Journal of Engineering Sciences
Vol 12, Issue 06, JUNE /2021
ISSN NO:0377-9254

STUDY OF CUSTOMER SEGMENTATION USING k-MEANS CLUSTERING AND RFM MODELLING

Saurabh Patil[1], Hasnath Khan[2], Sachin Mehta[3] and Prof. Umakant Mandawkar[4]

[1,2,3] *B. Tech, Computer Sciences and Engineering, Sandip University, Nashik, India*

[4]*Asst. Prof, Computer Sciences and Engineering, Sandip University, Nashik, India*

***Abstract*— The technique of separating consumers into distinct categories depending on specific characteristics is known as customer segmentation. But why segment customers? Well, every business is based on manufacturing specific products, and for every product developed there is a defined group of customers. Targeting those defined customers will assure guaranteed sales to the organizations. like all the other essential processes used in business strategies, customer segmentation is also an important step. This integration helps organizations to communicate with a specific group of customers based on their current interests and needs. This method not only helps to target customers but also helps to channelize communication to reach these targeted audiences. Organizations can improve their product qualities, customer services and establish customer relations. Customer segmentation allows organizations to focus clearly on the most profitable customers. Proposed projects on customer segmentation so far are only valid till integrating them into their respective categories, once done the obtained data is then used for further analysis. This project helps validate our data obtained after segmentation of customers into various required groups. To check the correctness of the segmented data and to confirm our theory based on further analysis.**

Keywords*— *k-means, un-supervised learning, segmentation, cluster

1. INTRODUCTION

In last few years a large growth on business sector has been recorded. Business is setting new goals daily and trying their hard to achieve them. This has led to rise of a competitive environment in business sector. It does not matter whether business is small or large, it is competing with others. But the problem is many of the competing businesses are not getting success. There are many reasons why a business might be failing but according to us one of the main reasons for businesses failure is "Companies choosing to avoid learning their customer". Every company has potential, but they fail to understand their market. In short companies fail to Segment the market.

Solution to this issue is understanding what Customer segmentation (aka Market Segmentation).

Customer segmentation can be explained as a game where a kid separates balls, cubes based on their shape or colors.

In simple language customer segmentation is segregating customers, market on different criteria and grouping them based on similar characteristics.

Why now to Use Customer Segmentation?

Today's market is growing at very high speed, so are the customers. The smartphone revolution has brought the customer community closer. Almost everything is shifting on online platform, increasing their reach to large group of customers. Also, customers are accepting this change happily. Each of the customer is generating a large amount of data too. So, why the companies should stay behind? Companies should also transform the way they are working and use available resources to help themselves grow.

Most of the company's goals can be achieved by using customer segmentation.

How would it profit companies?

For example, let's consider company start to use customer segmentation. Suppose company grouped its customer based on geography. Now company has the preview which product is most appreciated at which location. Using this information, company can now plan its advertisement campaign's, strategies and much more. Indirectly it will add profit to business.

3. Literature Review:

Marketing techniques are entirely based on mutual consumer-retailer relationship. One way to increase profits is to determine customer

requirements, through communication with the consumers. Communicating with consumers on personal level is practically not an easy task and without building communication, marketing disasters are inevitable. To tackle this problem retailers can communicate through the data generated by consumers. Retailers can group their consumers according their habits and later on develop business strategies according to it.

Customer segmentation is a manner to enhance communique with the consumer, to realize the desires of the consumer hobby in order that suitable communique may be built. Also, strategies of Customer Segmentation may be categorized into Simple technique, RFM technique, Target technique, and Unsupervised technique. In the aggressive marketplace of e-commerce, the hassle of figuring out capacity consumer is gaining increasingly more attention. This paper proposes the solution to identify potential customers using RFM analysis tool and k-means algorithm.

One of the important steps for customer segmentation is clustering. A MATLAB implementation of k-Means clustering set of rules for consumer segmentation primarily based totally on records accrued from mega enterprise retail outfit, has a purity of 0.95 indicating 95% accurate segmentation of the customers.

4. Methodology

4.1. Methodology Flowchart

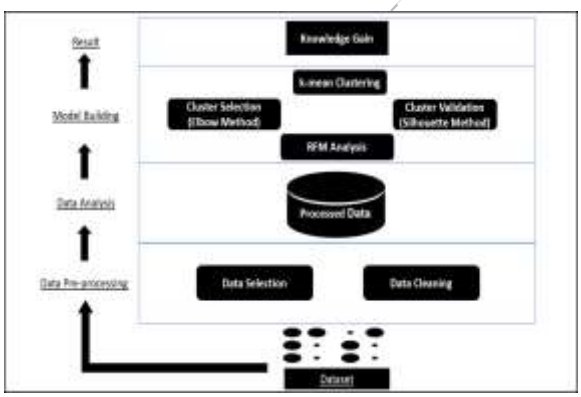

Fig. 1 Flowchart

4.2. Dataset

The data set used to implement clustering and k-means algorithm was collected from UCI Machine Learning Repository. The data set contains 541909 instances and 8 attributes. The attributes of data set consist of Invoice no, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country. The data is in raw format.

4.3. Visualising Dataset

Dataset contains data in raw format, hence it can contain anomalies like negative values, missing values. Visualising dataset help in gaining insight of data in it. Using information of data, we can pre-process data according to our needs. From Table I. we can see that there are certain null values in two columns:1] Description (0.27%) 2] CustomerID (24.93 %)

Table I (Data Description)

Sr no	Column	Non-Null Count	Data Type
1	InvoiceNo	541909	object
2	StockCode	541909	object
3	Description	540455	object
4	Quantity	541909	int64
5	InvoiceDate	541909	datetime64[ns]
6	UnitProce	541909	float64
7	CustomerID	406829	float64
8	Country	541909	object

From Table II. We can see that there are negative values in two columns;
1] Quantity
2] UnitPrice

Table II (Data Preview)

	Quantity	UnitPrice	CustomerID
count	541909	541909	406829
mean	9	4	15287
std	218	96	1713
min	-80995	-11062	12346
max	80995	38970	18287

4.4. Data Pre-Processing and Preparation

From above table we can infer that the data contains some missing and negative values.

Table III (Data Count Before and After)

	Data Count
Before (Pre-processing)	541909
After (Pre-processing)	397924

Data description after pre-processing is shown in below table.

Table IV (Data description after Pre-processing)

	Quantity	UnitPrice	CustomerID
count	397924	397924	397924
mean	13	3	15294
std	180	22	1713
min	1	0	12346
max	80995	8142	18287

Outliers can skew the dataset. They should be removed to have a normalized dataset for k-means clustering. Rescaling of the variables in dataset is important to have a comparable scale.

4.5. RFM (Recency Frequency Monetary)

Recency Frequency Monetary model is an analysis tool which help organizations to identify their valuable customers. This model work on the purchasing history of the customer. There are three important factors: 1] Recency 2] Frequency 3] Monetary.

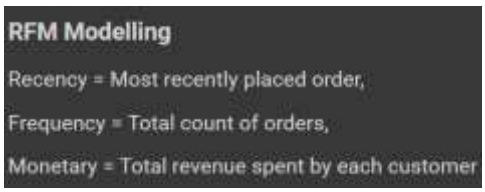

Fig. 2 RFM reference

Table V (RFM statistics)

	R (Recency)	F (Frequency)	M (Monetary)
count	4339	4339	4339
mean	91	91	2053
std	100	228	8988
min	0	1	0
max	373	7847	280206

Table V. shows the statistics of the columns.
RFM Results: Using RFM analysis data was split into four tiers as shown in below Table VI.

Table VI (Final Count)

Tier Level	Value Count
Tier 1	1263
Tier 2	1324
Tier 3	982
Tier 4	770

Tier 1 has the group of customers who are regular.
Tier 2 has the group of customers which spends medium amount of time and money.
Tier 3 has the group of customers who spends less amount of time and money. Company must focus on how to gain their interest.
Tier 4 has the group of customers who rarely purchase product.

4.6. k-means clustering

k-means clustering model is one the vastly used model for clustering. Being unsupervised learning algorithm, it has many applications. It requires the number of cluster's to be formed. The optimal number of clusters can be found by different methods. One of them is Elbow method.

4.6.1. Elbow method

To determine the optimal number of cluster's required for k-means algorithm elbow method can be used. Elbow method generates plot which has an elbow like curve. The point where elbow is formed is taken as the optimal value of k.

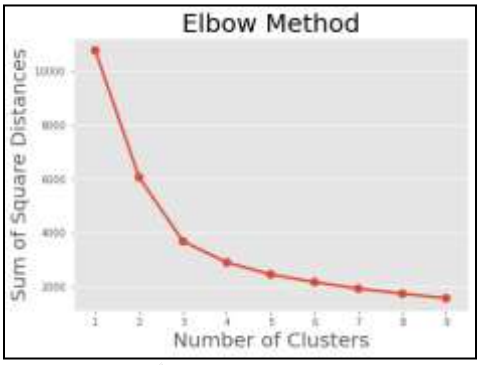

Fig. 3 Graph representing the number of clusters using elbow method.

The elbow is formed at cluster number three. Hence, the number of clusters to use in k-means clustering are k = 3.

4.6.2. Validation of number of optimal clusters

Validation of the number of clusters is important. Silhouette method can be used to validate the consistency of the data. The silhouette value checks the similarity level of an object with its own cluster in comparison to other clutters. The value of silhouette is always between -1 to +1, the cluster which has value close to +1 is valid. [5]

Fig. 4. Validation of number of clusters using silhouette score

The silhouette score is max at 3 number of clusters. Hence, optimal number of clusters are 3.

Conclusion

Customer in Tier 1 have high frequency rate and high spending. Being the most active company can plan some schemes to encourage them. Customer in Tier 2 have good recency in

comparison to other customers. As the recency rate is good, company should plan their campaigns to on board this customer. Customer in Tier 0 are least active customers. Company can avoid this customer.

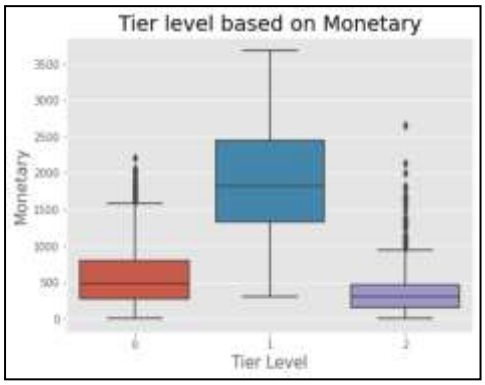

Fig. 5 Graph for Tier level based on monetary

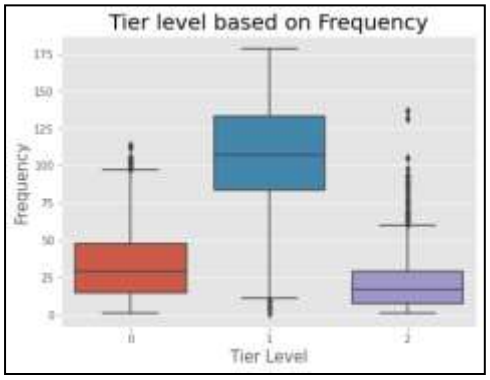

Fig. 6 Graph for Tier level based on Frequency

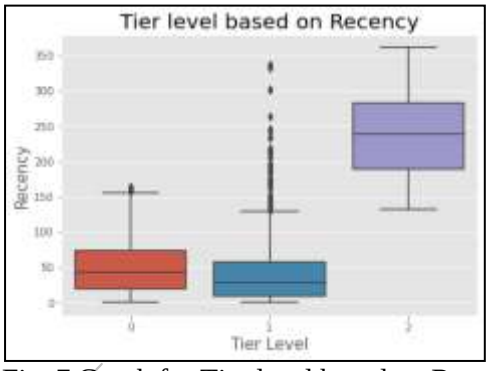

Fig. 7 Graph for Tier level based on Recency

REFERENCES

[1] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

[2] R.C. de Amorim, C. Hennig (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors". Information Sciences. 324: 126–145. arXiv:1602.06989. doi:10.1016/j.ins.2015.06.039.

[3] Leonard Kaufman; Peter J. Rousseeuw (1990). Finding groups in data : An introduction to cluster analysis. Hoboken, NJ: Wiley-Interscience. p. 87. doi:10.1002/9780470316801. ISBN 9780471878766.

[4] Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems. 52 (2): 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377. S2CID 40772241.

[5] Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4), 415-430.

[6] Tkachenko, Yegor. Autonomous CRM Control via CLV Approximation with Deep Reinforcement Learning in Discrete and Continuous Action Space. (April 8, 2015). arXiv.org: https://arxiv.org/abs/1504.01840

[7] Yeh, I-Cheng, Yang, King-Jang, and Ting, Tao-Ming, "Knowledge discovery on RFM model using Bernoulli sequence," Expert Systems with Applications, 2009.

[8] Robert L. Thorndike (December 1953). "Who Belongs in the Family?". Psychometrika. 18 (4): 267–276. doi:10.1007/BF02289263.

[9] Williamson, D & Parker, RA & Kendrick, Juliette. (1989). The box plot: A simple visual method to interpret data. Annals of internal medicine. 110. 916-21. 10.1059/0003-4819-110-11-916.

[10] Bhaya, Wesam. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences. 12. 4102-4107. 10.3923/jeasci.2017.4102.4107.

[11] Li, Youguo & Wu, Haiyan. (2012). A Clustering Method Based on K-Means Algorithm. Physics Procedia. 25. 1104-1109. 10.1016/j.phpro.2012.03.206.

[12] Wei, Jo-Ting & Lin, Shih-Yen & Wu, Hsin-Hung. (2010). A review of the application of RFM model. African Journal of Business Management December Special Review. 4. 4199-4206.