SEGMENTATION FOR SERVICE PROVIDERS USING MACHINE LEARNING

K.C.Bhanu[1], Dr.P.Uma Maheswari Devi[2]
[1]Research Scholar, Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry
[2]Associate Professor , Department of Commerce and Management Studies
Adikavi Nannayya University ,Rajahmundry
bhanu1605@gmail.com,umadevi_4@yahoo.com

ABSTRACT

As the global population rises at an alarming rate, business of all kinds will need to begin segmenting their clientele. It's growing in importance in today's corporate world. This aids businesses in understanding their product positioning, which in turn helps them retain and benefit from their consumer base. The next idea to be examined is customer segmentation, which involves categorizing a business' clientele into subsets defined by shared traits and preferences. One of the most important tools available is machine learning, which uses a variety of algorithms to uncover previously unseen patterns in data that can then be used to inform decisions. An unsupervised machine learning clustering technique called "k-means" is used to carry out this strategy.

Keywords: K-mean, Elbow method, WCSS.

INTRODUCTION:

Basically, customer segmentation is the process of dividing customers into subsets based on those distinguishing characteristics. The primary objective of customer segmentation is to collect data about individual consumers and analyze the resulting patterns. Common types of segmentation include geographical (dividing customers by region), demographic (dividing customers by age and gender), behavioral (dividing customers based on purchase and usage patterns), and psychographic (dividing customers based on attributes and values). Machine learning is a collection of techniques used to construct mathematical models and derive inferences from data. The K-Means clustering algorithm, based on unsupervised learning, is used to categorize customers into distinct groups.

LITERATURE REVIEW:

Elaheh et,al.,(2021) Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques Using preexisting online travel evaluations on TripAdvisor, this research hopes to learn about travelers' decision-making processes when selecting eco-friendly hotels. Therefore, a strategy combining segmentation and the strategy for Order of Preference by Similarity to Ideal Solution (TOPSIS) was created to classify tourists according to their reviews and rank the relevance of green hotel features. The information originated from TripAdvisor reviews of eco-friendly hotels in Malaysia written by actual guests. The majority of the segments surveyed indicated that a good night's sleep was an important consideration while choosing an eco-hotel.

Cormac et,al.,(2017), Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers. Due to the intense competition in the telecoms industry, mobile providers want a business intelligence model that will allow them to attract and retain customers at the lowest possible cost. Marketing strategies can be improved with the use of machine learning technology. Customer segmentation is another area where data mining techniques might be useful. This paper's objective is to examine the C.5 algorithm in the context of naïve Bayesian modeling for the aim of behavioral profiling of telecommunications consumers based on billing and socio-demographic factors. The findings have been put into practice experimentally.

In order to better sell to their customers, businesses often classify them into subsets based on shared traits. Marketers can target specific subgroups more effectively with the help of segmentation. Make marketing materials that appeal to specific audiences and share them with them. Choose the most effective method of reaching each consumer subset, such as radio, social media, etc.

METHODOLOGY:

Machine learning:

Machine learning is a branch of artificial intelligence (AI) concerned with teaching computers to learn and predict or act without being given explicit instructions. Machine learning is the application of statistical methods and mathematical models to computer systems so that they can develop or learn on their own based on inputs.

The primary objective of machine learning is to create algorithms with the ability to learn and adapt from data automatically, without the need to be specifically coded for each task or scenario. To accomplish this, models are built that can identify connections and trends in the data and produce reliable outcomes as a result of these analyses.

Among the many machine learning algorithms available today are:

First, there's supervised learning, which involves training an algorithm with data that has already been annotated with labels. The algorithm learns to generalize from the labeled instances and provide the right output based on the incoming data.

Second, we have unsupervised learning, in which the algorithm is given unlabeled data and instructed to discover hidden patterns or structures. The objective is to find patterns or associations in the data where none existed before.

Third, we have reinforcement learning, in which an agent learns to make decisions or execute actions in response to a reward signal through interaction with its environment. The agent is taught by the environment, which provides it with reinforcement or correction.

Among the many domains where machine learning has found success are

One use of machine learning is in the area of image and speech recognition, where it may be taught to recognize and classify images, recognize objects within images, and transcribe audio into text.

To facilitate applications like chatbots, language translation, sentiment analysis, and text summarization, machine learning is used to construct models that interpret and generate human language (natural language processing).

For example, sales forecasting, stock market prediction, and consumer behavior analysis can all benefit from the use of machine learning algorithms to evaluate huge datasets and anticipate or forecast future results.

Fourth, medical diagnostics can benefit from the application of machine learning techniques through the detection of patterns or signs of disease in patient data.

5. Self-driving cars: Machine learning is essential in the development of autonomous vehicles because it allows the automobile to see and comprehend its environment, make judgments, and safely drive.

Data Collection

Information is gathered through data collection when questions are asked and results are measured in relation to specific changes made to a known system. Collecting data is an integral aspect of research across many disciplines, from the hard sciences to the humanities to business. All data gathering efforts should be directed at accumulating sufficient proof for analysis and, ultimately, invention.

K-Means

To classify an unlabeled dataset into distinct groups, K-Means Clustering can be used as an Unsupervised Learning technique. Here K indicates the number of clusters that must be constructed ahead of time; for example, if K=2, only two clusters will be created, if K=3, only three, and so on. Each cluster has a centroid that the algorithm uses to make decisions. This algorithm's primary goal is to find the smallest possible sum of distances between data points and their respective clusters.

RESULTS:

The data set includes exchanges that occurred in the years 2021-2023. There are 10 characteristics in this data set. They are as follows: 1. invoice number 2. customer id 3. gender 4.age group 5. category 6.quantity 7.price 8.method of payment 9. invoice date and 10. shopping center.

Shoppers between the ages of 18 and 68 buy a wide variety of goods, including drinks, books, garments, cosmetics, electronics, footwear, and mementos.

Category /Age	Beverage	Books	Clothing	cosmetics	Electronic Gadgets	Shoes	Souvenir
18-24	0.124306	0.121624	0.121720	0.117866	0.116096	0.123605	0.117623
24-38	0.280287	0.275487	0.280287	0.281446	0.290758	0.278007	0.288217
38-48	0.199242	0.196148	0.202645	0.199174	0.203946	0.200909	0.197706
48-58	0.197977	0.207452	0.196107	0.203718	0.193146	0.196569	0.194786
58-68	0.198188	0.199288	0.199241	0.197797	0.196054	0.200909	0.201668

We have compared the things purchased by customers at various shopping centers.

Category/Shopping Mall	Beverage	Books	Clothing	Cosmetics	Electronic Gadgets	Shoes	Souven
Brook Field Mall	0.12673	0.095544	0.102215	0.102066	0.092874	0.107036	0.10622
Elante Mall	0.049611	0.049378	0.049527	0.048616	0.051241	0.049532	0.04500
Emaar Square Mall	0.046836	0.047371	0.050338	0.052642	0.049631	0.049631	0.04941
Fun Public Mall	0.151810	0.151144	0.152729	0.154524	0.145206	0.145206	0.01476
LuLu Mall	0.052250	0.051766	0.050136	0.051241	0.048834	0.048834	0.04801
Phoneix Mall	0.200135	0.205741	0.200748	0.203563	0.201615	0.201615	0.18963
Spencer Plaza	0.196887	0.203332	0.198341	0.200225	0.199560	0.201017	0.20744
Square Mall	0.048190	0.050582	0.050136	0.051000	0.046837	0.045645	0.05221
Z-Square Mall	0.100846	0.092734	0.097228	0.096569	0.097478	0.099761	0.10302
Zorlu Center	0.050761	0.052389	0.051064	0.050603	0.050040	0.051724	0.05141

We compared the amounts spent by men and women at several malls.

Gender/ Shopping Mall	Female	Male
Brook Field Mall	0.103292	0.100488
Elante Mall	0.049578	0.049156
EmaarSquare Mall	0.047779	0.049256
Fun Public Mall	0.150314	0.151845
LuLu Mall	0.049427	0.051307
Phoneix Mall	0.200094	0.201151
Spencer Plaza	0.200161	0.198049
Square Mall	0.050704	0.048305
Z-Square Mall	0.098753	0.097736
Zorlu Center	0.049897	0.052708

Elbow Method:

The elbow approach is a strategy used in data analysis and machine learning to find the best possible breakdown of a dataset into groups, or clusters. Unsupervised learning techniques, like k-means and other clustering algorithms, make extensive use of this technique.

The steps of the elbow method are as follows:

Select a cluster number (k) between 1 and : To begin, pick a range of numbers to use for the number of clusters. You can begin with k=1 and increase it till you reach some value.

Step 2: Cluster the data using the selected number of clusters (k) using a clustering method such as k-means.

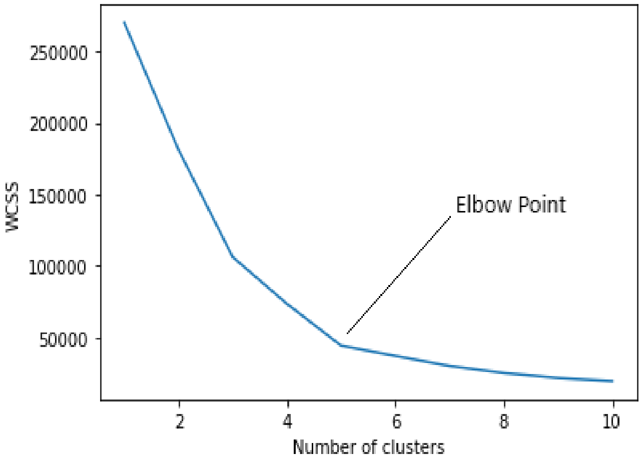
In order to determine how far off each data point is from the cluster centroid, we can use the sum of the squared distances (SSE) formula. The SSE quantifies the distance from the data points to the cluster centers.

Create a line chart or line plot showing the SSE values against the number of clusters (k) in a given dataset.

Find the point where the SSE values decline less dramatically with each new cluster (the "elbow" point) by examining the line plot. In most stories, the plot bends at this moment, creating an elbow shape.

6. Decide how many clusters to use; this is a trade-off between the complexity of the model (the number of clusters) and how well it fits the data, shown by the "elbow" point on the plot. Take the amount of clusters represented by the elbow as your starting point.

When employing the Elbow technique, we change K, the number of clusters, from 1 to 10. We are computing the WCSS (Within-Cluster Sum of Squares) for a variety of different K values. To calculate WCSS, add together the squares of everyone's distances from the cluster center. The WCSS plotted against K resembles an Elbow shape. With an increasing number of clusters, the WCSS value will begin to drop. The highest WCSS value occurs at K = 1. When we examine the graph closely, we observe that it will undergo a sudden transformation at a specific point, giving it the appearance of an elbow. From here on forth, the line begins to converge toward the X-axis. Here we have the optimal value of K, often known as the ideal number of clusters.



WCSS, which stands for "Within-Cluster Sum of Squares," is a statistic used to assess the efficacy of clustering in unsupervised learning algorithms like k-means clustering. By adding together the squared distances between each data point and its centroid within a cluster, WCSS calculates the compactness or cohesion of the clusters.

Following are the procedures for determining WCSS:

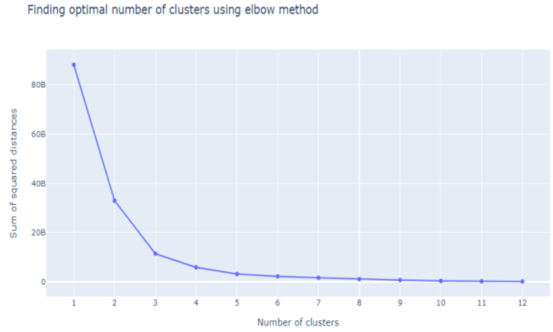
- Select a clustering parameter, k, for your study.
- Second, divide the data into groups using a clustering method like k-means.
- Third, determine the SSE between each data point and the cluster centroid for each cluster.
- The WCSS is calculated by adding the SSE values for each cluster together.

The WCSS can be computed mathematically as follows:

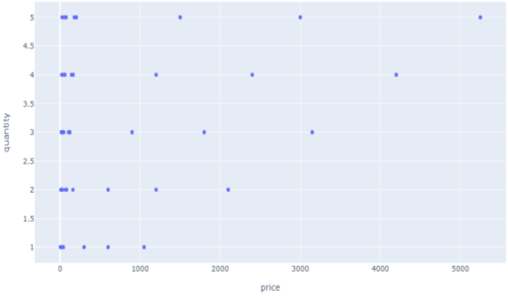
With a WCSS of = $\sum((x_i – c_i)^2$

where x is a data point inside a cluster
Cluster center is denoted by the letter c.

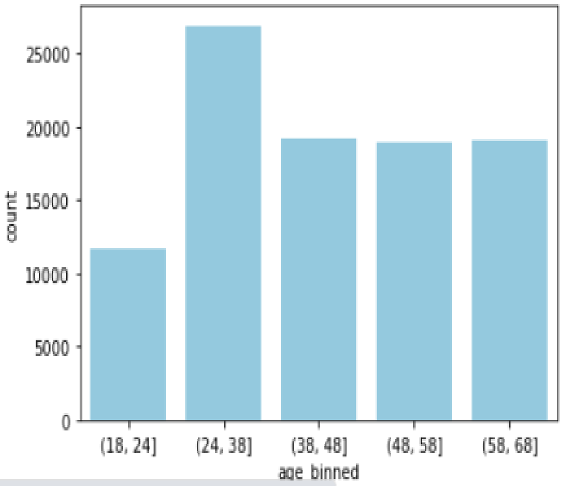
Clustering works by attempting to reduce the WCSS to a minimum. If the WCSS is low, then the data points are clustered closely together and are closer to the centers of those clusters. It should be stressed, however, that the absolute value of WCSS does not tell us very much. Its primary function is as a relative metric by which various clustering solutions with variable cluster sizes may be compared.



One of the most fundamental representations in cluster analysis is a graph in which the X-axis represents the number of clusters and the Y-axis represents the sum of squared distances between each pair of clusters. Data points are clustered using cluster analysis, and the total of squared distances is affected directly by the number of clusters. As we read the graph from left to right, we see that the sum of squared distances decreases as the number of clusters grows. Each data point tends to be closer to its own cluster centroid as more clusters are used, resulting in a smaller variance. More clusters may not appreciably reduce the sum of squared distances after a certain point on the graph, where the reduction in sum of squared distances slows down. The "elbow point" is a key criteria for establishing the suitable number of clusters; it is also known by another name. Choosing the right number of clusters is crucial for collecting relevant patterns in the data while avoiding overfitting. Decisions in cluster analysis can be greatly aided by consulting a graph depicting the number of clusters as a function of the sum of squared distances between them.



The x-axis of the price-quantity graph depicts the cost of the product or service, while the y-axis shows the quantity that is being purchased or sold. According to the law of demand, the quantity sought and the price drop together as we travel from left to right along the graph. This inverse relationship between price and quantity indicates that when prices drop, buyers are more likely to make bulk purchases. Moving from right to left on the graph, however, we see that as the price increases, the amount required often falls, in accordance with the concept of diminishing returns. The law of supply, which states that a rise in price usually leads in a rise in quantity provided by producers, may also be shown graphically using this graph. The connection between consumer demand and producer supply can be better understood through the visual depiction of price and quantity.



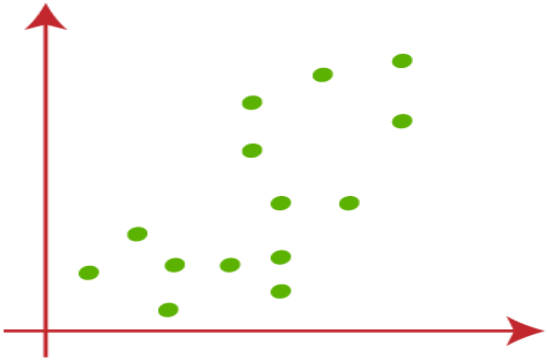
The age distribution of the population may be seen clearly in the bar chart by looking at the X-axis, which shows the age bins (18–24), (24–38), (38–48), (48–58), and (58–68), and the Y-axis, which shows the number of people in each age bin. The (24, 38) bracket stands out among these age groups due to its notably larger size. This finding implies that there is a disproportionately big group of people between the ages of 24 and 38. This age group's predominance may indicate a number of things, including a disproportionate share of the labor force or a shift in other demographic tendencies. Market targeting, public policy, and social research are just few of the many sectors that depend on an accurate understanding of these age distribution shifts. The bar chart provides an easy-to-understand visual depiction of the age-binned data, drawing attention to the significant role that the (24, 38) age group plays in the demographic makeup of the population as a whole.

K Means

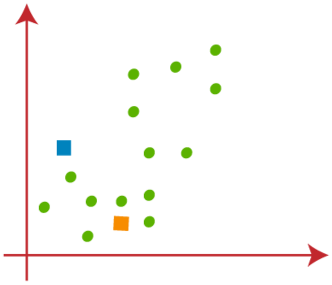
Here are the basic steps of the K-Means algorithm and how they work:

- First, decide how many clusters there will be by picking a value for K.
- Second, pick any K of the points or centers at random. Other than the input dataset, it can be anything.
- Third, place each data point within one of the K clusters based on its proximity to the data's centroid.
- Fourth, you will determine the variance and relocate the cluster centroids.
- Fifth, after rearranging the data such that it is once again centered on the new centroid for each cluster, repeat the third step.
- Sixth Step: Return to Step 4 if a Reassignment Occurs; Otherwise, FINISH.
- Seventh step: Finishing up the model.
- Cases in point of visualisation:

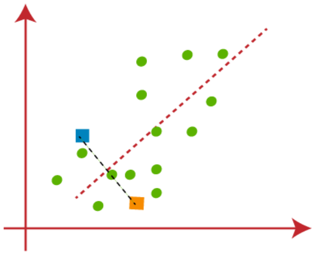
We have two separate numbers, M1 and M2. Scatter plot x-axis is displayed:



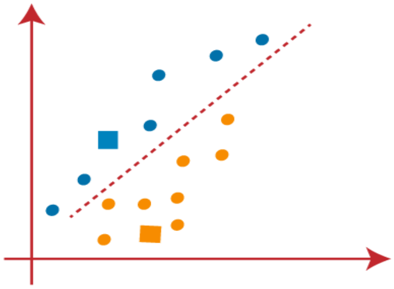
Consider a set of k clusters. In other words, k=2 classify the dataset into subsets. Therefore, we must pick a random k points or a random centroid to construct a cluster. It might be picked from inside the dataset or arbitrarily. Those are the two blue dots that we've decided to identify as K1centroid and orange as K2 centroid.



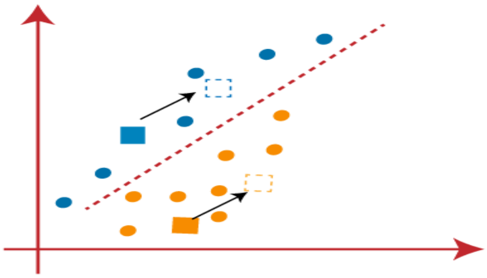
Each point on the scatter plot is now linked to the k-point or centroid that it is most closely associated with. The distance between the two spots can be found by drawing a median line between their centers.



As can be seen in the image above, the left-hand side contains points near to the K1 centroid, while the right-hand side has points close to the K2 centroid. And have them use blue and orange to fill up every space. Evidenced below:



We need to pick a new centroid to locate the nearest cluster point.



K-means clustering relies on forming extremely efficient clusters to get its desired results. However, determining how many clusters to use is a difficult process. We explore the best approach to determining the optimal cluster size, or value of K, out of the several options available.

Software for Customer Segmentation:

Customers can be broken down into subsets using a variety of criteria, including demographics (such as age and gender) and purchase history. HUBSPOT is customer-segmentation software.

HubSpot's marketing software, dubbed HUBSSPOT, includes modules for SEO, social media promotion, content curation, web analytics, landing pages, and customer service.

Contact lists allow you to organize your clientele into distinct groups. Event-based segmentation is another feature of this program. If your business is hosting a seminar, Event-Based segmentation can help you find and notify potential attendees.

Conclusion:

This study proves that segmenting customers based on behavioral characteristics is a better solution for the current customer segmentation problem, and K-means clustering is identified as a good choice for this task using data collected from customers at a shopping mall using features such as quantity, price, category, and payment method.

References:

1. Yadegaridehkordi, E., Nilashi, M., Nasir, M. H. N. B. M., Momtazi, S., Samad, S., Supriyanto, E., & Ghabban, F. (2021). Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. *Technology in Society*, *65*, 101528.

2. Dullaghan, C., & Rozaki, E. (2017). Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers. *arXiv preprint arXiv:1702.02215*.
3. Sharaf Addin, E. H., Admodisastro, N., Mohd Ashri, S. N. S., Kamaruddin, A., & Chong, Y. C. (2022). Customer mobile behavioral segmentation and analysis in telecom using machine learning. *Applied Artificial Intelligence*, *36*(1), 2009223.
4. Regmi, S. R., Meena, J., Kanojia, U., & Kant, V. (2022, April). Customer Market Segmentation using Machine Learning Algorithm. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1348-1354). IEEE.
5. Yuping, Z., Jílková, P., Guanyu, C., & Weisl, D. (2020, March). New methods of customer segmentation and individual credit evaluation based on machine learning. In *"New Silk Road: Business Cooperation and Prospective of Economic Development"(NSRBCPED 2019)* (pp. 925-931). Atlantis Press.
6. M Hesand M Tiveb, A Babaniac. 2014. Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. Procedia - Social and Behavioral Sciences 51 (2014), 1327—1332
7. R. Srivastava, "Identification of customer clusters using rfm model: a case of diverse purchaser classification", International Journal of Business Analytics and Intelligence, vol. 4, no. 2, pp. 45-50, 2016.
8. H. Valecha, A. Varma, I. Khare, A. Sachdeva and M. Goyal, "Prediction of consumer behaviour using random forest algorithm", 2018 5th IEEE Uttar Pradesh section international conference on electrical electronics and computer engineering (UPCON), pp. 1-6, 2018.
9. K. Khalili-Damghani, F. Abdi and S. Abolmakarem, "Hybrid soft computing approach based on clustering rule mining and decision tree analysis for customer segmentation problem: Real case of customercentric industries", Applied Soft Computing, vol. 73, pp. 816-828, 2018.
10. A. Vattani, "K-means exponential iterations even in the plane," Discrete and Computational Geometry, vol. 45, no. 4, pp. 596-616, 2011.
11. [5]. I.S.Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.
12. [6]. Gaurav Gupta and Himanshu Aggarwal International Journal of Machine Learning and Computing, Vol. 2, No. 6, December 2012 10.7763/IJMLC.2012.V2.256 Improving Customer Relationship Management Using Data Mining (20)
13. [7]. VandanaKorde and C NamrataMahender," Text classification and classifiers:A survey", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012