K-means Algorithm for Segmentation of Customers

THORAINELLORE MANJULA1, GUDAMSETTY RAJESH2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT_ To identify and categorise customer trends, the method of customer segmentation is used. The K-means clustering technique is used to classify consumers and find potential clients. The marketing team will benefit greatly from these findings, as they will be able to target these clients and adjust their marketing strategy accordingly, which could lead to an increase in profit. Product price, marketing campaigns, feature selection, and prioritisation can all benefit from customer segmentation. An example of unsupervised machine learning is used in this research, where we provide unlabeled data and the system discovers previously undiscovered patterns. The categorization of data that requires classification can be accomplished through the use of unsupervised machine learning. An unsupervised machine learning technique known as clustering is used to divide our data into "clusters." As a result of the project's calculations, summaries, and charting of pertinent graphs, we are able to see the relationships between the group of clients and their interests.

1.INTRODUCTION

Data mining [1] techniques have become increasingly popular over the past several years as a result of the increasing rivalry in the business world and the vast amount of historical data that can be mined from an organization's database. Extracting data patterns and presenting them in a format that may be used for decision support is the goal of data mining.

An understanding of consumer behaviour can help marketing managers re-evaluate and prepare for the expansion of their customer-facing strategy. There are clustering approaches that consider data sets like tuples of data. Objects within a cluster are comparable to one another, but not to other objects in the same cluster, hence they are grouped together.

Splitting clients into groups known as customer segments, each of which is comprised of customers with similar characteristics, is known as customer segmentation. Segmentation is based on many similarities, such as gender, age, interests, and other spending behaviours, which are relevant to marketing. It's critical to segment customers in order to tailor marketing campaigns to meet the needs of each group, as well as to assist in business decisions, identify products associated with each group and manage supply and demand for those products, as well

as to identify and target a potential customer base, forecast customer defections, and provide directions in finding solutions. Customers.

Use a data mining technique known as K-means clustering to find client subgroups in this article. The best clusters are found by employing the elbow approach.

However, there are various commercial advantages if best current consumer segmentation is done appropriately. It's possible, for example, that performing an activity to determine your best existing customer segmentation will have a direct impact on your business's performance. Making your entire product better is the first step. A comprehensive understanding of who your customers are and what they need can help you position yourself as the ideal organisation to meet their specific requirements. A happier customer and a superior performance in comparison to rivals are the expected outcomes. Additionally, the benefits extend beyond your main product offering, since any insights about your greatest clients will allow your business to provide superior customer service support, technical assistance, and any other products or services they provide are included here.
Improved product: Segmentation projects can help you generate more focused marketing messages that are customised to each of your best segments, resulting in greater quality inbound interest for your product.

Allowing sales organisations to chase greater percentage opportunities: The sales team will be able to increase its win rate, cover a larger area, and eventually increase revenues by focusing on the most profitable areas.

It's important to focus on the quality of your earnings, rather than the quantity of it. Customers that purchase from the wrong segment may have a greater churn rate or reduced upsell potential after the original purchase has been made. By avoiding bad clients and focusing on the good ones, you may improve your profit margins and build a more stable customer base.

There are a slew of other advantages to figuring out how to best serve your current clientele. This is the bottom line: Increasing sales to your most profitable customers will allow you to scale your business more efficiently and ensure that everything you do — from lead generation to new product development — revolves around the right things.

2.LITERATURE SURVEY

2.1 CUSTOMER CLASSIFICATION

Over time, the commercial world became more competitive because companies like these have to satisfy and attract their consumers' demands and wants in order to improve their enterprises. Identifying and addressing the wants and requirements of each customer in the organisation is a very challenging undertaking. That's because there are so many variables to consider when it comes to customer preferences. As things are, it is unethical to treat every consumer the same when doing business. Customer segmentation or market segmentation is a result of this problem, in which customers are separated into groups or segments based on characteristics of the market that are common to all of those categories. When it comes to customer segmentation, customers are divided into distinct groups.

2.2 BIGDATA

The field of Big Data has recently seen a resurgence of interest. is a word that expresses vast amounts of formal and informal data that cannot be examined using conventional methods and procedures With billions of data on customers, suppliers and operations and millions of sensors sent to the real world on devices like smartphones and automobiles, companies are able to sense, create, and communicate data in the real world. covers a wide range of disciplines such as traffic control, weather forecasting, disaster preparedness, financial fraud control, economic transactions as well as national security education and healthcare. Volume, variety, and speed are the three Vs that make up big data. Authenticity and worth are the remaining two Vs available, making it a total of 5V.

2.3 **DATA COLLECTION**

Data collection is the process of gathering and analysing information in order to answer relevant questions and evaluate the outcomes of a certain project. Regardless of the subject of study, data gathering is an essential aspect of the research process. The goal of any data collection effort is to find out more information.

2.4 CLUSTERING DATA

It is the act of grouping information in a dataset based on certain similarities. There are a variety of methods that can be used to analyse a dataset, depending on the circumstances. Nevertheless, there is no uniform algorithm for clustering, therefore the choice of proper clustering approaches becomes very important. The python sklearn module was used to develop three clustering methods in this paper.

2.5 K-Means

K- indicates that a classification algorithm is one of the most widely used ones. This approach relies on the centroid [5] where each data point is placed in one of the overlapping K clusters pre-programmed into the algorithm.. For this purpose, data clusters are produced that correlate to a hidden pattern, which provides insight into how an operation should be executed. We'll utilise the elbow approach for k-means assembly, which is one of many options.

3.PROPOSED SYSTEM

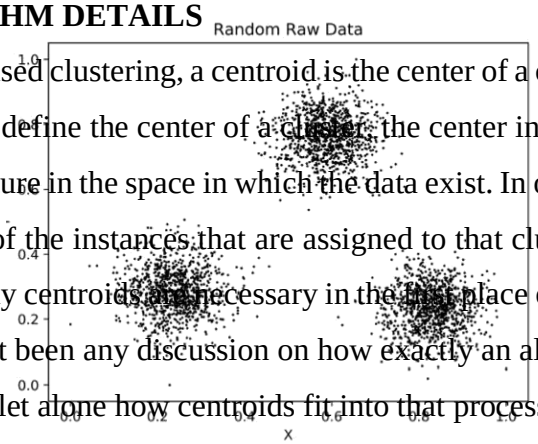
When we perform clustering in the dataset and divide people into groups (or) clusters, we want each cluster to share a common characteristic with the rest of the groups. Businesses can use this as an indicator to determine which marketing strategies will help them grow their business quickly. With the K-Means Machine Learning technique, this categorization may be done quickly and accurately. The more accurate a dataset is, the larger it is.

We begin by conducting exploratory data analysis to understand which attributes lead to the development of meaningful patterns. You put it another way, EDD helps to identify the right traits to focus on. After that, we'll use clustering to locate similar groups so that we may offer ads to them that will boost earnings for the company.

A kaggle dataset of 200 customer records is used for this task. The following characteristics (or traits) are included in the dataset.

- CustomerID
- Gender
- Age
- Spending Score
- Annual Income

3.1 K-MEANS ALGORITHM DETAILS



In the context of centroid-based clustering, a centroid is the center of a cluster of data. Although there are numerous ways to define the center of a cluster, the center in a k-means cluster is the arithmetic mean of each feature in the space in which the data exist. In other words, the centroid is the mean of the features of the instances that are assigned to that cluster. However, it might not be immediately clear why centroids are necessary in the first place or how to initially define them. After all, there has not been any discussion on how exactly an algorithm would go about grouping data into clusters, let alone how centroids fit into that process.

Fig 1: Random Raw Data

Immediately after looking at this figure, there are three things that are apparent. First, each data point exists in two dimensions: x and y. Although naming the axes x and y is convenient, it does not provide much insight into what the axes represent. So, if x and y are too simple, perhaps think of them as age and income or time between visits and average ticket. Regardless of what the axes' names are, the crucial point is that there are two dimensions. Second, the data is scaled between 0 and 1 in both dimensions.

Lastly, to a human eye, there appear to be at least three distinct clusters.In the figure, there are two features and although there are several thousand instances, plotting them all at once makes it easier to see the differences between each datum. Because of the low number of features and the ability to see all the data clearly, the human mind has little difficulty dividing up the data into clusters. [7] However, teaching a computer to perform the same task is slightly more difficult. For all the incredible tasks that a CPU can perform, it cannot visualize the data and divide it into nice groups like a human can. Therefore, it is reasonable to wonder how a computer would go about the task of clustering. In centroid-based clustering, the most popular algorithm is k-means. There are two important parts to the name: k and means. Here, k refers to the number of centroids (clusters) the algorithm will generate and "means" refers to what the centroids are: arithmetic means of the data. Roughly the K-Means algorithm can be broken up into four sections, each with their own important attributes.

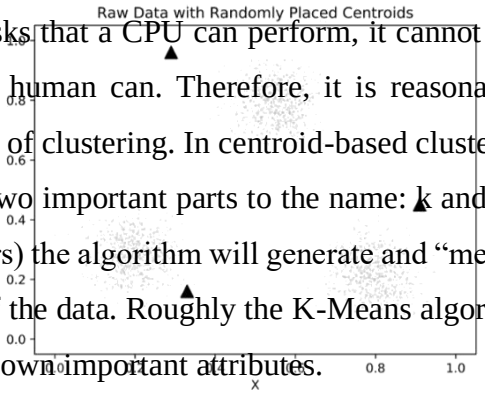


Fig 2: Raw Data with Centroids

Hence the K-Means clustering algorithm works by finding groups based on Euclidean distance, a measure of distance or similarity. The practitioner selects k groups to cluster, and the algorithm finds the best centroids for the k groups. The practitioner can then use those groups to determine which factors group members relate. For customers, these would be their buying preferences.

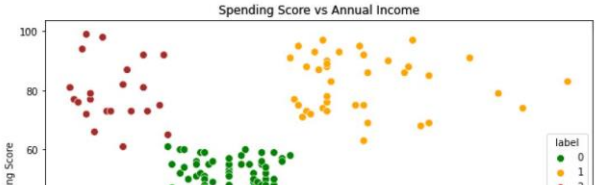


Fig 3: Example of Clustering

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. [8]The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

Algorithm

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centers.
3. Assign objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-Means is relatively an efficient method. However, we need to specify the number of clusters in advance and the final results are sensitive to initialization and often terminate at a local optimum. Unfortunately there is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk of over fitting.

4.RESULTS AND DISCUSSION

- Initially, a data set of 200 samples were taken. It has customer id , age , gender, annual income and spending score . Spending score is given to the customer by the mall based on their spending.
- Then, drop the customer id as it is just an id that uniquely identifies a row and has nothing to analyze with it.
- Then, the data has been cleaned. Here, columns with appropriate names are renamed and checked for null values.
- All the values were assigned numeric values, for instance ‘0’ and ‘1’ were assigned to male and female.
- To get a clear view, data is cleaned and analyzed.

- After analyzing the data, surprising results were made such as
    - There are more female customers than male customers
    - There are more customers in the age group 25 to 40, who visit the mall frequently.
    - The annual income of the customers of the mall ranges from 15K USD to 137K USD.
    - There are people with spending scores as low as 1 to as high as 99.

  from this data it can be concluded that there are several customers with varying needs.
- After analyzing, individual distributions of the variables, tried to find the relation among them.
- A correlation matrix was drawn and it has been found out that variables are not much correlated, but there is a negative correlation between age and spending score.
- Then, the relation between gender and spending score was drawn as there are more female customers. It was found that the mean spending score of male and female doesn't differ much even though there are more female customers. Hence, dropped the gender attribute as it is not showing any useful information for the further process.
- Then found the relation between age and annual income. It's found that people in their 30's,40's and 50's have higher income.
- While finding a relation between spending score and age, it is found that people of ages 40's or 50's spend more.
- Hence from this analysis, it is concluded that clustering between age and spending score, annual income and spending score would give us better results.
- Before clustering, K values should be selected for forming k number of clusters. They can't be assigned randomly. So, here the elbow method was used to find the value of 'K'.
- There are 2 measures to find the value of k. They are distortion and inertia.
    - Distortion is the average of the squared distances from the cluster's center.
    - Inertia is the sum of squared distances from the cluster centers.
- Here the inertia method was to plot the graph. After plotting the graph, we have to choose the value of k at the "elbow" that means the point after which the inertia starts decreasing in a linear fashion [10].
- For spending score vs annual income, it is found the value of k is 5 and we got 5 clusters such that there are people with less income and less spending score, more income and more spending score, less income and more spending score , more

income and less spending score , average income and average spending score. Also found that there are more people with average income and average spending score.

- Then again using the elbow method, found the optimum number of clusters for clustering spending score and age is 4.
- After clustering it was found that people from age groups 20 to 40 are spending more and average and people above 40 are spending average and low.

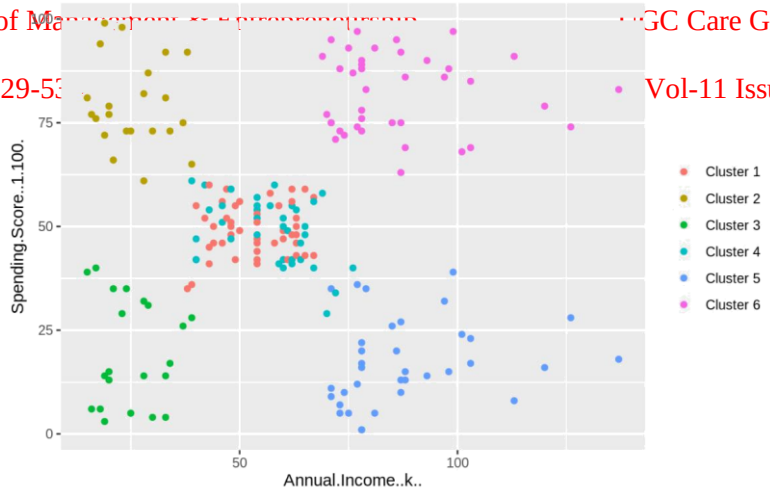


Fig 4: Clusters based on two attributes

- So by looking at the clusters we can conclude that the mall can improve its profits by:

1.Attracting people of the age group 20 to 40 by conducting some campaigns.

2.We can even see that people with high income are spending high and averagely. So the reason behind this might be customer dissatisfaction. Maybe due to less quality or no proper services these people are spending less. So by improving the quality and services the mall can get more profits.

And even the mall can target the customers with average income and average spending score. As the mall has more customers from this sector, having special campaigns for these people can increase the customer base and thereby increase the profit

5.CONCLUSION

It is possible to discover, prioritise, and target your best current customers by following this project's step-by-step method, but this does not ensure success. A successful process is one that takes into account all of the potential obstacles and difficulties that may arise during the various stages of implementation and constantly adjusts to the new information and feedback that may influence the final result.

As a last point, you cannot push this procedure into your firm. You won't get very far if you don't get buy-in from the people who will be affected by the best current customers segmentation procedure.

However, the influence it can have on every element of your business — marketing, product development, customer support, and so on — sales is enormous if you manage the finest current customer segmentation process properly. As a result, your company will be able to grow with greater predictability and efficiency thanks to a sharper focus on its customers and a more defined target market.

So in the long run you won't have to take on every customer who is prepared to pay for your product or service, allowing you to focus on a smaller subset of consumers who offer the most profitable opportunities and efficient use of resources. At this time, it might be the difference between great success and certain failure. That is crucial for every firm, of course.

REFERENCES

[1] Mobasher B, Cooley R, Srivastava J. Automatic Personalization Based on Web Usage Mining. Commun ACM. 2000;43(8).

[2] Al-Qaed F, Sutcliffe A. Adaptive Decision Support System (ADSS) for B2C ECommerce. 2006 ICEC Eighth Int Conf Electron Commer Proc NEW E-COMMERCE Innov Conqu Curr BARRIERS, Obs LIMITATIONS TO Conduct Success Bus INTERNET. 2006:492503.

[3] Cherna Y, Tzenga G. Measuring Consumer Loyalty of B2C e-Retailing Service by Fuzzy Integral: a FANP-Based Synthetic Model. In: International Conference on Fuzzy Theory and Its Applications iFUZZY.; 2012:48-56.

[4] Magento. An Introduction to Customer Segmentation. 2014. info2.magento.com/.../ An_Introduction_to_Customer_Segmentation..

[5] Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities.

[6] Griva, A., Bardaki, C., Pramatari, K., Papakyriakopoulos, D. (2018). Sales business analysis: Customer categories use market basket data. Systems Expert Systems, 100, 1-16.

[7] Hong, T., Kim, E. (2011). It separates consumers from online stores based on factors that affect the customer's intention to purchase. Expert System Applications, 39 (2), 2127-2131.

[8] Hwang, Y. H. (2019). Hands-on Advertising Science Data: Develop your machine learning marketing strategies… using Python and R.

[9] 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 ‖ Global Journal of Management and business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN:
0975-5853. Volume 12 Issue 1.

[10] 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.