Customer Segmentation Analysis Using RFM Model And K-Mean Clustering

S.Aruna[1], A.Karmehala[2]
[1, 2] Dept of Computer Science
[1, 2] Sri Kaliswari College (Autonomous) Sivakasi

Abstract- *Data Mining applied to the field of commercialization allows, among other aspects, to discover patterns of loyalty in customer, In today's business environment regardless of what type of industry we are in or what kinds of products and services that is sold, customers are the most important part of a business. Without the customer, sales doesn't happen. If the customers' views are not taken into an account, then it's likely the campaigns will not be successful. Hence classifying the right customers also matters a lot to make the products to be bought frequently. Companies follow different strategies to segment the customers. In this paper, RFM and K-means clustering is used to segment the customers.*

Keywords- Segmentation, Marketing, RFM, clustering, K-means,

I. INTRODUCTION

Customer Segmentation and Personalized Marketing provides a solution for marketing the available products to particular group of customers and also display the possible offers which can be provided to them that will yield profit to the company, retain their customers and also increase the sales. This is done by segmenting the customers into group of individuals who are similar in terms of gender, spending behavior, age and demography. Hence through achieving customer segmentation companies would be able to target the customers who has the specific needs and which helps in marketing the products to right group/customers. Customer segmentation is done by using RFM technique and K-means Clustering. Customer segmentation is the practice of dividing the customer into groups of individuals that are similar in specific ways relevant to marketing such as age, gender, interest and spending habits. There are different types of customer segmentation. They are:

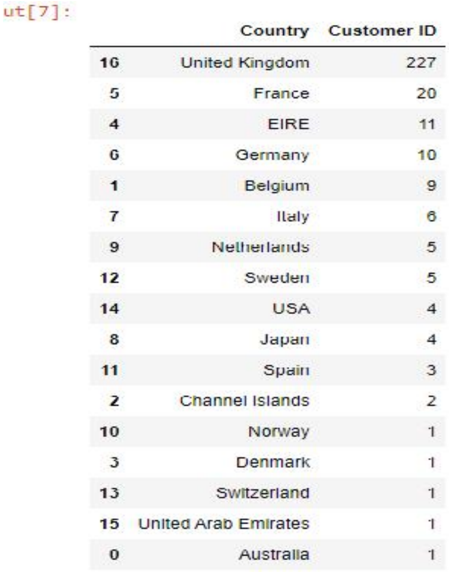
1) Geographic- based on a customer's location
2) Demographics- based on age, income, occupation and family size
3) Behavioral- based on the purchasing habits.
4) Psychological- based on customer's beliefs and values.

The data used for this classification depends upon the business decision. By differentiating their customer base, businesses can better target individuals and maximize sales, sell their products appropriately and provide more tailored shopping experiences. There are many benefits of Customer segmentation. They are:

1) Target the customer: By identifying the right customers for the product and selling to them by marketing to similar group of customers who would buy the products instead of spending in marketing to irrelevant groups.
2) Increase sales: By identifying similar groups and selling products to them gradually increase the sales for the company. This will help to have a long-term revenue from the customer.
3) Improve customer satisfaction and retention of the customers: By identifying and recommending the associated products bought by each group will result in customer segmentation as their needs where satisfied continuously so it helps to retain the customers.
4) Decrease the marketing cost: By classifying the customers into groups, the products could be marketed to specific group alone who has the probability of buying the products.

II. DATA PREPROCESSING

The dataset is on the details of taken from Kaggle. The data describes about the purchase made by the customers from different countries all over the word. The attributes in the dataset are: Customer ID, Gender, Age, Annual Income ($), Spending Score (1-100), Profession, Work Experience, Family Size, trans_date, Price, Invoice, Quantity and Country name. To see the customer distribution country wise, the customers are grouped according to the countries.



ut[7]:

	Country	Customer ID
16	United Kingdom	227
5	France	20
4	EIRE	11
6	Germany	10
1	Belgium	9
7	Italy	6
9	Netherlands	5
12	Sweden	5
14	USA	4
8	Japan	4
11	Spain	3
2	Channel Islands	2
10	Norway	1
3	Denmark	1
13	Switzerland	1
15	United Arab Emirates	1
0	Australia	1

Fig. 1 Count of customers country wise

From the fig. 1, it is clear that the majority of the customers are from United Kingdom. Hence the customer segmentation can be performed for United Kingdom. A data mining technique that involves transformation of raw data into an understandable format is called data pre-processing. The process of connecting and removing corrupt and inaccurate data is done by a process called data cleaning, also known as data cleansing. This process includes smoothing the noisy data, filling the missing values or resolving the inconsistencies in the data. The data cleaning is done by removing the missing values and negative values in the dataset.

III. METHODOLOGY

In this section we describe the two techniques used for customer segmentation.

A. Recency Frequency Monetary Analysis

It is a marketing technique which determines the best customers by analyzing the recency, frequency and monetary values called Recency Frequency Monetary (RFM) analysis.
Recency : How much time has elapsed since a customer's last activity or transaction

Frequency : How often has a customer made transaction
Monetary: How much a customer has spent for the product.

Customer segmentation [9] can be done with the help of RFM analysis by assigning RFM scores to each customer. RFM factors illustrate these facts:

1) The more recent the purchase, the more responsive the customer is to promotions.
2) The more frequently the customer buys, the more engaged and satisfied they are.
3) Monetary value differentiates heavy spenders from low-value purchasers.

The benefits of RFM analysis are to determine the high valued customers of the business, to target the customer who has the high chance for buying the products and to know the one- time customers of the business. There are three major groups of customers determined by RFM analysis. They are:

I. High RFM Customers:

They are the customers with high monetary, frequency and recency values. They add value to the business more and they are the promising customers where the company can market their products more.

II. Medium RFM customers:

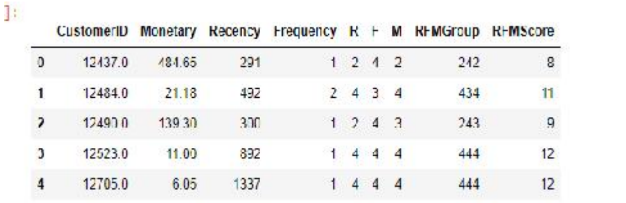
They are the customers with high recency, low frequency and monetary values. These groups of customers can be made to high RFM scored customers by using some marketing techniques.

III. Low RFM Customers:

They are the group of customers who are least engaged. They have low recency, frequency and monetary values. They could be brought to medium RFM level by providing new offers to them.

Calculation of RFM scores

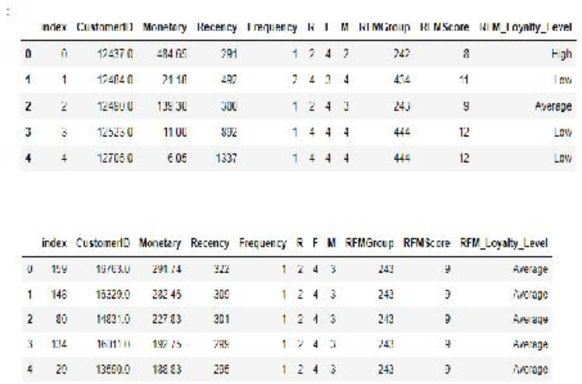
The details of each customer like Customer ID, Quantity and Unit price would be required to calculate the RFM scores. The day since the last purchase is used to calculate the recency, the total number of transactions is used for finding frequency and total money each customer has spent is calculated for monetary value. The next step is to create quartile such as 0.25, 0.50, 0.75. So that we can sub divide the set into 4 groups based on R, F and M values. In order to create segments with values 1, 2, 3 and 4, Rscoring and FM scoring is created. In Rscoring function, we assign value 1 for the lowest value of recency, because lower the value indicated the most recently visited. In FMscoring function, we assign value 1 for the highest value of frequency and monetary, because higher the value of frequency and monetary are tend to be promising customers. After applying both the functions, segmented RFM values are obtained. The next step is to calculate and add RFM-Group value column showing the combined concatenated score of RFM and also RFM-score value column to show the total sum of RFM_Group values. we can added the RFM-Group and RFM- score.



	CustomerID	Monetary	Recency	Frequency	R	F	M	RFMGroup	RFMScore
0	12437.0	484.65	291	1	2	4	2	242	8
1	12484.0	21.18	492	2	4	3	4	434	11
2	12490.0	139.30	300	1	2	4	3	243	9
3	12523.0	11.00	892	1	4	4	4	444	12
4	12705.0	6.05	1337	1	4	4	4	444	12

Fig. 2 RFM- Group and RFM-Score

The next step is to assign loyalty levels to the customers such as Very-High, High, Average and Low. After assigning, filtering can be done for each group and it can be stored separately for retrieval purpose. Fig. 3 shows the filtering of Average valued customers.



	index	CustomerID	Monetary	Recency	Frequency	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
0	0	12437.0	484.65	291	1	2	4	2	242	8	High
1	1	12484.0	21.18	492	2	4	3	4	434	11	Low
2	2	12490.0	139.30	300	1	2	4	3	243	9	Average
3	3	12523.0	11.00	892	1	4	4	4	444	12	Low
4	4	12705.0	6.05	1337	1	4	4	4	444	12	Low

	index	CustomerID	Monetary	Recency	Frequency	R	F	M	RFMGroup	RFMScore	RFM_Loyalty_Level
0	159	15703.0	291.74	322	1	2	4	3	243	9	Average
1	146	15329.0	282.45	309	1	2	4	3	243	9	Average
2	80	14831.0	227.83	301	1	2	4	3	243	9	Average
3	134	16011.0	197.75	299	1	2	4	3	243	9	Average
4	20	13600.0	188.83	206	1	2	4	3	243	9	Average

Fig. 3 Filtering of average valued customers

B. K-Means Clustering

K-Means clustering is an unsupervised learning technique used when the data is unlabeled in nature. It finds the groups based on the patterns present in the data and K is the variable that represents the number of clusters/ groups. Each data point is assigned iteratively to one of the K groups. This division is based on the features that are provided. Clustering happens based on feature similarity.

K-Means Clustering

Highlevel overview of K-Means algorithm:

- Vectorize feature values to define n-dimensional coordinates
- **Decide value of 'K'** -> K is the number of clusters
- K number of clusters are created with K centroids
- Each datapoint is assigned to nearest cluster
- New centroids are chosen based on Euclidian distance
- Data points are reassigned to nearest cluster
- The above steps are repeated untill there are no change.
- From the above explanation it can be observed that the value of **K** needs to be determined before hand.
- **Elbow Method** is used to find the optimal value of K

Algorithm

Step 1: Initialize k points randomly and these points are called means.
Step2: Categorize each item to the closest mean by using Euclidean Distance measure.
Step 3: Update the mean’s co-ordinates by calculating the average of the items categorized in the product menu.
Step 4: Repeat the steps for some iterations till the desired number of clusters are formed.
The number of clusters cannot be formed randomly. There are two main methods to determine the optimal number of clusters. They are

I. The Elbow Method

A naïve and commonly used method to determine the number of clusters is the Elbow method. The elbow method runs k-means clustering on the dataset for a range of values for k and then for each values of k computes an average score for all clusters. By default, the score is computed, the sum of square distances from each point to its assigned center. Fig. 4 is a graph depicting the optimal number of clusters that can be formed using elbow method. Here the elbow is at k=3, hence 3 is chosen as the number of clusters.

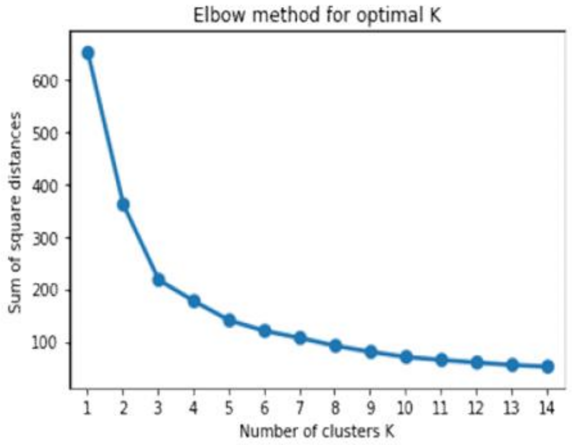


Fig.4 Elbow graph determining the number of clusters

II. The Silhouette Method

The quality of the cluster is measured by the silhouette method. A high value of the silhouette score indicates good clustering. It computes the silhouette score for all observations for different K values.

Normalization:

Normalization is used to scale the data of an attribute so that it falls in a smaller range, it is applied when the dataset has more deviation in it. Normalization is done using log transformation. Fig. 5, 6 and 7 is a graph depicting the recency, frequency and monetary values respectively after normalization.

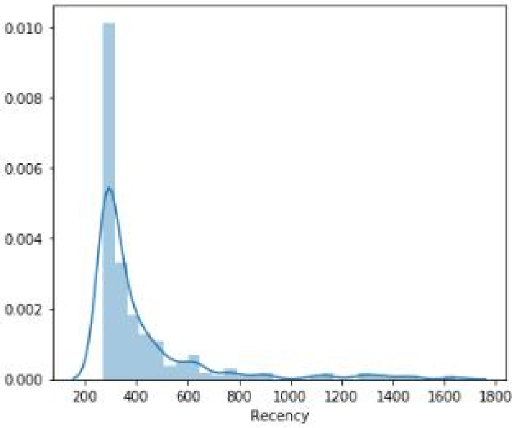


Fig.5 Recency values after normalization

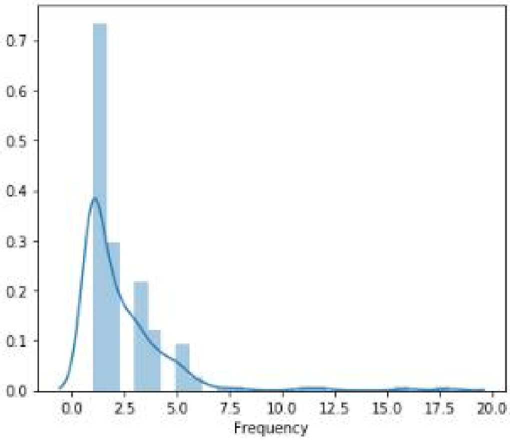


Fig.6 Frequency values after normalization

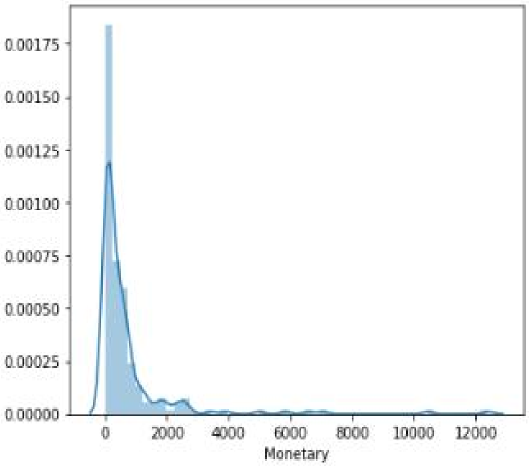
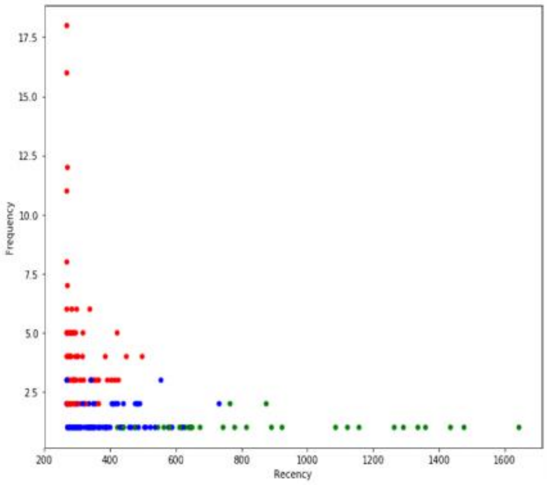


Fig.7 Monetary values after normalization

After normalization of the data, the k-means clustering is performed by having k=3 as the number of clusters. There are three clusters formed high, average and low with the color indications as green blue and red respectively. The clusters are formed after k-means clustering is performed. Fig. 8 is a plotted graph of the clusters.



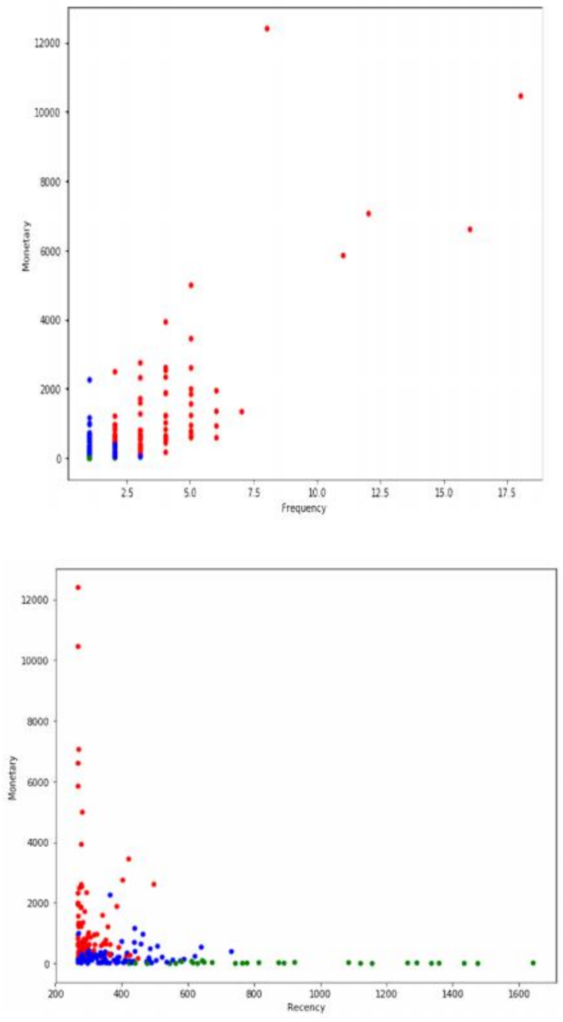


Fig. 8 Plotted graph of the clusters

IV. IMPLEMENTATION

The benefits customers, reduce marketing cost and customer satisfaction. Market basket analysis is the technique used for personalized marketing. It is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns, it is basically a filtering system that helps to predict and show the items that a user would like to purchase along with the product already purchased. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together. It works by looking for combinations of items that occur together frequently in transactions. as part of the user experience can increase the average order valuer. There are two types of market basket analysis and they are:

I. Predictive market basket analysis: Items purchased are considered to determine cross-sell (suggesting similar or complementary products or services).

II. Differential market basket analysis: Considers data across different stores, as well as purchases from different customer groups during different times of the day, month or year. These insights can lead to new product offers that drive higher sales.

V. CONCLUSION

Customer segmentation and personalized marketing segments or groups the customers and provide services for them according to their need, since product and services needs of individual customers differs and also would help the organization to increase sales and retain customers. The techniques used are K-means clustering which is used to segment the customers and market basket analysis which is used to find the associated products to give a the company's marketing problem by performing a targeted marketing. The future enhancement could be done by implementing with the dynamic dataset provided by the organization. It would add the real-time data and includes the dynamic recommendation in the website.

REFERENCES

[1] Ahmed, R. (2003). Benefit Segmentation: A Potentially useful technique of segmenting and targeting older consumers. International Journal of Market Research 45 (3), Retrieved on 04-11-2020 https://www.warc.com/fulltext/JMRS/78268.html.

[2] Ait Daoud, R. Bouikhalene, B. Amine, A. & Lbibb, R. (2015). Combining RFM model and clustering techniques for customer value analysis of a company selling online. 978-1-5090-0478-2/15, IEEE. [In text citation: Aid Daoud, Bouikhalene, Amine & Lbibb, 2015]

[3] P. P. Pramono, I. Surjandari and E. Laoh, "Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method," 2019 16th International Conference on Service Systems and Service Management (ICSSSM), 2019, pp. 1-5, doi: 10.1109/ICSSSM.2019.8887704.

[4] B. G. Muchardie, A. Gunawan and B. Aditya, "E-Commerce Market Segmentation Based On The Antecedents Of Customer Satisfaction and Customer Retention," 2019 International Conference on Information Management and Technology (ICIMTech), 2019, pp. 103-108, doi: 10.1109/ICIMTech.2019.8843792.

[5] E. Y. L. Nandapala and K. P. N. Jayasena, "The practical approach in Customers segmentation by using the K-Means Algorithm," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 344-349, doi: 10.1109/ICIIS51140.2020.9342639.