Using K-Means Clustering For Mall Customers Data

Ashutosh Kumar, Aman Agarwal, Vishal kumar
School of Computing Science and Engineering,
Galgotias University, Plot No. 2, Yamuna Expressway, Opposite Buddha International Circuit, Sector 17A, Greater Noida, Uttar Pradesh, India

1. Abstract

Every day we deal with a vast amount of data. To analyze those data is such a big task but also important. This paper demonstrates the segmentation of customers of e-commerce website based on a machine learning algorithm namely K-Means Clustering. We have collected the data of around 300 customers from an e-commerce website and their total yearly pay on an e-commerce website and then we will analyze it and found the number of optimum clusters. Now a day's every company is targeting their customers based on the available data and trying to find out the behaviors of customers like which kind of product most attracting them, what they like. Based on that customer's behavior companies are launching their new products. We will use the K-Means Clustering Algorithm a machine learning-based method to find the optimum number of clusters and analyze the underlying client segments supported by the information provided. In today's world, it is hard to get the customers behavior and categorizing the customers into different clusters and categories' customers supported their human ecology and shopping for behavior. . This could be a crucial side of client partition that permits the business to higher target their customers increase in selling's and modify their products according to their targeted customers and launching new products.

Keywords: K-Means Clustering, Customer Segmentation, Machine Learning

2. Introduction

Customer segmentation is the method of portioning the customers into small groups based on common behaviors so that companies can target each group very effectively and appropriately. In marketing, companies do segment customers based on the data that may include:

- Age of the customers
- Gender of the customers
- Marital status (Yes/No)
- Location of the customers (Rural, Urban)
- Life stage of the customers (Married Single, Working, Jobless, Retired, etc)

Segmentation enables Companies to Increase their marketing efforts to different targeted audience subsets. Those efforts can help to both product development and communication with the customers. For Customer segmentation a company needs to gather some related information – data – about customers and then need to analyze it to identify patterns or information in the data that can be used to create decision making and forming segments or clusters.

Few of the data can be collected from the company's purchasing website, information like – job related, address, purchased-products, annual pay, for example. Few of it might be collected from how the customers got to know about the company and how they opened your website or how they entered your system. The other ways to gather the information are face-to-face interviews or a phone call interview, Company surveys based on that we can identify the behavior of customers or a set of focus groups which may companies are focusing on.

3. Literature Review

3.1 Clustering

Clustering is a commonly used data analysis technique that is used to get to know about the structure of the data or in another way we can say that clustering is all about the grouping of the similar that like data in the same group is similar to each other and the data in other group is somehow different. Clustering is most commonly used in customer segmentation; where we use it to find out the customers that are similar to each other whether it is based upon behaviors or the clustering of documents based upon contents. Clustering is considered as an unsupervised learning technique where we don't have the labeled data and we don't know anything before.

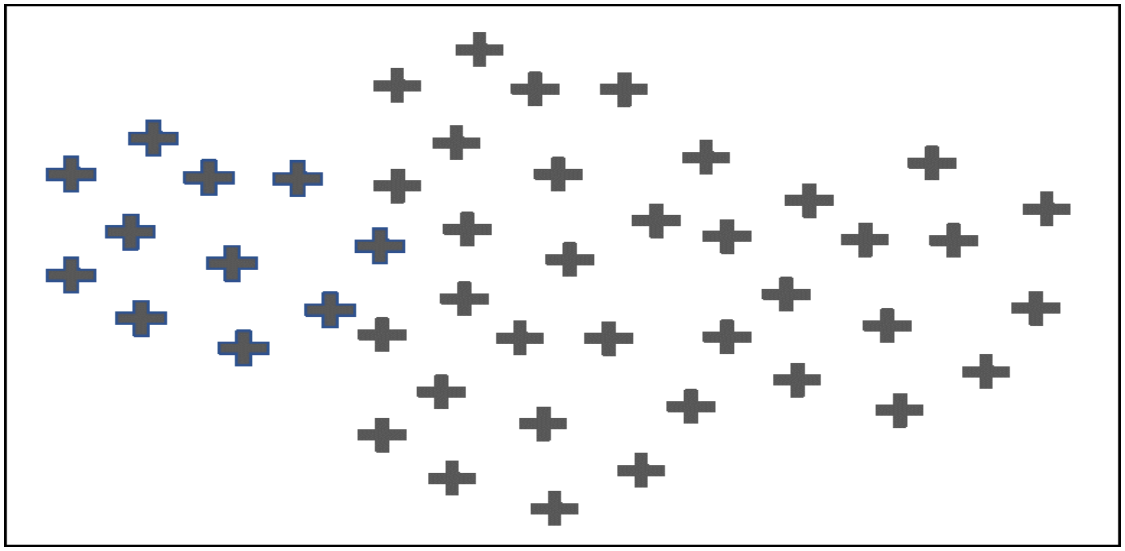


Fig1: Original Collected data sets

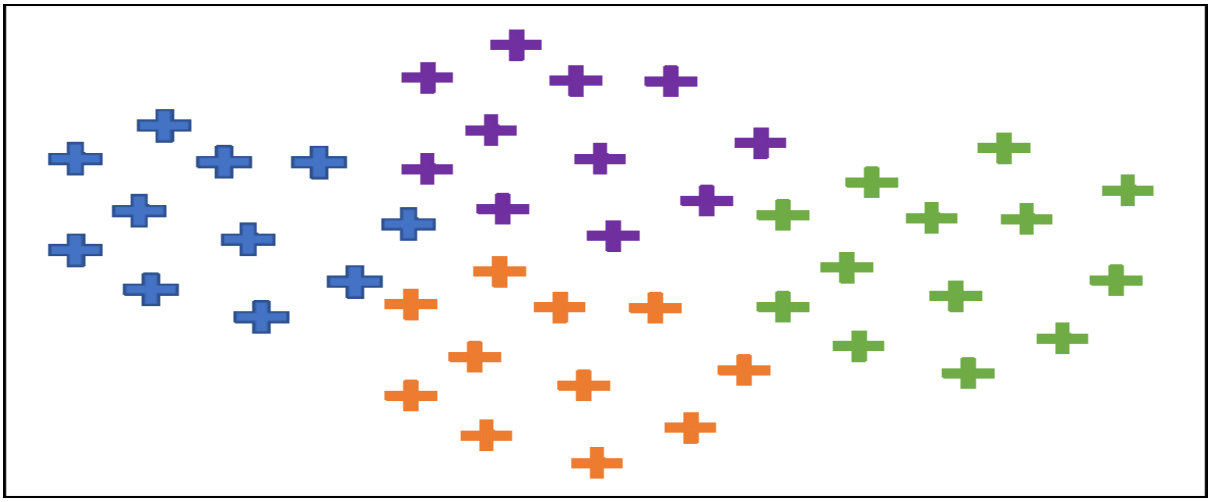


Fig2: Data sets after forming clusters

3.2 K-Means clustering Algorithm

K-Means Clustering is an algorithm that is based on Unsupervised Learning, where the data is not labeled and we group them. Here K is the number of defined clusters that need to be initialized in the process, and if there is K=2, then

there will be two clusters, and if there is K=3, then there will be three clusters and it goes on.

It is a Kind of iterative approach where we divide the unlabeled data into k different clusters where each dataset belongs to a group that has a similar property. It is Centroid based algorithm, A centroid-based partitioning technique uses the centroid of a cluster, Ci, to represent that cluster.

Steps of K-Means Clustering Algorithm:-

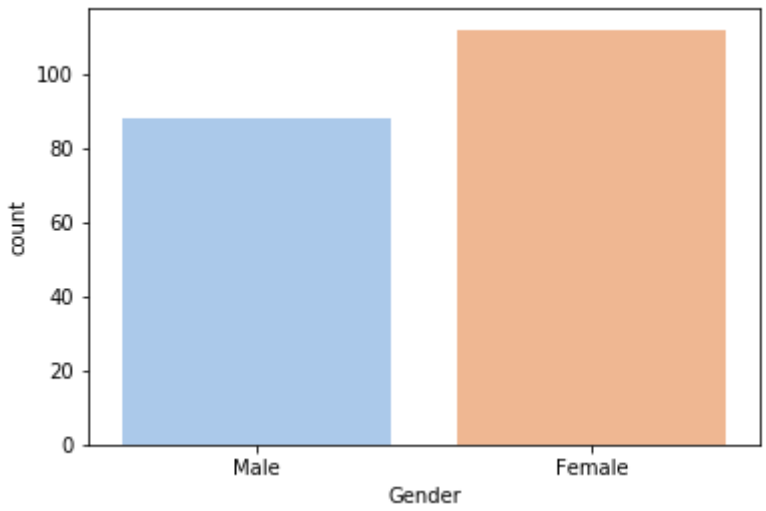
1. Determine the no. of clusters k based on data and business requirements
2. Assign K number of centroid which is selected from the dataset (each centroid represents a cluster).
3. Every data point is assigned to the nearest centroid (Usually assigned based on the Euclidean Distance).
4. Centroids are re-calculated
5. If data points re-assigned then repeat step4.
6. If not, final clustered Generated.

4. Methodology

The data set used to perform the clustering and k-means algorithm is the Mall_Customers.csv.It was collected from the shopping mall. It contains attributes that are CustomerId, Gender, Age, Annual- Income Spending Score (1-100).The dataset containing the data of 300 different customers.

4.1 Genders Count Visualization

```
##Visualizing Genders Count
sns.countplot(x="Gender",data=data,palette="pastel")
plt.show()
```

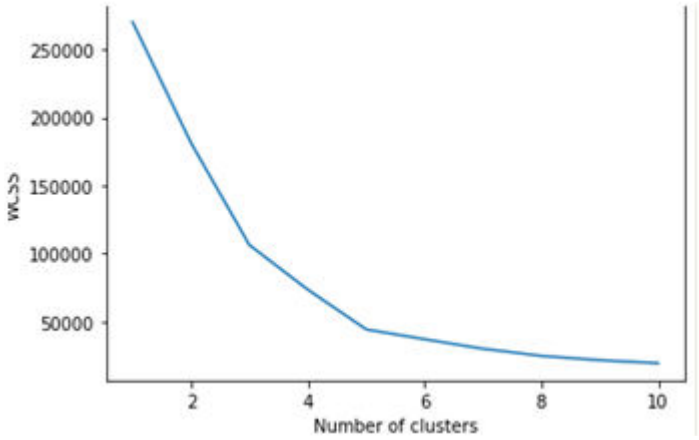


4.2 Elbow Method

The elbow method is the method to determine the optimal number of clusters in the K-Means algorithm. The elbow method plots the graph of the value of cost function calculated with the help of different values of K.

```
for i in range(1,15):
    kmean = KMeans(n_clusters=i,init="k-means++")
    kmean.fit_predict(X)
    wcss.append(kmean.inertia_)
```

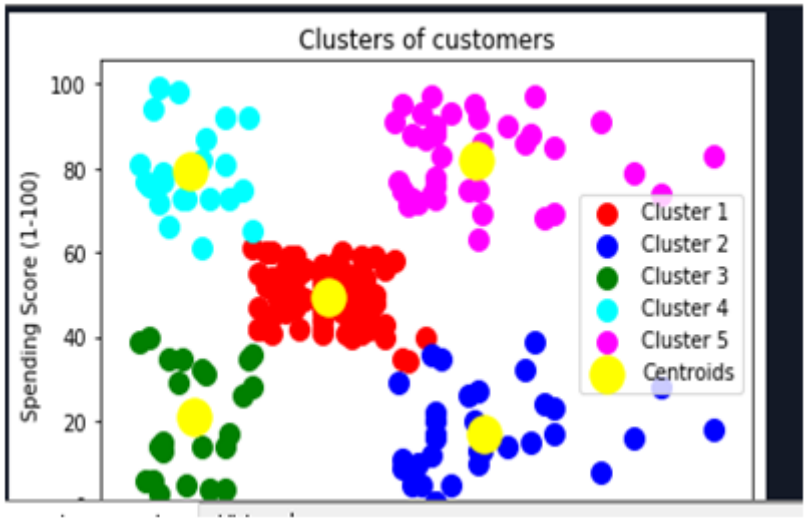
```
plt.plot(range(1,15),wcss)
plt.title('The Elbow Method')
plt.xlabel("No of Clusters")
plt.ylabel("WCSS")
plt.show()
```



We clearly obtained from the elbow plot we can have 5 clusters. i.e k=5

4.3 Obtained graph for cluster:-

```
#visualize the cluster
plt.scatter(X[y_hc == 0, 0], X[y_hc == 0, 1], s = 1
plt.scatter(X[y_hc == 1, 0], X[y_hc == 1, 1], s = 1
plt.scatter(X[y_hc == 2, 0], X[y_hc == 2, 1], s = 1
plt.scatter(X[y_hc == 3, 0], X[y_hc == 3, 1], s = 1
plt.scatter(X[y_hc == 4, 0], X[y_hc == 4, 1], s = 1
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```



5. Conclusion

From the above visualization, we can conclude that Cluster shows Medium monetary benefit, low yearly compensation, Cluster 2. Shows Low monetary benefit, low yearly compensation, Cluster 3 shows High monetary benefit, high yearly compensation Cluster 4 shows Low monetary benefit, high yearly compensation, and Cluster 5. Shows Medium monetary benefit, low yearly compensation.

6. References

[1] Wikipedia – Data Science https://en.wikipedia.org/wiki/Data_science

[2] Sowmya Vivek. Clustering algorithms for customer segmentation. https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac August 2018

[3] AgEcon SEARCH https://ageconsearch.umn.edu

[4] Jiawei Han, Micheline Kamber, Jian Pei "Data Mining Concepts and Techniques", Third Edition.

[5] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "The Basis Of Market Segmentation" Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.

[6] Medium –Data Science https://medium.com/@mudgalvivek2911/machine-learning-clustering-elbow-method-4e8c2b404a5d

[7] https://www.shopify.in/encyclopedia/customer-segmentation#:~:text=Customer%20segmentation%20is%20the%20process,Number%20of%20employees