International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Customer Analytics Using K-Means Clustering And Elbow Modelling With Product Associative Analysis Using Unsupervised Machine Learning

M.V.L Bhavani[1], Nemalidinne Rajamohan reddy[2],Thota Gayatri[3]
Shaik Roshan[4], V Chandra sekhar[5]

[1]*Sr.Asst.Professor, Department of Electronics and communication Engineering, Lakireddy Bali Reddy College of Engineering, AP, India, 521230.*
[2]*UG Students, Department of Electronics and communication Engineering, , Lakireddy Bali Reddy College of Engineering, AP, India, 521230.*

ABSTRACT: The feel of contemporary era is innovation, where everyone is involved into competition to be higher than others. Today's business run on the premise of such innovation having ability to enrapture the purchasers with the merchandise, however with such an oversized raft of merchandise leave the purchasers mazed, what to shop for and what to not and conjointly the businesses are puzzled regarding what section of shoppers to focus on to sell their products. This is often wherever machine learning comes into play, varied algorithms are applied for unravelling the hidden patterns within the knowledge for higher deciding for the long run. This elude idea of that phase to target is created unequivocal by applying segmentation. The process of segmenting the purchasers with similar behaviors into an equivalent phase and with totally different patterns into totally different segments and analyzing their purchasing patterns can be treated as customer analytics. Customer segmentation is carried out based on the RFM value. With RFM a firm can divide its customers into three segments such low, mid, high with subsequent implication of elbow modeling, and k-means clustering clubbed which product associative analysis to track the combination of products that the customers buy frequently.

Index Terms: RFM, K-Means

1. INTRODUCTION

In this modern era, maintaining and analyzing the customer data by the firms is one of the most typical jobs. Companies are investing a lot of time and wealth on analyzing their customer's behaviors in terms of response to their products and the constant amounts of monetary that they are putting on the products of the company. In this paper we are going to discuss the detailed procedure of customer segmentation based on the RFM analysis and the further extension of analysis like product associative analysis to identify the underlying patterns of customers purchase behaviors i.e.to have an analysis of combinations of items that occur together frequently in transactions using apriori algorithm. For this analysis we are going to take the data set of size nearly 540,000 from a U.K. based online retail stores

containing the transactions listed for a time period of almost one year. And for the effective results python programming is used for the analysis because python is best suited for data science projects.

2. METHODOLOGYMachine Learning:

Machine learning(ML), which is a part of Artificial Intelligence(AI) to study computer algorithms that develop automatically using information and data. Development machines create a model using sample data, known as "training data". Machine learning helps to deal with processing of large amounts of data with predefined algorithms and with support of machine learning libraries that are available in python.

Unsupervised Learning:
In unsupervised learning, the machine is trained using separate or unencrypted data and allows the algorithm to work on that data without any guidance. The task of the machine here is to collect random data according to similarities, patterns, and variations without previous data training. Unlike supervised learning, for unsupervised learning, there will not be any guidance no training will be provided by the machine. Therefore, the machine is restricted from accessing hidden structures to the data without the label itself.

Stages involved:
This analysis involves certain stages like data understanding, data cleaning, exploratory data analysis, RFM analysis, Elbow modeling, K-Means clustering and product associative analysis respectively.
The detailed explanation of these stages is as follows:

i. Data understanding and data cleaning involves the removal of failed or cancelled transactions and to remove the null values from the data set.
ii. Exploratory data analysis can be viewed as early stages of data analysis where we explore the data to view the number of transactions occurred in every month and every day. In addition to it creating the time cohorts in order to analyze the time of first transaction done by a particular customer. These time cohorts find their usage in calculating the customer retention rate in which we calculate the percentage of customers that the company had retained in that particular time frame by dividing the no.of transactions done by their unique customers on each month with the no.of unique customer transactions at the starting month in the data set.
iii. In RFM analysis we deal with the three major variables in data analytics i.e. Recency, frequency and monetary.RFM factors illustrate these facts:

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

We will calculate the recency by subtracting the last transaction date of the customer from the analysis date which will be at day after the end date of the date set. In the same way frequency and monetary values associated with each customer is calculated by the count of

no.of transactions done and the total amount spent on all transactions respectively.

Making RFM quartile: The RFM scores associated with each customer in the data set is rated over ascale of 1(highest rating) to 4(lowest rating) in order to make the customer data in terms of RFM segments[4] and each segment has a certain RFM score as depicted in the below table. So the entire dataset is now converted in terms of RFM values and the further analysis is done based on this.

Customer ID	Recency	Frequency	Monetary	R	F	M	RFM_Segment	RFM_Score
12346	326	1	77183	4	4	1	441	9.0
12347	2	182	4310.00	1	1	1	111	3.0
12348	75	31	1797.24	3	3	1	331	7.0
12349	19	73	1757.55	2	2	1	221	5.0
12350	310	17	334.40	4	4	3	443	11.0
…...	…	…	…	…	…	…	…	…

Fig.1. Table showing the assignment of RFM values and segments to each Customer Id

So the entire customer segmentation is carried out based on the RFM values, and the key K-Meansassumptions are:

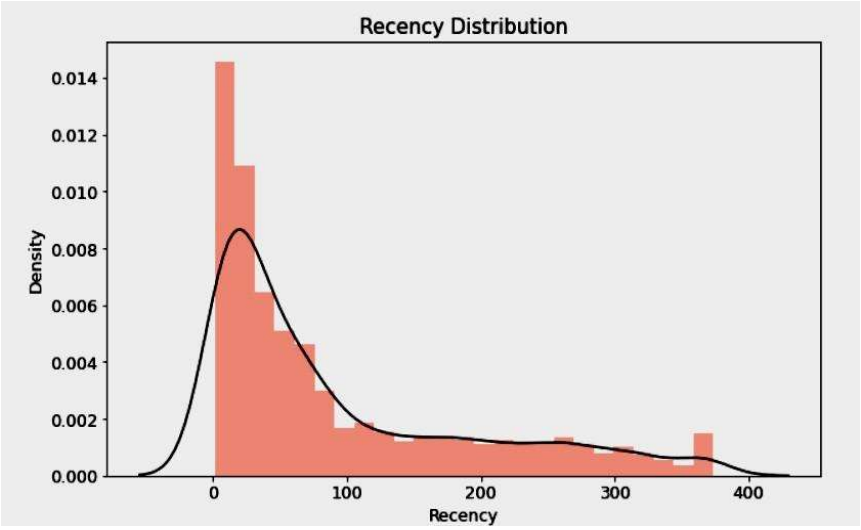
a) Symmetric distribution of variables i.e. R, F, M (not skewed)

b) Variables with same average values

c) Variables with same variance

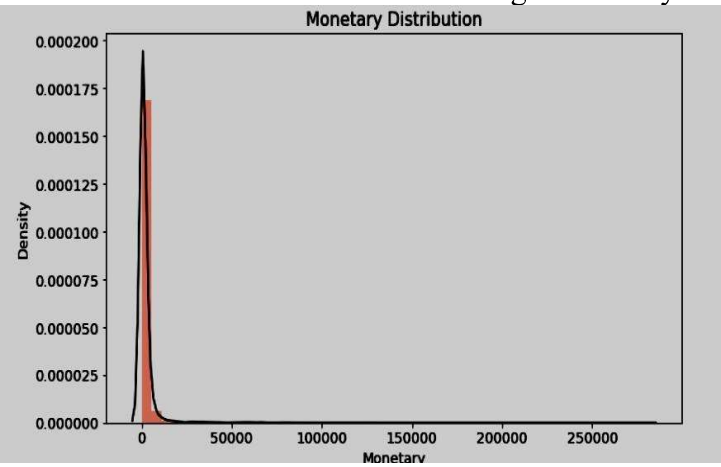
In order to ensure the values of skewness among the R F M values, Distributions of each variable is tedand the value of skewness is calculated from the mean, variance and standard deviation of each variable.

The respective distributions are depicted below:



International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Fig.2. Recency distribution



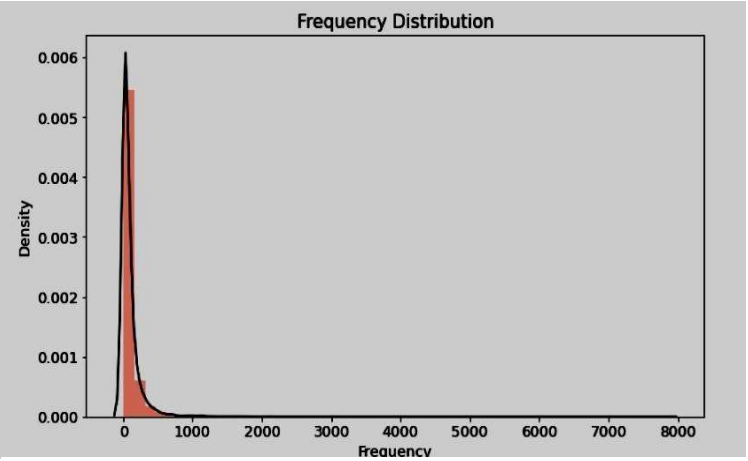
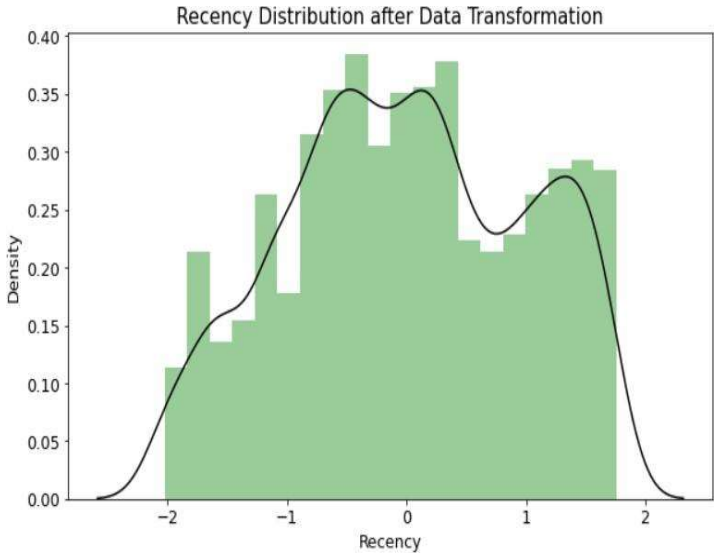


Fig.3. Monetary distribution

Fig.4. Frequency distribution

As we can see from the distribution plots these variables are not normally distributed and has high values of skewness. So in order to normalize the data we have to transform the variables using power transformer functions and standard scalar functions in order to transform the entire data into normal distribution or Gaussian distribution.
The respective distributions of the variables after data normalization or data transformation are depicted below as:



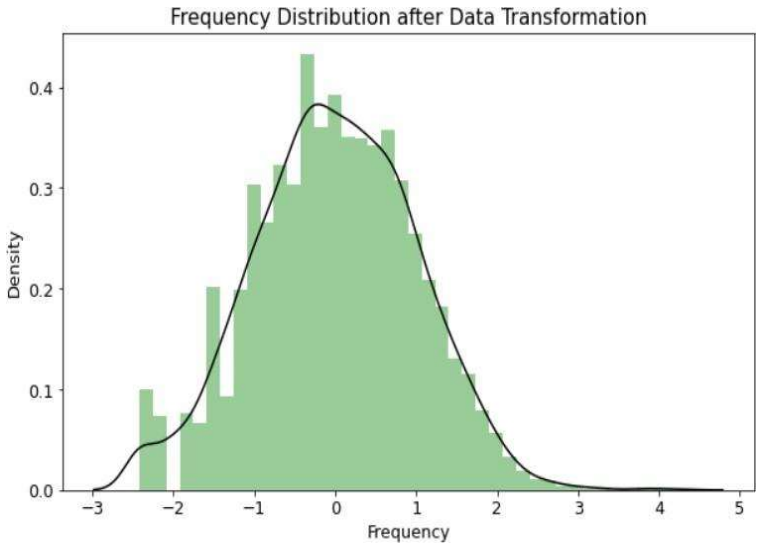
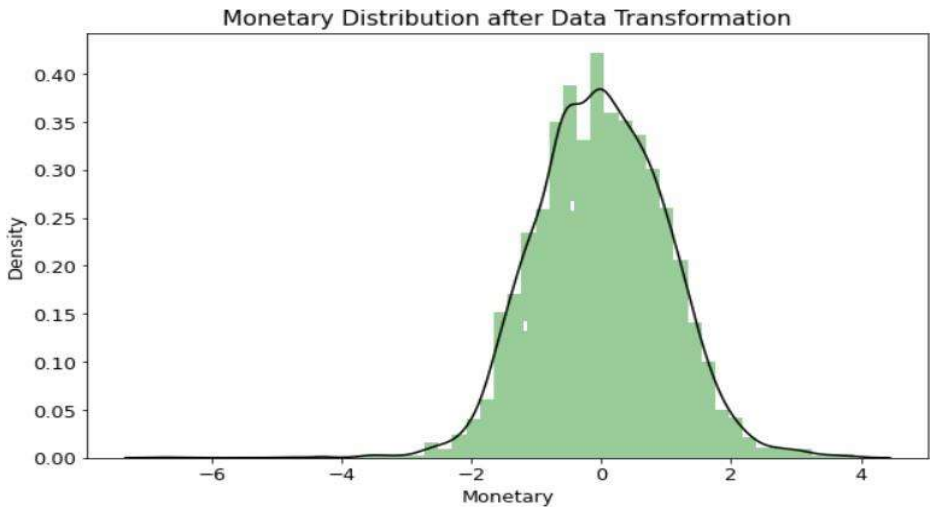


Fig.5

Fig6



As we can see that these variables are normally distributed after the transformation and the skewness has been shifted to values near to -0.05 to -0.02 or close to 0 which can be assumed that RFM data has normal distribution.

Elbow method: In determining the number of clusters elbow method is used extensively. The Elbow method runs K-means clustering on the data set for a range of values k (say 1-15) and for each value of k, calculates the sum of squared distances to the closest center[3]. The idea of elbow method is to choose k at sum of squared distances decreases abruptly. We will fit the rfm transformed data in the k-Means inertia function which will calculate the distances.

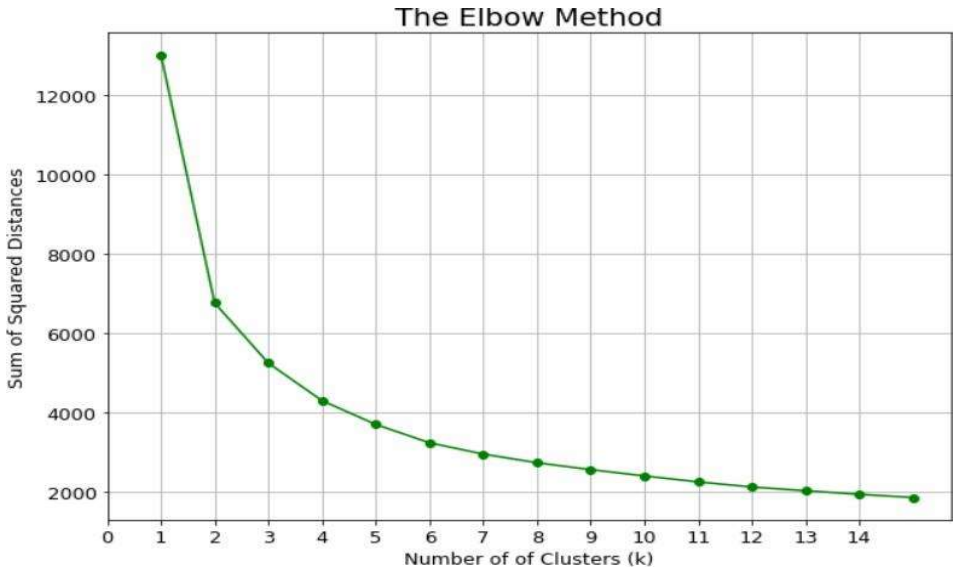


Fig.8.plot between the no. of clusters (k) and sum of squared distances to closest center

From the figure, we have to observe the point from where the sum of squared distances ceases to maintain drastic changes i.e. from the figure we can site that from k=4 there is not so much change in the values, so itcan be termed as elbow point and therefore no. of clusters is fixed as k=4.

K-Means clustering Algorithm: Here the conventional k-means algorithm is used for the clustering analysis and the RFM normalized data is used for training and fitting the model, resulting the 3 dimensional clustering model which is based on the r, f, m variables distribution.

The k-Means algorithm is as follows:

- At first the k centroids are randomly initialized then we've got to reckon the total of squaredistance between data points and every centroid.
- Assign every datum to the closest centroids
- calculate the centroids for the clusters by talking the mean of all data points that belong to

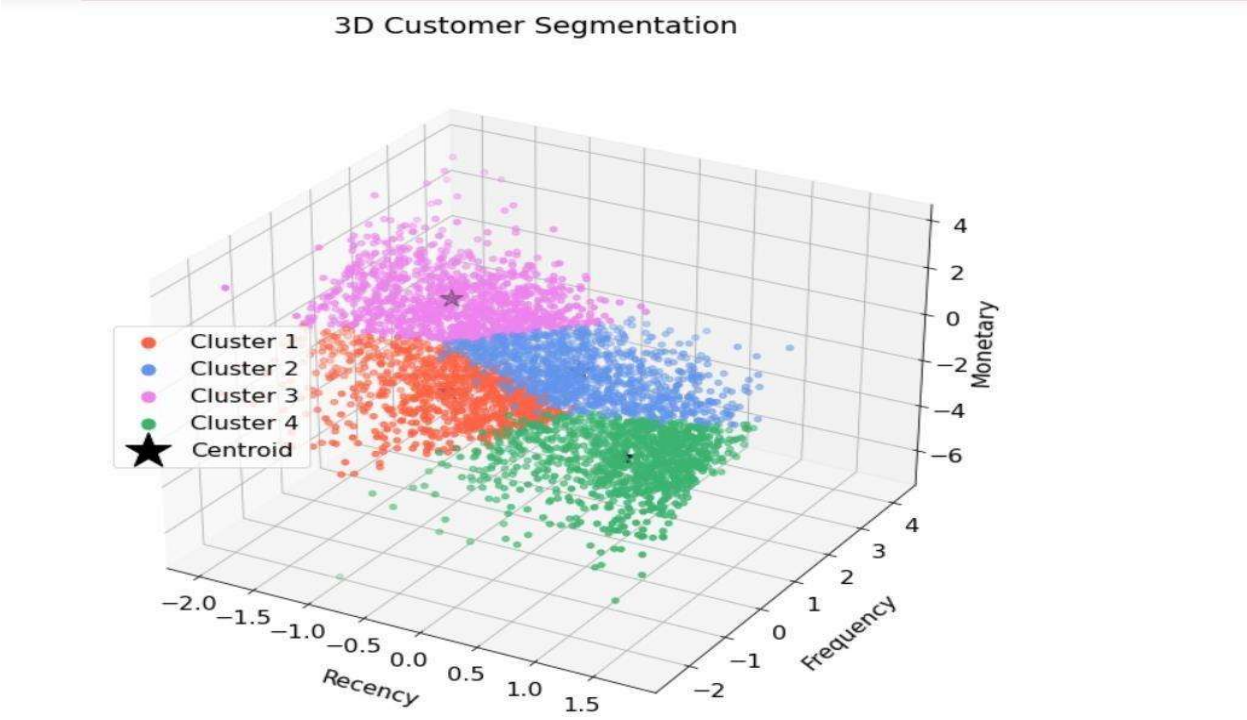


Fig.9. 3D Customer segmentation model

	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	30.4	27.0	441.2	962
1	95.7	74.9	1449.0	1195
2	14.1	248.7	6024.7	1067
3	217.8	15.3	293.4	1115

Fig.10. Overall clustered data analysis variables of different clusters

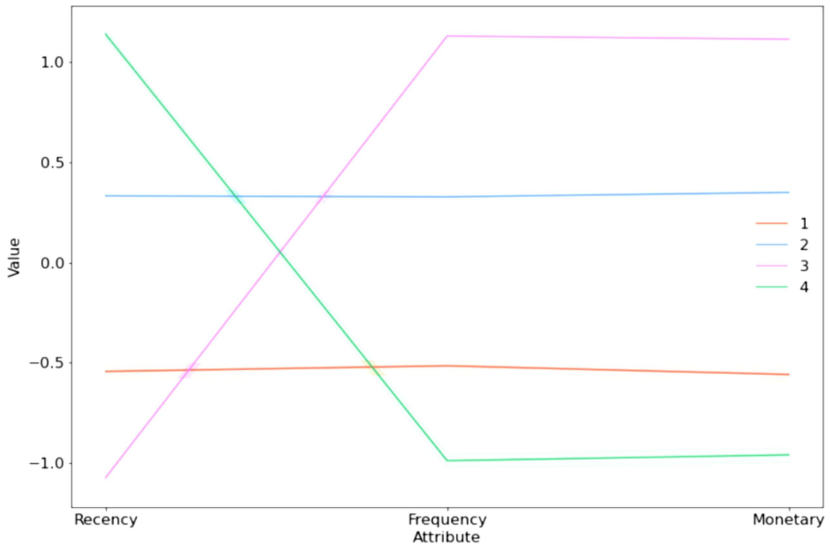


Fig.11. plot of standardized

From the analysis of clustered data, cluster 3 is our champion cluster which contains best customers, new buyers and heavy buyers. Cluster 4 is our risk cluster. At Risk Customers are customers who often buy and spend a lot of money, but haven't been shopping recently. Cluster 1 is our promising cluster. Promising customers are customers who have been shopping recently, but the frequency and amount of money spent in our stores is still small or below average. Cluster 2 is our lost cluster. Lost customers are customers who have not been shopping for along time, and the frequency and amount of money spent is also very low

Product Associative analysis: It works by looking for combinations of items that occur together frequentlyin transactions. This analysis has the objective of identifying products, or groups of products, that tend to occur together (are associated) in buying transactions[1]. This analysis involves the usage of Apriori algorithm which is frequently used in data mining techniques in order to find the hidden patterns in the larger data sets. It involves three major variables for the analysis they are lift, support and confidence. **Support**

It measures the percentage of item set occurrence in all transactions.

Confidence
Confidence measures how strong the association rule is. How often item Y appears in the purchasetransaction of item X.
Lift
Lift of the rule is defined as the ratio of observed support to the support expected in the case the elements of the rule were independent. Lift values > 1 are generally more “interesting” and could be indicative of a useful rule pattern.
The products that occur most frequently in pairs are termed as antecedents and consequents

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.029611	0.036759	0.024266	0.819473	22.293137
(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY...	(ROSES REGENCY TEACUP AND SAUCER)	0.024266	0.040723	0.020482	0.844059	20.726763
(ROSES REGENCY TEACUP AND SAUCER, PINK REGENCY...	(GREEN REGENCY TEACUP AND SAUCER)	0.023004	0.036759	0.020482	0.890339	24.221015

Fig.12. Result of product associative analysis

3. CONCLUSION

The entire customer analysis is done by using the predefined procedures with the implication of RFM normalizeddata in to the algorithms on the data set taken.

.

4. REFERENCES

[1] Han, J., Kamber, M., and Pei, J.: 'Data mining: Concepts and techniques. Morgan Kaufmann series in data management systems' (Morgan Kaufmann, 2011. 2011)

[2] 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

[3] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar “Telecom customer segmentation based on cluster analysisAn Approach to Customer Classification using k-means”, IJIRCCE,Year: 2015.

[4] LIN Sheng, XIAO Xu, "A method of telecom consumer market segmentation based on the RFM model," Journal of Harbin Institue of Technology, Vol 38,No 5, pp758-760,2006