IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006160|

Customer Segmentation Based on RFM Model and K-Means Clustering

Girish Kumar Singh [1], Sushil Krishna Patil [2], Dr.S.P.Godse [3]

B.E. Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India[1]

B.E. Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India [2]

Assistant Professor, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India [3]

ABSTRACT: In today's world, people are doing online transactions using smartphones. People buy from e-commerce websites, watch online movies, use social media. Various websites of the same kind are available in the market. So, each business wants to keep customers stick to their websites. They want to make more profit by building better relationships with their customers. This can be done by using customer segmentation. Customer segmentation is the process of analyzing customer's previous transactions to know about their likes and behavior. Using this, it becomes easier to provide services to targeted people. Only those things can be shown to the user which they are interested in. This study represents customer segmentation using the Recency, Frequency, and Monetary (RFM) model. K-means clustering is used to study the behavior of the customer, which was implemented using Python programming language. Tkinter is used to implement the user interface. Users can easily view all the results of this system.

KEYWORDS: Customer Segmentation, K-means Clustering, RFM Model, Python, Recency, frequency, Monetary.

I. INTRODUCTION

People are using and buying online services in enormous demand. Many websites are available which provide the same kind of services. There is a big competition in these websites to provide the best possible services to customers to earn more profit. Over the past few decades, e-commerce websites, social media websites have grown drastically. These websites are generating a large volume of data continuously. By analyzing these data, we can get information related to customer wants, likes, and behavior. Using this information, businesses can provide services to customers according to their likes. This analysis helps the business to understand their customers and take appropriate decisions. Businesses can make more profit if they focus on old customers and their behavior.

We can develop machine learning models to understand purchasing patterns in transactions made by customers. Using these patterns, we can understand customers' behavior and take the necessary decisions to attract them. We can serve them with better services and make more profit. Using customer segmentation, businesses can communicate with various customer subsets. They can establish a better relationship with customers and focus on more profitable customers depending upon their Recency, Frequency, and Monetary (RFM) values. We can use the RFM model to calculate how recently the customer made a purchase, how often he has made purchases, and how much revenue was generated from the customer. Depending upon these values, we can segment and focus on the customers from whom we can get more profit. K-means clustering can be used to segment customers depending on their behavior. Businesses can serve the same services to customers having the same likes and behavior.

Paper is organized as follows. Section II previous work done in the field of customer segmentation. Proposed algorithm is explained in detail in Section III. Section IV presents experimental results. Finally, Section V presents conclusion.

II. RELATED WORK

Various clustering algorithms can be used to understand customer behavior. These are K-means clustering, Fuzzy C-means clustering, and hierarchical clustering. The recency, Frequency, and Monetary (RFM) model can be used

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006160|

for customer segmentation. This analysis is useful for businesses to take necessary decisions and earn more profit [1]. We can also use advanced k-means, k-means, and agglomerative clustering algorithms for customer segmentation along with the RFM model. The comparative study and implementation of these three algorithms are provided in [2]. K-means clustering and RFM model can be also used for customer segmentation [3]. The [4] used the K-means clustering and RFM model to analyze the Retail Customer Churn.

III. PROPOSED SYSTEM

a) System Architecture –

In Fig. 1, we can see that we first perform data preprocessing on the dataset. After that, we calculate the values for recency, frequency, and monetary. Then we apply K-means clustering and RFM segmentation. At last, we can make different marketing strategies using this analyzed data.

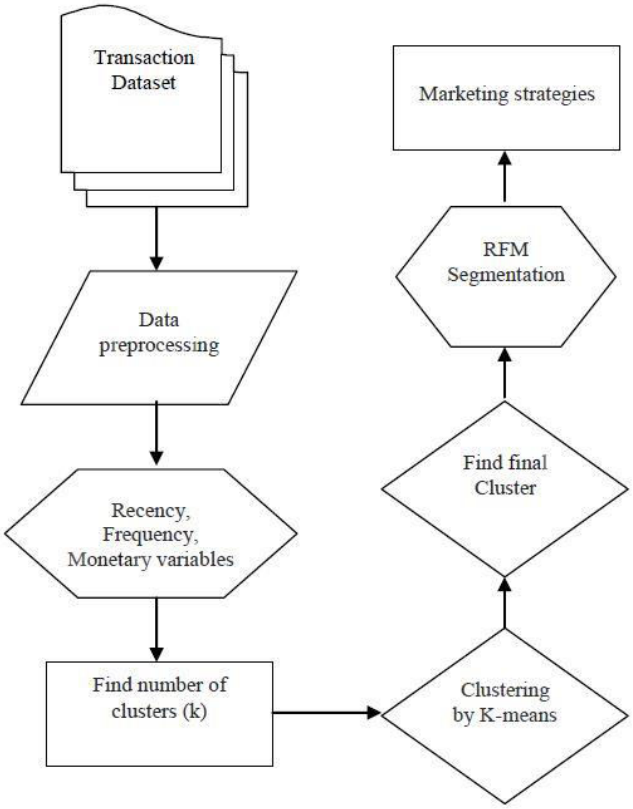

Fig.1. System Architecture for Customer Segmentation using RFM model and K-means Clustering

b) Dataset –

We have used Online Retail Data Set which is available on UCI Machine Learning Repository [6]. It contains transactions from 1 December 2010 to 9 December 2011. It contains 541909 observations, 22190 transactions, 4372 customers, and 3684 products. This dataset contains 8 features: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. We have provided this dataset as an input to our system.

c) Data Preprocessing –

Before using a raw dataset, as discussed in the previous subsection, we performed data preprocessing techniques on it. First, we search for missing values from the dataset. We found 1454 missing descriptions and 135080 missing Customer IDs in the dataset. We can see that 24.92% of the observations didn't have any customer ID. Such observations are not useful for analysis. Hence, we dropped all such rows having missing values. Then we search for duplicate entries present in the dataset. We found 5225 such entries. We dropped all the rows, having duplicate entries. After removing missing values and duplicate entries, we have 401604 observations.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006160|

We performed individual feature exploration. We found that Online Retail Company has customers from 37 countries. Out of these countries, 90% of customers belong to the United Kingdom. So, we considered only those observations whose customers are from the UK. Then we remain with 356728 observations.
We counted the number of unique customers, products, and transactions. The dataset contains 19857 such unique transactions. Then we checked for canceled orders. The customers who purchased only one product and canceled the order do not contribute to the company's revenue. These observations contain the prefix 'C' before their Invoice Number. We checked those entries and found that 3208 transactions were canceled out of 19857 such transactions. So, we dropped these observations. Then we checked for a negative value of quantity. Discounted price also contributes to negative values. So, we dropped all such observations.
We extracted keywords from descriptions and plotted a graph of the top 50 keywords that were appearing in most of the descriptions. See Fig. 2.

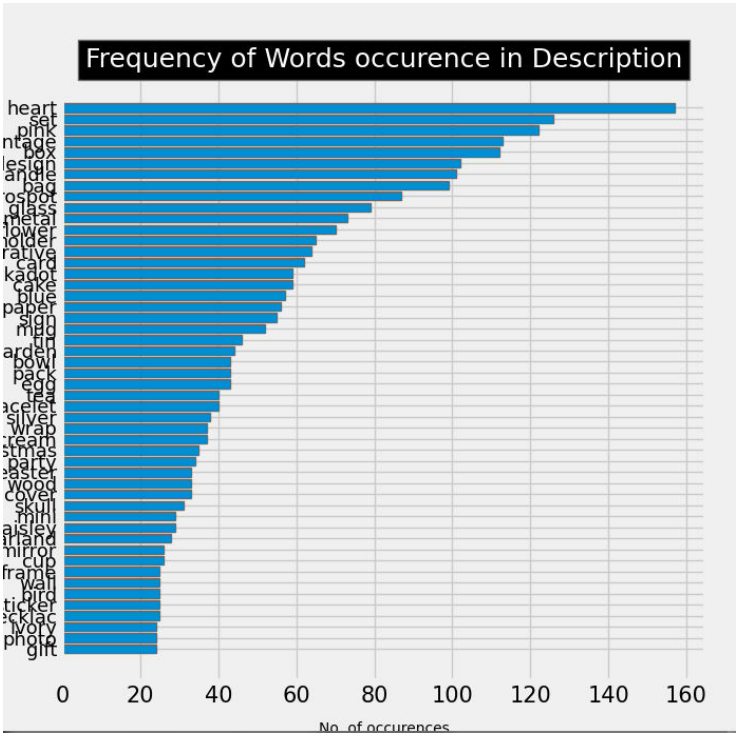

Fig. 2. Frequency of Words Occurrence in Description

We calculated the basket price, which is the total amount of revenue generated by the products purchased by the customer in one transaction. We have used a formula for calculating revenue:
Revenue = (Quantity Purchased – Quantity Canceled) * Unit Price
Then we plot the pie chart for the basket price range. See Fig. 3.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006160|

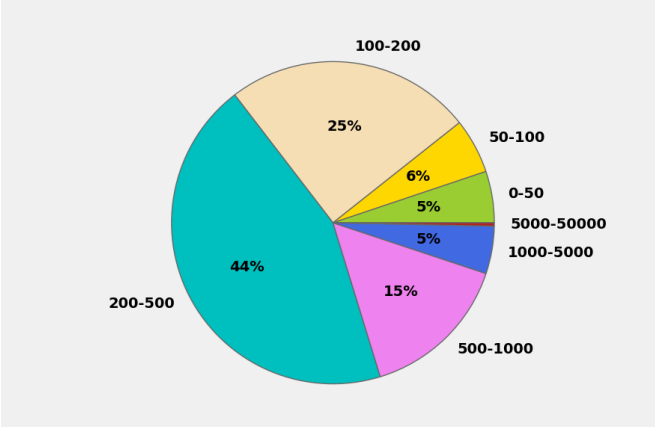

Fig. 3. Basket Price Range

We have used one-hot encoding to convert categorical data into numerical data. We have taken a list of keywords using which we have created the group of products that contains the keywords from that list. If the keyword is present, then we put value 1 otherwise 0 for that product. In this way, we have created a matrix with keywords as rows and products as columns. After encoding, we got a dataset with 3845 observations.

d) Recency Frequency Monetary (RFM) Model –

After data preprocessing, we calculated the values of recency, frequency, and monetary for each customer. Recency means how recently did the customer purchase any product. If the customer has purchased recently, that means he may purchase again as compared to those customers who haven't purchased recently.

Frequency means how often did the customer purchase any product. Higher the frequency, higher is the possibility that these customers respond to new offers.

Monetary means total revenue generated by the customer through their purchases. Customers who have made many purchases contribute more revenue to the business.

e) K-Means Clustering–

We have applied K-Means clustering in two methods. First by selecting features using values of recency, frequency, and monetary. Second, by selecting features by using product categories. The RFM feature selected for our analysis has three scales: Recency having 0-373, Frequency having 1-205, and Monetary having 2.9 – 259657. See Fig. 4.

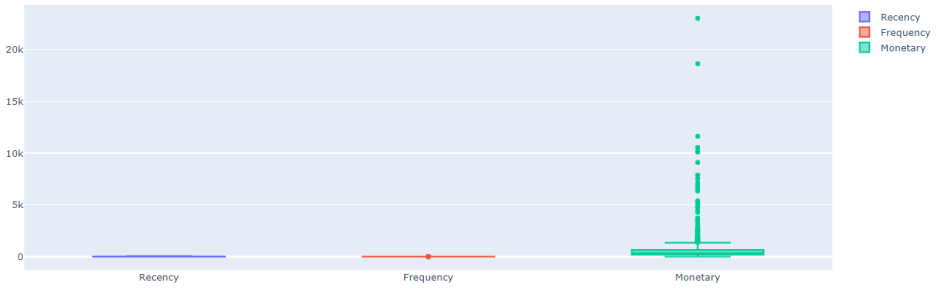

Fig. 4. Clustering using RFM

K-means is the iterative unsupervised learning algorithm. In this algorithm, we calculate centroids to form clusters. This algorithm works in the following steps:

1. For K clusters, define K centroids that are far away from each other.
2. Calculate the distance of each object with respect to the centroid and group them into clusters using this distance.
3. Now based on the elements of each cluster, calculate a new centroid.
4. Repeat the same method to form the clusters of elements based on the new centroid.
5. At every step, the centroid changes, and elements move from one cluster to another.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006160|

6. Do the same process till no element is moving from one cluster to another i.e., till two consecutive steps with the same centroid and the same elements are obtained.

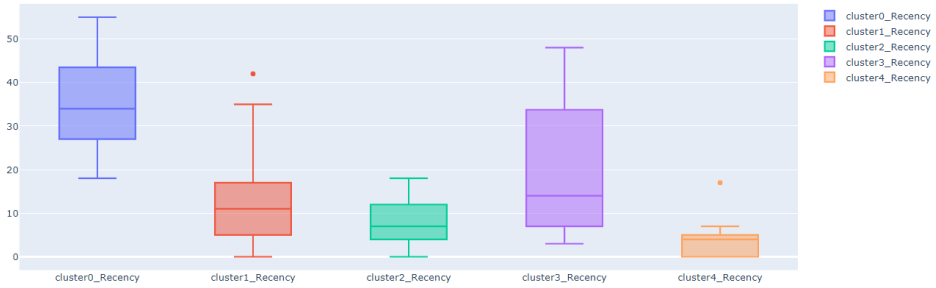

Fig. 5. Recency values for the clusters

Here, to find the optimum number of clusters, we have used the Elbow method and Silhouette Score method. We got K=5 using both the methods. Using this K, we prepared a model using K-means clustering. After preparing the model, we evaluated the model by plotting recency, frequency, and monetary values for the cluster. See Fig. 5, Fig. 6, Fig. 7.

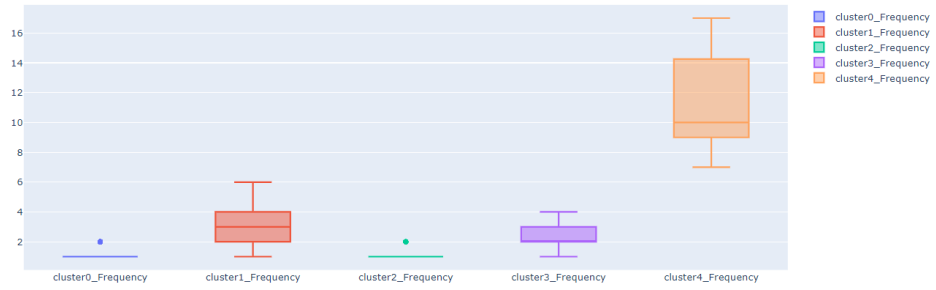

Fig. 6. Frequency values for the clusters

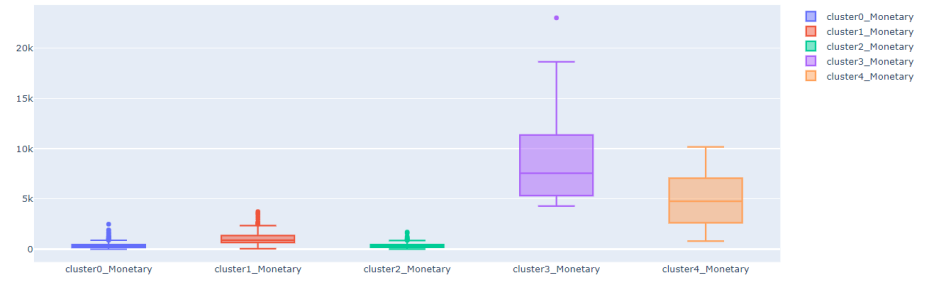

Fig. 7. Monetary values for the clusters

f) User Interface –

We have used Tkinter to implement the user interface for this system. The user interface helps to connect with the system. User can easily get the results for customer segmentation using this user interface. We have created a login system. We allotted only two roles who can access this system. They are Admin and Sales Team Members. Admin can see all the results of segmentation using RFM and K-means clustering. He can view the list of products having a high tendency of purchasing. Also, he can view top customers from the selected cluster. See Fig. 8 a).

A sales team member can view the list of products having a high tendency to purchasing. Also, he can view top customers from the selected cluster. Using these results, sales team can take decisions to make more efficient marketing strategies. See Fig. 8 b).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

|DOI:10.15680/IJIRSET.2021.1006160|

IV. EXPERIMENTAL RESULTS

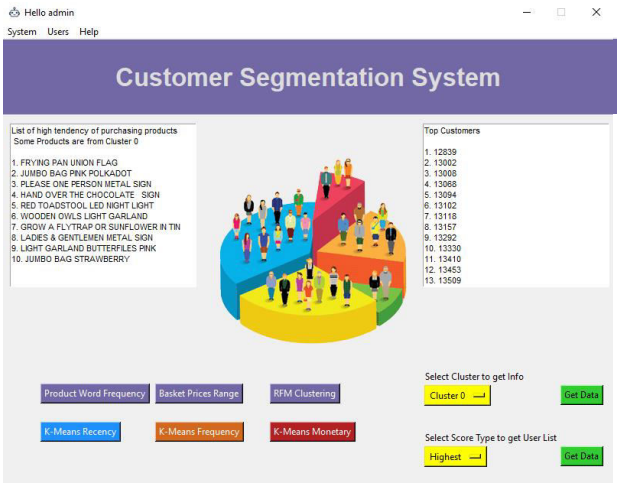

Fig. 8 a) User Interface for Admin

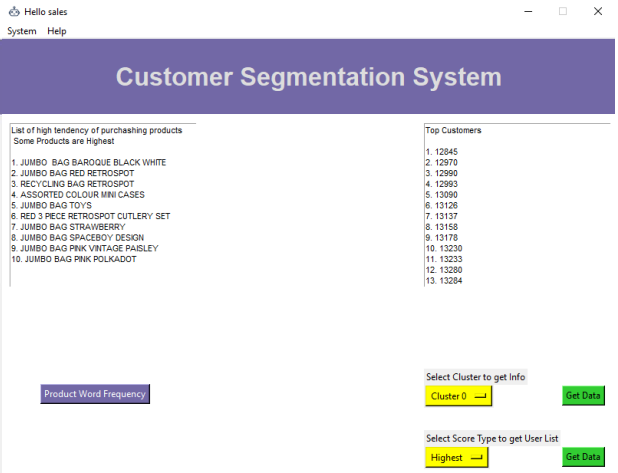

Fig. 8 b) User Interface for Sales Team Member

V. CONCLUSION

Analyzing customers' behavior, their likes, attracting new customers, and maintaining existing customers by providing the best services is the requirement for each business. Customer segmentation allows the business to make the necessary decisions to earn profit.

In this project, we have used the RFM model to segment the customers depending upon their previous transactions. We have created the clusters using their purchasing patterns.

By using these clusters, businesses can make wise decisions and apply marketing strategies to improve their business.

REFERENCES

[1] O. Idowu, A. Annam, E. Rangarajan, S. Kattukottai, "Customer Segmentation Based on RFM Model Using K-Means, Hierarchical and Fuzzy C-Means Clustering Algorithms", ResearchGate, August 2019.

[2] S. H. Shihab, S. Afroge, S. Z. 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),2019.

[3] I. Maryani , D. Riana, R. D. Astuti, A. Ishaq, Sutrisno, E. A. Pratama, "Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm", IEEE, 2018.

[4] N. Bagul, Prof. P. Surana, P. Berad, C. Khachane, "Retail Customer Churn Analysis using RFM Model and K-Means Clustering", International Journal of Engineering Research & Technology (IJERT), March 2021.

[5] P. Anitha, M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm", ScienceDirect, 2019.

[6] Dr Daqing Chen, Online Retail Data Set, https://archive.ics.uci.edu/ml/datasets/Online+Retail

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details