RESEARCH ARTICLE OPEN ACCESS

Customer Segmentation Approach for Finding Loyal Customers Using RFM and K-Means Clustering Technique

[1]Ms Sarika Rathi, [2] Asst. Prof. Vijay S. Karwande

1(Department of Computer Science &Engineering , ME Student in Everest College of Engineering & Technology , Aurangabad , Maharashtra,IndiaEmail:rathisarika11@gmail.com)
2(Department of Computer Science &Engineering, HOD in Everest College of Engineering & Technology, Aurangabad, Email:hodcse@eescoet.org)

Abstract:

Now a day's competition is huge and all companies are shifting ahead with their personal exclusive techniques. We must use this information and take right decision. Absolutely each and every human beings isn't like one another and we don't recognize what customer buys or what their choices are. However, with the assist of system learning technique you can possibly type out the statistics records and may locate the goal group by making use of numerous algorithms to the dataset.

The goal of this project is to analyse and segment the customers of an e-commerce company by using the RFM approach of given data. This will enable the e-commerce company to optimize their retention and acquisition strategies. Many businesses get most of their revenue from their 'best' or regular customers. Since the resources that a company has, are limited, it is crucial to find these customers and target them. It is equally important to find the customers who are dormant/are at high risk of churning to address their concerns. For this purpose, here we are trying to build an unsupervised learning model which can enable company to analyse their customers via RFM (Recency, Frequency and Monetary value) approach.To analyse this approach and proper grouping of customer segmentation. We are also focusing on all the cancelled orders which contain negative quantities are removed from the data. As expected, customers with the lowest RFM scores have the highest Recency value and the lowest frequency and monetary value, and the vice-versa is true as well. So, here we are trying to create segments within this score range of RFM_Score 3–12, by creating categories in our data and segmenting them as 'Good customers', 'Best customers', 'Loyal customers', and 'Lost customers' .

Keywords — **Clustering, Elbow Method, K-Means Algorithm, Customer Segmentation, Visualization,Good customers, Best customers Loyal customers, and Lost customers.**

I. INTRODUCTION

In recent times the competition is massive and lot of technologies came under consideration for powerful increase in sales era. For every business the most essential issue is data. With the assist of grouped or ungrouped statistics, we can carry out some operations to discover customer interests. One of this is Data mining beneficial to extract facts from the database in a human readable layout. However, we won't recognize the real beneficiaries inside the entire dataset.

Customer is always the first priority of every business; it has also proven many times that customer-oriented organizations are successful and ever growing in cooperate world. This was figured out by many organizations and they are trying to implement customer centric approach as their work criteria. To get into the shoes of the customers and try to merge according to the new trend followed by customers is constantly in generating huge profit.

For a small company, the customer base is usually quite small and individually targetable. But, as a business grows in size, it will not be possible for the business to have an intuition about each and every

customer. At such a stage, human judgments about which customers to pursue will not work and the business will have to use a data-driven approach to build a proper strategy.

For a medium to large size retail store, it is also imperative that they invest not only in acquiring new customers but also in customer retention. Many businesses get most of their revenue from their 'best' or high-valued customers. Since the resources that a company has, are limited, it is crucial to find these customers and target them. It is equally important to find the customers who are dormant/are at high risk of churning to address their concerns.

Many businesses get most of their revenue from their 'best' or regular customers. Since the resources that a company has, are limited, it is crucial to find these customers and target them. It is equally important to find the customers who are dormant/are at high risk of churning to address their concerns. For this purpose, companies use the technique of customer segmentation.

As it is well known by marketers, customers have various kinds of needs and wants. Companies have used several segmentation criteria and techniques to better identify and understand customer groups and provide preferable products and services to them in order to satisfy these different needs and wants. Also, segmentation is important that the company can create profitable segments and react to the selected segment based on its competitive advantages. However, many marketers have difficulty in identifying the right customer segments to organize marketing campaigns.

II. LITERATURE SURVEY

In recent research we have found that a case study of using data mining techniques to segment customers for an IT solution is presented. The objectives of this research are to construct a customer segmentation model based on 3 customer demographics and purchase behaviors and to help business better understand its customers and support their customer-centric marketing strategy. Regarding to the customers demographic data and RFM values generated from purchase behaviors, customers have been segmented using the K-means clustering technique into numerous groups based on their similarity, and the profile for each group is identified based on their characteristics. Accordingly, recommendations are provided to the business on marketing strategy and further analysis.

In today's business competition, customers are the main focus of the company to maintain its excellence. Companies must plan and use clear strategies in serving customers. [1][2]. The company's primary focus is not on how to get new potential customers but how to sell more products to the existing customers because the cost that companies must in curt acquire new customers is much more expensive than to retain existing customers.[3]

The model used by the researcher is RFM (Recency,Frequency, Monetary) commonly used to perform the last visit time grouping, visit frequency, and revenue obtained by the company [4]. There as on why continuing to use the RFM model is that it is easy to use and quickly implemented in companies, and in addition RFM is easily understood by managers and marketing decision makers[5].

RFM model is extremely useful in ceitbuilds customer segmentation model effectively. However, simplicity threatens the power of RFM and the models need to be made and be improved by a manual process. Moreover, RFM models cannot confront with changes in the business and managers should handle them by adhoc decisions. In this paper, we found the best definitions for R, F and M to have a dynamic RFM model and also using K-Means in order to propose R+FM model which builds customer segmentation model dynamically.[6]

There are 4customer categories that demand company to give different service to customer. RFM technique is based on three simple customer attributes, namely Recency of purchase, Frequency of purchase, and Monetary value of purchase. The values of recency, frequency and monetary are combined to form RFM scores. For example, in a five category ranking system, there are about 125 possible RFM scores and the highest RFM score is 555. RFM scores clearly shows the categories of

different consumers. The best customer search with the highest RFM scores. In this paper, the ranking1-4 isused to evaluate the customer retention.[7] Educational e-commerce products have strong social attributes. Therefore, constructing a new parameter “C” of social group dimension can help enterprises measure the social radiation breadth of customers which realizes the rapid promotion of products through customers' spontaneous grouping behavior. Based on the improved customer segmentation RFMC model, the appropriate customer clustering method is determined according to the actual needs of the enterprise. In this paper, the K-means clustering method is selected and the final value of “K” (the number of customer groups) is determined according to the elbow method, and clustering analysis is carried out on this basis.[8]

P. Anitha et al. applied the k-means algorithm to the RFM model to evaluate the buying behaviour of users in different regions [9]. Siti Monalisa et al. applied the RFM model to property insurance, customer investment, telecommunications services , healthcare , and FMCGindustries [10]. 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) . 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.[11]
A key consideration for a company would be whether or not to segment its customers and how to do the process of segmentation. This would depend upon the company philosophy and the type of product or services it offers. The type of segmentation criterion followed would create a big difference in the way the business operates and formulates its strategy.

A. Why Customer Segmentation

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes. The groups should be homogeneous within them selves and should also be heterogeneous to each other. The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable.

Customer segmentation has a lot of potential benefits. It helps a company to develop an effective strategy for targeting its customers. This has a direct impact on the entire product development cycle, the budget management practices, and the plan for delivering targeted promotional content to customers. For example, a company can make a high-end product, a budget product, or a cheap alternative product, depending upon whether the product is intended for its most high yield customers, frequent purchasers or for the low-value customer segment. It may also fine-tune the features of the product for fulfilling the specific needs of its customers.In a world where everyone has hundreds of emails, push notifications, messages, and ads dropping into their content stream, no one has time for irrelevant content.

Finally, this technique can also be used by companies to test the pricing of their different products, improve customer service, and upsell and cross-sell other products or services.

In this research, a case study of using data mining techniques to segment customers for an IT solution is presented. The objectives of this research are to construct a customer segmentation model based on 3 customer demographics and purchase behaviors and to help business better understand its customers and support their customer-centric marketing strategy. Regarding to the customers demographic data and RFM values generated from purchase behaviors, customers have been segmented using the K-means clustering technique into numerous groups based on their similarity, and the profile for each group is identified based on their characteristics. Accordingly, recommendations are provided to the business on marketing strategy and further analysis.

III. CASE STUDY

In the given analysis, I am going to use the Online Retail Data Set, which was obtained from the UCI Machine Learning repository. The data contains information about transnational transactions for a UK-based and registered non-store online retail. In this study the database used is the data collected from the transaction as much as 5,37,979 customers id and with 12 different attributes are there.Table 1is an example of a database.Where as Table II shows attributes used for dataset.

TABLE I
SAMPLE DATA SET FOR ANALYSIS

CustomerID	Item Code	InvoieNo	Date of purchase	Quantity	Time	price per Unit	Price	Shipping Location	Cancelled_status	Reason of return	Sold as set
4355	15734	398177	29-10-2017	6	3:36:00 PM	321	1926	Location 1			
4352	14616	394422	05-10-2017	2	2:53:00 PM	870	1740	Location 1			
4352	14614	394422	12-10-2017	2	2:53:00 PM	933	1866	Location 1			
4352	85014B	388633	22-08-2017	3	2:47:00 PM	623	1869	Location 1			
4352	15364	394422	10-10-2017	2	2:53:00 PM	944	1888	Location 1			
4349	14618	397122	27-10-2017	1	12:43:00 PM	256	256	Location 1			
4343	15364	368432	13-02-2017	-4	2:46:00 PM	922	-3688	Location 1	TRUE		
4341	85014B	377109	14-05-2017	3	9:22:00 AM	677	2031	Location 1			
4341	85014A	377109	12-05-2017	3	9:22:00 AM	692	2076	Location 1			
4341	85014B	390217	07-09-2017	6	2:47:00 PM	670	4020	Location 1			

TABLE III
ATTRIBUTES OF SAMPLE DATA SET

Column Name	Description
CustomerID	Unique idetifier for each Customer
Item Code	Unique id for each product
InvoieNo	Unique id for each purchase
Date of purchase	Date on which the purchase was made
Quantity	Number of items bought for each product
Time	Time at which the purchase was made
price per Unit	Price of single unit of item purchased
Price	total purchase price
Shipping Location	Delivery Loaction
Cancelled_status	Status of Cancellation
Reason of return	Reason for return of product
Sold as set	Was the product sold with another product/ Offer

A. *Data Set Details*

This data set consists of 537979 observations and 12 variables.

The data set also consist of some duplicate entries & Null values.

Provided data set is between the period 02-12-2016 to 19-12-2017 (Approx 1 year).

B. **Exploring the data:**

Before diving into insights from the data, duplicate entries were removed from the data. The data contained 9 duplicate entries. Again the Null values and the empty columns were dropped. After processing remaining data in dataset is 4324 customers with 3637 products abd 18305 transactions. This means that each product is likely to have multiple transactions in the data. There are almost as many products as customers in the data as well. When we try to find data based on transactions for a different locations , From Figure 1 it is observed that 85% quantity are deliverd to Location36. Therefore, for the purpose of this analysis, I will be taking data corresponding quantity to location 36.

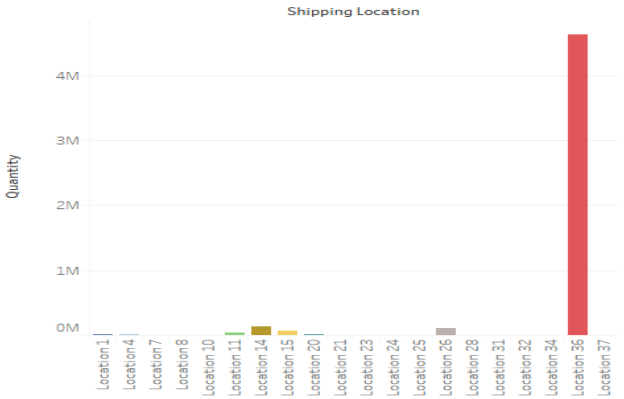

Fig. 1 Shippling Location36 quantity

IV. RFM SEGMENTATION

RFM stands for Recency, Frequency, and Monetary. RFM analysis is a commonly used technique to generate and assign a score to each customer based on how recent their last transaction was (Recency), how many transactions they have made in the last year (Frequency), and what the monetary value of their transaction was (Monetary).

RFM analysis helps to answer the following questions: Who was our most recent customer? How many times has he purchased items from our shop? And what is the total value of his trade? All this information can be critical to understanding how good or bad a customer is to the company.

After getting the RFM values, a common practice is to create ‘quartiles’ on each of the metrics and assigning the required order. For example, suppose that we divide each metric into 4 cuts. For the recency metric, the highest value,

4, will be assigned to the customers with the least recency value (since they are the most recent customers). For the frequency and monetary metric, the highest value, 4, will be assigned to the customers with the Top 25% frequency and monetary values, respectively. After dividing the metrics into quartiles, we can collate the metrics into a single column (like a string of characters {like '213'}) to create classes of RFM values for our customers. We can divide the RFM metrics into lesser or more cuts depending on our requirements. We create quartiles on this data as described above and collate these scores into an RFM_Segment column. The RFM_Score is calculated by summing up the RFM quartile metrics. Following figure shows values of RFM with respect to RGMgroup and RFMScore with irsRecency,Frequency and Monetary value.

CustomerID	Recency	Frequency	Monetary	R	F	M	RFMgroup	RFMScore
2.0	4	149	459974.0	1	1	1	111	3
3.0	77	24	218956.0	3	3	1	331	7
4.0	19	64	158562.0	1	2	2	122	5
5.0	311	16	41976.0	4	4	3	443	11
6.0	37	73	156400.0	2	2	2	222	6

Fig -2: RFM scores and quartiles

The RFM_Score values will range from 3 (1+1+1) to 12 (4+4+4). So, we can group by the RFM scores and check the mean values of recency, frequency, and monetary corresponding to each score.

As expected, customers with the lowest RFM scores have the highest recency value and the lowest frequency and monetary value, and the vice-versa is true as well. Finally, we can create segments within this score range of RFM_Score 3–12, by manually creating categories in our data: Customers with an RFM_Score greater than or equal to 3 can be put in the 'Top' category. Similarly, customers with an RFM_Score between 5 to 9 can be put in the middle category, and the rest can be put in the 'Low' category. Let us call our categories the 'General_Segment'. Analyzing the mean values of recency, frequency, and monetary.

V. K-MEANS CLUSTERING

In this section, we are going to preprocess the data for K-means clustering.K-means is a well-known clustering algorithm that is frequently used for unsupervised learning tasks. I am not going into details regarding how the algorithm works here, as there are plenty of resources online.

For our purpose, we need to understand that the algorithm makes certain assumptions about the data. Therefore, we need to preprocess the data so that it can meet the key assumptions of the algorithm, which are:

1. The variables should be distributed symmetrically
2. Variables should have similar average values
3. Variables should have similar standard deviation values

Figure shows histograms after checking these three assumption by building histograms of Recency, Frequency, and MonetaryValue variables using the seaborn library:

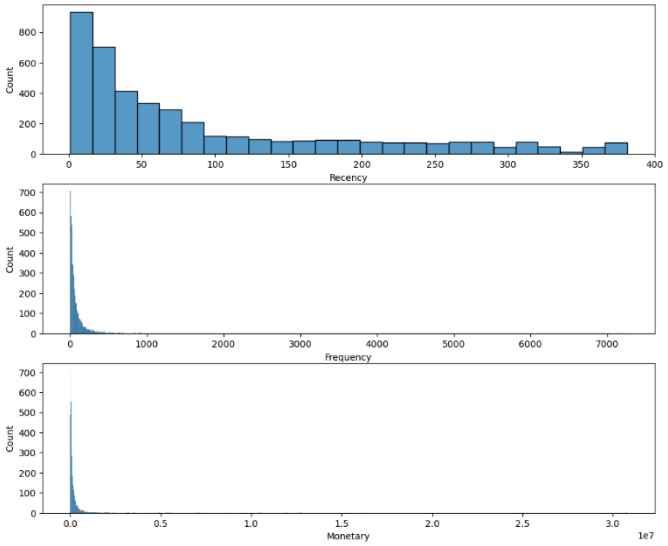

Fig -3: distribution of Recency, Frequency and MonetaryValue variables.

Lastly we will build multiple clusters upon our normalized RFM data and will try to find out the optimal number of clusters in our data using the elbow method. For each cluster, I have also extracted information about the average of the intracluster sum of squares through which we can build the elbow plot to find the desired number of clusters in our data. From the elbow plot, we can identified that the optimal number of clusters is 3 or 4.

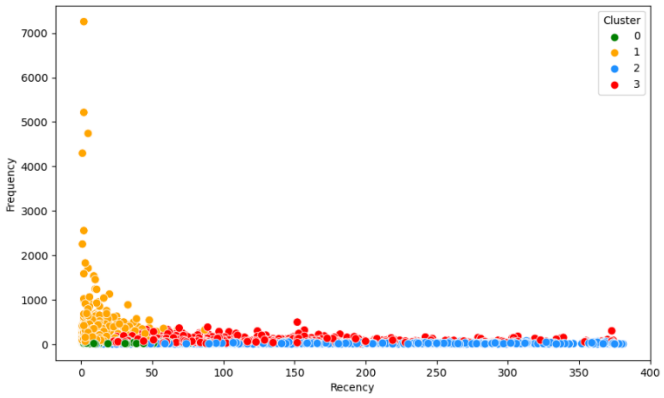

Fig-4: Final plotted cluster with four segments

Firstly we have find out number of customers in each segment. Out of 4322 customer ID 456 customers are Best Customers with RFM value as 3,1222 are Loyal Customers with RFM value as 4 to 6 ,1373 are Good customer with RFM value as 7 to 9,1271 are lost customers with RFM value greater than 9.

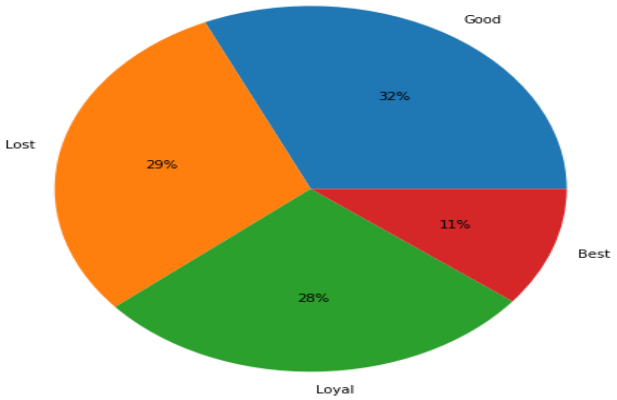

Fig-5: Pie Chart Showing Customers in Category

From the above analysis, we can see that there should be 4 clusters in our data. To understand what these 4 clusters mean in a business scenario, we should look back the table comparing the clustering performance of 3 and 4 clusters for the mean values of recency, frequency, and monetary metric. On this basis, let us label the clusters as 'New customers', 'Lost customers', 'Best customers', and 'At risk customers'.

Below is the table giving the cluster interpretation of each segment and the points that a company is recommended to keep in mind while designing the marketing strategy for that segment of customers.

Cluster	Type of Customers	RFM Interpretation	Recommended Action
0	New Customers	Customers who transacted recently and have lower purchase frequency,with low amount of monetary spending	Need to handle with care by improving relationships with them. Company should try to enhance their purchasing experience by providing good quality products and services, and customer care services.
1	Lost Customers	Customers with the least monetary spending and the least number of transactions. Made their last purchase long ago.	These customers may have already exited from the customer base. The company should try to understand why they left the system so that it does not happen again.
2	Best Customers	Most frequent spenders with the highest monetary spending amount and has transacted recently.	Potential to be the target of new products made by a company and can increase company revenue by repeated advertising. Heavy discounts not required.
3	Customers at Risk	Customers who made their transaction a while ago and made less frequent and low monetary purchase.	At high risk of churning. Need to be addressed urgently with focused advertising. May perform well if discounts are provided to them. Company should fing out why they are leaving.

Table -3: Final Cluster Table

VI. CONCLUSIONS

In this paper, we have proposed Customer segmentation and definition of proper strategies for each segment can provide tremendous returns for companies. In this way, there are various models of implementing customer segmentation. Some of these methods are RFM, customer value matrix, CLV and data mining methods. But it must be considered that there is great value to keeping things simple, especially for small and medium sized businesses. Methods that are derived from complex statistical modeling techniques can provide useful information for experts but are hard to implement for these businesses and are likely to present a challenge to the development and implementation of strategies. We Proposed, Frequency and Monetary method which also known as RFM method with KMeans algorithm has been used for customer segmentation on dataset available. Customer data and their attributes were mined in order to perform customer segmentation and consequently defining proper and useful strategies for having a better view of company customers and their behaviors and also increasing its profitability. Also company can recognize and classify an important or less important potential customer to set up proper marketing plan for those particular customers.The clustering algorithms used for segmentation of our data were k-means. Finally, the detail description and specification of all segments found in our case study were presented and based on their specifications, some useful strategies were proposed.

REFERENCES

1. *Ina Maryani,Dwiza Riana,Rachmawati Darma Astuti,Ahmad Ishaq,Sutrisno,Eva Argarini Pratama,"Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm",2018.*

2. *Ina Maryani,Dwiza Riana,"Clustering and Profiling of Customers Using RFM For Customer Relationship Management Recommendations",2018.*

3. *Pornwatthana Wongchinsri and Werasak Kuratach,"A Survey - Data Mining Frameworks in Credit Card Processing", IEEE, 2016.*

4. *Mohammadreza Tavakoli,Mohammadreza Molavi,Vahid Masoumi,Majid Mobini,Sadegh Etemad and Rouhollah Rahmani,"Customer Segmentation and Strategy Development based on User Behavior Analysis, RFM model and Data Mining Techniques: A Case Study",IEEE 15th International Conference on e-Business Engineering (ICEBE),2018.*

5. *Sabbir Hossain Shihab,Shyla Afroge and Sadia Zaman Mishu,"RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A Comparative Study",International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.*

6. *P. Anitha and Malini M. Patil,"RFM model for customer purchase behavior using K-Means algorithm",Journal of King Saud University Computer and Information Sciences,34,1785–1792,2022.*

7. *Siti Monalisa,Putri Nadya and Rice Novita,"Analysis for Customer Lifetime Value Categorization with RFM Model",Science Direct,The Fifth Information Systems International Conference, 2019.*

8. *Ms.Chaithra S,Mr.Hiffzull Rahman K H and Mr. Khalifa Abdul Musavvir,"Customer Segmentation using RFM analysis",IRJET,Volume: 08 Issue: 07,July 2021.*

9. *Zheng Shi,Zheng Wen and Jin Xia,"An Intelligent Recommendation System based on Customer Segmentation",International Journal of Research in Business Studies and Management Volume 2, Issue 11, November 2015.*

10. *A. Joy 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,33,1251–1257,2021.*

11. *Shreya Tripathi,Aditya Bhardwaj and Poovammal E,"Approaches to Clustering in Customer Segmentation",International Journal of Engineering &Technology, 7,3.12,802 -807,2018.*

12. *Priya P and Dr BH Chandrashekar,"Customer Relationship Management (CRM) for Library Software",IJCRT,Volume 8,Issue 5,May 2020.*

13. *Cristina Ledro, Anna Nosella and Andrea Vinelli,"Artificial intelligence in customer relationship management: literature review and future research directions",Journal of Business & Industrial Marketing 37/13,48–63,2022.*