary
0	12346	326	2	0.0
1	12347	40	182	4310.0
2	12348	76	31	1797.24
3	12349	19	73	1757.55
4	12350	311	17	334.4

3.3.2 Data Normalization

Normalization fits all values into the new range by rescaling the data from the original range to a new estimation between 0 and 1. To do normalization, must be aware of or be ablee to correctly estimatee the max and min observable values. The data you have could be used to approximate these numbers.

Attributes are usually normalized to lie in a given estimation, typically from 0 to 1, multiplying all values by the highest value encountered or deleting the lowest value and dividing by the distance between the max and min values.

An expression for normalizing a value is:

$$y = (x - min) / (max - min) \quad (1)$$

We can see that the final value won't fall between 0 and 1 if an x value is supplied that is beyond the range of the lowest and maximum values. Prior to creating predictions, you might look for these observations and either exclude them from the dataset or restrict them to the predefined maximum or minimum values.

Table 4. Normalization

CustomerID	Recency	Frequency	Monetary
12346	0.4677187948350072	7.404281155364032e-06	0.0
12347	0.05738880918220947	0.0013401748891208897	0.002976175350340722 8
12348	0.10903873744619799	0.00222212843466092093	0.001241044405254376

CustomerID	Recency	Frequency	Monetary
12349	0.027259684361549498	0.0005531082431862103	0.0012136373519701477
12350	0.44619799139167865	0.00011846849848582451	0.0002309125376227233

3.4 Visualization of Data

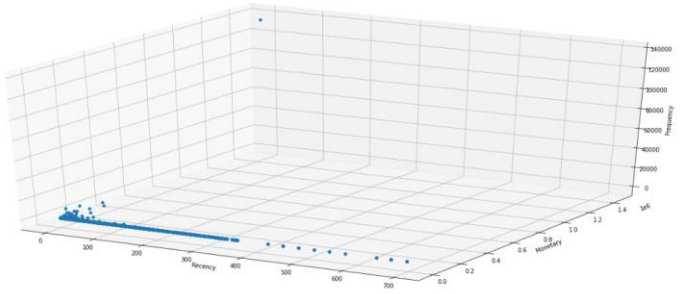

Figure 3. Consists of Outlier

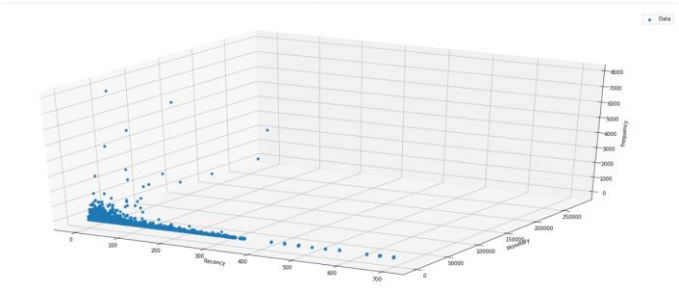

Figure 4. Removed Outlier

Here from above Figure 3 & 4, we can easily understands that firstly data consists of outliers, after finding that we had removed outliner.

4. Visualize Optimum number of clusters (Validity Indexes)

In this research method, four validity indicators are being captured to calculate the ideal numbers of cluster.

4.1 Elbow Method

For getting the ideal no. of clusters in k-means clustering or any other cluster analysis algorithm, apply the elbow procedure. The elbow procedure was applied for getting the ideal number of clusters in k-means clustering or any other cluster analysis algorithm. The average distortion was decreased and the instances moved as the elbow approach estimated the cost function using given range values of k which number of clusters increased.

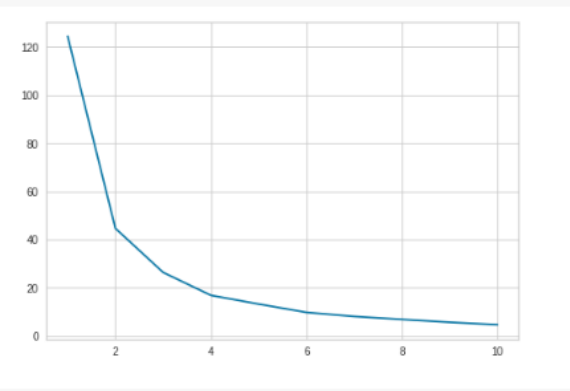

Figure 5. Elbow Method: Cost with various number of clusters

4.2 Davies Bouldin score

The information provided describes an estimation of cluster similarity as a function of intra-cluster dispersion and cluster disconnection.

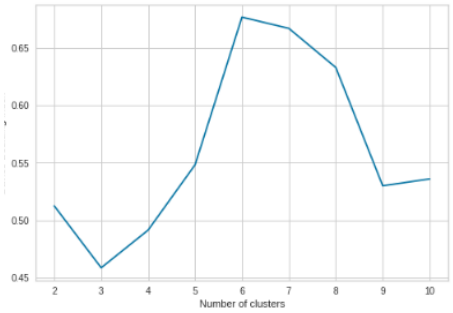

Figure 6. Davies Bouldin Index with various number of clusters

4.3 Calinski harabasz score (CHI Index)

By dividing the sum of the dispersion within each cluster, the sum of the dispersion between each cluster, the Variance ratio basis, also known as the Calinski-Harabasz index, is computed (where the dispersion is the sum of squared distances).

Calculating the inter-cluster dispersion or the sum of squares between groups is the first stage (BGSS).The weighted sum of squared distances between a cluster's centroid and the centroid of the whole dataset is what is referred to as the inter-cluster dispersion in CH (barycenter).

The formula to calculates the between group sum of squares is

$$BGSS = \sum_{k=1}^{K} N_k \times ||C_k - C||^2 \quad (2)$$

N_k: Quantity of observations for k-each cluster

C_k: cluster’s k ccentroid

C: dataset's ccentroid (barycenter)

K: total amount of clusters

$$\mathbf{CH} = [(\text{BGSS}/\ \text{WGSS})] \times [(N-K)\ /\ (K-1)]^{2} \qquad (3)$$

Between-Group Sum of Squares, or BGSS (between-group dispersion)

Within-group sum of squares (WGSS) (within-group dispersion)

N: total observations made

K: overall cluster count

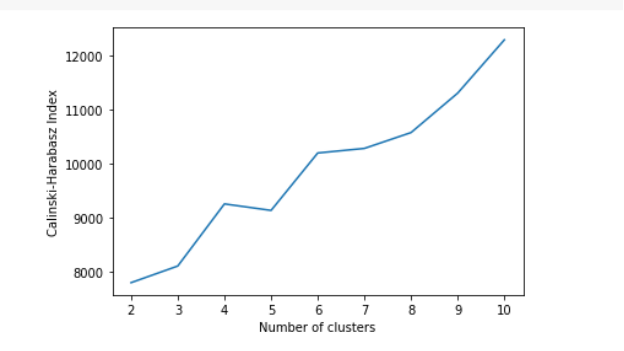

Figure 7. CHI Index with various number of clusters

As we can observed k=3 is having descent score

4.4 Silhouette score

The technique of analyzing and verifying consistency within data clusters is Silhouette score analysis. The coherence of an object within its own cluster in relation to other clusters is measured.

The Silhouette Coefficient is calculateed for each sample using the mean intra-clusterer range (xc) and the mean nearest-clusterer range (yc).

$$\text{SC(for i)} = (\ yc\ (i) - xc\ (i)\ )\ /\ (\ \text{maximum between}\ \{\ (\ xc(i), yc(i)\ )\ \}^{2} \qquad (4)$$

- xc(i) avg. dissimilarity between each object in a cluster and all the other objects

- yc(i) is avg. dissimilarity of i^{th} object with all objects in the closest cluster

score = 0.681

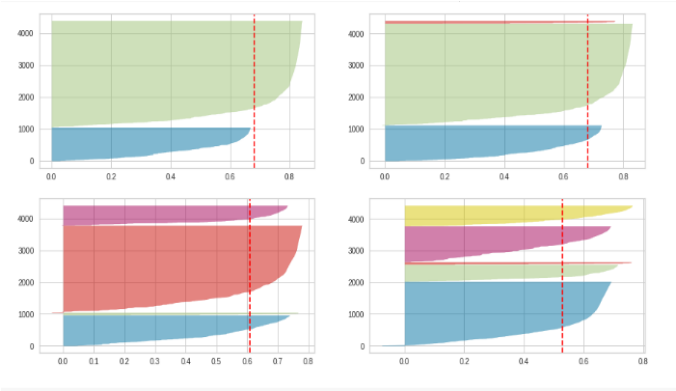

Figure 8. As we can observed k=3 is having descent score

5. Methodology & Results

From the real transactions dataset, we got all RFM Values and applied MinMaxScalar to that dataset. After getting optimal number of cluster, we have used K-Means clustering and get optimized clusters. The following steps are followed to implement our proposed system.

Table 5. Steps to be followed

Step 1	**:**	Clustering Algorithm
Step 2	**:**	Find Recency, Frequency, and Monetary Values for each data index.
Step 3	**:**	Apply normalization/standardization to newly created RFM Data frames.
Step 4	**:**	Find optimal no of cluster
Step 5	**:**	Apply hyper-parameters to K-means algorithm
Step 6	**:**	Optimized cluster as output.

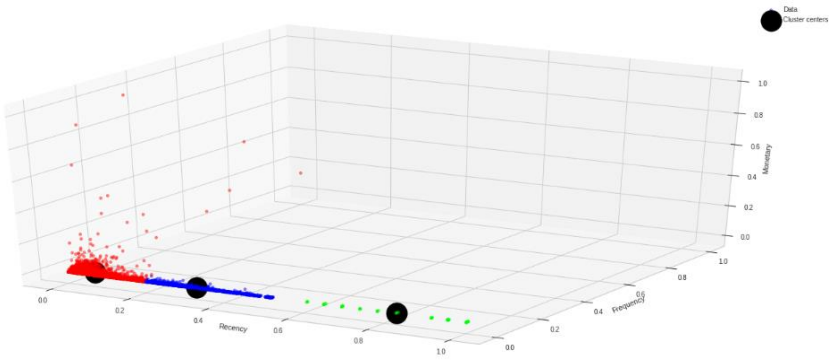

Figure 9. Optimized Clusters

As we can see that, above figure 9 shows optimized clusters as compared earlier clusters.

6. Cluster Quality

Uses the Euclidean distance (2-norm) as the measurement of distance between the two points.

The points in matrix X are arranged as dimensional row vectors.

dist = cdist(XA, XB, 'euclidean')

Table 6. Each data point's distance from its corresponding cluster centres using Euclidean Method

index	0	1	2
0	0.1376857116427518	0.40550872676173766	0.36272620239342174
1	0.27372771237839794	0.012063726856371294	0.7734665727006491
2	0.22110715892066063	0.04775578950313045	0.7214216288341161
3	0.30291705899568083	0.03578875827304983	0.8032255312072237
4	0.11611790610304712	0.38391484157273964	0.38423957012206195
5	0.22553953724036085	0.04237020639247372	0.725782706253438

Table 7. Variance comparison

Old Variance	**0.79**
New Variance	**0.37**

The Table 6 shows the variance values before and after applying our proposed system. The variance values is reduced, which shows that our algorithm played a vital role.

7. Conclusion

The essential idea of RFM analysis is that datas are split into three clusters of the ideal chunk. With the use of k-Means approach, this work has established a technique for RFM analysis. According to the assessment results, the optimal no. of clusters for the k-Means approach utilized in the RFM study is 3 clusters, with a variance value of 0.370.

8. Acknowledgement

We would like to convey our profound thanks to Head of the Department, Dean of our School for providing this opportunity and best environment.

9. References

[1] Eugene Wong, Yan Wei. (2018). Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction. International Journal of Retail & Distribution Management, ISSN: 0959-0552.

[2] Md Monir Hossain, Mark Sebestyen (2020) ... A Large-scale Data-driven Segmentation of Banking Customers. 2020 IEEE International Conference on Big Data (Big Data).

[3] Rendra Gustriansyah, Nazori Suhandi, Fery Antony (2020). Clustering optimization in RFM analysis based on k-means. Indonesian Journal of Electrical Engineering and Computer Science Vol. 18, No. 1, April 2020, pp. 470~477, ISSN: 2502-4752.

[4] Ismail Utku Sayan; Melike Demirdag; Guven Yuceturk; Sare Melek Yalcinkaya (2022). A Review of Customer Segmentation Methods: The Case of Investment Sector. 2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI).

[5] Yichen Xiao (2022). Hybrid Model for Customer Segmentation Based on RFM Framework. 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP).|

[6] Agneeswaran, V. S. (2012). Big-Data :Theoretical, Engineering and Analytics Perspective. In Big Data Analytics (pp. 8-15). Berlin Heidelberg: Springer.

[7] Brodie, R. J., Hollebeek, L. D., Juric, B., & Ilic, A. (2011). Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research. Journal of Service Research, 14(3), 252–271.

[8] Chen, H., Chiang, R., H., L., & Storey, V., C. (2012). Business Intelligence and Analytics: From Big Data to big impact, MIS Quaterly 36 (4), pp.1165-1188

[9] 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), 4176–4184.

[10] Gupta, R., Gupta, H., & Mohania, M. (2012). Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?. In Big Data Analytics (pp. 42-61). Springer Berlin Heidelberg.

[11] Ye, L., Qiuru, C., Haixu, X., Yijun, L., & Guangping, Z. (2013). Customer Segmentation for Telecom with the k-means Clustering Method. Information Technology Journal, 12(3), 409-413.

[12] Woo, J., Bae, S., & Park, S. (2005). Visualization method for customer targeting using customer map. Expert Systems with Applications, 28(4), 763–772.

[13] Pillai, J., & Vyas, O. P. (2012). CSHURI – Modified HURI algorithm for Customer Segmentation and Transaction Profitability, 2(2), 79–89.

Tables

Table 1. Online Retail Dataset

Table 2. Data cleaning techniques

Table 3. RFM Values for each customer

Table 4. Normalization

Table 5. Steps to be followed

Table 6. Each data point's distance from its corresponding cluster centres using Euclidean Method

Table 7. Variance comparison

Figures

Figure 1. RFM

Figure 2. Architecture

Figure 3. Consists of Outlier

Figure 4. Removed Outlier

Figure 5. Elbow Method: Cost with various number of clusters

Figure 6. Davies Bouldin Index with various number of clusters

Figure 7. CHI Index with various number of clusters

Figure 8. We can observed k=3 is having descent score

Figure 9. Optimized Clusters