IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Customer Segmentation Using Machine Learning: A Comprehensive Research Study

Yash Parab,Jugal Dave

ABSTRACT:

Nowadays Customer segmentation became very popular method for dividing company's customers for retaining customers and making profit out of them, in the following study customers of different of organizations are classified on the basis of their behavioral characteristics such as spending and income, by taking behavioral aspects into consideration makes these methods an efficient one as compares to others. For this classification a machine algorithm named as k means clustering algorithm is used and based on the behavioral characteristic's customers are classified. Formed clusters help the company to target individual customers and advertise the content to them through marketing campaigns and social media sites which they are really interested in.

Keywords - Machine learning, Customer segmentation, K-means algorithm.

1.INTRODUCTION

Today many of the businesses are going online and, in this case, online marketing is becoming essential to hold customers, but during this, considering all customers as same and targeting all of them with similar marketing strategy is not very efficient way rather it's also annoys the customers by neglecting his or her individuality, so customer segmentation is becoming very popular and also became the efficient solution for this existing problem.

Customer segmentation is defined as dividing a company's customers on the basis of demographic (age, gender, marital status) and behavioral (types of products ordered, annual income) aspects. Since demographic characteristics do not emphasize the individuality of customers because the same age groups may have different interests, behavioral aspects are a better approach for customer segmentation as it focuses on individuality and we can do proper segmentation with the help of it.

.

2.LITERATURE REVIEW

[1] A solution is proposed to distinguish the customers group into two groups named as premium and standard with the help of machine learning methods named as NEM, LiRM and LoRM

[2]. Tushar Kansal, Suraj Bahuguna, Vishal Singh, Tanupriya Choudhury. "Customer Segmentation using K-means Clustering", International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS).2018, In this paper customer segmentation on Telecom customers is achieved by using information such as age, interest, etc. with the help of cluster analysis method.

3.Objectives

The purpose of segmenting customers is to determine how to correlate to customers in multiple segments to maximize customer benefits. Perfectly done customer segmentation empowers marketers to interact with every customer in the best efficient approach.

4.Technical Introduction: -

K-means Clustering Algorithm

K-means Clustering is a clustering Algorithm in which we are given data points with its data set and features and the mechanism is to categorize those data points into clusters as per their similarities. The algorithm forms K clusters based on its similarity. To calculate the similarity Kmeans uses Euclidean distance measurement method.

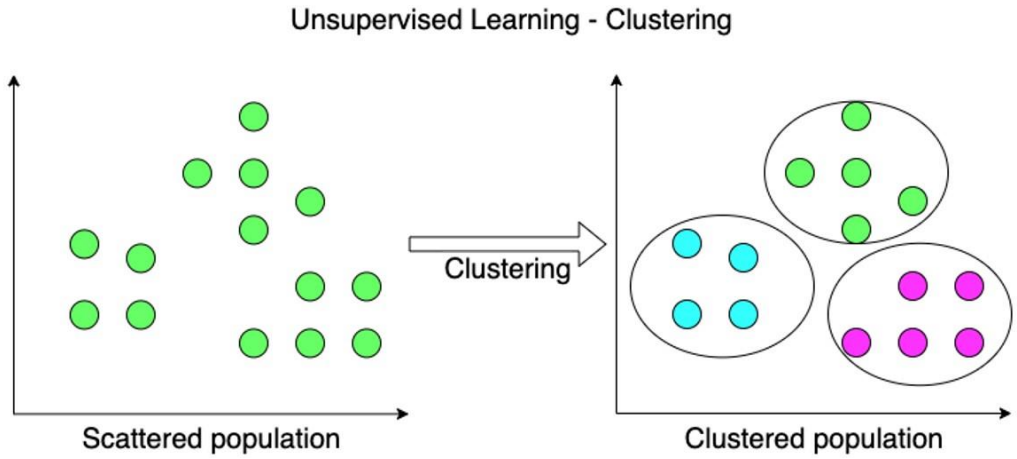

Figure 1. Flow of operation

Steps

1. In the first step, we randomly initialize k points.
2. K-means classifier categorizes each data point to its nearest mean and rewrites the mean's coordinates.
3. Iteration is continuing up till all data points are classified.

The Following Analysis was done by using the Anaconda Jupyter NoteBook and Python 3.x and some Python packages for editing, processing, analyzing, and visualizing information.

5.Propose Model :

A) Import packages and data:

To begin, we import the necessary packages to do our analysis and then the xlsx (Excel spreadsheet) data file.If you want to follow up with the same data, you have to download it from UCI. For this example, I place the xlsx file in the folder (directory) where I present Jupiter's notebook.

B) Data cleaning:

After importing the package and data, we will see that the data is not as helpful as that, so we need to clean and organize this data in a way that we can create more actionable insights.

C) Normalize the data:

The K-means area unit is sensitive to the scale of the information used, such as clustering algorithms, so we would like to normalize the data to make sure all dimensions are treated equally. In other words, we want each column to contribute the same impact on the distance. Note that normalization is done on each column separately (rather than on each row).

D) Select the optimal number of groups:

We are ready to run cluster analysis. But first, we need to find out how many groups we want to use. There are several approaches to selecting the number of groups to use, but I am going to cover two in this article:

(1) the silhouette coefficient,

(2) the elbow method.

E) Silhouette (clustering):

The silhouette refers to how to interpret and validate consistency within data structures. This method shows a diagram of how well each item is organized. The value of a silhouette is a measure of how something is more similar in its collection (combination) than other groups (partitions). The silhouette goes from -1 to +1, where a higher value indicates that an object matches its collection properly and is compared to neighboring groups. If several objects have a high value, the integration configuration is appropriate. If most points have a value or a negative value,the coordinate system may have too many or too few clusters.The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.Now that we know a whole lot of silhouettes, we use code to find the right number of groups.

F) Elbow criterion method (with the sum of squared errors) (SSE):

The idea behind the elbow method is to run a k-mean correlation in the data given for the k value (num_clusters,e.g. k = 1 to 10), and for each k value, calculate the sum of the squared errors (SSE). is.
Then, adjust the SSE line for each k value. If the line graph looks like a hand - a red circle (in the form of an angle) below the line of the line, the "elbow" on the hand is the correct value (collection value).[6] Here, we want to reduce SSE. SSE usually falls to 0 as we go up k (and SSE is 0 where k is equal to the number of data points, because where each data point has its own set, and there is no error
between it and its trunk) .

The objective is therefore to select a smaller value of k,which still has a lower SSE, and the cone usually represents where it begins to return negatively with increasing.

6.ANALYSIS / INTERPRETATION OF STUDY

In this analysis we include Mall Customer data is an interesting dataset that has hypothetical customer data. It puts you in the shoes of the owner of a supermarket. So on this basis of the data, we have to divide the customers into various groups.

The data includes the following features:

1. Customer ID

2. Customer Gender

3. Customer Age

4. Annual Income of the customer (in Thousand Dollars)

5. Spending score of the customer (based on customer behavior and spending nature)

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Figure 2. Mall Customer Dataset

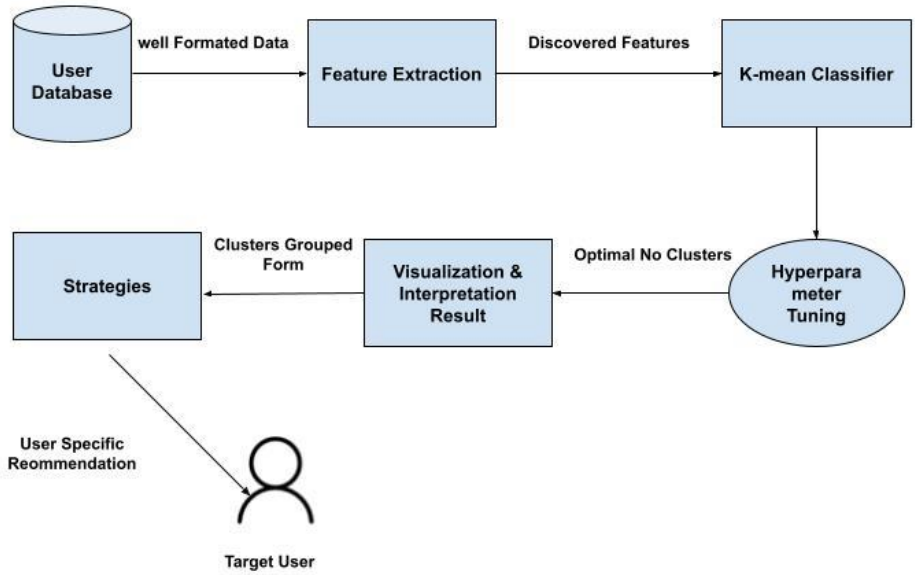

Figure 3. Flow of Analysis operation

Exploratory Data analysis (EDA)

Exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

Following Are Some statistical insight of mall customer case data:

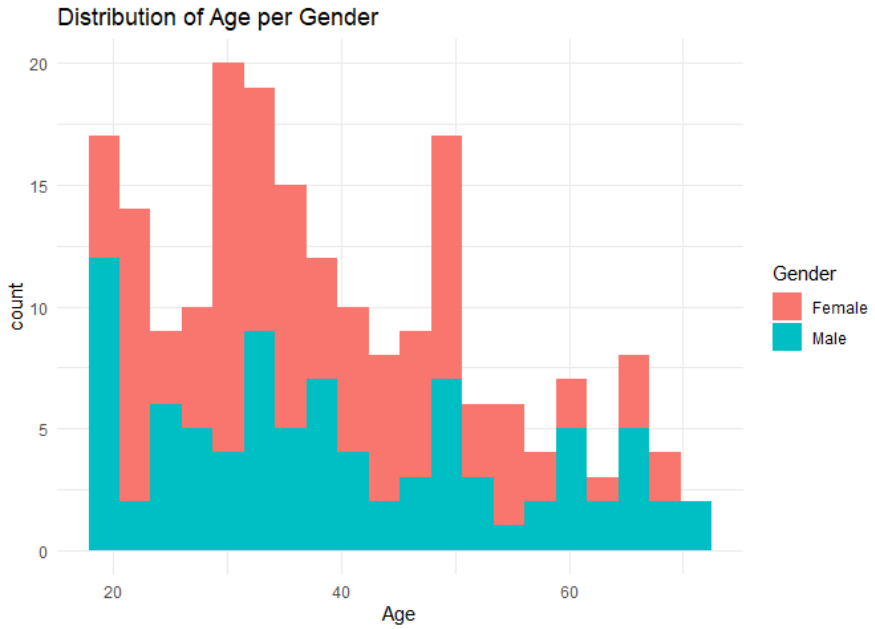

Figure 4. Distribution of Age Based on Gender

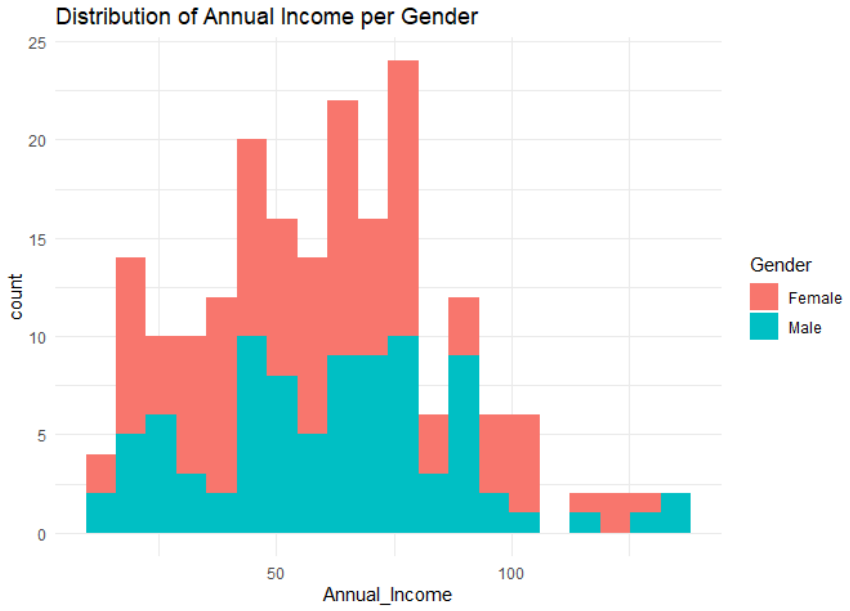

Figure 5. Distribution of Annual Income Based on Gender

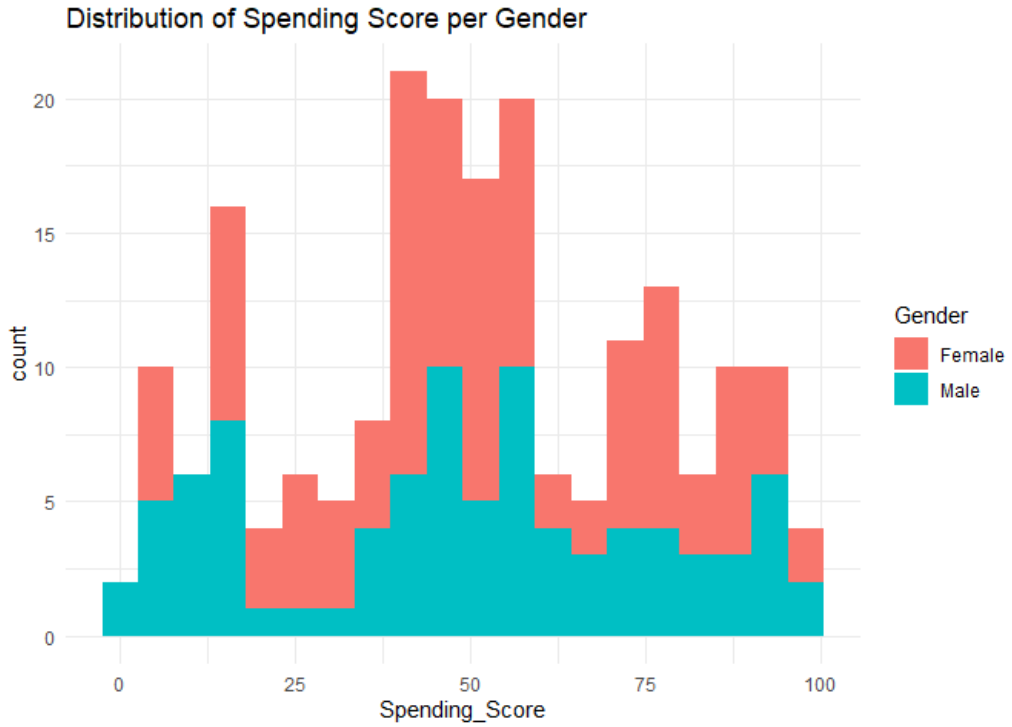

Figure 6. Distribution of Spending Score Based on Gender

6. Results

After analysis of data and classifying customers with features annual income and spending score, we got clusters of customers & with formed clusters marketing team form strategies for customers specific recommendation to make value out of them in figure 4.

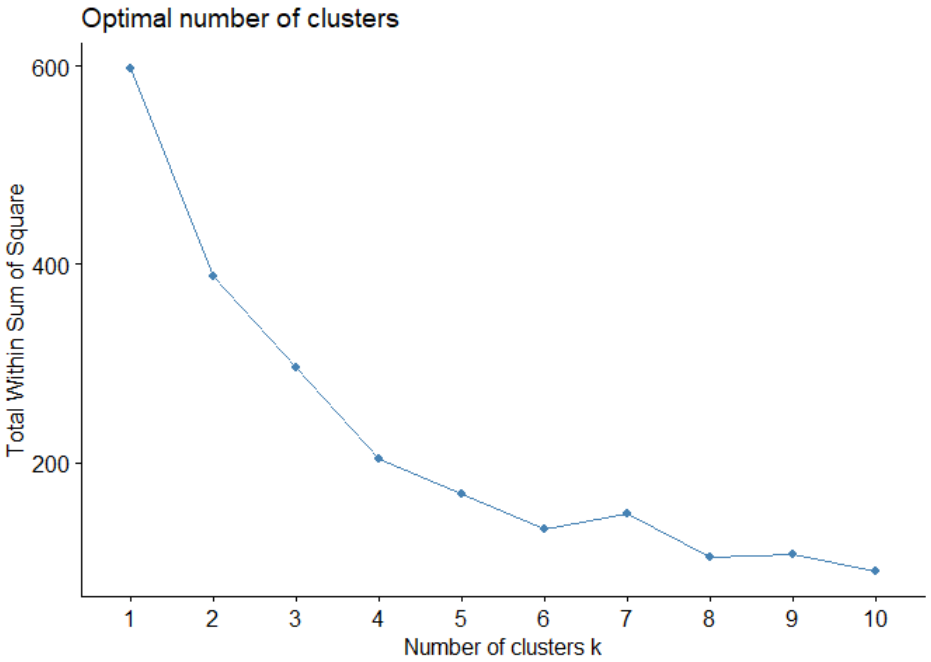

Figure 7. Elbow Method Chart

```
Silhouette Score for 4 Clusters: 0.4114
Silhouette Score for 5 Clusters: 0.3773
Silhouette Score for 6 Clusters: 0.3785
Silhouette Score for 7 Clusters: 0.3913
Silhouette Score for 8 Clusters: 0.3810
```

Fig. 8 Silhouette Score

Cluster 4 had the most complete silhouette fit, indicating that 4 may be the best number of clusters.(**Figure 7**)

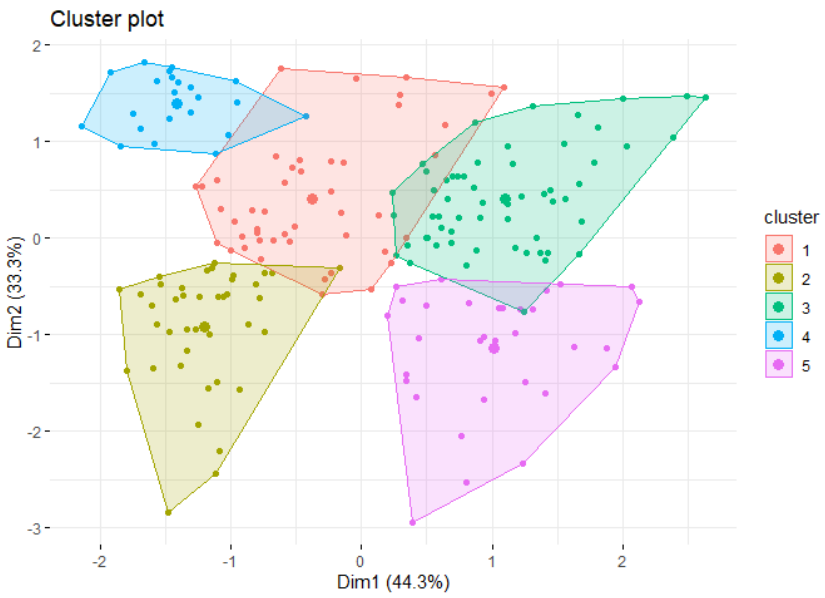

Fig. 9 K-mean Clusters Charts

Profiling:

- Cluster 1: This group consists of **young adults** with **medium** annual income and spending score.
- Cluster 2: This group consists of **middle-aged adults** with **high** annual income and spending scores.
- Cluster 3: This group consists of **old-aged adults** with **medium** annual income and spending score.
- Cluster 4: This group consists of **young adults** with **low** annual income and **high** spending scores.
- Cluster 5: This group consists of **middle-aged** adults with **high** annual income and **low** spending scores.

Based on the Cluster Chart Result above, we can build marketing strategy as follows:

- We know from exploratory data that female customers are higher than male customers. We could create a marketing campaign targeting the male customers.
- Cluster 1 have medium annual income and cluster 3 have high annual income, but they have medium in spending score. So, we could give some promotions to encourage them to purchase more.
- We could give special treatment or loyalty programs to customers in cluster 2 and 4 which are customers who have high spending scores.
- Cluster 5 have high annual income but a low spending score. It consists of middle-aged adults. We could do marketing research to identify their needs, wants, and demands, so that they will increase the purchases.

7.Drawback of System

1. Marketing will become expensive.
2. Because of having less no. of customers in a segment the problem of limited production occurs.

8.CONCLUSION

This paper presented an implementation of the k-Means clustering algorithm for customer segmentation using data collected from an online retail outfit. Our model has partitioned customers into mutually exclusive groups, Five clusters in our case. This will be useful for applying further data mining strategies and the derived insights are helpful in decision making for the business wings.

9.REFERENCES

[1] Sukru Ozan, "A Case Study on Customer Segmentation by using Machine Learning Methods", IEEE, Year: 2018.
[2] Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar “Telecom customer segmentation based on cluster analysis An Approach to Customer Classification using k-means”, IJIRCCE,Year: 2015.
[3] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu Electrical/Electronics and Computer Engineering Department, University of Uyo, Akwa Ibom State, Nigeria “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services”, IJARAI,Year: 2015.
[4] Potharaju, S. P., Sreedevi, M., Ande, V. K., & Tirandasu, R. K. (2019). Data mining approach for accelerating the classification accuracy of cardiotocography. Clinical Epidemiology and Global Health, 7(2), 160-164.
[5] Yogita Rani and Dr. Harish Rohil “A Study of Hierarchical Clustering Algorithm”, IJICT, Year: 2013.
[6] Omar Kettani, Faycal Ramdani, Benaissa Tadili “An Agglomerative Clustering Method for Large Data Sets”, IJCA, Year: 2014.
[7] Snekha, Chetna Sachdeva, Rajesh Birok “Real Time Object Tracking Using Different Mean Shift Techniques–a Review”, IJSCE, Year: 2013.SulekhaGoyat“The basis of market segmentation: a critical review of literature”, EJBM, Year: 2011.

[8] Potharaju, S. P., Sreedevi, M., & Amiripalli, S. S. (2019). An Ensemble Feature Selection Framework of Sonar Targets Using Symmetrical Uncertainty and Multi-Layer Perceptron (SUMLP). In Cognitive Informatics and Soft Computing (pp. 247-256). Springer, Singapore.
[9] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, Intelligent Classification and Clustering Of Lung and Oral Cancer through Decision Tree and Genetic Algorithm, International Journal of Advanced Research in Computer Science and Software Engineering,2015
[10] Tanupriya Choudhury, Vivek Kumar, Darshika Nigam, An Innovative and Automatic Lung and Oral Cancer Classification Using Soft Computing Techniques, International Journal of Computer Science and Mobile Computing,2015
[11] H. Mehta, V.S. Dixit and P. Bedi," Refinement of recommendations based on user preferences".
[12] H. Mehta, S.K. Bhatia, V.S. Dixit and P. Bedi," Collaborative personalized web recommender system using entropy-based similarity measure".
[13] Rivedi, A., Rai, P., DuVall, S. L., and Daume III, H. (2010, October). ´Exploiting tag and word correlations for improved webpage clustering in Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 3-12). ACM.
[14] Potharaju, S. P., & Sreedevi, M. (2017). A Novel Clustering Based Candidate Feature Selection Framework Using Correlation Coefficient for Improving Classification Performance. Journal of Engineering Science & Technology Review, 10(6).
[15] Domavicius, G., and Tuzhilin, A. (2015). Context-aware recommender systems. In Recommender
systems handbook (pp. 191-226). Springer US.

[16] K. Windler, U. Juttner, S. Michel, S. Maklan, and E. K. ¨Macdonald, "Identifying the right solution customers: A managerial methodology," Industrial Marketing Management, vol. 60, pp. 173 –186, 2017.
[17] R. Thakur and L. Workman, "Customer portfolio management (cpm) for improved customer relationship management (crm): Are your customers platinum, gold, silver, or bronze?" Journal of Business Research, vol. 69, no. 10, pp. 4095 – 4102, 2016.