Customer Segmentation and Behavior Analysis in a Retail Mall: A Study Using K-Means Clustering

K. Dileep

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract**—** *The retail industry is increasingly reliant on data-driven approaches to enhance customer experience and optimize business operations. This research paper explores the application of the K-Means clustering algorithm to segment customers based on their demographic and spending behavior. We analyze a dataset of mall customers, which includes gender, age, annual income, and spending score. This data, being unlabeled, presents an unsupervised learning challenge. Our study aims to uncover meaningful customer segments and provide insights into customer behavior.*

I. INTRODUCTION

Clustering is the most widely recognized approach to partitioning or assembling a given game plan of models into disjoint gatherings. This is finished so much that models in a comparative gathering are vague and plans having a spot with two unmistakable packs are remarkable. Clustering has been a for the most part focused on issue in a collection of purpose spaces [1]. Bundle assessment relies upon various kinds of articles' incongruities and uses distance capacities' rules to make model portrayal [2][7]. Whether or not the gathering is really have an impact is rest with the course sort of model person vectors. If the responsibilities of touches of vectors is clustered and test spots in a comparable get-together are engaged and test bits in different social events are far away, it will be easy to use distance capacities to portray the bits, which will very far make estimations in a comparative social event be tantamount and experiences in different social affair be exceptional. The eigenvector of the whole model plan social event can be treated as spots which course in feature space. The distance capacity between spots could go probably as the extent of closeness of models. According to the area of spots' distance, the activity can be used to orchestrate plans.

II. CLUSTERING

Pack assessment could be apportioned into different evened out gathering and nonhierarchical grouping methodologies. Occasions of moderate methodologies are single linkage, complete linkage, typical linkage, center, and Ward. Nonhierarchical strategies consolidate kmeans, flexible kmeans, kmedoids, and soft gathering. To sort out which computation is extraordinary is a part of the kind of data open and the particular inspiration driving examination. In more obvious way, the strength of gatherings can be analyzed in reenactment studies [3][4]. The issue of picking the "best" computation/limit setting is an irksome one. A good grouping computation ideally should convey packs with specific nonoverlapping limits, but an ideal separation can not customarily be achieved eventually.

III. K-MEANS COMPUTATION

K-Means computation considering isolating is a kind of bundle estimation. This computation which is independent is regularly used in data mining and model affirmation [4]. Focusing on restricting bundle execution record, square-mix-up and screw up measure are underpinnings of this computation [5][6]. To search for the optimalizing result, this estimation endeavors to find K divisions to satisfy a particular norm. From the outset, pick a couple of bits to address the fundamental gathering focal points(usually, we pick the chief K model touches of pay to address the hidden bundle point of combination); moreover, collect the overabundance model spots to their focal spots according to the proportion of least distance, then, we will get the basic portrayal, and if the request if silly, we will change it(calculate each gathering focal concentrations later on), underscore repetitively till we get a reasonable portrayal.

K-means clusreting adopts an iterative strategy to play out the grouping task. The functioning strides of this calculation are as per the following-

Stage 1: Pick the number K of groups.

Stage 2: Select indiscriminately K focuses, the centroids (not fundamentally from our dataset).

Stage 3: Appoint every information highlight the nearest centroid in light of euclidian or manhattan distance. That structures K bunches.

Stage 4: Figure and spot the new centroid of each group.

Stage 5: Reassign every information highlight the new nearest centroid. On the off chance that any reassignment occurred, go to stage 4.

It quits making or streamlining groups assuming either the centroids have balanced out importance no new reassignment of information focuses happens or the calculation arrives at the characterized number of cycles.

As the K-implies calculation works by taking the distance between the centroid and important pieces of information, we can instinctively comprehend that the larger number of groups will decrease the distances among the places. For that, we plot the quantity of bunches k and the Inside Group Amount of Squares(WCSS). The plot would seem to be an elbow, likewise with a rising number of bunches after a specific point, the WCSS begins to balance out and will in general go lined up with the even hub. So we will take that point after which the plot will in general become comparative.

IV. EXPERIMENTAL RESULTS

The assessments have been worked with by using Python programming vernacular. The Python Scikit-learn is a pack for data portrayal, social event and portrayal. We have considered the Mall_Customer from Kaggle dataset [8], where the subtleties of all clients in a mail are recorded. In this dataset consists of 200 instances and 5 attributes. The highlights are the class of the customers (Male or Female), their age, yearly pay and spending score on a size of 1 to 100. The information are unlabeled that is there is no result segment like in a relapse or characterization dataset. Thus, the issue falls in the unaided class. There are many distance measurements that can be utilized with grouping calculations. It relies upon the kind of information you are utilizing. The default distance metric for sklearn bunching is the Euclidian distance. The worth of k is exceptionally vital for ideal results from the calculation. There are a few procedures to pick the ideal incentive for k, we have chosen Elbow Strategy. In this experiment, we will carry out the elbow strategy to track down the ideal incentive for k.

Presently, our errand is to track down various gatherings among the clients as indicated by their yearly pay and spending score. We should apply K-implies bunching and see what occurs. The experimental results are shown in the figure-1 and figure-2.

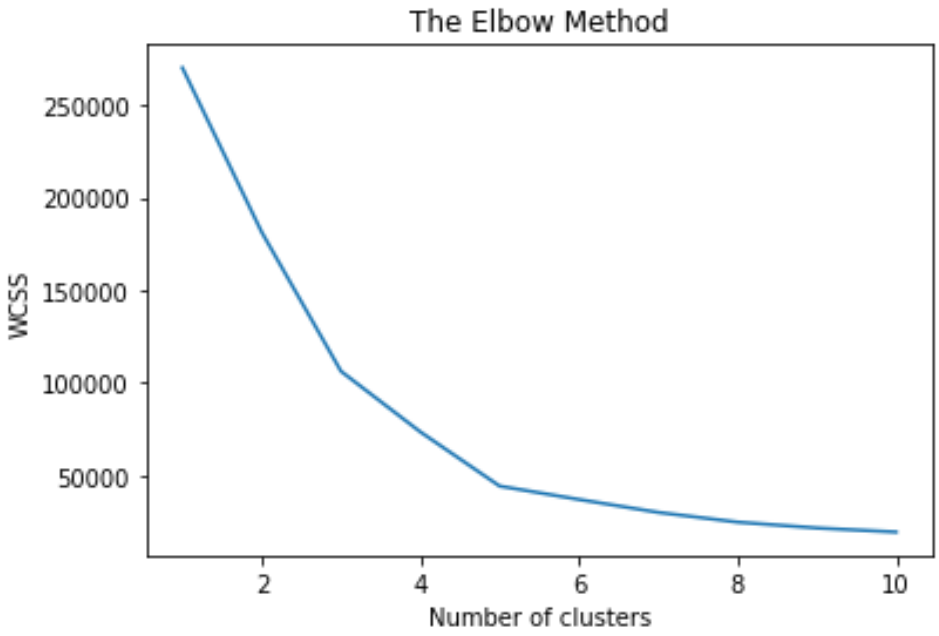

FIGURE-1: Experimental Results for choosing Optimal K value

From the above chart figure-1, we can see that the ideal incentive for k is 5(the place where the grouping going lined up with the rising number of bunches).

FIGURE-2: Experimental results K Clusters

4.1 Results

Our analysis using the K-Means clustering algorithm identified distinct customer segments within the mall dataset. We determined the optimal number of clusters to be four, based on the Elbow method and Silhouette score.

The four customer segments are as follows:

High Income, High Spending (Cluster 1): This segment comprises customers with high annual incomes and high spending scores. These individuals are likely to be high-value customers and may represent a target group for premium products or personalized marketing strategies.

Low Income, Low Spending (Cluster 2): Customers in this segment have low annual incomes and low spending scores. They may be price-sensitive shoppers who prioritize budget-friendly options. Retailers can tailor promotions and discounts to attract and retain these customers.

High Income, Low Spending (Cluster 3): This segment includes customers with high annual incomes but relatively low spending scores. Understanding the factors influencing their low spending can help retailers devise strategies to encourage these customers to spend more.

Moderate Income, Moderate Spending (Cluster 4): Customers in this segment exhibit moderate annual incomes and spending scores. This group represents a balance between income and spending, and retailers can target them with promotions and products that align with their preferences.

4.2 Discussion

The results of our customer segmentation provide valuable insights for mall management and retailers. By categorizing customers into distinct segments, businesses can tailor their marketing efforts, product offerings, and customer service to better meet the diverse needs of their clientele.

For the "High Income, High Spending" segment, retailers can focus on offering premium products, exclusive experiences, and loyalty rewards to further enhance their spending. Personalized recommendations and targeted marketing campaigns can strengthen customer engagement within this group.

The "Low Income, Low Spending" segment presents an opportunity to attract budget-conscious shoppers. Retailers can offer cost-effective products, discounts, and loyalty programs to capture their attention and boost their spending.

Understanding the "High Income, Low Spending" segment is crucial for retailers. By identifying the reasons behind their restrained spending, businesses can work to improve the in-store experience, enhance product offerings, or develop incentives to encourage these customers to spend more.

The "Moderate Income, Moderate Spending" segment represents a diverse group of customers. Retailers should aim to maintain the satisfaction of this segment by consistently meeting their needs and preferences. Targeted promotions and product recommendations can help maintain their loyalty.

V. CONCLUSION

The application of K-Means clustering to the mall customer dataset has yielded actionable insights into customer behavior and preferences. Retailers can use this information to refine their strategies, improve customer satisfaction, and optimize revenue generation. Future research could involve exploring additional features and employing more advanced clustering techniques for even finer-grained customer segmentation.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G Ravi Kumar, K Tirupathaiah and B Krishna Reddy, "Client Churn prediction of banking and fund industry utilizing machine learning techniques",IJCSE,Volume-7,Issue- 6,PP:842-846, 2019

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J.Han and M.Kamber, Data Mining concepts and Techniques, the Morgan Kaufmann series in Data Management Systems, 2nded.San Mateo, CA; Morgan Kaufmann, 2006.

[5] N. Michael, "Artificial Intelligence - A Guide to Intelligent Systems",2nd edition, Addison Wesley, 2005.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[7] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, "A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[8] https://www.kaggle.com/code/tochelle1/mail-customer-segmentation