Customer Segmentation using RFM analysis

Ms. Chaithra S[1], Mr. Hiffzull Rahman K H[2], Mr. Khalifa Abdul Musavvir[3]

[1]Assistant professor, Dept. of CSE, Yenepoya Institute of Technology, Moodibidri, India-574225
[2,3]Students, , Dept. of CSE, Yenepoya Institute of Technology, Moodibidri, India-574225

---***---

Abstract- *Customer performs a transaction process on a daily basis. There are 56,799 transactions in the process, which generates a lot of data. The goal of this research is to execute consumer segmentation using a data mining approach based on the RFM (Recency, Frequency, and Monetary) model and clustering techniques. Based on the RFM model, a total of 56,799 transactions were performed. Based on the RFM model, 910 best Customers were identified from 56,799 transactions that were conducted. We also looked at clusters, and found that Cluster 2 had 16,335 customers, Cluster 3 had 11,342 customers, Cluster 4 had 341 customers, Cluster 5 had 294 customers, and Cluster 6 had 782 customers. The findings of this study can be used by a corporation to determine customer categories, and then the business determine how to sustain customer engagement.*

Keywords: Data Mining; RFM Model; Cluster Analysis; Customer Segmentation;

I. INTRODCTION

Customers are the company's key priority in today's commercial competition to preserve its greatness. In order to serve customers, businesses must create and implement clear strategy [1]. Because the expense of acquiring new consumers is far higher than the cost of retaining existing customers, the company's primary focus is not on how to gain new potential customers but on how to sell more products to existing customers. [2]. Data can be obtained in the business based on previous data, so the data, such as transaction data from each agent, will continue to grow. The transaction process of agents in a server generates a large amount of data in the form of transaction profiles. This will happen to the company on a regular basis. The accumulation of agent transaction data will impede the search for information on that data. [3]. Data mining is a term used to describe the process of collecting and analyzing information. Data mining is a type of knowledge finding data that involves extracting information that is useful, previously unknown, and hidden from data. [4]. The unknown or concealed information can be discovered by processing the data so that it is beneficial to the business agent based on the quantity of available agent transaction data. [4], For example, information on the grouping of agent data has the ability to deliver the company the maximum profit, which will assist companies in making product marketing selections. The RFM (Recency, Frequency, Monetary) model was utilized by the researcher to execute the last visit time grouping, visit frequency, and revenue obtained by the company. [5]. The reason for continuing to utilise the RFM model is that it is simple to use and apply in businesses, and it is also simple to understand by managers and marketing decision makers. [6].

The findings of this study can be used in the business as a decision support system to map customers and identify potential customers..

II. LITERATURE REVIEW OF RFM MODEL

Some prior studies have used RFM to examine sales data, such as [8], where online sales were analyzed and the results were divided into eight groups. Cluster 7 has the highest RFM value in the entire cluster when compared to the other clusters. [7]'s work gives information for e-commerce entrepreneurs, allowing them to understand each type of customer. Then [8] utilized RFM to determine the worth of a customer at an airline. According to the findings of the study, there are four types of customers who require companies to provide distinct services.

In addition, the study [1] used RFM to process transaction data from exhaust sales, which were then clustered to classify the company's client types.

Recency of purchase, Frequency of purchase, and Monetary value of purchase are three simple customer attributes that the RFM approach is founded on. RFM's goal is to forecast future consumer behavior (and so make better segmentation decisions) [9]. As a result, it is vital to convert customer behaviour into a "number" that can be used consistently. In this example, the researcher sought to conduct the experiment utilising RFM Variable on a dataset of sale transactions with a large amount of data. Thousands of transactions are made each month. After the data is mapped using RFM variables, each client is categorised so that the organisation may know the category of each customer from the process.

IV. REVIEW OF CLUSTER ANALYSIS

Data mining is a technique for extracting and identifying usable information and associated knowledge from huge databases using statistics, mathematics, artificial intelligence, and machine learning approaches. Data mining is a type of knowledge finding data that involves extracting useful, unknown, and hidden information from data [4].

Data mining seeks to find a pattern or relationship that can provide important information [10]. Data mining is looking for a relationship between two or more variables in one dimension [10].

II. A CASE STUDY

The sales data from Nine Reload Server was used in this case study. There is a lot of data stacking at the organisation, with thousands of transactions per month. You can image how time-consuming it would be to manually analyze each piece of data. The researchers attempted to evaluate 56,799 client transactions in total. The proposed model for determining the profitable client is depicted in Figure 1, which displays the stages involved in determining the profitable customer.
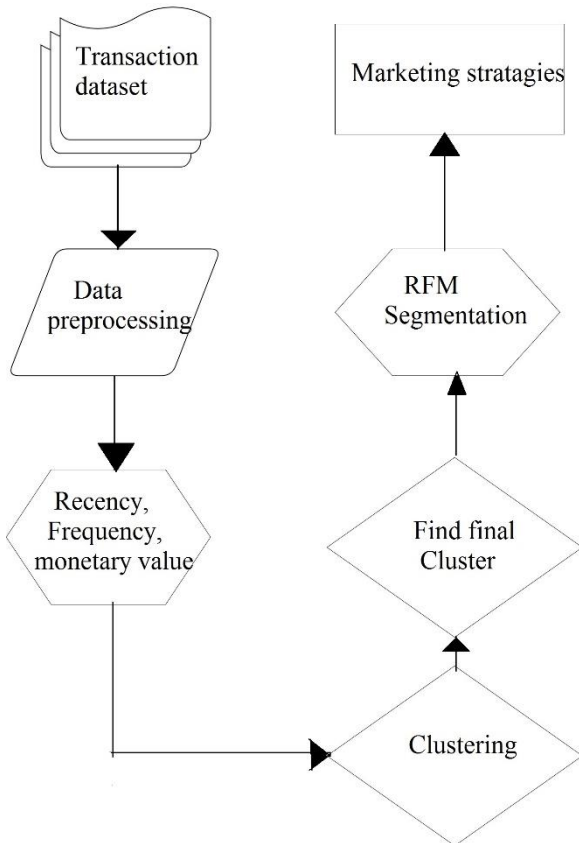

Fig 1. Framework for Customer Segmentation based on RFM analysis

In this study the database used is the data collected from the transaction as much as 56,799 sales transactions. Table 1 is an example of a sales transaction database.

Table 1. TRANSACTION DATASET

RetailStore	CustomerID	TransactionID	InvoiceDate	ProductID	Cost
1	1001	1ID10051	01-01-2019	1	144.75
2	1002	2ID10052	01-01-2019	2	97.25
2	1002	2ID10052	01-01-2019	3	47.25
2	1002	2ID10052	01-01-2019	4	197.25
3	1003	3ID10053	01-01-2019	5	216
3	1003	3ID10053	01-01-2019	6	137.25
1	1004	1ID10054	01-01-2019	7	253.25
2	1005	2ID10055	01-01-2019	8	147.25
3	1006	3ID10056	01-01-2019	9	192
3	1006	3ID10056	01-01-2019	10	101
3	1006	3ID10056	01-01-2019	10	101
3	1006	3ID10056	01-01-2019	11	202
3	1006	3ID10056	01-01-2019	12	101
3	1006	3ID10056	01-01-2019	13	242
2	1007	2ID10057	01-01-2019	5	216
2	1008	2ID10058	01-01-2019	14	197.25
2	1008	2ID10058	01-01-2019	15	147.25
2	1009	2ID10059	01-01-2019	16	197.25
2	1009	2ID10059	01-01-2019	17	197.25
2	1010	2ID10060	01-02-2019	18	210.75
2	1010	2ID10060	01-02-2019	5	197.25
2	1010	2ID10060	01-02-2019	19	106
3	1011	3ID10061	01-02-2019	20	252
1	1012	1ID10062	01-02-2019	20	199.75
1	1012	1ID10062	01-02-2019	20	199.75
1	1013	1ID10063	01-02-2019	21	142.25

Data Preparation

The database structure will be prepared at this point to make the mining process easier. Selection, pre-processing, and data transformation are the three key components of the preparation process. This procedure also includes the selection of qualities that are appropriate for the data mining procedure. Table 2 shows the attributes that were utilised.

Table 2. ATTRIBUTES USED

FIELD	INFORMATION
Customer ID	Used to specify the customer
Invoice Date	The date of Customers Transaction is used to model Recency and Frequency
Cost	To model the monetary attribute

In line with the RFM variable, the whole data transaction. The total of 56,799 transactions are then selected by RFM variable. Table 3 displays the dataset according to the characteristics of recency, frequency, and monetary value

Table 3. The Description of Recency, Frequency and Monetary

CustomerID	recency	frequency	monetary_value
1001	763	1	144.75
1002	309	16	2162.75
1003	763	2	353.25
1004	321	13	2673.50
1005	763	1	147.25
1006	763	6	939.00
1007	760	2	432.00

This study used a sales transaction history dataset with 56,799 transactions to determine criteria weighting based on recency, frequency, and monetary variables.

To decide the data that can be used, the overall data available on the transaction dataset must be selected first. Table 4 shows how the weighting was separated into four scales/scores.

Table 4. Decision Table

Weight	R (*Recency*)	F (*Frecuency*)	M (*Monetary*)
4	rare	lowest	fewest
3	long	low	few
2	regular	regular	normal
1	shortest	highest	many

Once the scale is found, the next step is to transform its data on the existing scale. Table 5 shows the sample data transformed.

Table 5. Example R-F-M values of some customers after data preprocessing

CustomerID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile
1001	763	1	144.75	4	4	4
1002	309	16	2162.75	2	1	1
1003	763	2	353.25	4	4	3
1004	321	13	2673.50	2	1	1
1005	763	1	147.25	4	4	4

After all transaction data is transformed into numeric form, to be able to group these data into several clusters needs to do some steps

1. In this study the existing data will be grouped into six clusters based on RFM score as shown in table 6.

Table 6. RFM score

CustomerID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
1001	763	1	144.75	4	4	4	444
1002	309	16	2162.75	2	1	1	211
1003	763	2	353.25	4	4	3	443
1004	321	13	2673.50	2	1	1	211
1005	763	1	147.25	4	4	4	444

2. The clusters are made by considering scores of all three metrics that is recency, frequency and monetary. Based on weights from 1 to 4 of RFM.

3. By clubbing together RFM scores six customers are made which is shown in table 7. Which further can be used for marketing strategies which is also explained in the Table.

Table 7. RFM METRICS

Segment	RFM	Description	Marketing
Best Customers	111	Bought most recently and most often and spend the most	No price incentives, new products, and loyalty programs
Loyal Customers	X1X	Buy most frequently	Use R and M to further segment
Big Spenders	XX1	Spend the most	Market your most expensive products
Almost Lost	311	Haven't Purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost customers	411	Haven't Purchased for long time, but purchased frequently and spend the most	Aggressive price incentives
Lost cheap customers.	444	Last purchased a long ago, purchased few and spent little	Don't spend too much trying to re-acquire

In the below Table 8, Table 9, Table 10, Table 11, Table 12 and Table 13, there are samples of customers that are in clusters in which the data can be utilized by the company.

Table 8. Best customers

CustomerID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
2708	41	191	33096.50	1	1	1	111
5553	34	174	32740.50	1	1	1	111
1758	33	229	25505.50	1	1	1	111
4786	58	27	25150.75	1	1	1	111
4684	42	140	21468.75	1	1	1	111
...	...	...	...	...	...	...	...
9589	40	9	1065.75	1	1	1	111
5173	154	7	1053.25	1	1	1	111
9024	74	8	1048.75	1	1	1	111
7823	131	7	1045.75	1	1	1	111
5089	38	8	1041.50	1	1	1	111

Table 9. Loyal customers

recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
471	8	1038.25	3	1	2	312
361	7	1033.00	3	1	2	312
399	7	1030.75	3	1	2	312
344	7	1028.50	3	1	2	312
354	7	1023.75	3	1	2	312
467	7	1023.00	3	1	2	312
453	8	1019.25	3	1	2	312
327	8	1013.25	3	1	2	312
453	7	1011.25	3	1	2	312
355	7	994.50	3	1	2	312
486	8	992.75	3	1	2	312
385	8	990.25	3	1	2	312
468	7	980.75	3	1	2	312
360	7	980.75	3	1	2	312
476	7	965.25	3	1	2	312
471	7	957.00	3	1	2	312
432	10	954.25	3	1	2	312
389	8	953.00	3	1	2	312
336	7	933.25	3	1	2	312
378	8	923.25	3	1	2	312
344	8	915.50	3	1	2	312
403	7	913.00	3	1	2	312
379	7	901.00	3	1	2	312
333	7	895.75	3	1	2	312

Table 10. Big Spenders

CustomerID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
7609	178	5	5624.00	2	2	1	221
7472	241	6	5148.50	2	2	1	221
11250	317	4	4148.50	2	2	1	221
3344	229	4	3999.00	2	2	1	221
8093	199	4	3149.00	2	2	1	221
...	...	...	...	...	...	...	...
7162	183	6	1053.50	2	2	1	221
3718	197	5	1053.25	2	2	1	221
8446	175	6	1048.50	2	2	1	221
8519	168	6	1042.75	2	2	1	221
4228	294	6	1042.25	2	2	1	221

Table 11. Lost cheap customer

CustomerID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
4173	512	2	347.00	4	4	4	444
2936	631	2	346.75	4	4	4	444
4361	493	2	344.50	4	4	4	444
1587	714	2	344.50	4	4	4	444
3161	611	2	344.50	4	4	4	444
...	...	...	...	...	...	...	...
1340	737	2	114.50	4	4	4	444
1338	737	1	99.75	4	4	4	444
1214	747	1	97.25	4	4	4	444
2108	687	1	47.25	4	4	4	444
2547	658	1	42.25	4	4	4	444

Table 12. Almost lost

CustomerID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
10635	328	27	37593.25	3	1	1	311
4984	362	8	13498.00	3	1	1	311
2143	488	7	10123.25	3	1	1	311
2365	444	57	7996.25	3	1	1	311
5111	375	38	7652.50	3	1	1	311
...	...	...	...	...	...	...	...
10650	350	9	1086.75	3	1	1	311
10795	359	9	1064.25	3	1	1	311
5980	334	11	1052.00	3	1	1	311
5002	416	8	1046.50	3	1	1	311
10648	370	8	1043.50	3	1	1	311

Table 13. Lost customers

CustomerID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
2774	496	50	12037.00	4	1	1	411
1612	546	61	11740.50	4	1	1	411
2384	628	58	10330.50	4	1	1	411
2005	514	65	8451.25	4	1	1	411
1591	630	13	8172.00	4	1	1	411
...	...	...	...	...	...	...	...
3010	625	8	1063.00	4	1	1	411
3847	542	7	1058.50	4	1	1	411
1223	602	8	1051.50	4	1	1	411

CONCLUSION

The major goal of this study was to use the RFM model to segment customers from a total of 56,799 transaction data sets, as well as perform clustering analysis..

910 consumers in cluster 1,16,335 customers in cluster 2, 11,342 customers in cluster 3, 341 customers in cluster 4, 294 customers in cluster 5, and 782 customers in cluster 6 are the results of this study.

It is assumed that the company will be able to make the best marketing strategy decisions if it understands the categories of each customer.

REFERENCES

[1] Maryani, Ina, and Dwiza Riana. 2017. "Clustering and Profiling of Customers Using RFM for Customer Relationship Management Recommendations." 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017, 2–7. https://doi.org/10.1109/CITSM.2017.8089258.

[2] Tama, Bayu Adhi. 2010. "Penetapan Strategi Penjualan Menggunakan Association Rules Dalam Konteks CRM." Jurnal Generic Vol. 5 (No.1):35–38.

[3] Hand, David J. 2007. "Principles of Data Mining." Drug Safety 30 (7):621–22. https://doi.org/10.2165/00002018-200730070-00010.

[4] Ramamohan, Y, K Vasantharao, C Kalyana Chakravarti, and a S K Ratnam. 2012. "A Study of Data Mining Tools in Knowledge Discovery Process." International Journal of Soft Computing and Engineering 2 (3):191–94.

[5] Wongchinsri, Pornwatthana, and Werasak Kuratach. 2016. "A Survey -Data Mining Frameworks in Credit Card Processing." 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2016. https://doi.org/10.1109/ECTICon.2016.7561287.

[6] Peiman Alipour Sarvari, Alp Ustundag, and Hidayet Takci. 2014. "Performance Evaluation of Different Customer Segmentation Approaches Based on RFM and Demographics Analysis." Kybernetes 43 (8):1209–23. https://doi.org/10.1108/K-01-2015-0009

[7] Rachid, et al. 2015. "Combining RFM Model and Clustering Techniques for Customer Value Analysis of a Company selling online." 2015 12th International Conference of Computer Systems and Applications (AICCSA) 2015,1-6.

[8] Liu Jiali and Du Hyung. 2010. "Study on Airline Customer Value Evaluation Based on RFM Model (2010)." 2010 International Conference On Computer Design And Appliations (ICCDA 2010) ,278-281

[9] Aviliani, U. Sumarwan, I. Sugema, and A. Saefuddin. 2011. "Segmentasi Nasabah Tabungan Mikro Berdasarkan Recency, Frequency, dan Monetary : Kasus Bank BRI." Finance and Banking Journal 13 (1):95–109.

[10] Kusrini Luthfi, Ema Taufiq. 2009. Algoritma Data Mining. Edited by Theresia Ari Prabawati. Yogyakarta: C.V Andi OFFSET.