Innovations

Analysis of Customers Purchasing Behavior with RFM Model Using Machine Learning

Kavitha[1] and Dr. S. Manikandan[2]

[1]Research Scholar, Research & Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India.
[2]Research Supervisor, Professor & Head, Department of IT, K.Ramakrishnan College of Engineering, Trichy, Tamil Nadu, India.
Corresponding Author: **Kavitha**

Abstract

Customers buying behavior has been predicted in modern techniques such as predictive analytics such as the RFM model and clustering algorithms have helped a lot to entice a customer. In the competitive world, business people need to predict the pulse of customer shopping behavior with predictive analytics it is possible to see what a customer like more and shows interest in a particular product. The objective of this research paper is to retain loyal customers and attract new customers. In this paper, the implementation of a clustering algorithm to find the frequently purchased products has been conducted to analyze the purchases of a customer. Initially, the purchases have been studied by grouping the customer based on their country, name of products, and quantity of products on their past purchases, after that we get loyal customers for improving sales and profit.

Keywords: *1.Predictive Analysis, 2.RFM model, 3.Clustering Algorithm*

Introduction

Customer profiling data is segmented based on their past purchases like their country name, name of product and quantity of product, and invoice date, those attributes are categorized based on their past purchases this leads increase sakes, profit and making a marketing strategy.[1] Abbasimher and Shabani have proposed a new methodology that handles the issue of the dynamic behavior of customers over time. A new methodology is executed based on time series clustering to extract foremost behavioral patterns of customers over time. This methodology is implemented using bank customers' transaction data which are in the form of time series data. The data include the recency (R), frequency (F), and monetary (M) attributes of businesses that are using the point-of-sale (POS) data of a bank. This data was obtained from the data analysis department of the bank. Findings After carrying out an empirical study on the acquired transaction data of 2,531 business customers that are using the POS devices of the bank, the dominant trends of behavior are discovered using the proposed methodology. [2]AhamedHosainiand Kuldeep Chand Rojhe proposetheir paper this an extensive review on the influencing factors on consumers behavior and their buying decision-making process in marketing.

Marketing starts and finishes with the consumer hence, consumer purchasing decision taking shows how well the organizations' marketing strategy suits marketing demand. Consumer behavior includes the emotional procedures which consumers experience in understanding their requirements. Discovering patterns to rectify these requirements, taking buying decisions for example, whether to purchase goods and services and if so, which types of brands and where, interpreting tips, making plans, and executing these plans for example, by engaging in comparison shopping

or real buying of products, modern and professional marketing staffs try to know consumers and their responses, therefore, analyses the essential traits of their behavior. [3]Megha Grover and RishitaGoyal's Consumer reviews may influence overall product sales and help a customer in purchasing decisions. This research paper has been conducted with the perspective of finding the correlation between consumer reviews and consumer purchasing behavior. This study aims to disentangle the effect of different online reviews on consumers' purchasing behavior and intentions.
[4]N.Jain and Ahuja proposethe research study aims to segregate online consumers into diverse consumer segments based on of their online shopping behavior. In this paper, the diverse stages of the consumer decision-making process have been discussed and this study explores only three important phases of consumer behavior in impacting the pre-purchase decision.[5]P.Anitha and Malini Patel to execute and apply the scientific approach using the K-Means algorithm, and thereal-time transactional and retail datasets are analyzed. Spread over a specific duration of business transactions,the dataset values and parameters provide an organized understanding of customer buyingpatterns and behavior across various regions.

[6]MD Uncles and GR Dowling execute and apply the scientific approach using K-Means algorithm, the real time transactional and retail dataset are analyzed. Spreads over a specific duration of business transactions, the dataset values and constrains provide an organized understanding of the customer buying patterns and attitudes across various regions. [7]T.Turcineck J.Stastny and Motickya their study concluded that the application of cluster analysis on the number of such attributes is possible, but not all types of methods are suitable for this purpose. However, the results need to consult an expert in consumer behavior, if the results are relevant. This type of data would be useful to test other approaches such as the creation of association rules. The research on consumer behavior in the food market has performed the analysis of data by following the methods of cluster analysis: DBSCAN, K-means, and Hierarchical methods.

2. Materials and Methods

There are many clustering algorithms to select from and no single best clustering algorithm for all cases. Instead, it is a good idea to explore a range of clustering algorithms and different configurations for each algorithm.This article demonstrates the concept of segmentation of a customer data set from an e-commerce site using k-means clustering in python.We will use the K-means clustering algorithm to derive the best number of clusters and understand the core customer segments based on the data provided.

K-means Algorithm is one of the best segmentation algorithms which is used for predictive analysis or customer profiling. The K-Means algorithm helps to find the number of clusters based on the input given by the user. It's a well-known algorithm because it takes only one input at the same time performs faster in a larger set of data, also reduce the unclassified data from the given dataset

Fig.1depicts the framework for the proposed methodology, this methodology divides the work into three major parts.

Phase 1: Data preprocessing

Data preprocessing is used to prepare and clean the data for our model, which means cleaning and formatting the data. It helps to check the data quality in terms of accuracy, completeness, consistency, timeliness, believability, and interpretability.Data preprocessing for data cleaning, data integration, data reduction, and data transformation

Phase 2: Analysis of RFM

The completion of phase 1,is to analyze the recency which means recent transactions, then frequency which means how frequently they purchased then monetary which means how much amount spent for their all transactions. Finally, create a variable that is used for reference as a date one day before the last transaction.

Recency:
The number of days a customer made the last purchase before the reference date purchase. More recently the customer has transacted with the shop.
Frequency: A customer how often interacted with the shop during a particular period.
Monetary: It indicates how much amount a customer has to spend to purchase a product in a particular period.
Phase 3: b) K-Means algorithm is applied using Euclidean distance metric to partition the customers for RFM values. K-Means is used double to analyze the amount got for Recent and Frequent trades as mentioned below:
i) To partition the customers based on the amount generated with recent transactions.
ii) To group the customers on the amount

Phase 4: Clusters Evaluation
Let K=number of clusters. Using the Elbow method to identify the optimal clusters based on the value. After the analysis, compare the recency of the sale with sales amount and sales frequency with sales amount from one cluster to another cluster respectively. This helps in identifying the group of customers having the highest sales recency, sales frequency, and sales amount.

2. b) Methods
Clustering using the K-means algorithm is a method of unsupervised learning used for data analysis. This algorithm identifies 'K' centroids from the dataset 'D' and assigns the no overlapping data points to each of the nearest clusters. The intra-cluster distance is maximum compared to the inter-cluster distance in the K-means algorithm. Since it is an iterative approach, data points are moved to different clusters, based on the calculation of the centroid. As per the pseudo algorithm shown in figure 2, the mathematical model for the manual calculation of silhouette for an object is given below. Consider K-clusters of which each cluster holds variable objects. Since K-Means is applied twice in the present experiment, objects are clustered based on customer transaction data for recencyvs monetary and frequency vs monetary values.
K= {{(p1, q1), (p2, q2)….(px, qx)},{(p1,q1),(p2,q2)…(py,qy)},……{(p1,q1),(p2, q2)…(pz, qz)}}
Where, K= number of clusters, (p,q)=object in a cluster.

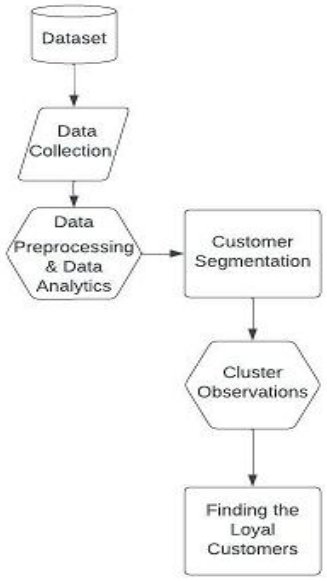

Fig 1: Framework for Proposed Method

Identify any point, for example {p1, q1} in cluster 1. Objects in a cluster represent RFM values. Calculate the average distance from {p1,q1}to all objects of the same cluster (intra distance value ai). Calculate the average distance from {p1,q1}to the objects of other clusters as given in equation 1.

$$\sqrt{\sum_{i=1}^{n}(p1-pi)+}$$ (Q1-Qi)2 ----------(1)

Repeat the same procedure for other clusters and find the minimum average distance from {p1,q1} of cluster 1 to cluster 2,3…. m (y(i)).

Fig 2: K- Means Algorithm

Input Taken: D: Dataset with 'n' instances K: Total number of clusters. Output: Dataset partitioned into 'K' clusters Algorithm: 1. Take arbitrarily 'K' random points from D as the cluster centers. 2. Repeat and reassign each object to the clusters based onthe calculation of mean value using Euclidean Distance measure. 3. Revise the cluster means by recalculating the mean value of each cluster 4. Until there **is** no change in the clusters obtained

2. c) Data set and Description: The description of the dataset is shown in Table 1. The dataset has been collected from Kaggle Repository; it has 541909 instances and 8 attributes which are described in the following table.

Table 1: Description of the Dataset

S.No.	Attribute Name	Type of the Attribute	Description of the Attribute
1.	Invoice Number	Nominal	6 Digit number is assigned for each transaction
2.	Stock Code	Nominal	5 Digit number is assigned to a distinct product
3.	Description	Nominal	Name of the product
4.	Quantity	Numeric	Quantity of each product per transaction
5.	Invoice date	Numeric	Date and time of each transaction generated at the time of the transaction
6.	Unit Price	Numeric	Product price per unit of transactions
7	Customer Id	Numeric	5 digit number assigned to each customer
8	Country	Character	Name of the Country

3. Experimental Results

The data set has been from Kaggle, it has 541909 instances and 8 columns, using python code to drop out duplicates. In this dataset having 16 countries, most UK people used this for purchasing, so keep only UK shoppers also checked missing values.
Having negative values, and then converting the date string field into date-time type. Introduce a new column namely Total amount, this column has the entries multiplying Quantity and unit price, after cleaning the dataset having the 354345 instances 9and 9 columns including the total amount into the dataset.
So, ignore missing values of the Customer ID and Description, also validate any negative value in quantity and unit price, and filter out records.

Table 2: Missing Values of attributes

InvoiceNo	0
StockCode	0
Description	1454
Quantity	0
InvoiceDate	0
Unit Price	0
CustomerID	135080
Country	0

The five records of the dataset are

Table 3: After including the total amount

Ins No.	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalAmount
0	536365	85123A	White Hanging Heart T-Light Holder	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	White Metal Lantern	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	Cream Cupid Hearts Coat Hanger	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	Knitted Union Flag Hot Water Bottle	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	Red Woolly Hottie White Heart.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

3.a) RFM Modeling

Recency = Latest Date - Last Invoice Data, Frequency = count of invoice no. of the transaction(s), Monetary = Sum of TotalAmount for each customerSet Latest date 2011-12-10 as last invoice date was 2011-2-09. This is to calculate the number of days from recent purchase Latest Date = dt.datetime(2011,12,10) Create RFM Modeling scores for each customer. Now convert the invoice date into anint, then rename the column names likely invoice date into recency,Invoice No into Frequency, Total Invoice Date into Recency, InvoiceNo into Frequency then TotalAmount into Monetary now the data index will become

Table 4: DescriptiveStatistics of Recency

count	4339.000000
mean	92.041484
std	100.007757
min	0.000000
25%	17.000000
50%	50.000000
75%	141.500000
max	373.000000

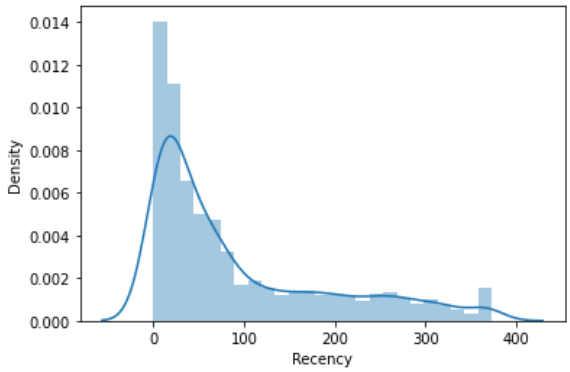

Fig 3: The Recency of Distribution Plot

Table 5: Descriptive Statistics of Monetary

count	4339.000000
mean	2053.793018
std	8988.248381
min	0.000000
25%	307.245000
50%	674.450000
75%	1661.640000
max	280206.020000

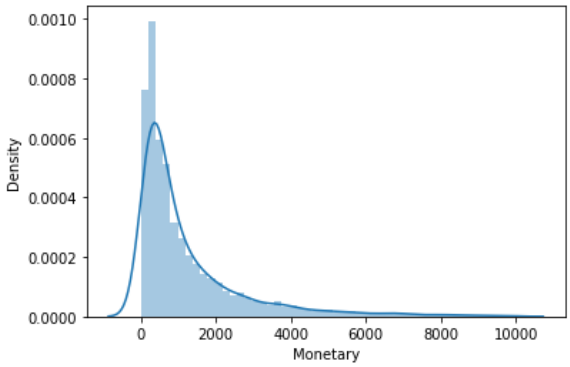

Fig 4: Monetary value less than 10000

Table 6: Descriptive Statistics of Frequency

count	4339.000000
mean	91.708689
std	228.792852
min	1.000000
25%	17.000000
50%	41.000000
75%	100.000000
max	7847.000000

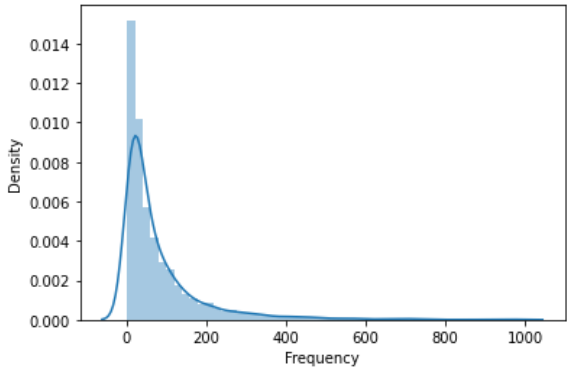

Fig 5: Frequency less than 1000

Using quintiles, subdivide the dataset into three segments based on the requirements for RFM is 0.25, 0.5, 0.75.Then, convert the quintiles into the dictionary, the following quintiles.

The quintiles are

{'Frequency': {0.25: 17.0, 0.5: 41.0, 0.75: 100.0},

'Monetary': {0.25: 307.24499999999995, 0.5: 674.4499999999998, 0.75: 1661.64},

'Recency': {0.25: 17.0, 0.5: 50.0, 0.75: 141.5}}

Created two functions for assigning a level to R, F, M such as 1,2,3,4. The first function has to assign 1 to the lowest value of recency, which means that the customer is more engaged with the specific brand. The second function is for frequency and monetary assigns 1 for the highest value.

```
defr_score(x):
if x <= quintiles['Recency'][0.25]:
return 1
elif x <= quintiles['Recency'][0.50]:
return 2
elif x <= quintiles['Recency'][0.75]:
return 3
elif
return 4
```

Fig. 6 Algorithm for Recency

```
deffm_score(x, c):
if x <= quintiles[c][0.25]:
return 4
elif x <= quintiles[c][0.50]:
return 3
elif x <= quintiles[c][0.75]:
return 2
else:
return 1
```

Fig. 7 Algorithm for Monetary and Frequency

Now, we are calculating and adding R, F,M segment value columns in the existing dataset to show the R,F,M segment values by using the above two functions r_score and fm_score functions, the following table shows the first five customers of the data set after segment and calculate the values.

Table 7: Values of R, F, M

Customer Id	Recency	Frequency	Monetary	R	F	M
12346.0	325	1	77183.60	4	4	1
12347.0	2	182	4310.00	1	1	1
12348.0	75	31	1797.24	3	3	1
12349.0	18	73	1757.55	2	2	1
12350.0	310	17	334.40	4	4	3

Then, calculate and add RFMGroup value column showing combined concatenated score of RFM, calculate and add RFMScore value column showing total sum of RFMGroup values. The RFM score is assigned to each customer.

Table 8: Values of RFM group and RFM Score

Customer Id	Recency	Frequency	Monetary	R	F	M	RFM Group	RFM Score
12346.0	325	1	77183.60	4	4	1	441	9
12347.0	2	182	4310.00	1	1	1	111	3
12348.0	75	31	1797.24	3	3	1	331	7
12349.0	18	73	1757.55	2	2	1	221	5
12350.0	310	17	334.40	4	4	3	443	11

Assign the loyalty level to each customer such as Platinum, Gold, Silver, and Bronze. Based on the RFM score, if the RFM score is less the customer would be more loyal customers, so assign them a loyalty level inPlatinum; if the RFM score value is high the loyalty level is Bronze.

Table 9: Loyalty Level of each customer

Customer Id	Recency	Frequency	Monetary	R	F	M	RFM Group	RFM Score	RFM_Loyalty_Level
12346.0	325	1	77183.60	4	4	1	441	9	Silver
12347.0	2	182	4310.00	1	1	1	111	3	Platinum
12348.0	75	31	1797.24	3	3	1	331	7	Gold
12349.0	18	73	1757.55	2	2	1	221	5	Platinum
12350.0	310	17	334.40	4	4	3	443	11	Bronze

Now, validate the data, based on the RFM group which is equivalent to 111, and sort the values based on the monetary in ascending order, the following table depicts the first five records of the index

Table 10: Loyalty Level index based on the higher level

Customer Id	Recency	Frequency	Monetary	R	F	M	RFM Group	RFM Score	RFM_Loyalty_Level
14646.0	1	2080	280206.02	1	1	1	111	3	**Platinum**
18102.0	0	431	259657.30	1	1	1	111	3	**Platinum**
17450.0	8	337	194550.79	1	1	1	111	3	**Platinum**
14911.0	1	5677	143825.06	1	1	1	111	3	**Platinum**
14156.0	9	1400	117379.63	1	1	1	111	3	**Platinum**

3. b) K-Means Algorithm

In the Elbow method, we are varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters.

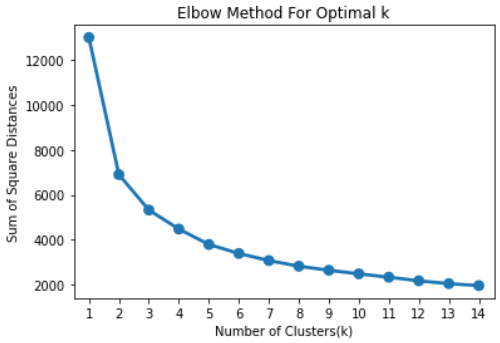

Fig 8::Elbow Method

The point at which the elbow shape is created is 3, that is, our K value or an optimal number of clusters is 3. Now let's train the model on the dataset with many of clusters 3.

Perform K-Mean Clustering or build the K-Means clustering model, and find the clusters for the observation given in the dataset

Table 11: Clusters for the Observation

Customer Id	Recency	Frequency	Monetary	R	F	M	RFM Group	RFM Score	RFM_Loyalty_Level	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	Silver	1
12347.0	2	182	4310.00	1	1	1	111	3	Platinum	0
12348.0	75	31	1797.24	3	3	1	331	7	Gold	1
12349.0	18	73	1757.55	2	2	1	221	5	Platinum	1
12350.0	310	17	334.40	4	4	3	443	11	Bronze	2

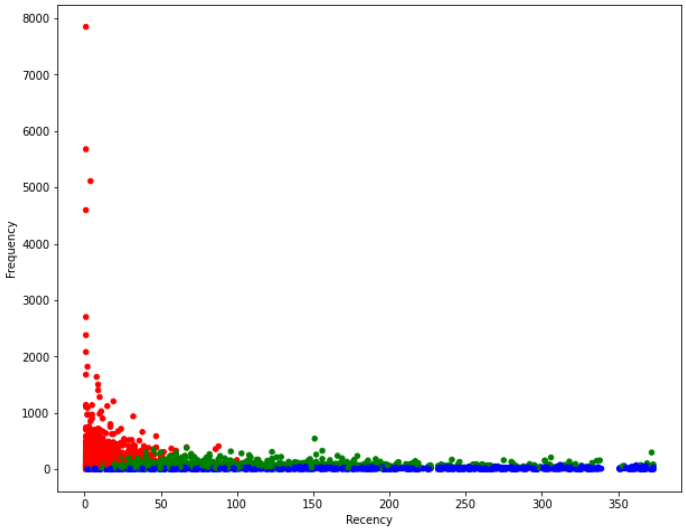

Fig 9: Visualization of 3 clusters

Table 12: Assigning colors

Customer Id	Recency	Frequency	Monetary	R	F	M	RFM Group	RFM Score	RFM_Loyalty_Level	Cluster	Color
12346.0	325	1	77183.60	4	4	1	441	9	Silver	1	Green
12347.0	2	182	4310.00	1	1	1	111	3	Platinum	0	Red
12348.0	75	31	1797.24	3	3	1	331	7	Gold	1	Green
12349.0	18	73	1757.55	2	2	1	221	5	Platinum	1	Green
12350.0	310	17	334.40	4	4	3	443	11	Bronze	2	Blue

Table 8 depicts the colors of three different clusters, cluster 0 depicts the Red color they are our most loyal group, cluster 1 indicates the Green color, and cluster 3 shows the Blue color, based on the colors we can identify graphically our potential customers. Loyalty levels are designated by bronze, silver, gold, and platinum.Green and Blue clusters are scattered more in the graph.

4. Conclusions

Customer retention is a major task for both online and physical enterprises. In the present work, the RFM model is implemented for synthetic datasets, to analyze customer segmentation. Also, clusters are evaluated using the Elbow method for the K-Means clustering algorithm with a different number of clusters. Based on the Elbow method, the Sales Recency, Sales Frequency, and Sales Monetary can be analyzed and an optimal solution is found.Also, found the loyal customers based on the different loyalty levels. According to the loyalty of customers the marketing department makesstrategies to retain potential customers and entice new customers.

5. References

1. *H. Abbasimehr and M. Shabani, "A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers," Kybernetes, vol. 50, no. 2, 2021,*
2. *R. Goyal, "A Study On Consumer's Buying Behaviour Based On Customers' Online Reviews."(www.rdias.ac.in)*
3. *N. Jain and V. Ahuja, "Segmenting online consumers using K-means cluster analysis," Int. J. Logist. Econ. Glob., vol. 6, no. 2, p. 161, 2014,*
4. *M. D. Uncles, G. R. Dowling, and K. Hammond, "Customer loyalty and customer loyalty programs," Journal of Consumer Marketing, vol. 20, no. 4–5. 2003,*
5. *D. T. Nguyen, V. T. Pham, D. M. Tran, and D. B. T. Pham, "Impact of service quality, customer satisfaction and switching costs on customer loyalty," J. Asian Financ. Econ. Bus., vol. 7, no. 8, 2020,*
6. *Jayendra Sinha (USA), Jiyeon Kim (USA) "Factors affecting Indian consumers' online buying behavior, Innovative Marketing, Volume 8, Issue 2, 2012*
7. *Dr. Sankar Rajagopal, Enterprise DW/BI Consultant ,Tata Consultancy Services, Newark, DE, USA," Customer Data Clustering Using Data Mining Technique", International Journal Of Database Management Systems (Ijdms) Vol.3, No.4, November 2011 (www.ijdms.in)*
8. *E.W.T. Ngai ,*, Li Xiu , D.C.K. Chau," Application of data mining techniques in customer relationship management: journal homepage: (www.elsevier.com/locate/eswa).*
9. *Aditya Kumar Gupta & Chakit Gupta,"Analyzing customer behavior using data mining Techniques: optimizing relationships with customer" International Journal Of Management Insight Vol. VI, No. 1; June, 2010*
10. *Krishna R.Kashwan, Member of IACSIT, and C.M.Velu, "Customer Segmentation using Data mining Techniques" Vol.5, No 6, December 2013.*
11. *Dattatray V.Bhate, M.Yaseen Pasha, "Analysing target customer behavior using datamining techniques for e-com.(www..ijircst.org)*
12. *Mohammad Ali Farajian, Shahriar Mohammadi, "Mining the Banking Customer Behavior Using Clustering and Association Rules Methods", International Journal of Industrial Engineering & Production Research, December 2010, Volume 21, Number 4pp.. 239—245(www.ijiepr.iust.ac.ir)*
13. *Belsare Satish and Patil Sunil, "Study and Evaluation of user's behavior in e-commerce Using Data Mining", (www.isca.in).*
14. *Jayendra Sinha (USA), Jiyeon Kim (USA), "Factors affecting Indian consumers' online buying behavior", Innovative Marketing, Volume 8, Issue 2, 2012.*
15. *R.Deiva veeralakshmi, "A study on online shopping behaviour of customers", International journal of scientific research and management (ijsrm) ISSN (e): 2321-3418(www.ijsrm.in)*
16. *Murray, Paul W., Agard, Bruno, Barajas, Marco A., 2017. Market segmentationthrough data mining: a method to extract behaviors from a noisy data set.Comput. Ind. Eng. 109, 233–252.(www.elsevier.com)*
17. *Nguyen, Dung H., de Leeuw, Sander, Dullaert, Wout E.H., 2018. Consumer behaviourand order fulfilment in online retailing: a systematic review. Int. J. Manage. Rev.20 (2), 255–276.*
18. *Patak, Michal et al., 2014. The e-pharmacy customer segmentation based on theperceived importance of the retention support tools. Procedia-Soc. Behav. Sci.150, 552–562.(www.sciencedirect.com)*
19. *Qadadeh, Wafa, Abdallah, Sherief, 2018. Customers Segmentation in the InsuranceCompany (TIC) Dataset. Procedia Comput. Sci. 144, 277–290.(www.sciencedirect.com)*

Corresponding Email: ***[1]kavithapandian.s@gmail.com, [2]smk76dgl@gmail.com***