International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

Credit Profile of E-Commerce Customer

Kirti Maheshwari[1], Ria Khapekar[2], Anmol Bahl[3], Kunal Bhatia[4], Prof. Amol Lachake[5]

[1,2,3,4]Dr. D Y Patil School of Engineering and Technology, Dept. of Computer Engineering, Pune.

Abstract - *Creating a positive and negative credit profile for ecommerce customers to minimize the loss incurred by the companies using RFM strategy and machine learning. One of the most popular approach of customer segmentation is based on RFM (Recency, Frequency & Monetary) strategy which helps to form clusters based on their behavior using various clustering algorithms. There are various clustering algorithms, one of the most efficient and appropriate algorithms that can be implemented is K-means algorithm. A comparative analysis of k means and advanced k means concluded that advanced k means would likely be an appropriate option. The study also tells that in respect to the intra cluster distance and inter cluster distance, advanced k means gives better result than the standard k means. Therefore, we have used advanced k means.*

Key Words: **Cluster Analysis, K-means, RFM, CRM,**

1. INTRODUCTION

Due to the advancement in technology, new business are coming up everyday, so it has become more important for the old businesses to stay in the market using different marketing strategies. Nowadays, with the growth of internet across the globe people have started doing things online. From buying stuff to making payment, everything has become digital. Therefore, e-commerce companies make a drastic impact to the economy of the country. Though digital world has made things easier for the man kind, it is also facing losses on the other hand. One of the possible losses that the e-commerce face is about the unwanted cancellations that the customers make. There are times when customers cancels the order after the product has been dispatched from the warehouse. This incurs loss to the company. This research is related to design a method to minimize the losses by identifying the genuine and the fraud customers.

Customer segmentation is one of the way which helps to segment customers based on the similar pattern into same clusters hence providing easiness to handle the large customer base. This segmentation can help to influence the market directly or indirectly as it opens up the paths for company to visualize the type of customer or their needs, it also allows company to find their target customers and minimize the losses.

RFM model for customer segmentation is used for the analysis of customer behavior [16]. Using the RFM variable i.e. recency frequency monetary and couple of other variables credit points will be allotted to each and every customer. Clustering has been proven most effective way to carry out customer segmentation. And by using advance k – means clustering algorithm, customers will be divided into categories. If the customer falls into the worst category, COD and EMI options will be blocked whereas if it falls into a good category, it shall be given some delivery benefits.

1.1 Literature Survey

The literature review and related works on this RFM strategies give an overall idea of how RFM based customer segmentation have been used and implemented over the past years and how clustering algorithms have been implemented for customer segmentation. RFM model for customer segmentation is very vast and widely applied model for the analysis of RFM is a simple and very effective framework to analysis a particular customer on the basis of customer behavior. Which nowadays has become very important for the ecommerce companies for building CRM i.e., Customer relationship management. RFM states recency, frequency and monetary respectively. Recency means how often or latterly a customer has bought a service or product, whereas recency means how regularly a customer buys a product and monetary means up to how much a customer would like to spend to buy a product. RFM factors generally illustrate the following: The more regular and latter customer is buying, the more responsive to the benefits and promotions for that customer. The more frequently a customer buys, the more engaged, happy and satisfied he is and monetary value differentiates between premium spenders and low-value order or service purchasers [2].

These values of recency, frequency and monetary are combined to form RFM scores or credits. For example, in a 4 category ranking system, there can be different RFM score for customer behavior. It has been used by many researchers for segmentation and mining of transactional data [1].

Different customer so according to the range defined by particular company different customers will fall into different ranges. RFM along with other attributes combined clearly shows the categories of different consumers. The best customers are chosen with the highest RFM scores. In this paper, the ranking 1-4 is used to evaluate the customer retention.

For past twenty years, researchers have used RFM model to implement classification, and making predictions using segmentation. A. Fisher, O. Etzion, and S. Wasserkrug classified customers using their lifetime values and probability [3].

What was performed by [4] provides information for e-commerce entrepreneurs, so they can know from each category of customer. And then furthermore make prediction on it.

Then [5] also used RFM to recognize customer value at airlines customer. And From the result of research, there were 4 customer categories that demand company to provide different service to respective customers which fall into that category.

Cheng, C. H. & Chen, Y. S. suggested a data mining model to presume the loyalty of the customer. The customer segmentation model includes RFM analysis and A-priori algorithm. Their idea was to develop and implement an algorithm which will generate RFM patterns of purchase data of the customer [6].

C. Cheng, Y. Chen, C. Lai, C. Hsu and H. Syu exposed another segmentation model to show a classification of two-stage clustering. This model implemented clustering of patients and tried to optimize health care services [7].

Another model was proposed by A. Dursun and M. Caber to cluster hotel customers. Loyal customers, lost customers, new customers, promising customers, highly potential customers are identified by this model [8].

M. Sebastian and K. Nazeer put heads together about some of the major cons of k-means. It generate various clusters for different initial centroids [9].

S. Na, G. Yong and L. Xumin talked over the drawbacks of k-means and then put up an advanced k-means algorithm. The old standard k-means clustering calculates the difference between all the data objects and each of the cluster centroids in every iteration. Where in improved k-means it is not required to calculate all distances in each iteration, thus successfully reduces the total running time [10].

Another agglomerative hierarchical clustering method was first implemented by Day and Edelsbrunner back in 1984 [11].

A popular algorithm called AGNES was introduced by kaufman and Rousseeuw in 1991 [12].

In this paper, researchers transformed the traditional K-Means algorithm based on the triangle inequality theorem, to improve the efficiency of K-means clustering algorithm. Implementation of the improved algorithm on Iris Data and comparison of the performance between the two algorithms. Which results showed the effectiveness of the improved method. They used the improvised algorithm to carry out customer segmentation on the customer dataset which was employed from a company in communication industry.[13]

Furthermore the study also used RFM to process the transaction data of exhaust sales which were then clustered to categorize the customer type of the company. Every month, there are thousands of transactions and based on that the recency , frequency and monetary was calculated for the particular customer or transaction ID. [14]

To be more precise about the type of customer, it is necessary to translate consumer behavior in "number" so that it can be used all the time. In this case the researcher intended to do the test by using RFM Variable on the dataset of credit sale transaction where the amount of the data was very huge. [15]

1.2 Methodology:

Calculating the negative and positive credit profile of the customers considering various factors including RFM so that the company focuses on genuine customers to give out the benefits, thus reducing the losses caused by it.

Step 1: Data Cleaning and Transformation

We will have the database of all the customers that are making transactions. This database will also have some redundancies which is going to be transformed into another sub dataset which is actually been provided to the machine learning algorithm.

Step 2:

Credit point calculation: The total credit score has been decided by RFM strategy along with the mode of payment, valid / invalid returns which will completely describe an individual customer.

Total Credit score = Recency credit + Frequency + Monetary + credit on payment method + credit on return on valid/invalid reasons

Recency credit: These are the points that will reflect how recently a customer is making orders in an interval of time. Say for eg: if a customer orders some product in every 7 days it will be given a higher credit point (say 0.5), if it orders some product in a time gap of 8 to 14 days, it will be given a slightly less credits (say 0.4), and so on.

Table -1: Recency Credit allotment

RECENCY	Code	Credit
First 7days	1	0.05
8 – 14 DAYS	2	0.04
15 – 30 DAYS	3	0.03
31 – 60 DAYS	4	0.02
61 – 100 DAYS	5	0.01

Frequency: It is given by number of orders/ 90

Monetary: credit points will vary with the value of the product bought by the customer.

Table -2: Credit Point allotment.

Price Range	Credit Points
<1001	0.2
100-499	0.25
500-999	0.35
1000-9999	0.5
5000-9999	0.7
10000-50000	0.95
>50000	1.5

Credit on payment method: Whether the customer is paying through shopping wallet, credit/debit card, net banking or COD will also add up to the total credit score.

Credit on return on valid/invalid reasons: There might be some genuine reasons of returning the product, so some reasons have been classified into company's fault and customer's fault.

If it's the fault of company, no points would be deducted from the credit score, but if the fault is customer's, some points will be deducted from the credit score (say 0.5).

Now there might be the question of when to add or subtract the credit points of the customers. Probably, there might be three possible cases.

Case1: If the order is cancelled before it has been dispatched from the warehouse, there will be no change in the credit points.

Case2: If the order has been dispatched and has been delivered to the customer, the credit points for that product will be added to the existing credit score of the customer.

Case3: If the order has been cancelled after the product has been dispatched from the warehouse, the credit points will be deducted from the existing credit score of the customer.

2. ALGORITHM:

K-means Clustering Algorithm:

The K-Means clustering algorithm is a partition-based cluster analysis method [17]. It is used in solving various clustering problem particularly for large datasets.

The algorithm has two different parts. In first part, K number of centers are selected randomly and K is initially fixed. In second part, every data object is taken to the closest center [18]. In order to choose the initial value of K, self-organizing map can be used.

Self-organizing map (SOM) is a technique of data visualization that helps to understand the high dimensional data by mapping it to a lower dimensional space. It also represents clustering concepts by grouping similar things together. By implementing the self-organizing map the value of K can be found that will be fed into thek-means algorithm [19].

To calculate the distance between two data points (center point of the cluster and data object. Euclidian distance measure is used in general. The criterion function used in k-means is as follows:

$$E = \sum_{i=1}^{k} \sum_{x \in Ci} |x - x_i|^2 \quad (1)$$

Here E is called as sum of squared error. The distance function used in this equation is the Euclidean distance. The Euclidean distance d (xi, yi) has the following form:

$$d(x_i, y_i) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2} \quad (2)$$

The K-means algorithm calculates cluster centers iteratively as shown

(1) Initialize K centers locations (1 c ,..., k c) using random sampling.

(2) Assign each i x to its nearest cluster center k c .

(3) Calculate new k c centers as:

$$c_k = \frac{\sum_{x_i \in F_k} x_i}{|F_k|} \quad (2)$$

F_k the number objects in the k^{th} cluster

(4) Repeat steps 2 and 3 till the cluster centers remain the same.

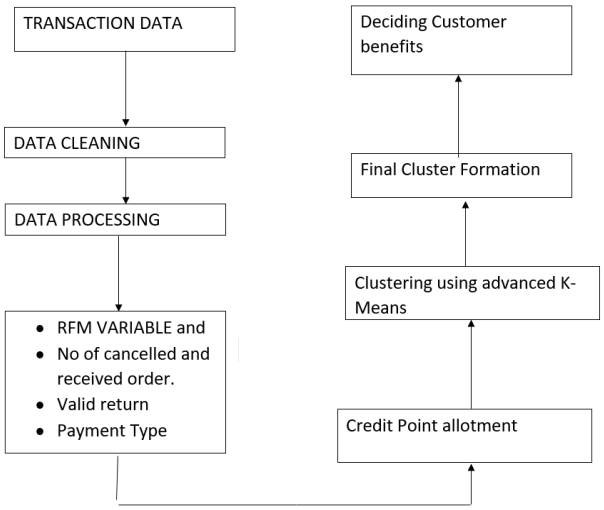
Advanced K-means Clustering Algorithm: h

The enhanced version of k-means has almost the same structures as standard k-means. However, it has some additional features. The algorithm has two data structures to store the labels of clusters and the distances between each data object to the closest centers in every iteration. In the next iteration, this information will be used.

In each iteration, the distance between every object and the new center is calculated. If the distance is less than or equal to the earlier distance of the object to its cluster center, then the object will remain in its cluster. There is no necessity to compute the distances to the remaining clusters. However, if the distance is greater than previous value, then the distance between the object and all the remaining cluster centroids are calculated. After that, the labels and the distances are updated. In this way, the total running time will be reduced.

The algorithm of improved k-means is almost the same as standard k-means except some extra steps shown. This algorithm is applied for RFM based customer segmentation approach and thus allot credit points to customers and further categories them into respective clusters.

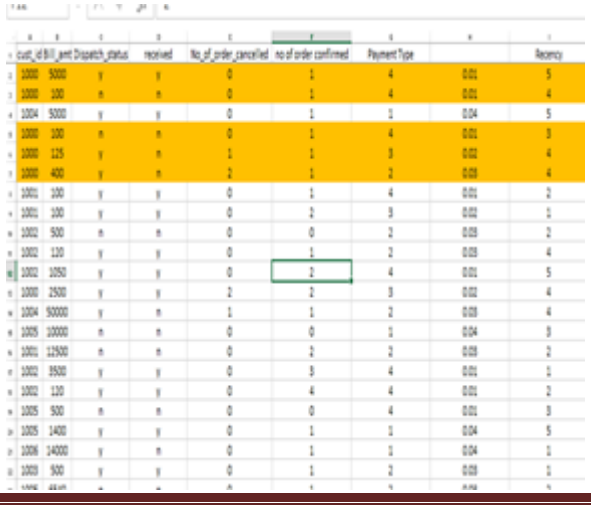
3. PROPOSED MODEL:

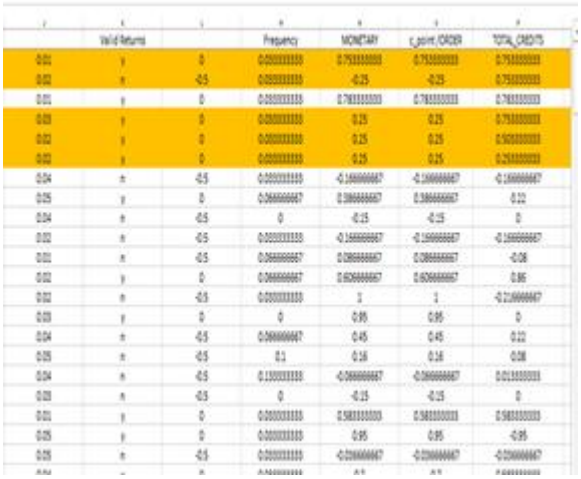


3.1 Experimental Analysis

Dataset Description

Table 3-Transaction details of all the customers





The total credit score has been calculated using the below formula specified in the calculation.

Table 4 - Last transaction detail of a particular customer

	A	B	C
1	cust_id	No_of_order_cancelled	no of order confirmed
2	1000	2	2
3	1004	1	2
4	1001	0	2
5	1002	0	4
6	1005	0	2
7	1006	1	1
8	1003	0	1
9	1010	1	0
10	1014	1	0
11	1045	2	5
12	1036	2	3

D	E	F	G	H
c_point /ORDER	RATIO	TOTAL_CREDITS		
0.606666667	1	0.86		
-0.043333333	0.5	-0.26		
0.45	0	0.22		
-0.066666667	0	0.013333333		
0.806666667	0	1.39		
0.45	1	-1.4		
-0.036666667	0	-0.036666667		
0.95	1	-0.95		
0	1	-0.35		
0.436666667	0.4	0.84		
1.11	0.666666667	0.133333333		

This customer dataset when provided to the machine learning algorithm will provide us with 4 clusters, namely, excellent customers, good customers, average customers and worst customers. Based on the cluster in which a particular customer falls, actions will be taken.

As in, excellent customers will be granted with emi and cod options with some delivery benefits, good customers will be granted emi and cod options, average customers will be given cod options, whereas worst customers will be blocked from any kind of perks.

The company might even want to set a time quantum of say 6 months, which means if any customer is in the worst customer cluster, it would be set back to an average customer so that the company doesn't losses its customer.

3. CONCLUSION

Thus, we have studied about the possible approach towards making of a credit profile of a e-commerce customers. In this era of advancement in technology, there is a lot of competition arising. All the multinational e-commerce giants are aiming to increase their profit, decrease their losses by increasing reach and also minimize the transportation cost. Our system will give these MNC's a upper hand on those particular individual customer who are trying to mess around with the companies. The e-commerce companies will able to differentiate between genuine and fraud customers through this machine learning system in which they only have to use the pervious datasets and not much changes have to be made in their current system.

4. REFERENCES

[1] P. Spring, P. Leeflang and T. Wansbeek, Journal of Market-Focused Management: The Combination Strategy to Optimal Target Selection and Offer Segmentation in Direct Mail. Kluwer Academic Publishers, 1999, pp. 187-203.

[2] P. Fader, B. Hardie and K. Lee, "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis", Journal of Marketing Research, vol. 42, no. 4, pp. 415-430, 2005.

[3] O. Etzion, A. Fisher, and S. Wasserkrug, "e-CLV: a modelling approach for customer lifetime evaluation in e-commerce domains, with an application and case study for online auctions," IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004. EEE 04. 2004, 2004.

[4] Rachid, et al. 2015. "Combining RFM Model and Clustering Techniques for Customer Value Analysis of a Company selling online." 2015 12th International Conference of Computer Systems and Applications (AICCSA) 2015,1-6.

[5] Liu Jiali and Du Hyung. 2010. "Study on Airline Customer Value Evaluation Based on RFM Model (2010)." 2010 International Conference On Computer Design And Appliations (ICCDA 2010) ,278-281

[6] C. Cheng and Y. Chen, "Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications, vol. 36, no. 3, pp. 4176-4184, 2009.

[7] Y. Chen, C. Cheng, C. Lai, C. Hsu and H. Syu, "Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment", Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221, 2012.

[8] A. Dursun and M. Caber, "Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis", Tourism Management Perspectives, vol. 18, pp. 153-160, 2016.

[9] K. Nazeer and M. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering WCE 2009, July 1 - 3, vol. 1, 2009.

[10] S. Na, L. Xumin, and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010.

[11] W. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods", Journal of Classification, vol. 1, no. 1, pp. 7-24, 1984.

[12] W. Fox, L. Kaufman and P. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis.", Applied Statistics, vol. 40, no. 3, p. 486, 1991 .

[13] Xiaoping Qin, Shijue Zheng, Ying Huang, Guangsheng Deng,"Improved K-Means algorithm and application in customer segmentation",2010 Asia-Pacific Conference on Wearable Computing System(IEEE).

[14] Maryani, Ina, and Dwiza Riana. 2017. "Clustering and Profiling of Customers Using RFM for Customer Relationship Management Recommendations." 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017, 2–7. https://doi.org/10.1109/CITSM.2017.8089258.

[15] Ina Maryani 1, Dwiza Riana 2, Rachmawati Darma Astuti 3, Ahmad Ishaq4, Sutrisno 5, Eva Argarini Pratama6 ,"Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm",2018 3rd International Conference On Informatics and computing(ICIC).

[16] P. Spring, P. Leeflang and T. Wansbeek, Journal of Market-Focused Management: The Combination Strategy to Optimal Target Selection and Offer Segmentation in Direct Mail. Kluwer Academic Publishers, 1999, pp. 187-203.

[17] Y. Chen, C. Cheng, C. Lai, C. Hsu and H. Syu, "Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment", Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221, 2012.

[18] A. Fahim, A. Salem, F. Torkey and M. Ramadan, "An efficient enhanced k-means clustering algorithm", Journal of Zhejiang UniversitySCIENCE A, vol. 7, no. 10, pp. 1626-1633, 2006.

[19] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map", IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 586-600, 2000.