International Journal of Mechanical Engineering

Analysis of RFM Customer Segmentation Using Clustering Algorithms

[1]Dharmaiah Devarapalli, [2]Ayinavilli Sowjanya Virajitha, [3]Geddam Sai Veera Venkata Satya Sunanda, [4]Amudalapalli Sri Sravya, [5]Boddu Tharuna Keerthi, [6]Allada Poulami Devi

[1,2,3,4,5,6]Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women(A), Bhimavaram, West Godavari, Andhara Padesh, India, 534202

Abstract - In the fast-emerging world, we discover different scenarios where sellers are worried about finding new customers along with maintaining old ones. So, to make them convenient with the concept of shopping trends, a new kind of analysis based on customers purchasing is introduced, i.e., dividing the customers into different segments based on their similarities and differences within those segments in order to better serve their customers and this analysis is called Customer Segmentation. We used both k-means and DB Scan clustering techniques to train the model, and the program separated the datasets into clusters based on the recency, monetary, and frequency values of the customers in the datasets. After completion of training phase customers data, we perform data visualization for better understanding and we get the recency, frequency and monetary mean values of customers in different segments. In this model they can get the recency, frequency, monetary values of that particular customer by giving customer ID. The fundamental aim of customer segmentation is to improve the advertising and marketing performance of an e-commerce platform and this can be very a good deal.

Index Terms - k-Means, DB Scan, Recency, Monetary, Frequency.

INTRODUCTION

‘Customer segmentation’ is the division of a company's customers into groups so that clients in the same category share some characteristics [2]. The purpose of customer segmentation is to determine how to relate customers within a segment in order to maximize each customer's value to the business [7].

‘Customer segmentation’ gives marketers the ability to reach out to each and every customer [22]. Customer segmentation aids marketers in accurately identifying distinct categories of customers [18][6]. These clients are classified into groups based on demographics, behavior, gender, and other factors. Customer segmentation is one of the Unsupervised Learning applications [4].

Companies can find percentages of similar groups of individuals and sales using clustering algorithms, which results in higher profits for the company and also aids marketing [3]. We will employ the k-means and db scan clustering algorithms in this machine learning project, which are important methods in unsupervised learning [14]. However, if consumer segmentation is done correctly, there are significant business benefits [10]. For instance, a best current customer segmentation exercise can have a significant impact on your operating results [5].

Increasing the overall quality of the product

Understanding who will buy your product and what they will use it for can help your company stand out as the greatest option for their needs. As a result, the challengers' satisfaction and performance will improve [3]. The advantages extend beyond product beneficence since these insights about your greatest clients will allow your business to make money.

Focusing your marketing message

If we undertake a client segmentation project while upgrading the product, we will be able to improve the product. It enables you to create more important marketing messages that are tailored to each of your best segments, ensuring a higher product quality.

C. Enabling your sales team to look at deals with higher percentages

Taking up less time on the customers who order fewer amounts of products and orders the products least frequently and giving more time to the customers whose purchasing behaviors are great will help to increase the win rate for your sales teams and also the revenues will also be increased

D. Getting higher quality revenues

Businesses that cater to the wrong type of customer may be pricey to sell and maintain, with a high turnover rate and few prospects for up selling after the original purchase. As a result, avoiding poor consumers and focusing on great ones is more important. By avoiding these customers, you may increase your profit margins and keep your customer base consistent.

Problem Definition

Customers fall into a variety of categories, each with their own set of criteria. As a result, applying the same method and advertising to all customers isn't ideal [2]. It is vital for a firm or organization to identify client groups and understand the similarities and differences within those segments in order to better serve their customers.

Create and transmit marketing messaging that will resonate with certain customer groups but not others [4]. Choosing the most effective communication medium for the segment, which may include emailing, social media, radio advertising, and other options, depending on the segment [5].

Improve customer service by improving client connections and focusing on the most profitable customers.

Customer Segmentation and Analysis

Increasing the number of libraries in the system.
Exploration of data
Data visualization
Clustering
Analysis of clusters
Producing a graph depicting the customer's boundary values.

METHODOLOGY

A. Unsupervised Learning:

Unsupervised Learning is one of the machine-learning approaches. Models aren't supervised using a training datasets in unsupervised learning. In the context of the data available, models discover hidden patterns. The learning pattern is comparable to that of the human brain while learning new things in unsupervised learning. Unsupervised Learning can be defined as a Machine Learning technique where the models are trained using an unlabeled dataset and are allowed to work on data without any oversight [7]. In supervised learning, we can utilize classification or regression right away because we have input and output data [21].

Unsupervised Learning's major objective is to discover the data set's Elemental structure, cluster it according to its similarity, and represent it in a compact way.

B. K-Means Algorithm:

This algorithm splits data into K clusters. A cluster is a subset of data points that belong to a group with similar features. It aspires to divide into clusters that are as similar as possible while remaining unique. It assigns data points to a cluster by computing the distance between the points and the centroid using the sum of squared distance. The centroid is calculated by taking the average of all the data points in the cluster. The data points are more comparable when there is less variance within clusters [26].

The Kmeans algorithm operates like this:

Give K, which is the number of clusters into which the data should be split, where k is a key parameter in the k-means method.
Randomize the centroids by shuffling them, then alter the value of the centroid by picking K data points at random.
Change the data points' cluster assignments until they are stable.
Calculate the distance between each data point and each of the centroids, and then relocate each data point.
The cluster with the least distance is the one to point to.
Add all the data points in each cluster together to find the centroid [28].

C. DB Scan Algorithm:

Density-Based Clustering algorithms are one type of unsupervised learning method. A high point density area in the data space is isolated from other clusters by succeeding low point density regions, according to this method for finding data clusters. It's a simple method for classifying items based on their density. It locates data clusters, including noise and outliers, in massive datasets [28].

DB SCAN focused on two key parameters:

• *minPts***:** This is the smallest number of points a cluster must have to be deemed dense in an area.
• *eps ()***:** If any point's distance from our point is less than or equal to this, it will be added to a cluster.

After the DBSCAN clustering is complete, there will be three categories of points:

• **Core** — A core is a point that is surrounded by at least m points that are smaller than or equal to n.
• **Border** – A boundary is a point that has at best one Core point that is n distance away from it.
• **Noise** — A point that is not a Core point and not a Border data point, and has a distance of lower than n. [19]

D. Hierarchical clustering:

The Hierarchical clustering method works by collecting data into a cluster tree. Hierarchical integration begins by treating all data points as a separate collection.

Then, repeat the following steps:

1. Identify 2 groups that can be very close, too
2. Combine the top 2 comparable groups. We need to continue with these steps until all the collections are put together.

In Hierarchical Clustering, the goal is to produce a hierarchical series of constructed clusters. The Dendrogram clearly represents this arrangement and is a twisted tree that describes how things come together or how a cluster divides.

D. Sklearn:

Scikit-learn is a python library which has several machine learning and modeling capabilities, such as classification, regression, and clustering. Scikit-learn have a number of features, and a few of them are useful for understanding supervised and unsupervised learning methods, as well as feature extraction [16].

E. Pandas:

A panda is a software program that makes it simple and straightforward to work with tabular and multi-dimensional data. Series, a one-dimensional data structure, and Data Frame, a two-dimensional data structure, are the two basic data structures used by pandas. It's commonly employed in a variety of contexts, especially statistics and engineering. Pandas are mainly used for merging and joining data sets , handling missing data , and provides size mutability etc..[17].

F. Matplotlib:

Matplotlib is a library that helps create an interactive environment by displaying figures in a number of forms. It's compatible with IPython shell, web application servers, and a variety of GUI toolkits.

Data set Description

A. Data Gathering:

Data gathering generally involves collection of data from various sources. In this project we had taken the dataset from Kaggle [1] which consists of various details of customers' i.e., Online Retail Dataset which consists of eight fields and 541880 records. Some records of the data set.

The fields in the dataset are:

1. Invoice No:
   The invoice number is a 6-digit integral number assigned to each transaction separately. This code indicates a cancellation if it starts with the letter c.
2. Stock Code:
   Product (item) code, Nominal is a 5-digit integral number assigned to each unique product.
3. Quantity:
   Quantity is a numeric value that represents the quantity of each product (item) per transaction.
4. Invoice Date:
   The day and time that each transaction was generated are represented by the Invoice Date, which is a numeric value.
5. Unit Price:
   Unit Price is a numeric value that represents the price of a product per unit in sterling.
6. Customer ID:
   Customer ID is a 5-digit nominal value that is assigned to each customer individually.
7. Country:
   The Country is the name of the country in which each customer resides.

B. Data Prepossessing:

"Data processioning" is the approach of modifying the raw data to make it suitable for a machine learning model. During the creation of a machine learning model, data prepossessing is the first most significant step, because we may not come across clean and formatted data always. In our case, we have removed the data which has quantity less than 0 and also converted to the type of Invoice Date column to calculate recency, we have also added the field Sale as a product of quantity and unit price [27][25].

C. RMF Approach:

In our project, we calculated the RMF values of each customer where R denotes recency, M denotes Monetary, and F denotes Frequency [23].

Recency: How recently does the customer buy?

Monetary Value: How much does he spend?

Frequency: How frequently does he buy?

We have to look into the invoice dates to calculate recency. Since our last invoice date is 12/01/2010, we have to consider it as the most recent one; we have to subtract each day from the day after to calculate the other ' recency '.

We have to simply add the invoice numbers of each customer to calculate recency.

Designing the model:

K-MEANS: We used the “K - Means algorithm” in designing model. The main parameter of the “K - Means algorithm” is the ‘k’ value which represents the number of clusters the data should be segmented. We need to choose the right number of clusters for that we have two methods: The Elbow method (it takes some random values and represents in a graph by that we can get the best possible k value) and the Silhouette score method (the value which has the highest score will be taken as best k value) using these two we find out the best k value is 3 and trained the model with k value as 3 [24].

DB SCAN: We have also used the “DB-Scan algorithm” in designing the model. The key parameters of DB-Scan are eps and MinPts. Where eps is the environment around the data point, i.e. , if the distance between two points is lower or equal to epochs then they can be considered as neighbors we have taken that as 0.8 and MinPts is the number of points within Eps radius we have taken that as 4 with metrics as Euclidean.

E. Segmentation:

“Data Segmentation” is the method of considering the data and dissecting and organizing similar data based on the chosen parameters so that we can use it more efficiently within marketing operations. Examples of Data Segmentation can be:

Age

Customers Vs Prospects

Therefore, we have segmented the customers based on their recency, monetary, and frequency values which gives better results and helps the company to get profits and to sell their products.

F. Data Visualization:

“Data visualization” depicts the representing information and data in graphs including diagrams such as graphs, maps, charts. “Data visualization” tools provide a fast and accessible way to see and analyze outliers, data trends, and data patterns. Therefore, we have to use matplotlib to represent graphs after training the model according to the clusters segmented using the model here.

G. Implementing the model:

The version is carried out the use of the net interface. If a business enterprise has a much less quantity of orders on a specific day then they could get the info of every patron the use of the interface and advocate the goods in line with that. If there may be any event then the worker can area reductions primarily based totally at the whole evaluation of all of the clients via way of means of dividing them into clusters the use of their recency, monetary, and frequency values.

ALGORITHM 1: Pre-processing and Training on Dataset

```
INPUT: Online retail Dataset
OUTPUT: Trained Model
STEP 1:   Loads dataset.
STEP 2:   Removing the orders in which quantity < 0.
STEP 3:   Transforming the type of the field Invoice Date from object to Date Time
STEP 4:   Adding a field named sale by multiplying quantity and unit price.
STEP 5:   Calculating recency, monetary, and frequency values and applying the k-means algorithm with the best suitable value
          of k.
STEP 6:   Save the model for future use.
```

ALGORITHM 2: Deployment of Customer Segmentation

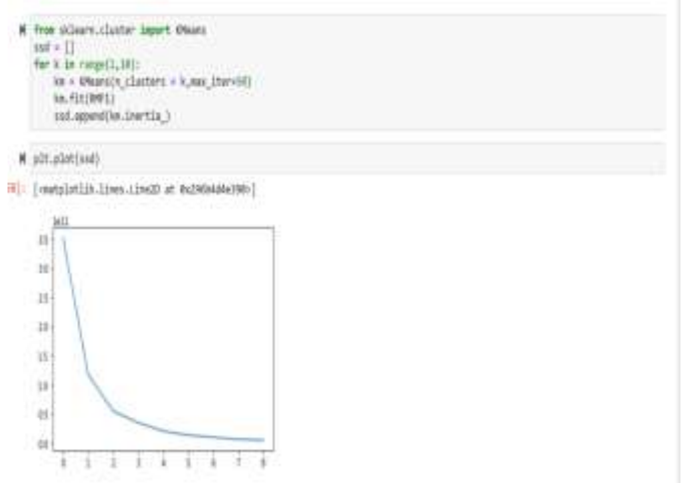
```
INPUT:    Customer Id.
OUTPUT: Recency, Monetary, and Frequency values of that particular user.
STEP 1:   Get the Id given by the customer.
STEP 2:   Search for the recency, monetary, and frequency values of that particular customer.
STEP 2.1: If customer id is present in the dataset:
              Return recency, monetary, and frequency values.
Else:
Return Customer Id is invalid.
```

RESULTS

A. Training the Model:

We have used k-means algorithm for clustering the k value is determined by elbow method as show in the below figure 4.1.1.

Step 1: In the initial step right after prepossessing we had trained the model using some random K values for K-means clustering and we had represented them using a graph. Training with the best value of k determined by the elbow method in the below fig 2.

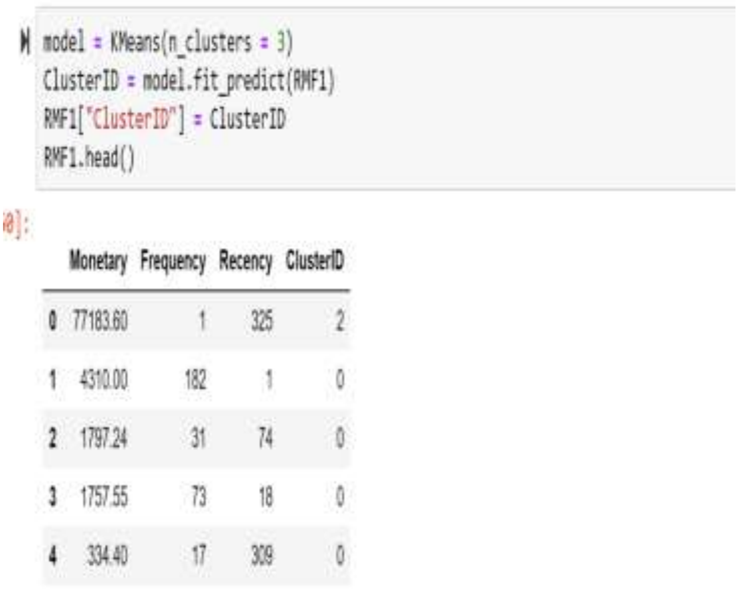


```
from sklearn.cluster import KMeans
ssd = []
for k in range(1,10):
    km = KMeans(n_clusters = k,max_iter=50)
    km.fit(RMF1)
    ssd.append(km.inertia_)
```

```
plt.plot(ssd)
```

```
[<matplotlib.lines.Line2D at 0x2960d4de390>]
```

Fig. 1. Elbow method



```
model = KMeans(n_clusters = 3)
ClusterID = model.fit_predict(RMF1)
RMF1["ClusterID"] = ClusterID
RMF1.head()
```

	Monetary	Frequency	Recency	ClusterID
0	77183.60	1	325	2
1	4310.00	182	1	0
2	1797.24	31	74	0
3	1757.55	73	18	0
4	334.40	17	309	0

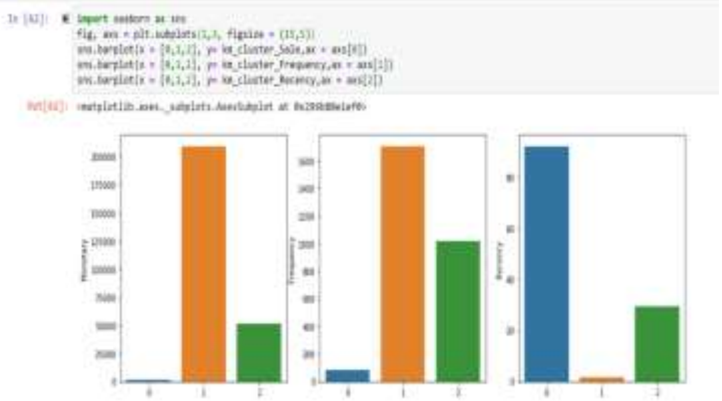
Fig. 2. Training the model

Step 2: By analyzing the above graph we came to the conclusion that trains the model with a k-value of 3 gives us the most accurate results and training the model. The Recency, Monetary, and Frequency values after training the model [22].

B. Result Analysis

Complete data Analysis

If the company wants to know the type of users which categories purchases what kind of products then the company can perform clustering on the previous sales data of the company and analyze the recency, frequency and monetary values of customers so that the management of company can change their marketing and discount sales accordingly. Therefore below is the representation of different clusters.



```
In [42]: import seaborn as sns
         fig, axs = plt.subplots(1,3, figsize = (15,5))
         sns.barplot(x = [0,1,2], y= km_cluster_Sale,ax = axs[0])
         sns.barplot(x = [0,1,2], y= km_cluster_Frequency,ax = axs[1])
         sns.barplot(x = [0,1,2], y= km_cluster_Recency,ax = axs[2])
```

```
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x2958d8e1af0>
```

Fig. 3. Cluster Analysis

The mean of the Monetary, Regency and frequency values of the clusters.

ClusterID	Monetary	Receny	Frequenccy
0	1480.453253	91.965645	83.36095
1	209342.33400	1.600000	1705.600000
2	57188.600769	34.692308	1164.500000

The above values mentioned in the table are the mean of the monetary, recency and frequency values of the clusters.

Hence here we can conclude that concentrating on the customers in cluster with high monetary and frequency values and less recency value gives the company with more profits.

Therefore, if the employee wants to recommend any particular user then they can get the details of the particular user through the interface and the complete customer analysis will be shown in figure according to the clusters. Along with this we have included a feature where they can get the recency, frequency and monetary values of the customer by giving customer id as shown if the customer id is 12346 then

The RFM Values

Name	Values
Recency	771836
Frequency	1
Monetary	325

If the customer id is invalid then it shows that the customer id is invalid

CONCLUSION

The fundamental aim of this undertaking client segmentation is to growth advertising and marketing performance through directing attempt especially closer to the detailed phase in a way regular with the traits of that phase. Where the performance of segments may be completed through device mastering unsupervised algorithms that is K-approach Clustering. And this may be very a good deal useful for the advertising and marketing of products.

REFERENCES

1. Daqing C., Sai L.S, and Kun G., Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining (2012), Journal of Database Marketing and Customer Strategy Management.
2. *K.* Jarrod Millman and Michael Aivazis. Python for Scientists and Engineers(2011). Computing in Science & Engineering.
3. Werner Reinartz,Manfred Krafft,and Wayne D.Hoyer, The Customer Relationship Management Process:Its Measurement and impact on Performance, Journal of Marketing Reasearch,293 Vol. XLI(August 2004),293-305).
4. Darll Rigby,Frederick F.Reichheld,Avoid the Four Perils of CRM,Harvard Business Review, 2002 ,(1):101-109.
5. KOTLER, P. & ARMSTRONG, G. 2006." Principles of Marketing ". New Jersey: Prentice-Hall.
6. STRYDOM, J., JOOSTE, C. & CANT, M. 2008. Marketing Management. Cape Town: Juta.
7. M. J. A. Berry, and G. S. Linoff (2000). Mastering Data Mining. New York: Wiley.
8. J.A.Hartigian "Clustering Algorithms".Wiley.
9. J.A.Hartigan and M.A.Wong (1979) "A K-Means Clustering Algorithm",Applied Statistics ,Vol.28,No.1,p100-108.
10. Cooil, B., Aksoy, L., & KeininghamT. L. (2008), 'Approches to customer segmentation, Journel of Relationship marketing,6(3-4)9-39
11. Bhatnagar, Amit;Ghose, S.(2004) , 'A latent class segmentation analysis of e-shoppers', journal of 758-769 Business Research 57,
12. Rogers,S.& Girolamo,M.(2016),A first course in Machine Learning , Second Edition ,Chapman & Hall/CRC
13. Wagstaff,K.,Cardie,C.,Rogers,S.&Schrodl,S.(2001),Constrained k-means clustering with background knowledge ,in 'Proceedings of the Eighteenth International Conference on Machine Learning ',pp.555-584
14. T. L. (2008), 'Approaches to customer segmentation', Journal of Relationship Marketing 6(3-4), 9–39
15. Jiawei Han,Michelline Kambar,Data mining:conception and technology,Bejjing:Mechanic Industry Publish,2002.
16. Agnes Niam,and Paul Bottomley,Cluster analysis procedures in the CRM era,International Journal of Market Research, Vol 45 Quarter 2 2003.
17. Koh Hian Chye,Chan Kin Leong Gerry,Data mining and customer relationship marketing in the banking industry,Singapore Management Review,2002;24,2

18. Claudio Marcus,A practical yet meaningful approach to customer segmentation,Journal of customer marketing, Vol, 15 No.5 1998,pp 494-504.
19. Andrew Banasiewicz,Acquiring high value,retainable customers,Database Marketing & Customer Strategy Management,2004, Vol.12,1,21-31.
20. Jon Kleinburg,Christos Papadimitriou,Prabhakar Raghavan,Segmentation Problems,Journal of the ACM,Vol.51,No.2,March 2004,pp.263-280.
21. Khodabandehlou, S., Zivari Rahman, M.: Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. J. Syst. Inf. Technol. 19(1/2), 65–93 (2017)
22. Khajvand, M., Zolfaghar, K., Ashoori, S., Alizadeh, S.: Estimating customer lifetime value based on RFM analysis of customer purchase behavior: case study. Procedia Comput. Sci. 3, 57–63 (2011)
23. Wei, J.T., Lin, S.Y., Wu, H.H.: A review of the application of RFM model. Afr. J. Bus. Manage. 4(19), 4199–4206 (2010)
24. I. S. Dhillon and D. M. Modha, “Concept decompositions for large sparse text data using clustering,” Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.
25. Hosseini, S.M.S., Maleki, A., Gholamian, M.R.: Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Syst. Appl. 37(7), 5259–5264 (2010)
26. T. Kanungo, D. M. Mount, N. S. Netanyahu, C.D. Piatko, R.Silverman, and A.Y.Wu, “An efficient K-means clustering algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.
27. Subramanya, K. B., Somani, A.: Enhanced feature mining and classifier models to predict customer churn for an E-retailer. In 2017 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence, pp. 531–536. IEEE, January 2017
28. MacKay and David, “An Example Inference Task: Clustering,” Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292, 2003.

.