Customer segmentation using K - Means Algorithm

Arjun Bijili Gopakumar[1], Gautam Krishnan[2], Nitha L Rozario[3], Remijius Brian Nazreth[4], Pooja Anilkumar[5]

[1,2,3,4,5]*Department of Computer Science, Marian Engineering College, Kazhakuttam, India*

***Abstract* - Customer Segmentation is often said as the subdivision of a market into customer groups that can share similar characteristics or features. It can be a powerful means to identify unsatisfied customer needs. In today's world businesses run based on innovation and having the ability to capture the fascinated attention of customers with a large variety of products. But such a large raft of products can leave the customers confused, what to buy and what not to. Even the companies are nonplussed about what section of customers to target to sell their products. This is where machine learning comes into play. Various algorithms are applied to unravel hidden patterns to make data for better decision-making for the future. Most customer segmentation approaches based on customer value fail to look for the factor of time and the trend of value changes in their analysis. Here, we classify customers using the RFM model and K-means clustering method. An assessment of changes over several intervals of time is administered and customer value for each time interval is calculated. A python program has been developed and the program is being trained using the K means Algorithm by applying MinMax scaler for normalization onto a dataset having shopping details and trends of customers in a shopping mall. This research is based on the time and trend of customer value changes for improving the accuracy of predictions and will be based on the past behavior of customers.**

***Index Terms* – Customer segmentation, K means Algorithm, RFM model.**

I.INTRODUCTION

The development of customer-oriented behavior in business, more attention has been paid to customers and their needs. These important factors are put together to make higher profits Customer value analysis is an analytical structure for interpreting customer behavior from the vast source of otherwise meaningless data. There are different definitions for customer value analysis, technique where the business will take all the available information regarding the market and is trained using machine learning algorithms. This is a really vital piece of activity for any kind of business. It also shows you ways well if you've got wiped out the market using your current marketing techniques. It shows the progress and results of the past and predict the future. Once you are an investor or maybe an entrepreneur you would like to understand what you're getting yourself into and would like to possess all the info to your goal or vision for the corporation. For this reason, you would like to try to do a marketing analysis. Coming to what is customer segmentation, Customer segmentation is the process of dividing customers into groups with similar characteristics or features. Some basic customer segmentation includes demographic, geographic, behavioral and psychographic.[1-2] Segmentation of potentially profitable customers, whom we call good customers, becomes significantly important One among the most effective customer segmentation models based on customer value is that the RFM model, which was introduced by Bauer and later developed by Hughes. In recent studies, the RFM model has been adopted in several industries and under various conditions by adding extra parameters. Then cluster it using K means algorithm which is an unsupervised machine learning algorithm. This vision can help businesses develop new strategies or verify and correct their current strategies.

II. LITERATURE SURVEY

In the early 90s, the birth of radio in the 1920s increases the reach of advertisers, and marketing ideas start to look for the best way to use this new technology, although at this point it is mainly to make customers aware that a product exists. The Great Depression puts a stop to mass production, and companies must now focus on selling their existing stock. Sales are not as easy as they used to be, and companies start looking for professionals who will increase sales; some do this through aggressive and unethical means including false advertising, which is later regulated.

The 1970s were a great time for new marketing ideas, and they mark the birth of synergy marketing. Synergy is defined as “the interaction of elements that when combined produce a total effect that is greater than the sum of the individual elements.” Synergy marketing is what happens for example when a new animated movie comes out, McDonald’s offers the toys for the film with the Happy Meal, Disney parks create an attraction related to the film, and the film manages to name drop McDonald’s.

In the late 90s, the Internet changed the world as we knew it, and marketing professionals quickly started testing new marketing ideas on this exciting new medium. One digital advertising strategy that went on to be universally loathed is “spam”, the Internet equivalent of flyers and a disruptive and disrespectful approach to marketing. A much smarter marketing idea born in this decade is SEO (Search Engine Optimization), which tries to rank a product or service at the top of Google or yahoo’s search results to give the seller an edge.

With the new millennium comes another huge milestone for marketing: the birth of social media. The Internet turned out to be a double-edged sword for marketers, because while it afforded unprecedented access and information about potential customers, it also allowed those customers to filter or block advertising, as well as to compare and shop around in ways that were not possible before. Marketing now becomes about catering to customers’ needs and desires, and about building relationships of trust.

Inbound marketing gathers force, as companies put forth valuable content that savvy customers seek out. The integration of smartphones to everyday life expands marketing opportunities, which now also include email marketing and mobile marketing campaigns. Tools such as HubSpot help marketing professionals stay on top of the many sides of their integrated marketing campaigns, while sticking to the usually tight marketing budget of a small company.

Now marketing uses artificial intelligence technologies to make automated decisions based on data collection, data analysis, and additional observations of audience or economic trends that may impact marketing efforts. AI is often used in marketing efforts where speed is essential. AI tools use data and customer profiles to learn how to best communicate with customers, then serve them tailored messages at the right time without intervention from marketing team members, ensuring maximum efficiency.

Some of the existing works on customer segmentation are:

i. Infiniti Research: Established in 2003, Infiniti Research is a leading market intelligence company providing smart solutions to address your business challenges. Infiniti Research studies markets in more than 100 countries to help analyze competitive activity, see beyond market disruptions, and develop intelligent business strategies. Companies, at present, devise customer segmentation models in a way that can resonate with the audience and make them feel that the brand is addressing them personally through their marketing efforts.[3]

ii. Web Hosting Canada: Web Hosting Canada (WHC) is a privately owned, Canadian technology and IT infrastructure company based in Montreal, Quebec. We’re a team of passionate web professionals providing Canadian businesses with the means to succeed online. Since 2003, WHC has set the highest standard for service reliability and security and is now trusted by tens of thousands of clients throughout Canada and abroad. It is among the fastest-growing web service providers in Canada. Web Hosting Canada is accredited by the Canadian Internet Registry Authority (CIRA), is an official cPanel, Cloudlinux, and SpamExperts partner and has consistently been ranked the #1 Canadian domain name provider by the .CA registry. Building models.[4]

III. BUILDING MODELS

A. RFM Model

To identify customer behavior, the well known method called recency, frequency and monetary (RFM) model is used to represent customer behavior characteristics. This approach models three dimensions of customer transactional data, namely recency, frequency and monetary, to classify customer behavior. [6]

The calculated RFM values are summarized to clarify customer behavior patterns. This study proposes using the following RFM variables:

- Recency (R): the latest purchase amount.

• Frequency (F): the total number of purchases during a specific period.
• Monetary (M): monetary value spent during one specific period.
So the RFM value can be calculated by the equation:
Value = R × F × M

B. Clustering using K means Algorithm

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycenter's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.[7]

C. Android Application development

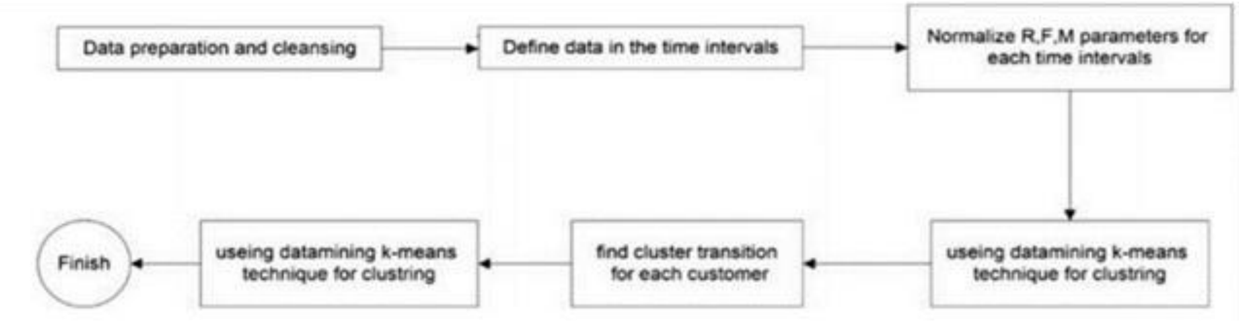
The Android SDK is a set of tools including IDE, Compiler, Debugger, etc to help develop applications for the Android Operating System. Android Studio is an IDE from Google and Jetbrains used to develop Android Applications. Kotlin is a general purpose, free, open-source, statically typed “pragmatic” programming language initially designed for the JVM (Java Virtual Machine) and Android that combines object-oriented and functional programming features. Various android libraries are used for specific purposes

D. Server Side development

Here in server side side development we use node js. Node.js enables us to participate in large-scale collaboration under agreed upon terms that no one company, person or entity can change or dictate. Node.js Foundation is a Collaborative Project at The Linux Foundation. Node.js is the fastest growing open source platform in the world used by 5 million developers. It is used for everything from web front and back end development to creating mobile, tablet, and desktop apps. It's also being used with a whole new set of API services and even more complicated IoT devices.[8]

IV. MODULES

Here we represent the modules and describe its algorithm. First, we need to split dataset into separate subsets based on the timestamp of the individual data points and then based on the RFM model calculate customer value for each subset. In doing so, we obtain a value trend for each customer. Finally, we segment customers by using K-means clustering technique based on the value trends of these customers to obtain their respective clusters and identify common characteristics.


Data preparation and cleansing
Define data in the time intervals
Normalize R,F,M parameters for each time intervals
useing datamining k-means technique for clustring
find cluster transition for each customer
useing datamining k-means technique for clustring
Finish

Our proposed model consists of the flowing six steps:
1. Data preparation and cleansing
2. Incorporating data into time intervals
3. Normalizing R, F and M parameters for each time interval
4.Using data mining K-means technique for clustering
5. Finding cluster transition for each customer
6. Setting trends

A. Data preparation and cleansing

The data is cleaned by removing irrelevant values, noisy and incorrect data as well as the fields of data inappropriate to our research and we prepare data structure for implementing RFM analysis like choosing R, F and M parameters. Approaches adopted for data cleaning are :

1. Remove Irrelevant Values
2. Get Rid of Duplicate Values
3. Avoid Typos or Errors

4. Convert Data Types if necessary
5. Take Care of Missing Values

B. Incorporating data into time intervals

As mentioned earlier, we try to include time into the RFM model. For this purpose, we split the time period into z equal time intervals.

T = ∑ Ti

T = { T1, T2....... Tz}

C. Normalizing R, F and M parameters for each time interval

To control the effect of each parameter on other parameters, we normalize R, F and M parameters between (0, 1). Here, because of the independency of time intervals in analysis, each time interval must be normalized separately. If the whole time is normalized and then split into time intervals, there would be some values that their influence exceeds time intervals and affect the normalization of whole data. Given the importance of the independent of analysis of each time interval, this will be inaccurate.

D. Using data mining K-means technique for clustering

After preparing data in Step 3 using the data mining K-means technique, we try to segment customers based on the RFM model. It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated, after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the centres are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.

Steps in K Means Algorithm:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

E. Finding cluster transition for each customer

After implementing RFM model for all time intervals, a changing value trend for each customer is obtained. It shows the schematic view of the value changing trend for customer Xi by calculating the customer cluster value for each interval {TI,T2,…,TZ}. The result is a set of data that shows the value changing trend of the customer throughout the analysis period. D(Tj,Xi) is the value of customer Xi in time interval Tj.

F. Setiing trends

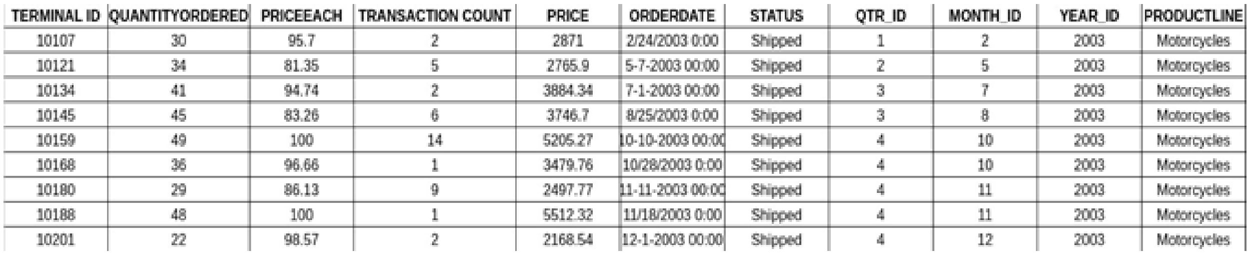
On the basis of the data set obtained in Step 5, we will use K- means technique to cluster the changing value trends of customers. In doing so, the customers with the same value changing trend will be in the same segment

V.EMPERICAL ANALYSIS

Here the data is cleaned by removing irrelevant values, noisy and incorrect data as well as the fields of data inappropriate to our research and we prepare data structure for implementing RFM analysis like choosing R, F and M parameters.

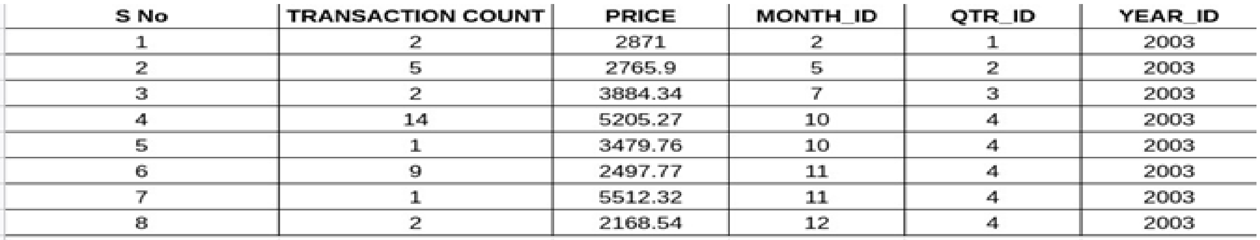
A.Data set

The dataset we used is a collection of sales data of a hypermarket in Argentina over 3 years ranging from 2003 to 2005. The dataset consits of 11 columns each having approximately 2500 records. The dataset initially contained NaN values which was taken care of in the preprocessing step. The whole data set was divided into sub data sets on the basis of year and quarter id.



TERMINAL ID	QUANTITYORDERED	PRICEEACH	TRANSACTION COUNT	PRICE	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLINE
10107	30	95.7	2	2871	2/24/2003 0:00	Shipped	1	2	2003	Motorcycles
10121	34	81.35	5	2765.9	5-7-2003 00:00	Shipped	2	5	2003	Motorcycles
10134	41	94.74	2	3884.34	7-1-2003 00:00	Shipped	3	7	2003	Motorcycles
10145	45	83.26	6	3746.7	8/25/2003 0:00	Shipped	3	8	2003	Motorcycles
10159	49	100	14	5205.27	10-10-2003 00:00	Shipped	4	10	2003	Motorcycles
10168	36	96.66	1	3479.76	10/28/2003 0:00	Shipped	4	10	2003	Motorcycles
10180	29	86.13	9	2497.77	11-11-2003 00:00	Shipped	4	11	2003	Motorcycles
10188	48	100	1	5512.32	11/18/2003 0:00	Shipped	4	11	2003	Motorcycles
10201	22	98.57	2	2168.54	12-1-2003 00:00	Shipped	4	12	2003	Motorcycles

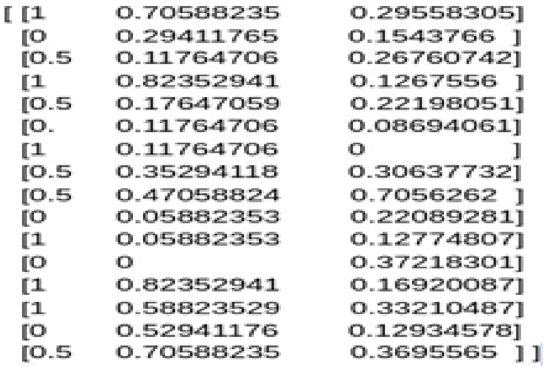
Fig. 1. Initial dataset



S No	TRANSACTION COUNT	PRICE	MONTH_ID	QTR_ID	YEAR_ID
1	2	2871	2	1	2003
2	5	2765.9	5	2	2003
3	2	3884.34	7	3	2003
4	14	5205.27	10	4	2003
5	1	3479.76	10	4	2003
6	9	2497.77	11	4	2003
7	1	5512.32	11	4	2003
8	2	2168.54	12	4	2003

Fig. 2. RFM table

B. Normalization



```
[ [1      0.70588235     0.29558305]
  [0      0.29411765     0.1543766 ]
  [0.5    0.11764706     0.26760742]
  [1      0.82352941     0.1267556 ]
  [0.5    0.17647059     0.22198051]
  [0.     0.11764706     0.08694061]
  [1      0.11764706     0                 ]
  [0.5    0.35294118     0.30637732]
  [0.5    0.47058824     0.7056262 ]
  [0      0.05882353     0.22089281]
  [1      0.05882353     0.12774807]
  [0      0                   0.37218301]
  [1      0.82352941     0.16920087]
  [1      0.58823529     0.33210487]
  [0      0.52941176     0.12934578]
  [0.5    0.70588235     0.3695565  ] ]
```

Fig. 3. Normalized Matrix

- Fig 1 depicts the initial dataset which had columns that were irrelevant for our prediction.
- The data as in the fig 2 represents the chosen R, F, M values along with the quarter and year. MONTH_ID, TRANSACTION COUNT, PRICE represents R, F and M respectively
- Initially the chose R, F and M columns were incomparable. We used MinMaxScaler of the Sklearn library to transform the data in all the three columns into values in the range of 0 to 1.
- Fig 3 represents the normalized R, F, M values in a matrix representation.

C. Elbow method and clustering

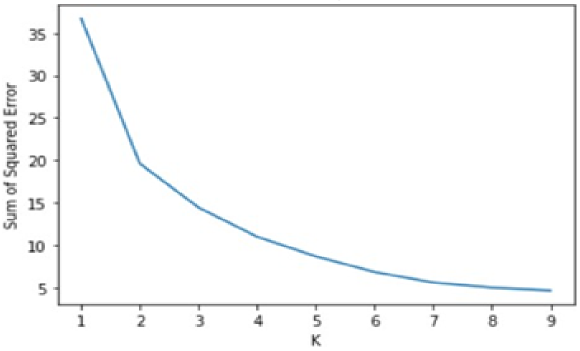


Fig. 4. Elbow graph

• The optimal number of clusters, K is determined using elbow method.
• The elbow graph uses the concept of sum of squares within clusters which defines the total variations within the cluster. The graph is plotted with the number of clusters K against the sum of errors, inertia.
• Fig 4 represents the elbow graph. The python library kneed returns the elbow of the graph.

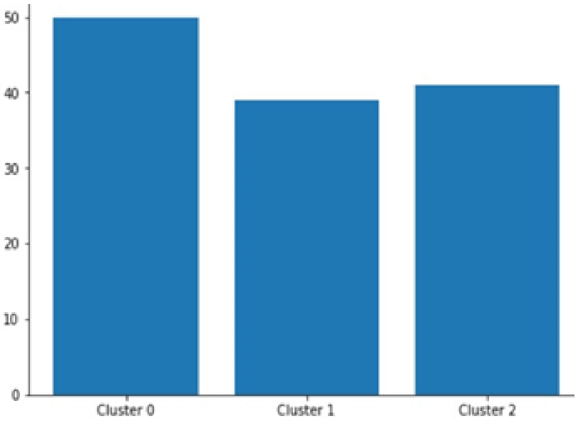


Fig. 5. Clusters

- The optimal number of clusters returned was 3. Fig 5 represents a bar graph for each quarter of the year with the cluster id against the size of the cluster.
- The target customers have been divided into three groups by customer subdivision model, Figure 5 shows the customers segmentation results.

Different approaches can be adopted for calculating the distance between the centroids :

Euclidean distance

Euclidean distance between two points in Euclidean space is the length of a line segment between the two points. It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem, therefore occasionally being called the Pythagorean distance.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

p, q = two points in Euclidean n – space
qi , pi = Euclidean vectors , starting from the orgin of the space.
n = n- space

D. Trends of clusters over the year

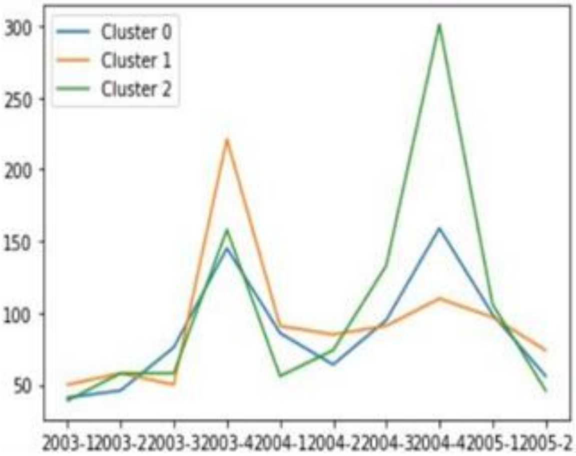


Fig. 6. Cluster size trend

- Figure 6 represents the change in the size of clusters 0, 1 and 2 over three years or ten quarters
- The graph depicts the huge spike in the size of the cluster 2 over the years in the range of the years 2004 3rd quarter and 2005 first quarter
- • Each cluster can represent any group of customers depending upon the bussiness model such as loyal customers, biased customers, less frequent customers etc.
- The trends can be used to device more effective methods to maximize the size of any desired cluster .

VI. CONCLUSION

Mining changes over different time periods provide appropriate conditions to make efficient strategies based on former behavior of customers. The basic RFM model does not consider time transition and given the lack of details about the changes in customer behavior, ineffective results are obtained .It is obvious that dealing with customers with decreasing or increasing value trends, and similar total values require different strategies. In the proposed model, we tried to use the positive aspect of the RFM model by incorporating time into our analysis with the aim of overcoming the shortcomings of the basic RFM model.

ACKNOWLEDGMENT

We would like to express our sincere and deepest gratitude to Dr. Ruby Abraham, Principal, Marian Engineering College for the help and facilities rendered towards the completion of this project. We wish to place on records our ardent and earnest gratitude to our project guide Ms. Nitha L Rozario, Assistant Professor, Dept. of Computer Science and Engineering. Her tutelage and guidance was the leading factor in translating our efforts to fruition. Her prudent and perspective vision has shown light on our trail to triumph. We are extremely happy to mention a great word of gratitude to Dr. P Jayaprakash, Head of the Department of Computer Science and Engineering for providing us with all facilities for the completion of this work. Finally, yet importantly, we would like to express our gratitude to our project coordinator Ms. Keerthi Krishnan, Associate Professor, Dept. of Computer Science for her valuable assistance provided during the course of the project. We would also Extend our gratefulness to all the staff members in the Department. We also thank all our friends and well-wishers who greatly helped us in our endeavor. Above all we thank GOD, the almighty, for his grace without which it would not have been possible to complete this work on time.

REFERENCES

[1] Monireh Hossein (2015) New approach to customer segmentation based on changes in customer value. Journal of Marketing Analytics 3(3).
[2] Customer segmentation using K means clustering by Abhinav Sagar, VIT Vellore. J. Clerk Maxwell,
[3] https://www.infinitiresearch.com/
[4] https://whc.ca/en
[5] https://link.springer.com/chapter/10.1007/978-3-030-60796-8_42
[6] RFM Value and Grey Relation Based Customer segmentation Model in the Logistics Market Segmentation XIONG Weiwen, CHEN Liang,

ZHANG Zhiyong, QIU Zhuqiang Department of Logistics engineering, School of Economics and Commerce, South China University of Technology SCUT Guangzhou,China

[7] A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[8] https://ieeexplore.ieee.org/abstract/document/8769171