Customer Segmentation Using K-Means Clustering in Unsupervised Machine Learning

Mohd Faraz Alam
Galgotias College of Engineering & Technology Greater Noida, India
farazalam925@gmail.com

Raushan Singh
Galgotias College of Engineering & Technology Greater Noida, India
kunalitian99@gmail.com

Sandhya Katiyar
Galgotias College of Engineering & Technology Greater Noida, India
san.katiyar@galgotiacollege.edu

2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 978-1-6654-3811-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAC3N53548.2021.9725644

Abstract—The new era's perspective is one of creativity, in which everybody is competing to be better than the others. Today's businesses are built on the potential of creativity to enslave consumers with goods, but with too many options, customers are left unsure of what to buy and what not to buy, and companies are focused on which segments of customers to target to sell their products. Various algorithms are employed to uncover hidden trends in data so that future judgments can be made with confidence. The uncertain definition of which segment to target can be resolved by employing segmentation. A python program was developed, and it was trained using a dataset with two features: 250 training samples taken from a local retail shop and a standard scalar. Both of these characteristics are the average of the amount of shopping done by consumers and the average of the customer's annual visits to the store. We will be able to classify the target customers using clustering based on demographic information (such as gender, age, income, and so on), geographical information, psychographics, and behavioral data.

Keywords: CRM, K-means Clustering, WSS, WCSS, Seaborn

I. INTRODUCTION

When more companies open every day, it is becoming increasingly necessary for existing businesses to employ marketing strategies in order to remain competitive. In today's world, the basic rule of current marketing is "change or die." Since the number of customers is growing, everyday it has become challenging for every customer. Data mining is critical when dealing with secret data trends in a company database. Customer segmentation is a data mining technique that divides consumers into clusters based on common features, making it easier for a business to handle a large number of clients.This sub-segment can impact directly or indirectly to marketing strategies.

II. RELATED WORK

A. This section reviews and discusses key ideas such as customer segmentation, CRM, and customer usability. The importance of these ideas is also emphasised.

A. CUSTOMER SEGMENTATION

As the market expands, the pace of rivalry between all commercial organizations is increasing. As a result, many companies are increasing their marketing expenses in order to acquire a competitive advantage. In this perspective, incorporating Information Technology (IT) into marketing plans stands out as a vital element in a modern business strategy. Client segmentation is a typical approach of splitting a customer base into externally separate and internally standardized segments, each of which is targeted based on its own set of criteria. Despite an increase in the gathering of consumer behavior data, many researchers are now focusing on grouping clients from transactional data [1]. In general, It is the technique of categorising customers into groups based on their preferences, characteristics, and purchase habits[6]. Marketing of various Businesses can be enhanced by these strategies which depend on the customer's preferences by studying and analyzing massive volumes of collected consumer data [3],[8]. According to [9] every business organization can maximize earnings if resources are appropriately allocated in order to cultivate the most loyal and valuable group of clients following customer segmentation and clustering.[4] The entire consumer base can be grouped and segmented into clusters based on their purchasing behaviors, spending time, and demographics [2]. As a result, rather than researching each customer individually, businesses can aggregate customers using approaches such as statistical approaches, conceptual clustering, and resilient clustering [5].

B. Customer Segmentation Methodology

As the market expands, the pace of rivalry between all commercial organizations is increasing. As a result, in order to acquire a competitive advantage, many companies are boosting their marketing spending [10],[11]. In this environment, incorporating Information Technology (IT) into marketing plans stands out as a key component of a modern marketing strategy. Customer segmentation is used by ecommerce industries and companies to target specific consumer groups with data and things that customers in that category are likely to find relevant. [2]. Client segmentation is a means of breaking a customer base into publicly separate and internally homogenous groups in order to design different marketing tactics for each category depending on their attributes. It is defined as the practise of grouping a company's customers based on their common interests, characteristics, and behaviours.

Identifying the potential customer base for selling the product
Implementing Clustering Algorithms to group the customer base
Selling product to the identified customer group

ISBN: 978-1-6654-3811-7/21/$31.00 ©2021 IEEE

B. Use of Customer Segmentation

When it comes to target marketing and customer segmentation, it appears that the two are inextricably linked. They're frequently interchanged. The term "target marketing" refers to the categorization of customers depending on the properties that the company wants to serve. In the context of a specific customer, It's been referred a "personal branding" strategy.In order to develop segmentation-based marketing strategies, there are three phases that must be followed. Customers in the targeted market are first divided into categories depending on their preferences. Second, The features of the segments are explored, as well as the numerous marketing methods that might be employed to target that specific group. Finally, required brand comparisons and customer behavior research for competing brands can be done.

C. Clustering Technique for Customer Satisfaction

Clustering, sometimes called cluster analysis, is an important aspect of data mining [12]. Data points in one cluster are more tightly linked to those in other clusters than data points in other clusters. The data points are grouped together using raw data properties to find correspondences [3], but the main purpose is to calculate the average. This is an iterative and repeating operation that entails looking for comparable qualities and patterns in vast amounts of raw data [12]. Data points are distributed to accomplish the intended outcomes after random data is searched for relevant information.

III. Clustering for Segmentation Purposes

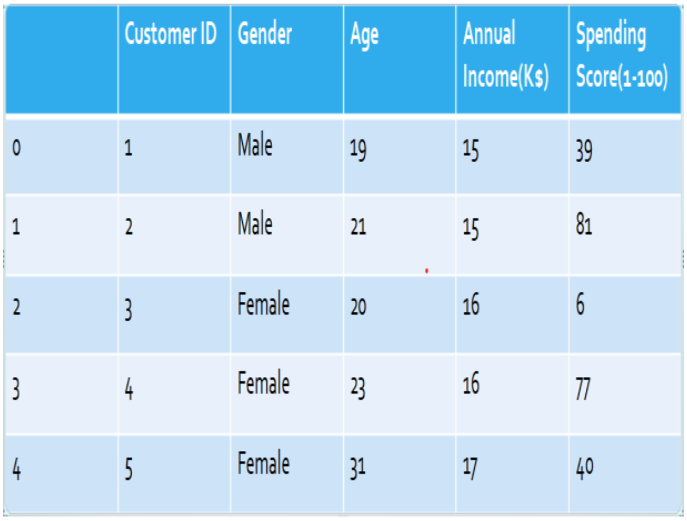
Clustering algorithms identify internally homogeneous and outwardly varied groups.Customers differ in their attitudes, desires, wants, and features, and clustering tactics are used to classify distinct consumer groups and segment Target marketing may be done more efficiently by grouping consumers with similar profiles into clusters. K-Means and Agglomerative Hierarchical Clustering [13], two of the most common hierarchical and non-hierarchical clustering methods used in consumer segmentation, utilise K-Means as part of their clustering process. [14] employed K-Means for customer segmentation on their dataset. Despite the fact that many people find the hierarchical clustering technique undesirable, [15][16] employed that how to apply clustering algorithms to supermarket transaction data in their research for intelligent consumer segmentation. K-means and Hierarchical Clustering are two approaches that can be used in consumer segmentation and are useful for clustering data. As a result, they will be the focus of our attention.

A. K Means Clustering

K-Means is very common clustering algorithm because it is simple and successful. M points in N dimensions are divided into K a priori specified groupings using the K Means algorithm (say, k centroids). These centroids should be properly positioned in order to achieve the best results, which may change if the centroids' positions are altered. As a result, they should be as evenly distributed as possible.After then, each data point is taken and associated with the nearest centroid until there are no more data points to be acquired. The goal of K-Means clustering is to reduce a specific objective function, which in this case is the squared-error. It's a measure of how far the data points are from their respective cluster centers. This algorithm's process never stops, but the relevance or best configuration cannot. The algorithm is also affected by how the initial random cluster centers are chosen. That's why it runs numerous times to lessen the effect, but even though it's iterative, it appears to work effectively for a large number of data points.

Figure 4: Using the Elbow approach for K-Means Clustering, find the optimal number of clusters. Figure 5: Clusters produced as a result of using K-Means Clustering to analyze the dataset. Figure 5 depicts a scatter plot of the clusters, with Annual Income displayed against the X-axis and Spending Score placed against the Y-axis. The data points under each cluster are coloured differently, and the centroids are also highlighted, as shown above. K-Means clustering is also applied in all of the major domains; for example, [17] utilized it to detect clusters in the retail business [16].

Head of our Dataset



	Customer ID	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

IV. Environment and Tools Required

Seaborn: It is a data visualization built on top of matplotlib and closely integrated with pandas data structures.

Application: Predominantly used for statistical graphics, Dataset oriented API, automatic estimation and plotting of linear regression plots, high-level abstractions for multi-plot grids.

Scikit-learn: It's perhaps the most helpful machine learning library in Python. Classification, regression, clustering, and dimensionality reduction are just a few of the useful methods in this toolkit for statistical modelling. Applications: Financial cybersecurity analytics product development,barcode scanner development, neuroimaging, medical modeling.

Pandas: It is a python library which uses data structures and operations for data manipulation and analysis.

Applications: Used in Big data, To access Big data python allows Pandas to connect with Hadoop and Spark.

Numpy: Numpy is a multi-purpose array processing library. It includes a high-performance multidimensional array object as well as utilities for manipulating them.

Application: It is used for complex mathematical operations, maintain minimal memory, broadcasting functions, Overcomes slower execution.

V. Visualizing the Dataset

Graph bar is used to illustrate the number of clients based on their spending scores in Fig:3. The vast majority of clients have a spending score of 41–60.

The bar graph is created to show the number of consumers based on their annual income in Fig:2. The bulk of clients earn between $60,000 and $90,000 each year.

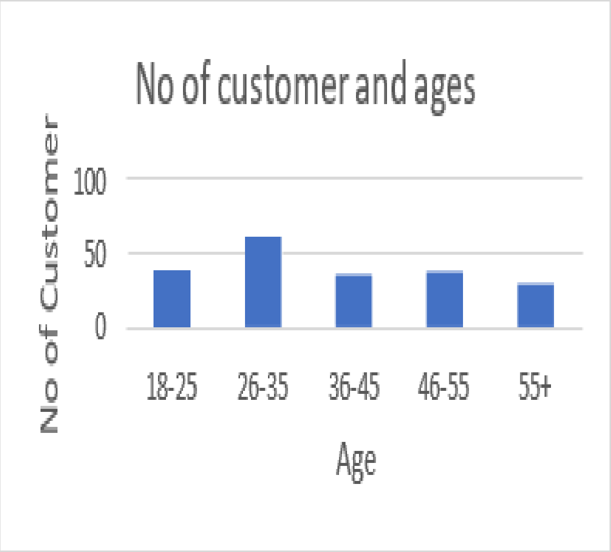


Figure 1.Graph between no of customers v/s their ages

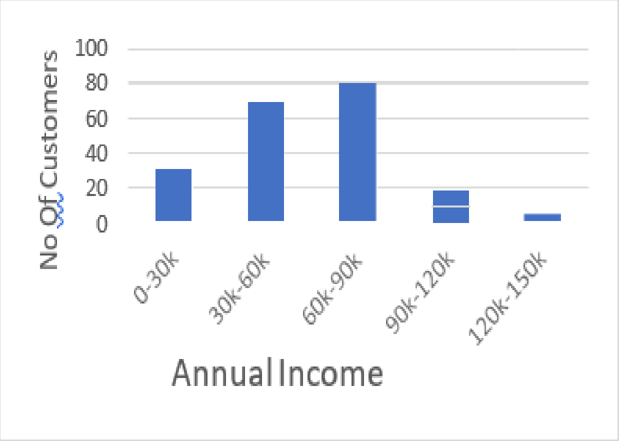


Figure 2. Graph between No. of clusters v/s Annual Income

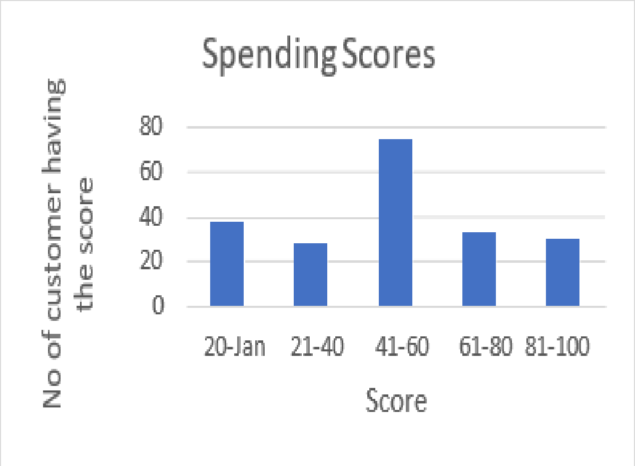


Figure 3. Graph between no of customers v/s spending score

Next, calculation of the optimal number of clusters by plotting Within Cluster Sum of Squares (WCSS) versus the number of clusters (K Value). WCSS calculates the sum of observations' distances from their cluster centroids using the formula.

$$WCSS = \sum (X_i - Y_i)^2$$

VI. The Elbow Method

Calculate the Cluster Sum of Squared Errors (WSS) for various k values and choose the k at which the WSS starts to drop. In the WSS-versus-k diagram, this is depicted as an elbow.The steps can be summarized as: Calculate K-Means clusters for various K values ranging from 1 to 10 clusters.

Calculate the total sum of squares within each cluster for each value of K. (WCSS). Draw the WCSS vs. K number of clusters curve. The number of clusters is often determined by the location of a bend in the plot.

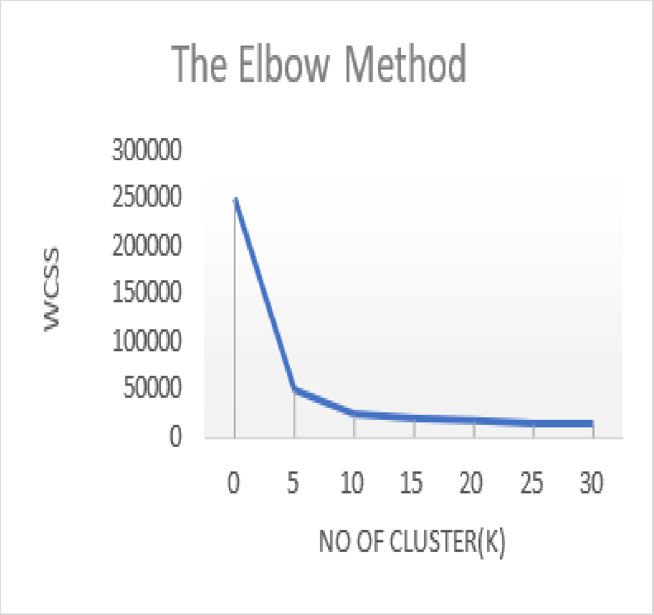
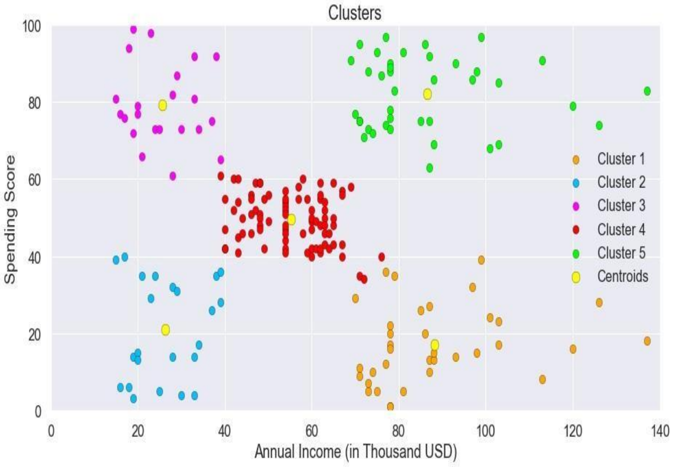


Figure 4 Graph between WCSS v/s No of clusters

Using the above elbow approach, the ideal value of K is discovered to be 5.

Finally, we created a plot to visualize the clients' spending score in relation to their annual income. As seen below, the data points are separated into five classes, each of which is represented by a different color.



CLUSTER BLUE – BALANCED CUSTOMERS

They make less money and spend less money. People with a low annual income and a low expenditure score are the sensible ones who know how to spend and save money. People from this cluster will be of little interest to the shops/mall.

CLUSTER ORANGE – PENNY CUSTOMERS

Earning a lot of money while spending little. These are likely to be the mall's primary objectives, as they have the capacity to spend money. As a result, mall officials will attempt to provide additional amenities in order to attract these customers and suit their needs.

CLUSTER RED – NORMAL CUSTOMERS

Customers with average earnings and spending habits will not be the shops' or mall's primary focus, but they will be considered, and other data analysis techniques may be utilized to boost their spending score.

CLUSTER PURPLE - SPENDERS

Customers in this category earn less but spend more. Because their annual income is low but their expenditure is considerable, they can be considered a prospective target consumer. The shops/malls may not be able to properly target these folks, but they will not be lost.

CLUSTER GREEN – TARGET CUSTOMERS

Target Customers that earn a lot of money and spend a lot of money. A target consumer with a high annual income and a high spending score. These individuals may be regular mall patrons who have been persuaded by the mall's amenities.

VII. RESULT

We looked at the five customer categories based on their Annual Income and Spending Score, which are said to be the best factors/attributes for determining consumer segments in a mall. Pinch Penny Customers, Balanced Customers, Target Customers, Spender, and the Ordinary Customer are among them.

We can enroll Target Customers in an alerting system that sends them SMS and emails on a daily basis about the deals and discounts available at the Mall; for the rest, we may send them blast SMSs once a week or once a month to inform them about our products. Similarly, we now have a better understanding of client behavior based on their Annual Income and Spending Score. Customers can be targeted with a variety of marketing methods based on Cluster Analysis. Clients with a high income and spending score are our target customers, and we always want to keep them since they provide the highest profit margin for our company. Customers with a high income and low spending score may be drawn to the Mall Supermarket by a large choice of products that meet their lifestyle needs.

Less Income Less Spending Score can be given further offers, and providing them offers and discounts on a regular basis will entice them to spend. We can also conduct a cluster analysis to determine what types of products clients are most likely to purchase and adjust our marketing efforts accordingly. More analytics on the same data set were not possible due to a lack of data.

VIII. CONCLUSION

Companies, malls, and supermarkets, as well as small businesses, should do a Market Basket Analysis. Companies will be able to target individual customers as a result of this. A customer segmentation model provides for the efficient distribution of clients into groups.

Cross- and up-selling opportunities are maximized through the use of marketing tools. When businesses send personalised communications to a group of customers as part of a marketing mix tailored to their needs, it's easier for them to make special offers to attract those customers to buy more products. By increasing customer service, consumer segmentation can also aid with customer loyalty and retention. Because of their customised nature, marketing materials that use customer segmentation are more valued and appreciated by the customer who gets them than impersonal brand messages that neglect purchase history or any type of customer interaction.

Finally, consumer segmentation is important because it allows businesses to stay ahead of the competition in specific market segments by identifying new products that are available or that potential customers might be interested in, as well as improving existing products to meet customer expectations.

References

[1] Xiaojun Chen, Yixiang Fang, Min Yang, Feiping Nie, Zhou Zhao and Joshua Zhexue Huang,“PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data” in IEEE Transactions on Knowledge and Data Engineering, IEEE, 2017, pp. 1-3.

[2] Jisun An, Haewoon Kwak, Soon- gyo Jung, Joni Salminen, Bernard J.Jansen, “Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data”, Springer, 2018, pp. 1-4

[3] A. Sharma, H. Jit and S. Sidhu, "Customer Relationship Management Using Clustering And Classification Technique", vol. 20, no. 5, pp. 67-72, 2018.

[4] M. Hosseini and M. Shabani, "New approach to customer segmentation based on changes in customer value", November 2018.

[5] W. Qadadeh and S. Abdallah, "Customers Segmentation in the Insurance Company (TIC) Dataset", Procedia Computer Science, vol. 144, pp. 277-290, 2018

[6] J. Qian and C. Gao, “The application of data mining in CRM”, in 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), Deng Leng, 2011, pp. 5202-5206.

[7] A. Ansari and A. Riasi, “Taxonomy of marketing strategies using bank customers clustering”, International Journal of Business and Management, vol. 11, no. 7, pp. 106-119, 2016.

[8] D. Liu and Y. Shih, “Integrating AHP and data mining for product recommendation based on customer lifetime value”, Information & Management, vol. 42, no. 3, pp. 387-400, 2005.

[9] A. Riasi, “Barriers to international supply chain management in Iranian flower industry”, Management Science Letters, vol. 5, no. 4, pp. 363368, 2015

[10] A. Riasi, “Competitive advantages of shadow banking industry: An analysis using porter diamond model”, Business Management and Strategy, vol. 6, no. 2, pp. 15-27, 2015.

[11] Q. Zhao and P. Franti, “Centroid Ratio for a Pairwise Random Swap Clustering Algorithm”, IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 5, pp. 1090-1101, 2014.

[12] T. Kanungo, et al., “An efficient k-means clustering algorithm: analysis and implementation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, 2002. [13] Y. Chen, et al., “Identifying patients in target customer segments using a two-stage clustering classification approach: A hospital based assessment”, Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221, 2012.

[13] G. Lefait and T. Kechadi, “Customer segmentation architecture based on clustering techniques”, in Fourth International Conference on Digital Society, Sint Maarten, 2010, pp. 243-248.

[14] M. Namvar, M. Gholamian and S. KhakAbi, “A two-phase clustering method for intelligent customer segmentation”, in International Conference on Intelligent Systems, Modeling and Simulation, Liverpool, 2010, pp. 215-219.

[15] D. Gaur and S. Gaur, “Comprehensive analysis of data clustering algorithms”, in Future Information Communication Technology and Applications. Dordrecht: Springer Netherlands, 2013, pp. 753 -762.

[16] S. Salvador and P. Chan, “Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms”, in 16th IEEE Conference.