International Journal of Scientific Research in Engineering and Management (IJSREM)

Volume: 07 Issue: 01 | January - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Customer Segmentation Using Machine Learning

N.Pavan kumar
(B.TECH Scholars, Department Of Computer Science and Engineering , SNIST - 501301,India)

A.Ashwini
(B.TECH Scholars, Department Of Computer Science and Engineering , SNIST - 501301,India)

G.Meghana Reddy
(B.TECH Scholars, Department Of Computer Science and Engineering , SNIST - 501301,India)

Under Guidance of

Dr. Rama Chandra

(Associate Professor Head, Department Of Computer Science and Engineering , SNIST - 501301,India)

Abstract – Both books and software on customer relationship management stress the necessity of customer segmentation. The most popular strategy for differentiating one consumer from another is to brand some of them as premium and the rest as basic. In this paper, consumer data that has been manually segmented by a corporation is examined. The study uses its real-time data regarding its customers' details to tackle the segmentation challenge. Machine learning techniques are used to find solutions to data management issues since they are effective in doing so. In in order to distinguish between premium and standard clients in a company's database, different classification methods are evaluated.

1. INTRODUCTION

Customer segmentation is the process of classifying your consumers based on a wide range of factors (for example grouping customers by age). It is a technique for businesses to comprehend their clients. Making strategic choices about product development and marketing is made simpler when one is aware of the variations among client segments. The possibilities for segmenting are limitless and largely depend on the volume of client data you have at your disposal. There are various consumer segmentation approaches, and they are based on four basic sorts of parameters

- **Geographic** customer segmentation uses the location of the user is the only important factor in segmentation. Several strategies can be used to do this. By nation, state, city, or zip code, you can create groups.

- **Behavioral** customer segmentation is based on previously recorded customer behaviours that can be used to anticipate future behaviour. For instance, popular brands or the times of year when people buy the most.

- **Psychological** segmentation typically include factors like personality traits, attitudes, or beliefs. Customer surveys were utilized to collect this information, which can be used to determine how customers feel.

Insights and patterns can be discovered by studying consumer data using machine learning approaches. Models using artificial intelligence are effective tools for decision-makers.

There are numerous machine learning algorithms, each of which is appropriate for a certain class of issues. The k-means clustering approach is one popular machine learning algorithm that is appropriate for client segmentation issues. DBSCAN, BIRCH, and other clustering methods are available as well.

2. RELATED WORK

By using customer segmentation, you may better personalize your marketing strategies to each consumer group's unique needs. Additionally, you may communicate with your customers more effectively by using this type of marketing. Here are some reasons why customer segmentation allows for efficient client communication.

- Need and Objective
- Scope of Research

2.1. Need and Objective

By using customer segmentation, you may better personalize your marketing strategies to each consumer group's unique needs. Additionally, you may communicate with your customers more effectively by using this type of marketing. Here are some reasons why customer segmentation allows for efficient client communication.

The practise of segmenting a company's customers into groups that demonstrate similarities among customers in each group is known as customer segmentation. In order to optimize each customer's value to the company, it is important to select how to interact with each category of customers.

2.2. Scope of Research

We will make use of university-provided personnel data from Kaggle in our study. In order to select the most accurate model out of all of them and compare their accuracy, we used K-means and a density-based clustering technique as our machine learning algorithms.

2.3 Proposed System

Shopping centre or malls frequently compete with one another to attract more consumers and so boost their revenues. Machine learning is already being used by numerous shops to complete this work. Realizing how machine learning may support such goals is astonishing. The retail centre take the data from their patrons and create ML algorithms to target the appropriate ones. In order to select the most accurate model, we are comparing the DBSCAN and K-means clustering techniques. This boosts sales while also improving the effectiveness of the complexes.

3. MODULE

Data Gathering module

This module deals with dataset collection to build a machine learning model.Here we have collected the data that is given by Wholesale customers.The dataset given wholesale customer is a fictional dataset that is made available on kaggle.

Cleaning the Dataset

We will clean the dataset provided by the wholesale customer in this module.

Data cleaning is the process of removing duplicate, corrupted, incorrectly formatted, inaccurate, or corrupted data from a dataset.

The processes for cleaning the dataset generally include deleting any duplicate or unnecessary observations, handling missing data, handling null values, etc.

Therefore, using the Label Encoder from Scikit-Learn, we will convert the categorical data in this instance into numerical (integer) data.

Training the Model

After training the model we will test the trained model using the test data so that we can find the accuracy of the model. As we have built models using different algorithms .We can pick the most accurate model out of all models tested depending on the accuracy so that we can predict the customers more accurately.

Predicting using the model

Using the most accurate model from the above module for every given input details of customers we will predict about the attrition of the customers.

.

4. SYSTEM DESIGN

Our current method is the K-means algorithm. K-means has both advantages and disadvantages. It cannot determine the median for all datasets, which is one of its limitations. Therefore, we suggested comparing k-means and dbscan in order to determine which method is superior for a particular data set.

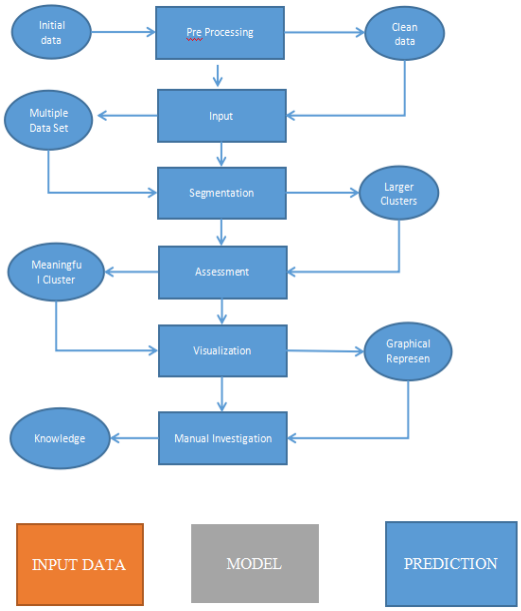


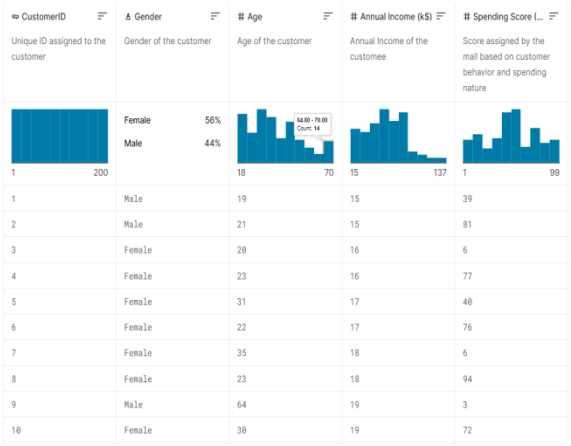
Fig1 : Architecture

The above figure contains a detailed working flow of an K-means algorithm. Firstly, input data was fed into the process in order to preprocess the step. Later on several data transformation methods were performed such as aggregation and smoothing. Now the number of multiple data sets were formed and formed some meaningful clusters which are useful for the research. Secondly, Visualization technique were used to better understand a result which may be in pictorial representation. Therefore, our expected output can be in graphical manner.

5. DATA SET

For customer segmentation, this article leverages a data collection of wholesale clients from the machine learning repository at the University of California, Irvine. This data collection, which relates to consumers of a wholesale distributor, includes information on the consumption of various goods by customers as well as their yearly spending. In all, there are 440 occurrences and 8 traits. The characteristics in the dataset indicate each customer's annual spending on various commodities, and each instance in

the dataset represents a separate customer. Due to their lack of significance to consumer purchasing patterns, the dataset's "Channel" and "Region" attributes have been left out in order to facilitate implementation. Additionally, rather than using anything else for segmentation, the focus of this study is on taking consumers' buying patterns into account. Consequently, this article will make use of additional six characteristics that are used to track client spending. This dataset is said to have been derived from a different, bigger dataset that was mentioned in a research, according to the University of California, Irvine machine learning repository. Table 1 displays a few examples of the utilized dataset.



CustomerID	Gender	Age	Annual Income (k$)	Spending Score (...
Unique ID assigned to the customer	Gender of the customer	Age of the customer	Annual Income of the customee	Score assigned by the mall based on customer behavior and spending nature
1 200	Female 56% Male 44%	18 70	15 137	1 99
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72

Fig 2 : Data Set

6. K-MEANS ALGORITHM

Since k-means is the most widely used algorithm of its kind, it has been chosen for implementation in this project among all centroid-based methods. Kmeans must choose the value of k and the distance metric to utilize before applying to a dataset. Several techniques may be used to estimate the correct total number of clusters, k. This study takes into account 2 to 5 client clusters. Despite the fact that they ultimately do not exhibit that many differences, Euclidean distance and Manhattan distance are employed here as distance measures for k-means. The "Feature Normalization" technique is used to scale all the values falling between -1 and 1.

K-Means is the most used partitional clustering algorithm. It was independently developed across various places in the 1950s and 1960s and quickly became quite popular due to its simplicity, ease of use, and numerous empirical successes (e.g. in business, medicine and science).

The K-Means method, sometimes called Lloyd's algorithm, consists of three key steps:

1. Use seed points to divide samples into beginning groups. Initial clusters will be formed by the samples that are closest to these seed points.

2. Determine the distances between the samples and the centroids of the groupings, then place the closest samples in each cluster.

3.Calculating freshly formed (updated) cluster centroids is the third stage.

Once the algorithm converges, repeat steps 2 and 3 once again.

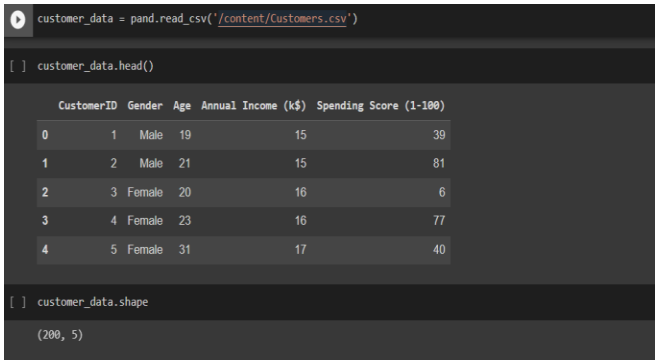
As mentioned earlier the goal of K-Means is to minimise the objective function (inertia) over all clusters.The objective function is defined as:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Fig 3: Mathematical Formula

The avove formula is used ro represent WCSS (Within cluster Sum of square) Which will be used in this paper and forms clusters based upon centroids and distance between the different clusters. In order to find an appropriate number of clusters, the elbow method will be used. In this method for this case, the inertia for a number of clusters between 2 and 10 will be calculated. The rule is to choose the number of clusters where you see a kink or "an elbow" in the graph.

7. IMPLEMENTATION



```
customer_data = pand.read_csv('/content/Customers.csv')

[ ] customer_data.head()
```

	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

```
[ ] customer_data.shape

(200, 5)
```

Fig 4: Loading data and viewing

```
[ ]  P = customer_data.iloc[:,[3,4]].values
```

Fig 5 : Using iloc function for choosing different set of data from data set.

```
wc = []

for i in range(1,11):
  kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
  kmeans.fit(P)

  wc.append(kmeans.inertia_)
```

Fig 6: Using WCSS method

```
[ ] kmeans = KMeans(n_clusters=5, init='k-means++', random_state=0)

    Y = kmeans.fit_predict(X)

    print(Y)

[4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
 3 4 3 4 3 4 1 4 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 2 0 2 1 2 0 2 0 2 1 2 0 2 0 2 0 2 0 2 1 2 0 2 0 2
 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0
 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2]
```

Fig 7 : Training Kmeans Clustering Model

8. RESULTS AND DISCUSSIONS

The K-means algorithm's success is dependent on the extremely effective clusters it creates. However, determining the ideal number of clusters is a difficult process. There are several other approaches to determining the ideal number of clusters, but in this article we focus on the best technique. The process is described below:

Elbow Approach

One of the most often used techniques for determining the ideal number of clusters is the Elbow approach. The WCSS value idea is used in this technique. WCSS refers to the total deviations within a cluster and stands for Within Cluster Sum of Squares.

In order to find an appropriate number of clusters, the elbow method will be used. In this method for this case, the inertia for a number of clusters between 2 and 10 will be calculated. The rule is to choose the number of clusters where you see a kink or "an elbow" in the graph

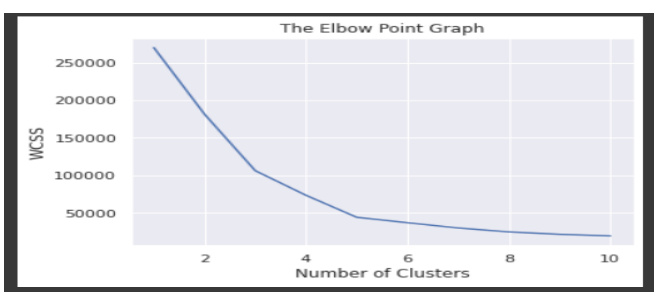


Fig 8: Elbow Graph

```
# plotting all the clusters and their Centroids

plt.figure(figsize=(8,8))
plt.scatter(X[Y==0,0], X[Y==0,1], s=50, c='green', label='Cluster 1')
plt.scatter(X[Y==1,0], X[Y==1,1], s=50, c='red', label='Cluster 2')
plt.scatter(X[Y==2,0], X[Y==2,1], s=50, c='yellow', label='Cluster 3')
plt.scatter(X[Y==3,0], X[Y==3,1], s=50, c='violet', label='Cluster 4')
plt.scatter(X[Y==4,0], X[Y==4,1], s=50, c='blue', label='Cluster 5')

# plot the centroids
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], s=100, c='cyan', label='Centroids')

plt.title('Customer Groups')
plt.xlabel('Annual Income')
plt.ylabel('Spending Score')
plt.show()
```

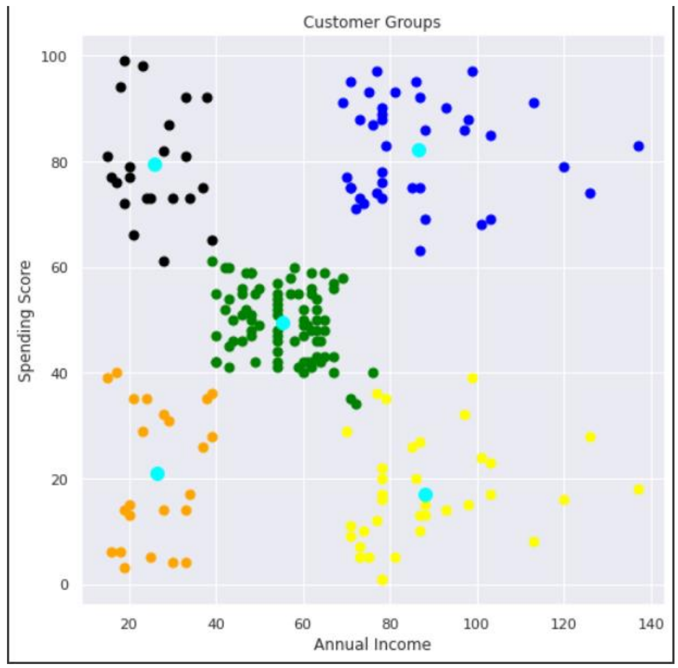


Fig 9: Visualization Technique

When compared to other groups, the black group has higher spending scores, which aids various organizations in supplying greater necessities for those with higher spending scores.

9. CONCLUSION

One of the most widely used clustering algorithms, K means clustering is frequently used by practitioners to begin clustering assignments in order to gain an understanding of the dataset's structure. Data points are organized using K means into discrete, non-overlapping groupings. One of the main uses for K means clustering is the segmentation of consumers to gain a deeper knowledge of them, which might then be utilized to boost the company's income.

10. REFERENCES

[1] Data Clustering - Charu.Agrawal , Chandan Reddy
[2] The element of statistical Learning Data Mining, Inference, and Prediction - Trevor Hastie, Robert Tibshirani, Jerome Friedman
[3] Data Clustering - Robert Haralick
[4] Data Mining and Knowledge Discovery Series - Vipin Kumar
[5] J. A. Hartigan. Clustering Algorithms. John Wiley and Sons, 1975.
[6] Trends in big data analytics - Kamballa K, Kollias G, Kumar V,
What to do when k-means clustering fails: A simple yet
Principled alternative algorithm - Raykov YP, Boukouvalas A,
Little MA.
[7] A new Initialization technique for generalization Lloyd iteration - Katsavounidis I, Kou J, Zang Z.
[8] Data set - https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python