Segmenting e-Commerce Customer through Data Mining Techniques

Ritu Punhani*, V.P.S Arora#, A. Sai Sabitha*, Vinod Kumar Shukla**

*Department of Information Technology, ASET, Amity University, India
#Sharda University, Greater Noida, Uttar Pradesh, India
**Department of Engineering and Architecture, Amity University, Dubai, UAE

ritupunhani@gmail.com, vpsarora@gmail.com, saisabitha@gmail.com, vshukla@amityuniversity.ae

Abstract: Business-to-business e-commerce is in much attention nowadays, mainly due to its growing use. In today’s world, it has become imperative for companies to segment their customers and thereby take required measures to survive against other companies. Since there exist a lot, each company must fulfill the demands of their users or they might lose them to other alternatives that exist. This report aims to analyse the customer data from two years: 2018 and 2019 of the company: Autofurnish.com and thereby recommend methods to increase customer influx and give suggestions on what mistakes should not be repeated by the company in future for better performance and sales. To analyse the database, RapidMiner tool has been used. RapidMiner is an open source predictive analytic software that gives support regarding data mining. It lets the user to build models based on their needs and gives solutions quickly. For this analysis, K-means algorithm clustering will be used. Clustering is dividing groups based on similarities and K means is one of the very commonly used methods to do so. In the software, data has been imported having information about customers which is then analysed to prove different results draw a contrast between two years as told before keeping customer segmentation in mind based on various attributes given in the dataset. Relationships between the features are identified as assess to company’s performance.

1. Introduction

In today’s world, the internet has become accessible beyond our imagination. And the people who use internet for online services is increasing every day marginally. Basically, a digital market is any kind of market that exists online. Few examples of digital market can be social media marketing, search engine optimization etc and e-commerce similarly serves as a platform to purchase and give variety of products using the internet. Customer segmentation enables organisations to become more client centric. Companies often segment according to statistical data related to age, gender, life stage etc.

RapidMiner is a tool that offers lots of different operators/ways to connect to our data. The data can be saved in various formats and then be worked upon for various results depending on the needs of the user. It is basically a data science software that brings artificial intelligence to our disposal. It helps in data preparation, machine learning, predictive analysis, data mining etc. This tool may be used for
business as well as commercial purposes like research, education, training etc. RapidMiner was earlier known as YALE which refers to yet another Learning Environment. Data Mining is becoming an increasingly important tool used to transform bare data into information and allows data handling. Another section of your paper

![Customer Segmentation](image1.png)

**Fig 1. Customer Segmentation**

2. **About the Company- ‘Autofurnish’**

Autofurnish is one of the major online sellers of automobile accessories and fittings in India. It is involved in manufacture, export, trade and supply of quality array of products. It’s a renowned manufacturer and wholesaler of all categories of car and bike accessories at wholesale rates. The company came into action in the year 2012 as first online store of automobile accessories in the country. Understanding the element that vehicles have become an important mode of transportation and this market is getting bigger with each day, this brand brings every possibility to acquire modern and most recent vehicle accompaniments to look over, be it inside or exterior accessories, one may discover every one of them in various variations to meet their particulars. People may question the necessity of online shopping of car accessories when we can go and see a product before purchasing. But the options one gets from a retail shop is quite limited and can be same which is already there on other cars/bikes available locally. Though internet shopping will in general give an edge undoubtedly, one can purchase items that may become pioneers for other people, it likewise gives the purchaser a more extensive view by expanding the rundown of potential alternatives.

![Autofurnish.com Home Page](image2.png)

**Fig 2. Autofurnish.com Home Page**
3. Customer Dataset

For this analysis we have taken two datasets of the customers from Autofurnish.com which belong to years 2018 and 2019. Each dataset has various attributes that are Organic searches, age, city and users, has over 600 rows. This data is based on various demographic, behavioural, psychographic fields and therefore helps in customer segmentation. The analysis is performed on each dataset individually using clustering and k-means algorithm. Then the results are compared to understand the customer behaviour and for targeting accurate groups of audience based on age group and demographic location. While drawing a comparison between two years we discover the patterns that might have been made for purchase by customers and the importance of customer segmentation in E-commerce.

The data has been retrieved using Google Analytics is a web examination service offered by Google that tracks site traffic. Google Analytics is a device that can assist with understanding advanced advertising viability. Google Analytics totals the information gathered from your site in numerous manners. We can use custom dimensions as keys for combining information from Google Analytics and other systems, as well as to enhance your reports with information that’s relevant to your analysis.

There are mainly 2 kinds of data that can be collected using Google Analytics:

1. User Acquisition Data (data about users before they visit a website)
2. User Behaviour Data (data about users when they visit the website)

Dataset for the year 2018

Dataset for the year 2019

4.0 Analysis of different component with RapidMiner

In the analysis we have designed a process various operators in the Rapid Miner Tool. These operators play different roles in accurately analysis the dataset and thereby comparing. The results may be stored in repository. There are five operators used in total which are as follow:

**Read Excel:** Read Excel reads data from an Excel file and it’s connected to the input terminal. This operator can peruse information from Excel 95, 97, 2000, XP, and 2003. In the chosen spreadsheet, each column must represent an attribute whereas each row must be an Example. One must make sure that the Excel File is delivered properly prior to building a process by means of it.

**Nominal to Numerical:** This operator changes the type of non-numeric attributes into a numeric type. In addition, it also maps all values of the attributes to numeric values. Values that are binary are mapped to 0 and 1. But the already existing numeric attributes of input database remains as it is.

**Normalize:** This operator mainly normalizes values of chosen attributes. Normalization is a process that’s used to scale values in order to fit them in a specific range. It is useful to draw a comparison between attributes that have different sizes. There are four ways provided to perform normalization.
**Clustering:** This is an operator that performs an important function called clustering using the k-means algorithm. Clustering refers to grouping together Examples that are like one another. K-means determines the values of k (number of clusters required) and assigns each Example to one of them.

**Performance:** For this analysis, Cluster Distance Performance has been used. This operator is used for performance assessment of centroid based clustering procedures. It delivers a list of performance criteria values based on the cluster centroids. The Cluster Distance Performance operator takes the centroid cluster model and clustered set as its input and evaluates the performance of the model based on cluster centroids. Two performance activities that are supported: Average within cluster distance and Davies-Bouldin index.

1. **Average within cluster distance:** The average within cluster distance is calculated by averaging distance between centroid as well as all Examples of a given cluster.
2. **Davies-Bouldin index:** An Algorithm that produces clusters with less intra-cluster distances and more inter cluster distances bear a low Davies Bouldin Index (DBI).

This analysis is based on Davies-Bouldin index.

![Fig 4. Designing Processes for year 2018](image)

Fig 4. Designing Processes for year 2018

![Fig 5. Designing Processes for year 2019](image)

Fig 5. Designing Processes for year 2019

### 4.1 Davies Bouldin Index

The Davies Bouldin Index was presented by David L. Davies and Donald W. Bouldin in 1979. It is a measurement for assessing algorithms involving clustering. This is an inner evaluation plot, where the assess of how well gathering has been done is made using amounts and highlights of the database. Given n-dimensional focuses, let Ci be a group of information focuses. Let Xj be an n-dimensional element vector allocated to cluster Ci. Here Ai is the centroid of Ci and Ti is the size of the group I. Si is a measure of scattering inside the cluster.

\[ S_i = \left( \frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^p \right)^{1/p} \]

As a component of the ratio of within the cluster scatter, a lesser value of Davies Bouldin Index will infer that grouping/clustering is done in an improved way. It shows the normal likeness between each bunch and its most comparative one, arrived at the midpoint of overall clusters. The likeness is portrayed as Si. This ensures no group should be similar. Therefore, best grouping method minimizes the value of Davies Bouldin Index. This has a job in deciding the k’s value in the k-means algorithm used. Empty clusters are ignored in calculation of DBI.
4.2 K-Means Algorithm

This calculation is an iterative calculation that segments the dataset as indicated by their highlights into K number of predefined non-covering unmistakable clusters or subgroups. The algorithm helps determine a set of k clusters and allocates every Example to one of them. Each cluster contains a set of alike Examples. The resemblance among Examples is built on distance measure between them. In the given algorithm the clusters are determined by the location of the centre in the n-dimensional space of n attributes in given dataset. And this site is known as a centroid. The algorithm begins with k points which are the centroids of given k potential clusters. All the Examples are allocated to their closest cluster which is characterized by measure type. The centroids are then recalculated by averaging over all Examples of a cluster. The system is repeated max run times with each time a fluctuated set of start focuses. The arrangement of groups is conveyed which has the negligible aggregate of squared separations of all Examples to their relating centroids. This algorithm has many applications as it has many advantages. Some applications may include:

1. Market Segmentation
2. Cluster Analysis
3. Vector Quantization
4. Used in Search Engines
5. Insurance Fraud Detection
6. Identifying Crime Prone Areas etc.

4.3 K-Value vs Davies-Bouldin Index for Analysis

While analysing data, one may look for meaningful groups/clusters. Clustering is dividing groups based on similarities. And K means is one of the most commonly used methods to do so. In order to perform the analysis, K needs to be calculated which indicates the number of clusters to be formed on running the processes. We have taken the value of K from 3-10 one by one. And calculated The Davies Bouldin index of each. The Davies–Bouldin index (DBI) is a measurement for assessing grouping calculations. The K value with least corresponding DBI is chosen to be appropriate for analysis. The Davies Bouldin values originally were negative as the tool returns value between 0-1, since by default these values are multiplied by -1 so that we can run a minimizer on it. The closer these values are to 0, the better they are and as they approach 1 the index is considered bad for analysis.

| K value | Davies Bouldin Index |
|---------|----------------------|
| 3       | 0.446                |
| 4       | 0.363                |
| 5       | 0.414                |
| 6       | 0.393                |
| 7       | 0.422                |
| 8       | 0.41                 |
| 9       | 0.449                |
| 10      | 0.402                |

Fig 6. K-value vs DBI of year 2018

| K value | Davies Bouldin Index |
|---------|----------------------|
| 3       | 0.411                |
| 4       | 0.424                |
| 5       | 0.458                |
| 6       | 0.429                |
| 7       | 0.382                |
| 8       | 0.372                |
| 9       | 0.452                |
| 10      | 0.477                |

Fig 7. K-value vs DBI of year 2019

For the year 2018 we see that when k=4, the Davies Bouldin Index is the least that is 0.363. Therefore, we use four as the number of clusters.
Fig 8. Davies Bouldin Indexes of year 2018

Fig 9. Davies Bouldin Indexes of year 2019

For the year 2019 we see that when k=8, the Davies Bouldin Index is the least that is 0.372. Therefore, we use 8 as the number of clusters.

5.0 CLUSTER MODELS

The year 2018 and 2019 as mentioned above will have four and eight clusters as represented below:

| 2018: 4 clusters | 2019: 8 clusters |
|-------------------|-------------------|
| Cluster 0: 604 items | Cluster 0: 80 items |
| Cluster 1: 6 items | Cluster 1: 3 items |
| Cluster 2: 1 item | Cluster 2: 6 items |
| Cluster 3: 32 items | Cluster 3: 1 item |
| Total number of items: 643 | Cluster 4: 3 items |
|                       | Cluster 5: 18 items |
|                       | Cluster 6: 6 items |
|                       | Cluster 7: 406 items |
|                       | Total number of items: 523 |

6.0 Analysis of the year 2018

The data for the year 2019 was divided into 8 clusters based on K-means algorithm considering Davies Bouldin Index. The dataset is distributed into these clusters as follows:
By seeing this graph, it is evident that cluster 0 has the greatest number of users and cluster 2 has least number of users.

6.1 using the aggregate function

The dataset of 2018 is added using Retrieve function. And the Aggregate operator is used.

**Aggregate**: The Aggregate operator creates a new Example Set from the input Example Set showing the results of the aggregation functions. Many aggregation functions are supported together with SUM, COUNT, MIN, MAX, AVERAGE and many other similar functions.
6.2 ABOUT THE FUNCTIONS

- Least only occurring: This function gives the name of city which has occurred the least.
- Mode: This function tells the city which has occurred the maximum number of times.
- Least: It tells the age which has recurred least number of times.
- Maximum: It tells the maximum value of selected attribute.
- Minimum: This function tells the minimum value of a selected attribute.
- Sum: This function tells the sum of all values in a attribute.
- Average: This function tells the average of all values that are there in an attribute.

Fig 13. Result of all the aggregate

City and Age:

|         | LEAST | MAXIMUM | MINIMUM | MAXIMUM |
|---------|-------|---------|---------|---------|
| CITY    | Agartala | Agra   | 65+     | 25-34   |

- From this table we understand that least active city is Agartala and most active is maximum.
- The minimum users belonged to the age group of 65+ and most users belong to age group of 25-34.
Organic searches:

|               | Minimum | Maximum | Sum    |
|---------------|---------|---------|--------|
| Organic searches | 19      | 47463   | 598367 |

From this table we infer that the total number of organic searches that have taken place is 598,367.

Users:

|       | Minimum | Maximum | Average | Sum     |
|-------|---------|---------|---------|---------|
| Users | 56      | 74395   | 1571.24 | 101038  |

From this table we infer that the max number of users is 74,395 and minimum users are 56. The sum of all users is 101,038.

7.0 Analysis of the year 2019

The data for the year 2019 was divided into 8 clusters based on K-means algorithm considering Davies Bouldin Index. The dataset is distributed into these clusters as follows:

By seeing this graph, it is evident that cluster seven has the greatest number of users and cluster 3 has least number of users.

7.1 Using the aggregate function
The dataset of 2019 is added using Retrieve function. And the Aggregate operator is used.

Fig 16. The various aggregate functions used for analysing dataset of year 2019

Fig 17. Result of all the aggregate

City and Age:

| CITY     | LEAST  | MAXIMUM | MINIMUM | MAXIMUM |
|----------|--------|---------|---------|---------|
| Ajman    | 65     | 25-34   | 60734   | 65+     |
| Agartala | 60734  | 25-34   | 65      | 25-34   |

- From here we infer that the city where maximum users are present is Agartala and the minimum users are from Ajman.
- From here we infer that age group of 65+ makes least visits whereas age group of 25-34 visits the website most.

Organic searches:

| Organic searches | Minimum | Maximum | Sum    |
|------------------|---------|---------|--------|
| 14               | 30710   | 345175  |

From this table we infer that the total number of organic searches that have taken place is 3,45,175.

Users:

| Users | Minimum | Maximum | Average | Sum    |
|-------|---------|---------|---------|--------|
| 65    | 60734   | 1701.369| 889816  |
From this table we infer that the max number of users is 60,735 and minimum users are 65. The sum of all users is 8,89,816.

8.0 Future work and Conclusion

On performing this analysis, we have concluded that year 2018 received more users that is 10,10,308 when compared to year 2019 which had 8,89,816 users. The maximum activity was from Agra during the year 2018 and age group of 25-34 was most active. Therefore, for the future years Auto furnish must focus on age groups of 25-34 that belong to Agra to increase the customer influx. And least on the age group of 65+. For future analysis, we will find which year yielded the highest revenue and how to focus on increasing the revenue.

9.0 References

[1] K. Tsiptsis and A. Chorianopoulos, “Data Mining Techniques in CRM: Inside Customer Segmentation”, John Wiley and Sons, Ltd., Publication, 2009.
[2] Farris, Paul W: Neil T bendle, Philip E Pfeifer, David J Reibsten, (2010)-Marketing metrics : The definitive guide to measuring marketing performance” Upper saddle river, New Jersey, Pearson education Inc. ISBN 0-13-705829-2.
[3] Hasan Ziafat, Majid Shakeri, (2014), Using Data Mining Techniques in Customer Segmentation Int. Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 9 ( Version 3), September 2014, pp.70-79
[4] R Punhani, S Batra (2014) Understanding cultural variations of e-commerce websites in a global framework, Global Journal of Finance and Management
[5] Thirunavukkarasu Kannapiran (Feb 2016) Analysis and Comparison Study of Data Mining Algorithms Using Rapid Miner, Research Gate, International Journal of Computer Science, Engineering and Applications (IJCEA) Vol.6, No.1, February 2016
[6] Peter Kolarovszki, Jirí Tengler, Margita Majercáková, (2016), “The new model of customer segmentation in postal enterprises”, Elsevier, 2016, pp. 121-123
[7] Fredricka Joyous Masih, Ritu Punhani (2017), Analysing E-Commerce website using Google Analytics Tool to enhance Customer Interaction, JIM S8I - International Journal of Information Communication and Computing Technology 5(1):280
[8] Masoud Karimi (Nov 2017), Data Analysis Using RapidMiner, Research Gate, DOI: 10.13140/RG.2.2.22316.36481
[9] Krzysztof Kubacki, Timo Dietrich, Sharyn Rundle-Thiele (Oct 2017), Segmentation in Social Marketing: Why We Should Do It More Often that We Currently Do, in book Segmentation in social marketing: Process, methods and application
[10] Camilleri, M. A. (2018). Market Segmentation, Targeting and Positioning. In Travel Marketing, Tourism Economics and the Airline Product (Chapter 4, pp. 69-83). Springer, Cham, Switzerland.
[11] M. A. (2018). Market Segmentation, Targeting and Positioning. In Travel Marketing, Tourism Economics and the Airline Product (Chapter 4, pp. 69-83). Springer, Cham, Switzerland
[12] AS Sabitha, R Punhani (2019), Identification of Potential Regions for Wind Power Development Using Data Mining, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)
[13] Ritu Punhani, Dr. V.P.S Arora, Dr. Sai Sabitha, Nikhil Vazirani (2019), Data Mining Technologies used for Identifying Brand-Switching Patterns of Customers in Telecom, pp 319-333. Springer, Singapore. Print ISBN 978-981-15-4450-7, Online ISBN 978-981-15-4451-4
[14] Kh Khaled Kalam (2020), Market Segmentation, Targeting and Positioning Strategy Adaptation for the Global Business of Vodafone Telecommunication Company, International Journal of Research and Innovation in Social Science (IJRISS) |Volume IV, Issue VI, June 2020|ISSN 2454-6186