International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 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*: : Customer segmentation, SVM, Random forest, RFM**

1. INTRODUCTION

Due to the advancement in technology, new business are coming up every day, 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].

This approach suggests that the mass market consists of some number of relatively homogeneous groups, each with distinct needs and desires.[10]

A different model proposed that the market segmentation is no longer limited to variables of customer behaviour characteristics, poses the afterwards market segmentation strategy based on attitude variables. Describes traditional K-means and SOFM cluster methods, proposes SVC(support vector clustering) algorithm to conduct market segmentation.[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]

The SVM is a learning machine algorithm, can reduce the segmentation error which caused by fast motion of the object. A hierarchical decomposed SVM binary decision tree is used for classification. Experimental results show that the algorithm is effective and robust.[15]

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. [16]

Using Random forest to investigate to find relationship between consumer behaviour to buy products on changing parameters such as environmental factor, organizational factor, individual factor and interpersonal factor[17].

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 , frequency and monetary are calculated as follow:

- Recency= latest date – last invoice date
- Frequency = count of invoice number of transactions
- Monetary = sum of total order amount

After the calculation of RFM variables, the whole range is divided into four segments using quantiles.

Q = [0.25 , 0.50 , 0.75]

Allotting credit score

- After dividing RFM in four quantiles, grant the credit score as follows:

For frequency and monetary

Table -1: Frequency and Monetary

Quantile	Credit
0.25	4
0.50	3
0.75	2
0.1	1

Table -2: For recency

Quantile	Credit
0.25	1
0.50	2
0.75	3
0.1	4

Table -3: For mode of payment

Mode of payment	Credit
Online payment	2
Cash on delivery	1

Table -4: For possible return case

PAYTM'S FAULT	**CUSTOMER'S FAULT**
Wrong item delivered	By mistake order placed
Defected item	No longer needed
Didn't order this	Got at cheaper rate

- if invalid return >5 then deduct 2 points from total credit score
- If invalid return >10 then deduct 3 points from total credit score

Classification based on the credit score

- Based on the individuals credit score, each will fall onto either of one category.
- The calculated credit score is again segmented into four quantiles and hence four classes are created as follows:

Table - 5: Classification based on the credit score

Loyalty level	**Category**	**Benefits**
0	Worst customer	No benefits and no COD
1	Average customer	COD option available
2	Good customer	COD and EMI available
3	Excellent customer	COD , EMI , offers , discounts and prime delivery benefits

2. ALGORITHM:

Support vector machine (SVM)

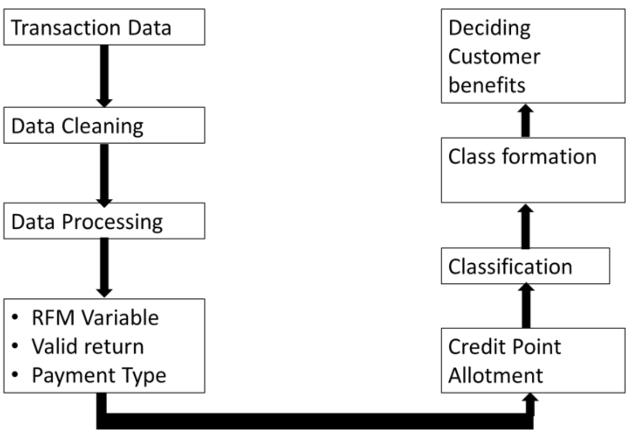
SVM is an algorithm that takes the data as an input and outputs a line that separates those classes if possible.
Suppose some given data points each belong to one of two classes, and the goal is to decide which class a *new* data

point will be in. In the case of support-vector machines, a data point is viewed as a p-dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a p -dimensional hyperplane This is called a linear classifier.
The accuracy of this algorithm comes out to be 87%.

Random forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.
The prerequisites for random forest to perform well are:
There needs to be some actual signal in our features so that models built using those features do better than random guessing.
The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.
The accuracy of this algorithm comes out to be 99%.

3. PROPOSED MODEL:



3.1 Experimental Analysis

Dataset Description

Dataset-1

	A	B	C	D	E	F
1	InvoiceNo	StockCode	Descriptio	Quantity	InvoiceDate	UnitPrice
2	536365	85123A	WHITE HA	6	12-01-2010	2.55
3	536365	71053	WHITE ME	6	12-01-2010	3.39
4	536365	84406B	CREAM CU	8	12-01-2010	2.75
5	536365	84029G	KNITTED U	6	12-01-2010	3.39
6	536365	84029E	RED WOO	6	12-01-2010	3.39
7	536365	22752	SET 7 BABI	2	12-01-2010	7.65
8	536365	21730	GLASS STA	6	12-01-2010	4.25
9	536366	22633	HAND WA	6	12-01-2010	1.85
10	536366	22632	HAND WA	6	12-01-2010	1.85
11	536367	84879	ASSORTED	32	12-01-2010	1.69
12	536367	22745	POPPY'S PI	6	12-01-2010	2.1
13	536367	22748	POPPY'S PI	6	12-01-2010	2.1
14	536367	22749	FELTCRAF	8	12-01-2010	3.75
15	536367	22310	IVORY KNI	6	12-01-2010	1.65
16	536367	84969	BOX OF 6 A	6	12-01-2010	4.25
17	536367	22623	BOX OF VI	3	12-01-2010	4.95
18	536367	22622	BOX OF VI	2	12-01-2010	9.95
19	536367	21754	HOME BUI	3	12-01-2010	5.95
20	536367	21755	LOVE BUIL	3	12-01-2010	5.95

G	H	I	J	K	L
CustomerI	Country			date	time
17850	United Kingdom			12-01-2010	08:26:00
17850	United Kingdom			12-01-2010	08:26:00
17850	United Kingdom			12-01-2010	08:26:00
17850	United Kingdom			12-01-2010	08:26:00
17850	United Kingdom			12-01-2010	08:26:00
17850	United Kingdom			12-01-2010	08:26:00
17850	United Kingdom			12-01-2010	08:26:00
17850	United Kingdom			12-01-2010	08:28:00
17850	United Kingdom			12-01-2010	08:28:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00
13047	United Kingdom			12-01-2010	08:34:00

Dataset-2 (classification)

	A	B	C	D	E	F
1	R	F	M	payment n	credits on	RFM_Loyalty_Level
2	1	1	4	1	-2	0
3	4	4	4	1	0	3
4	4	4	4	1	-3	2
5	2	4	4	2	-3	2
6	4	2	2	1	-2	1
7	1	1	1	2	-2	0
8	4	3	3	1	-3	1
9	4	1	4	1	0	2
10	4	2	2	2	-2	1
11	4	3	3	1	-2	2
12	4	3	3	2	0	3
13	1	1	2	2	-2	0
14	3	3	4	1	-2	2
15	1	1	1	2	-2	0
16	3	2	1	2	-2	1
17	2	2	2	1	0	1
18	1	2	2	2	0	1
19	4	4	4	1	-2	3
20	2	1	1	1	-2	0
21	1	3	2	1	0	1
22	4	4	4	1	0	3

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. CONCLUSIONS

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.
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 Applications (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. 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. 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).
10. 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.
11 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
12 J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map", IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 586-600, 2000.
13 W. Yinghui and L. Xilin, "Research on SVC Algorithm in Customer Segmentation of KIBS," 2010 International Conference on Intelligent Computation Technology and Automation, Changsha, 2010, pp. 128-131, doi: 10.1109/ICICTA.2010.543.
14 LUO Ji-ning, "Market Segmentation Research: Critical Review and Perspectives [J]", *Journal of Shandong University(Philosophy and Social Sciences)*, no. 6, pp. 44-48, 2003.
15 Tang Faming, Wang Zhongdong, Chen Mianyun. An Improved Multiclass Support Vector Machines Based on Binary Tree[J], Computer Engineering and Applications, 2005.(7):24-26
16 Torizuka, K. & Oi, H. & Saitoh, F. & Ishizu, S.. (2018). Benefit Segmentation of Online Customer Reviews Using Random Forest. 487-491. 10.1109/IEEM.2018.8607697
17 H. Valecha, A. Varma, I. Khare, A. Sachdeva and M. Goyal, "Prediction of Consumer Behaviour using Random Forest Algorithm," 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gorakhpur, 2018, pp. 1-6, doi: 10.1109/UPCON.2018.8597070.