INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078

(Volume 8, Issue 3 - V8I3-1351)
Available online at: https://www.ijariit.com

Segmentation of customers using Machine Learning algorithm

Souvik Biswas
biswasssouvik@gmail.com
Narula Institute of Technology, Kolkata, West Bengal

Neha Bhattacharya
nehabhattacharya450@gmail.com
Narula Institute of Technology, Kolkata, West Bengal

Ananya Samanta
99.samantaa@gmail.com
Narula Institute of Technology, Kolkata, West Bengal

Nitin Gupta
Sonu64dc@gmail.com
Narula Institute of Technology, Kolkata, West Bengal

Dr. Sangita Roy
roysangita@gmail.com
Narula Institute of Technology, Kolkata, West Bengal

ABSTRACT

The separation of the market into a discrete customer group of customers with comparable characteristics is known as customer segmentation. Client segmentation may be an effective tool for identifying unmet customer demands. The advent of a slew of new rivals and entrepreneurs has created a lot of friction among competing firms as they try to gain new customers while keeping the ones they already have. The k-means clustering technique is utilised in this research for this purpose. Segmentation of customers allows a business to personalise its interactions with consumers, much as we do in our daily lives. We discovered that each customer's behaviour and demands had similar features when we segmented them. Then they're grouped together to meet needs using diverse ways. Doing segmentation manually can be exhausting. That is why we have implemented Machine Learning Algorithm to do for us.

Keywords- *Machine Learning, K-means clustering, Customer Segmentation, Big Data, Prediction, sklearn, pandas, python.*

1. INTRODUCTION

Faced with product rivalry, businesses should mine customer resources to achieve targeted measures for different customers and provide the services they desire [1]. To give different types of consumers with varied marketing strategies and increase their happiness, the key to develop a company is to start with an customer analysis and see their demands and utilise customer segmentation as a tool to discover and evaluate various consumer groups in the system. Segmentation of customers is one of the most useful approaches in business analytics for analysing customers behaviour. Customers with comparable means, ends, and behaviours are placed together into homogeneous clusters utilising clustering algorithms [2]. Cluster analysis is a type of data mining algorithm that is mostly used in the study of information of data to notice distribution features in data sets in order to achieve specific goals. Several clustering algorithms, such as hierarchical clustering and density-based clustering have been offered in addition to K Means. Combining clustering algorithms can give good results than using each approach alone. Customers differ in their behaviour, requirements, wants, and traits, and the major purpose of clustering techniques is to identify distinct customer types and split the customer base into clusters of similar profiles so that target marketing may be carried out more efficiently. Certain parameters are considered while segmenting the customer. The clustering parameters can broadly be classified as geographic, demographic, psychographic and behavioural. Prediction of future customer demands and trends through customer segmentation and consumption behaviour, and planning of market for businesses, so as to accomplish the objective of appropriate service resource allocation and the most lucrative design of customer marketing programmes.

2. METHODOLOGY

2.1 Customer Segmentation

The process of segmenting the market into native groups is known as customer segmentation. Customer segmentation is a method of categorising customer groups based on various factors. This theory recommends segmenting customer information and consumption behaviour, as well as profit market planning for businesses, to analyse and anticipate future consumer consumption trends. Over the years, the commercial world has become more competitive, as organizations such as these have to meet the needs

and desires of their customers, attract new customers, and thus improve their businesses.[3] The task of identifying and meeting the needs and requirements of every customer in the business is very difficult. This is because customers can vary according to their needs, wants, demographics, size, taste and taste, features etc. This challenge has adopted the concept of customer segmentation or market segmentation, where consumers are divided into subgroups or segments, where members of each subcategory exhibit similar market behaviours or characteristics.[4]

2.2 Data repository

Data collection is the process of collecting and measuring information against targeted changes in an established system, which enables one to answer relevant questions and evaluate the results.[5] In all disciplines of study, including the physical and social sciences, the humanities, and business, data collecting is an important aspect of the research process. All data collecting is done with the goal of obtaining high-quality evidence that will allow the analysis to develop concrete and deceptive responses to the questions posed. The dataset used in this paper were collected from Kaggle.

2.3 Clustering

Clustering is the process of grouping information into a dataset based on some commonalities. There are several algorithms, which can be applied to datasets based on the provided condition.[6] However, no universal clustering algorithm exists, hence it becomes important to choose the appropriate clustering techniques. It is basically a grouping of items based on their similarity and dissimilarity. There are several types of clustering algorithms that may be used to efficiently classify data. In this paper, we have implemented K-means clustering algorithm using the Python sklearn library.

2.4 K-Means Clustering

The K-means clustering algorithm is a division-based clustering technique. To re-divide data items and re-update cluster centres, it uses a practical iterative method. The algorithm's core concept is to consider a collection of element objects and the number of clusters to be formed. In the first round, a random sample element is chosen as the cluster centre. After analysing the distance between additional sample items and the centre point, the clusters are split into groups based on the distance. The iterative operation of the above steps is repeated in each of the following rounds, and the average of the element objects obtained this time is used as the middle point of another round of clustering till the criterion that the clustering midpoint does not change in the iteration process is met.

3. COMPARING CLUSTERING TECHNIQUE

In order to determine which clustering method should be used in which case, it is necessary to examine the many clustering techniques described. The table of comparisons is as follows:

Table 1: Various Clustering Techniques

Criteria	Affinity Propagation	Density Based Clustering	Hierarchical Clustering	K-Means
Computation Speed	Low	Low	Low	High
Clustering Time	More	More	More	Less
Granularity	No	Yes	No	Yes
Effect on size of data	Not Good	Not Good	Not Good	Good
Handle Dynamic Data	Yes	Yes	No	Yes
Clustering Result Efficiency	High	Medium	Low	Medium

Each and every method for clustering has its own set of benefits and drawbacks depending on the circumstance. For segmentation, K Means is the most often used clustering method. The K Means need a difficult to anticipate starting number of clusters, which might alter clustering results. Hierarchical clustering requires no initial number of cluster conditions, has a high temporal complexity, and is best suited for small to medium-sized datasets. The density-based clustering approach may be used to discover arbitrary formed clusters; however, it is inefficient when there is a large density difference between datapoints. The Affinity Propagation approach does not require a starting cluster, and the clustering result efficiency is great which ensures high clustering efficiency and applicability from small to medium datasets.

4. MARKET SEGMENTATION

Market segmentation is the process of categorising clients into groups with comparable characteristics such as purchasing behaviour, lifestyle, and food preferences. Market segmentation is a core strategic developing marketing concept that involves categorising people into distinct groups based on their eagerness, purchasing potential, and interest in purchasing. The segmentation procedure is carried out based on people's similarity in numerous parameters linked to the product in question. The more exactly and properly segments are used by a company to target clients, more the successful the company is in the marketplace. The basic objective of market segmentation is to precisely estimate consumer wants and hence increase profitability by acquiring or manufacturing items in the right quantity at the right time for the right client at the best price. To achieve these demanding standards, market segmentation using the k-means clustering approach may be used to make suitable forecasting and planning decisions. Cluster analysis may be used to categorise items such as brands, goods, utility, durability, and simplicity of use. For example, in a positioning exercise, which brands are grouped together in terms of customer perceptions, or which towns are grouped together and in terms of income, qualification, and so on.

5. PROPOSED MODEL

5.1 Importing Packages and Data

We began by loading all of the necessary libraries and dependencies. Customer id, gender, age, income, and spending score are the columns in the dataset. The matplotlib, pandas, Numpy and sklearn libraries have been imported. Pandas and NumPy will be used to manipulate data, sklearn will be used for modelling, and matplotlib will be used to plot graphs and pictures. Then, to put the data into the pandas data frame after loading the library. We'll utilise pandas' read csv function to do this.

5.2 Data Cleaning

We'll find that the data isn't nearly as valuable after importing the package and data, so we'll need to clean and organise it so that we can extract more significant insights.

5.3 Plotting of Graphs

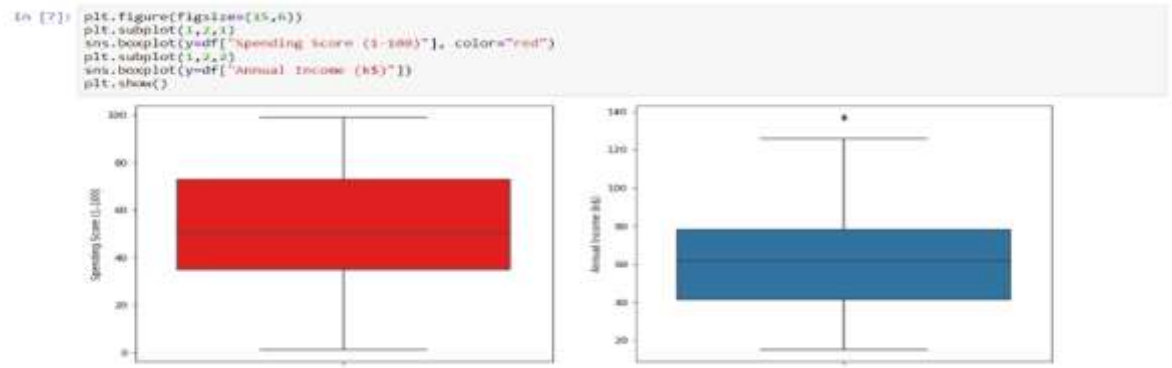

Chart 1: Box Plotting of Annual Income and Spending Score

To verify the distribution of male and female population in the dataset, we created a bar plot. The female population much outnumbers the male population. We also created a bar graph to examine the distribution of clients by age group.

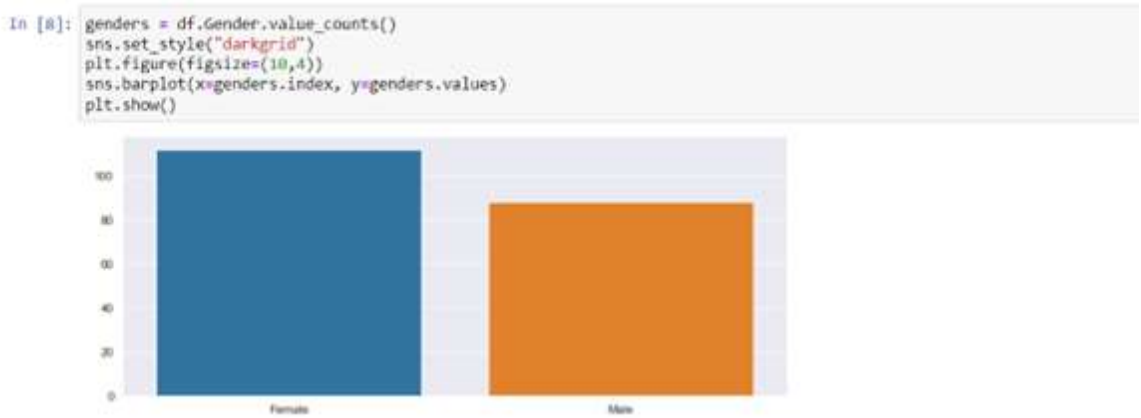

Chart 2: Bar Plot of Female and Male Customers

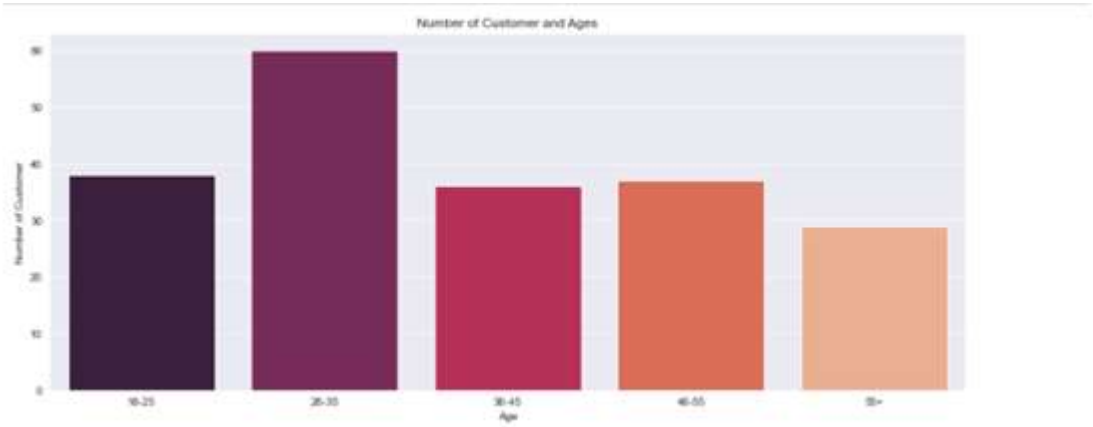

Chart 3: Bar Plot of Customer Distribution based on age

5.4 Select the optimal number of groups:

When using the k-means clustering method, one of the most important jobs is determining the ideal number of clusters. It's worth noting that while a k-means clustering model can converge for any value of K, not all values of K will result in the optimal model. We're going to use the elbow method.

5.5 Elbow Method

Calculate the Within Cluster Sum of Squared Errors (WCSS) for different values of k, and choose the k for which WCSS first starts to diminish.

The optimal K value is found to be 5 using the elbow method. We picked 5 since the inertia or sum of squared distance changes very little if the number of clusters is increased beyond 5.

The next stages can be summarised as follows:

1. Calculate K-Means clustering for various K values ranging from 1 to 10 clusters.
2. Calculate the total within-cluster sum of squares for each K. (WCSS).
3. Draw the WCSS curve versus the number of clusters K.
4. In most cases, the location of a bend in the plot is used to determine the proper number of clusters.

```
In [13]: from sklearn.cluster import KMeans
         wcss = []
         for k in range(1,11):
             kmeans = KMeans(n_clusters=k, init="k-means++")
             kmeans.fit(df.iloc[:,1:])
             wcss.append(kmeans.inertia_)
         plt.figure(figsize=(12,6))
         plt.grid()
         plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
         plt.xlabel("K Value")
         plt.xticks(np.arange(1,11,1))
         plt.ylabel("WCSS")
         plt.show()
```

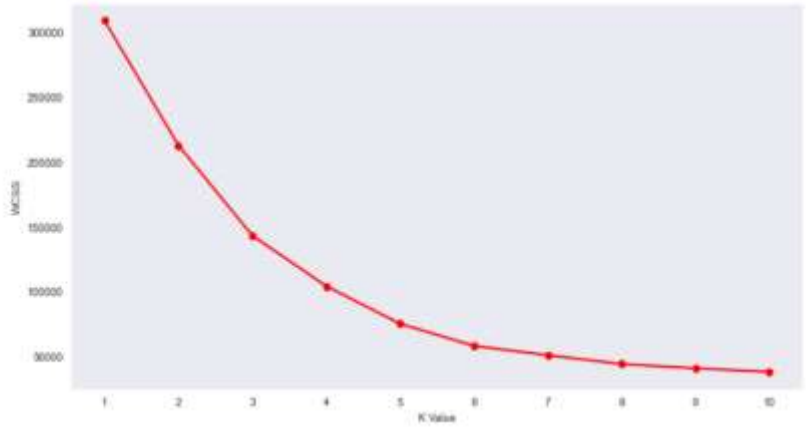

Chart 4: Graph Between K value and WCSS

5.6 Visualization

Finally, a 3D plot was built to show the consumers' spending scores in proportion to their annual income. The data points are separated into five groups, each of which is represented by a different colour in the 3D picture.

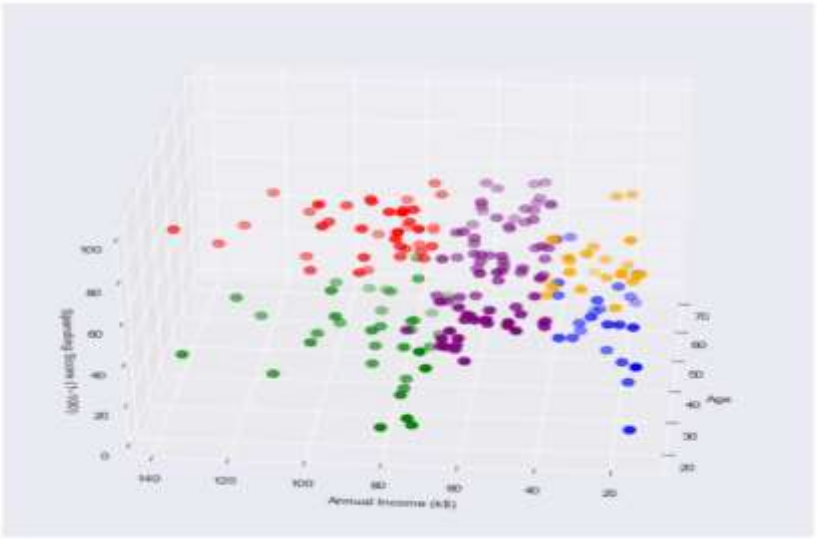

Chart 5: 3D Plot to illustrate the customers' spending scores versus annual income

6. CONCLUSION

When dealing with large volumes of data, companies must use more efficient clustering algorithms for customer segmentation. These clustering models must be capable of processing such a large volume of data. Each of the clustering approaches listed above has its own set of benefits and drawbacks. Using the aforementioned strategies, it was revealed that a hybrid way of combining algorithms can be beneficial depending on the circumstance and requirement, and applying the strategy correctly. To analyse,

execute, and process data with an appropriate grasp of the goals and apply the algorithm on a need-to-know basis, the clustering technique selection process would be lengthy. As a consequence, the organisation would benefit from recognising the distinct set of clients who improve earnings. It also assists businesses in sustaining client relationships and retaining customers by implementing various marketing methods.

7. REFERENCES

[1] Deng, Yulin, and Qianying Gao. "A study on e-commerce customer segmentation management based on improved K-means algorithm." Information Systems and e-Business Management (2018): 1-14.
[2] Tripathi, S., A. Bhardwaj, and E. Poovammal. "Approaches to clustering in customer segmentation." International Journal of Engineering & Technology 7.3.12 (2018): 802-807.
[3] Bhade, Kalyani, et al. "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018.
[4] Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.
[5] T. Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. Anu Manufactured cluster analysis using a new algorithm from structured and unstructured data. International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2.
[6] Jean Yan. - Big Data, Big Opportunities- Domains of Data.gov: Promote, lead, contribute, and collaborate in the big data era. 2013. Retrieved from http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf July 14, 2015.
[7] Sulekha Goyat. "The basis of market segmentation: a critical review of the literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online). Vol 3, No.9, 2011.
[8] Kansal, Tushar, et al. "Customer Segmentation using K-means Clustering." 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). IEEE, 2018.

BIBLIOGRAPHY

Souvik Biswas
Narula Institute of Technology, Kolkata, West Bengal, India

Neha Bhattacharya
Narula Institute of Technology, Kolkata, West Bengal, India

Ananya Samanta
Narula Institute of Technology, Kolkata, West Bengal, India

Nitin Gupta
Narula Institute of Technology, Kolkata, West Bengal, India

Dr Sangita Roy
Narula Institute of Technology, Kolkata, West Bengal, India

