COMPARTIVE ANALYSIS OF MACHINE LEARNING MODEL FOR CUSTOMER SEGMENTATION

Abstract :- *In the present competitive era, entrepreneurs are struggling to increase and retain their customer base. Behavioral-based customer segmentation assists in the identification of potential customers, their buying habits, and shared interests. This helps build an efficient strategy to increase customer base and product sales. In this paper, we compare the efficacy of four machine learning algorithms, namely, KMeans, DBSCAN, Agglomerative Clustering, PCA with KMeans, in performing behavioral-based customer segmentation. The machine learning algorithms divide customers into an optimal number of customer segments based on demographics i.e. annual income, age, spending score that will assist in exploring novel ways to increase marketing persona. The algorithms group customers with common interests by extracting and analyzing patterns in the available customer data. Our comparison shows that Agglomerative Clustering has the highest silhouette score of 0.6865 in performing behavioral-based customer segmentation in comparison to other models. In this work, we also use machine learning algorithms to draw conclusions from the analysis.*

KEYWORDS
Patterns, Clustering, Data Analysis, Segmentation

1. INTRODUCTION

Due to the increased usage of the internet for online marketing in recent years, customer data has exploded. Due to increased competition among peer firms, new objectives like maximising sales, profits, minimising costs, customer satisfaction, market satisfaction etc. are being set. However, because they do not learn from their clients, all firms fail. We can learn and comprehend the market and our consumers more effectively if we make the most use of the massive data available. Customer segmentation can help you solve this problem. In this method, customers are divided into 'n' number of groups based upon the above-mentioned traits, clustering means grouping the information based on similarities in the dataset. Customers belonging to a particular group have some common traits. Customers are grouped in a way that a customer belonging to a particular group shares a common interest with other customers of the same group [1]**.** This enables us to communicate effectively with various groups and enhances the likelihood of a customer purchasing the items. For instance, the business may use social media to market its brand among teens.

By analysing the emotion behind the reviews, this project leverages real-time data to create client categories. The corporation may now design its advertisement campaigns, tactics, and much more using the client information from the segments. It will benefit the company indirectly. This aids businesses in building stronger consumer relationships and improving their overall performance.

This paper presents a comparison of the performance of different machine learning algorithms used for segmentation of customers using the Sihoutte and Davis-Boulton scores. In the further sections we try to define customer segmentation, its use and analysing four different algorithms used.

2. Problem statement

Increase and retain of customers or clients has become a concern for the companies to gain a competitive edge in the market and gain profits. Customer segmentation, The practise of discovering a common attribute among consumers and categorizing them has become a part of marketing strategy, but the problem arises in selection of an appropriate and optimal algorithm that suits our customer data which gives the best results. Different data sets may have different efficient algorithm that gives best efficiency. This report gives insights on different algorithms to increase the segmentation efficiency and their comparison to finally choose the best efficient algorithm that suits our mall customer data set.

3. Background

Market segmentation is the actual process of identifying segments of the market and the process of dividing a broad customer base into sub-groups of consumers consisting of existing and prospective customers [2].

Every customer is different and every customer journey is different so a single approach often isn't going to work for all. This is where customer segmentation becomes a valuable process.

However, if best current customer segmentation is done correctly, there are various commercial benefits. A best current customer segmentation exercise, for example, can have a measurable impact on your operating outcomes by:

1. Improving the overall quality of your goods.
2. Keeping your marketing message focused.
3. Enabling your sales team to explore more high-percentage offers.
4. Increasing the quality of revenue.

KMeans model

K-Means clustering is an unsupervised machine learning method for categorising data into a set number of clusters. The letter "K" denotes the number of pre-set clusters that can only be generated. This centroid-based methodology pairs each cluster with a centroid. The underlying objective is to reduce the distance between each data point and its cluster centroid. The model divides unlabelled raw data into clusters and repeats the procedure until the best clusters are found.

DBSCAN MODEL

Density-based Spatial Clustering of Applications with Noise is a well-known unsupervised machine learning clustering approach. Dbscan's approach is based on the density threshold notion of cluster, which can be deduced from the name.

Two parameters establish the density threshold: eps (): the radius of the neighbourhood/circle, and minPts: the minimal number of neighbors/data points inside the radius of the neighbourhood.

K-MEANS USING PCA

K means is the simplest and the most popular unsupervised machine learning algorithm which tries to partition dataset iteratively into non overlapping subgroups.

AGGLOMERATIVE CLUSTERING

Agglomerative model is a hierarchical unsupervised machine learning algorithm that uses bottom-up approach for clustering.

5. LITERATURE REVIEW

Customer segmentation is an important management tool in CRM(Customer Relationship Management) literature. In

practice customer segmentation maximizes the customer satisfaction and hence improves company's profit significantly. It is also an active research area especially in industrial management literature [3] [4].

Segmentation started to become more widely accepted and used in the middle of the 20th century. Smith W.R. described in details segmentation and strategies that can be adapted by using segmentation. It is emphasized that segmentation is essentially a merchandising strategy, merchandising being used here in its technical sense as representing the adjustment of market offerings to consumer or user requirements [5].

Demographic segmentation is a process of splitting customer groups based on traits such as age, gender, ethnicity, income, level of education, religion, and profession [6] [7]. Our lifestyles have been continuously changing and so the data gets updated too often in attributes such as age, income etc. It is also a subjective topic as it doesn't provide any insights on needs and values of customers. This type of segmentation is not appropriate for fields like music, movie recommendation, online shopping etc.

Psychographic segmentation allows incredibly effective marketing by grouping customers at the more personal level by defining their hobbies, personality traits, values, life goals, lifestyles, and beliefs [3] [8]. It uncovers hidden motivation and attitudes of customers however customer actions, loyalty and other factors are not taken into consideration. Collection of this data is difficult and also requires complex setup process to get accurate data and sometimes may rely on assumptions.

Geographic segmentation allows many different kinds of considerations when advertising to consumers by grouping them based on their geographic location such as their country, region, city, and even postal code [3] [9]. However buying behaviour, needs or wants of customers cannot be interpreted as the needs of people living in the same region need not be same. Changing population and weather also makes this segmentation less effective.

Behavioural segmentation is perhaps the most useful of all e-commerce businesses as most of this data such as customers' spending habits, browsing habits, purchasing habits, loyalty to a brand, interactions with the branch, and previous product ratings can be gathered via the website itself [3] [10]. It enables us to understand customer needs and behaviors which help us to prioritize group of customers that have a common trait, to build brand loyalty and also to avoid wasting time on low spending customers .

6. METHODOLOGY

1. Firstly we import all the required libraries or modules (pandas, numpy, matplotlib, seaborn, sklearn, scipy).
2. Visualization of the mall data.

Dataset : Malls Customer data

Column	Non-Null	Count	Dtype
CustomerID	200	non-null	int64
Gender	200	non-null	object
Age	200	non-null	int64
Annual Income	200	non-null	int64
Spending Score	200	non-null	int64

Fig 1.0 dataset info

	CustomerID	Age	Annual Income	Spending Score
CustomerID	1.000000	-0.02676	0.977548	0.013835
Age	-0.026763	1.000000	-0.01239	-0.32722
AnnualIncome	0.977548	-0.01239	1.000000	0.009903
SpendingScore	0.013835	-0.32722	0.009903	1.000000

Fig 1.1 correlation in dataset

Pairwise correlation of all columns in the data frame.

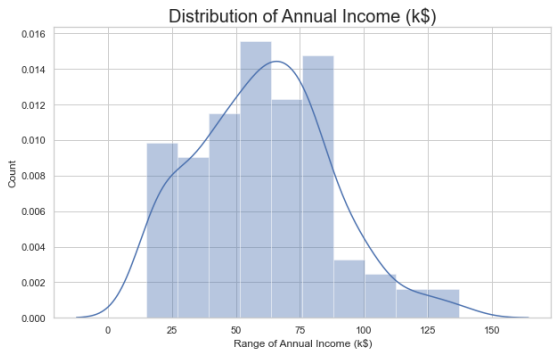

Fig 1.1.0 distribution of annual income

Most of the annual income falls between 50K to 85K.

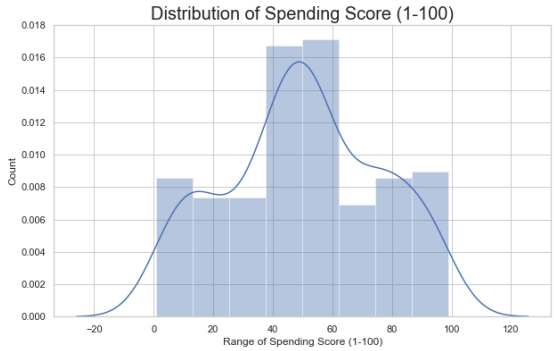

Fig 1.1.1 distribution of spending score

Spending Score is normally distributed.

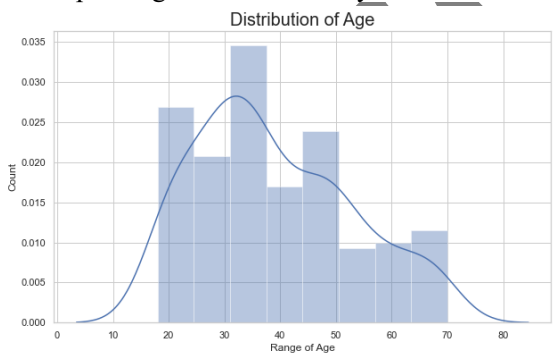

Fig 1.1.2 distribution of age

Most of the customers are within 25 to 40 years old with an average age of 38.5 and median age of 36 years.

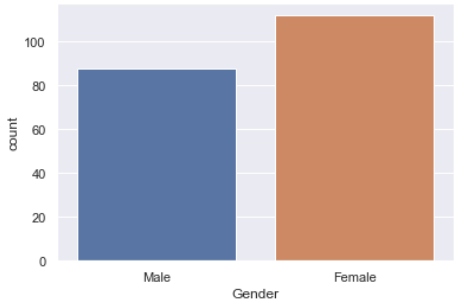

Fig 1.1.3 gender distribution

Both the bar chart and counts show that there are more females than male customers. It is clear from the exploratory data analysis that all of the factors have some form of link with the spending score. To create the clustering models, we will use all of the variables.

I. KMeans model

Algorithm

Step 1: Determine the number of clusters k.

Step 2: Pick k random points from the data to serve as centroids.

Step 3: Assign each point to the nearest cluster centroid.

Step 4: Calculate the centroids of freshly generated clusters again.

Step 5: Repetition of steps 3 and 4.

Algorithm 1: KMeans Model

input : Data Points D, Number of Clusters;
output : Data Points with Clusters ;
Result: Kpre-defined distinct non-overlapping subgroups
initialization;
Elbow method for optimal number of clusters;
for *k in range* **do**
    plot() //curve bends = optimal point;
    randomly intialize centroids ***c=c1,c2....ck***;
end
Provide the number of clusters to be assigned (k);
while *until centroid position does not change* **do**
    assignment step;
    for *each data point di* **do**
        find closest center $ck \in c$ ***to data point di***;
        assign $di -> ck$;
    end
    update the Centroid value;
end

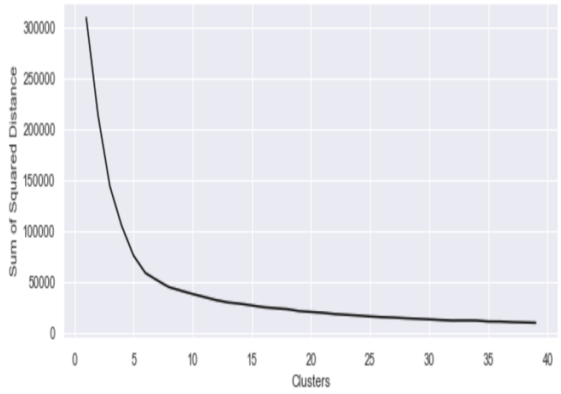

Fig 2.0 Elbow method plot

The elbow graph is shown in Figure 1.0, with the x-axis representing the number of clusters measured at the elbow joint point. Making clusters is very important at this stage since the value of WCSS abruptly stops decreasing. The decline in the graph is minor after 5, thus we pick 5 to be the number of clusters.

II. DBSCAN MODEL

Algorithm

1. First, an arbitrary point is selected from the dataset (until all points have been visited).
2. If there are at least 'minPoint' points within a radius of 'ε' to the point, we consider all of these points to be part of the same cluster.

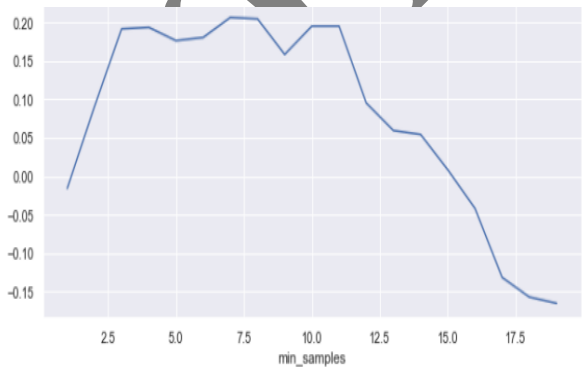

Fig 3.0 Min_samples curve for dbscan model

3. The clusters are then extended by repeating the neighbourhood computation for each surrounding point recursively.
4. Sort the data points into three categories: core points, boundary points, and noise points.
5. Discard the noise points.
6. Assign cluster to a core point.
7. Colour all the density connected points of a core point.
8. Colour boundary points according to the nearest core point.\

Core points: points that have sufficient number of neighbours with in the radius.
Boundary points: points that are within radius of a core point, but don't have sufficient neighbours.
Noise points : points other than core and boundary.

Algorithm 1: DBSCAN Model

input : Data Points D, radius threshold ϵ, minpts;
minpts-min number of points required in a cluster;
output : A set of Clusters ;
$\epsilon \rightarrow$ avg dist btw ***di*** and its nearest neighbour;
x-axis $\rightarrow$ avg distances , y-axis $\rightarrow$ data points;
(elbow of K-distance graph) ;
minpts $\rightarrow$ no. of dimensionality of dataset;
DBSCAN(*D* , ϵ , minpts){
for *each unvisted points* ***di*** $\in D$ **do**
  consider ***di*** *as visited;;*
  $X \leftarrow$ *GetNeighbours(d',*ϵ *);*
  if $|X| <$ *minpts* **then**
    consider ***di*** *as* ***noise*** *point;*
  end
  else
    $P \leftarrow \{di\}$
  end
  for *each data point* ***di*** $\in X$ **do**
    $X \leftarrow X \setminus d'$;
    if *d' is not visited* **then**
      mark x' as visited;
      $X' \leftarrow$ *GetNeighbours(d',*ϵ *);*
      if $|X'| >=$ *minpts* **then**
        $X \leftarrow X \cup X'$
      end
    end
    if *d' is not in any cluster* **then**
      $P \leftarrow P \cup \{d'\}$
    end
  end
end
}

Agglomerative Clustering (using PCA)

Agglomerative model is a hierarchical unsupervised machine learning algorithm that uses bottom-up approach for clustering.

Steps followed in our model:

1. Scaling the Data
   Scaling is used to make our data closer and reduce variance by converting them to values in the range of 0-1. We used Minmax scaler for our data where,
   $$Xsc = X - Xmin \,/\, Xmax - Xmin$$
2. Dimensionality reduction using PCA
   Principal Component Analysis(PCA) is an unsupervised learning algorithm that is used for the dimensionality reduction. It converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation as finding features that are correlated is time consuming.

   Here we are fitting the data points in 2 principle components (PCA1,PCA2) that are uncorrelated.

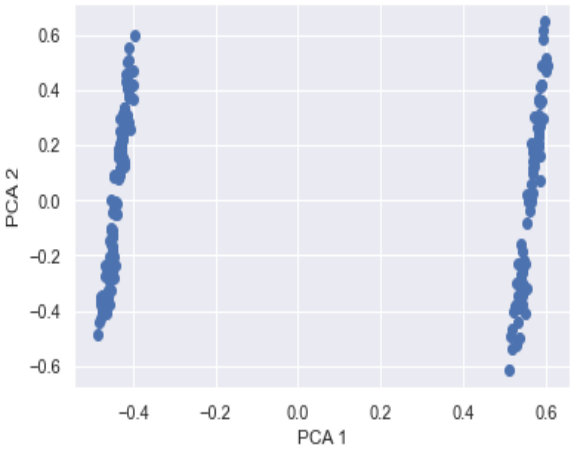

Fig 4.0 fitting data points in 2 Principle components

3. Make one cluster for each data point.
   We will have 200 clusters initially, one cluster for each data point.
4. Combing clusters that contain closest pair of elements
   Greedily we combine two clusters with closed distance as one using single linkage method
5. Repeat step 4 until there is only one cluster that contains all the data points
6. Visualize the grouping by creating a dendrogram and find the optimal number of clusters.

Algorithm 1: Agglomerative Clustering with PCA Model

input : Data Points D;
output : Data Points with Clusters ;
initialization;
C , C' ← n, Di ← {xi}, $i = 1, 2, ...n$;
Scaling the Data ;
$Xsc \rightarrow X - Xmin \,/\, Xmax - Xmin$;
Convert dataset from multidimensions to 2 dimensions;
pca → PCA(2) ;
data → pca.fit_transform(D);
while *C' equals C* **do**
  C'← C'−1;
  Find nearest clusters, example ***Di*** and ***Dj***;
  Merge ***Di*** and ***Dj***;
  return C clusters;
  for *each data point di* **do**
    find closest center ***ck*** ∈ ***c to data point di***;
    assign ***di*** → ***ck***;
  end
  update the Centroid value;
end
dendograms($PCA\_model, method \rightarrow$ ward);
scatterplot(clustering model);
Visualize the optimal number of clusters

K-MEANS USING PCA

Steps followed in our model:

1. Employ PCA for projecting into lower dimensional space.
   Initially our data set contained only few features, we are further trying to reduce it to 2 components. By reducing number of features we are also trying to reduce the noise
2. Determination of number of clusters in k means.

i. We run the algorithm for different number of clusters

ii. Calculate WCSS (Within Cluster Sum of Squares) considering different number of clusters.

iii. Plot WCSS against the number of components and using a heuristic, elbow method find the optimal number of clusters .

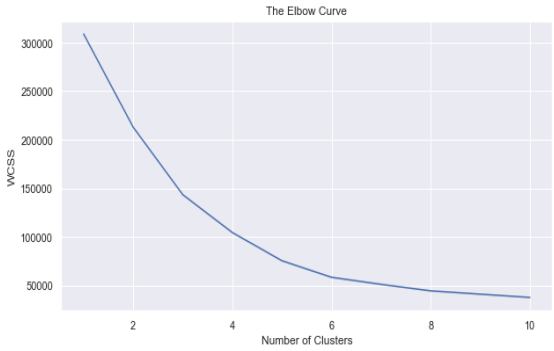

Fig 5.0

Number of possible clusters: 3, 4, 5 and 6 clusters.

Algorithm 1: KMeans with PCA Model

input : Data Points D, Number of Clusters;
output : Data Points with Clusters ;
Result: Kpre-defined distinct non-overlapping subgroups
initialization;
Convert dataset from multidimensions to 2 dimensions;
pca → PCA(2) ;
data → pca.fit_transform(D);
Calucate WCSS;
K→ elbow bend in (plot(WCSS , no. of clusters));
Provide the number of clusters to be assigned (k);
while *until centroid position does not change* **do**
  fit the model with the principal component scores. ;
  for *each data point di* **do**
    find closest center ***ck*** ∈ ***c to data point di***;
    assign ***di*** → ***ck***;
  end
  update the Centroid value;
end
ScatterPlot($PCA\_Scores, fitmodel$);
Visualize the optimal number of clusters

3. Creating the best scoring model employing the regular k means model with modified features.

K means Clustering analysis is done calculating the scores and centres for number of clusters ranging 3 to 7 and 2 principle components. The optimal values are selected based on the analysis.

4. Parameters with highest silhouette score are selected .

5. Visualizing the optimal clusters using scatter plot.

RESULTS AND DISCUSSION

1. Kmeans model

Fig 5.0 age vs spending score

Age is clearly the most important element in determining Spending Score, as seen in the graphs below. Younger folks, regardless of their annual income, tend to spend more.

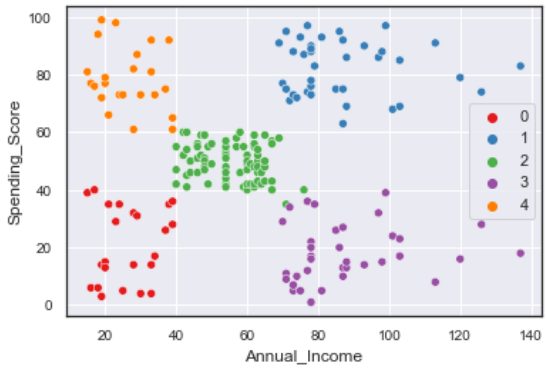

Fig 5.1 spending score vs annual income

The data in fig 5.1 clearly shows that 5 separate clusters have been produced. Customers in the red cluster have the lowest income and lowest spending score, while customers in the blue cluster have the highest income and highest spending score.

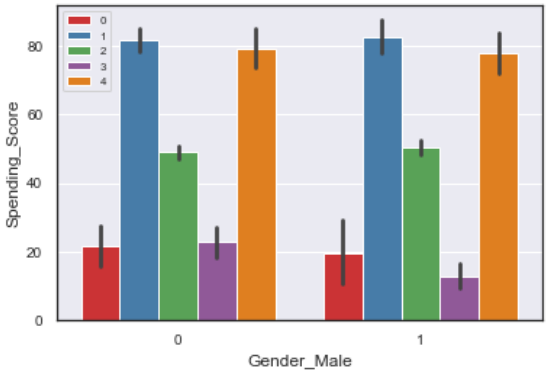

Fig 5.2 spending score vs gender male

From the plots ,it is evident that younger people tend to spend more. Focusing on their interests would make considerable benefit.
Cluster 0 has low spending score with low annual income.
Cluster 1 has high spending score with higher annual income.
Cluster 2 has an average spending score with average annual income.
Cluster 3 has low spending score with annual income just greater than average.
Cluster 4 has high spending score and high income with age groups lesser than that in cluster 1The silhouette coefficient is a statistic that measures how effectively a clustering technique works. It has a value between -1 and 1.
The sihoutte score is **0.45 (approx)**.
Davies Bouldin Score:**0.82 (approx)**.

2. DBSCAN model

Fig 6.0 spending score vs annual income

Fig 6.1 age vs spending score

Dbscan also shows that younger people spend more, irrespective of their annual income.

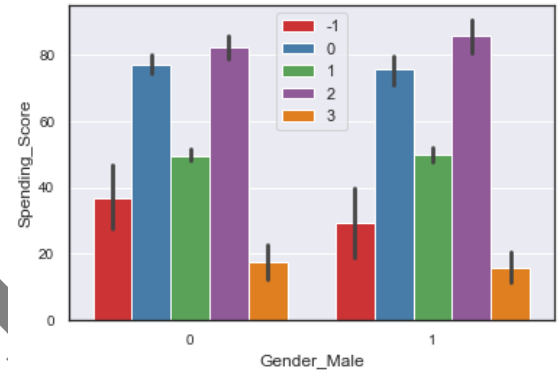

Fig 6.2 spending score vs gender male

The points that are marked -1 represents the noise points which does not belong to any cluster. Cluster 0 represents high spending score with low annual income. Cluster 1 represents average spending score with annual income less than average. Cluster 2 depicts customers with high spending score and annual income greater than average.
Cluster 3 depicts low spending score and average income customers. Similarly, DBSCAN reveals that age is the most important element to consider, with younger people spending more regardless of their annual income.
The sihoutte score is **0.20 (approx)**.
Davies Bouldin Score:**2.23 (approx)**.

3. Agglomerative with PCA model

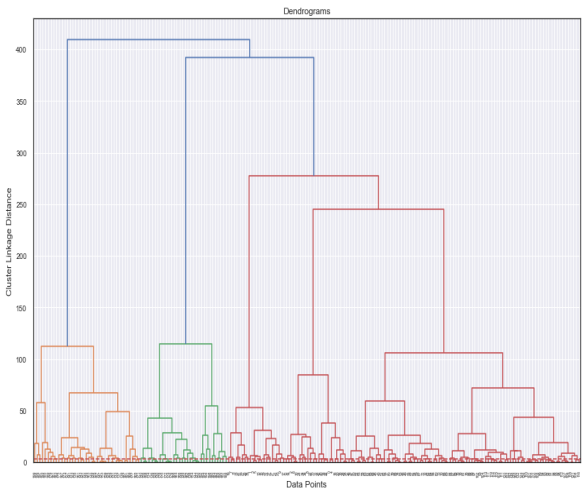

Fig 7.0 dendograms

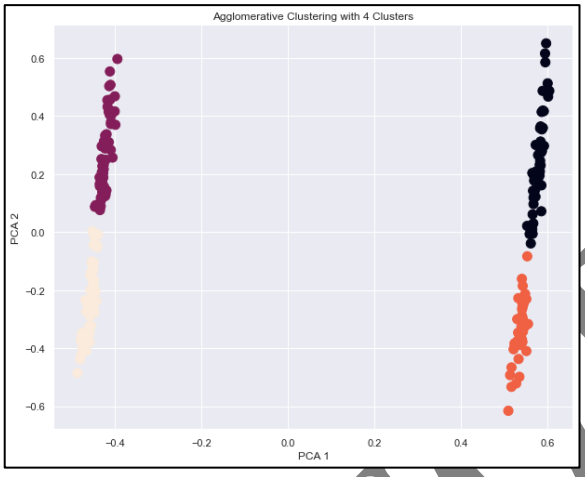

Fig 7.1 clusters with agglomerative

Fitting the updated data into an agglomerative model and constructing a dendrogram, which is a tree-like structure used to readily understand the relationship between characteristics and to retain each step as a memory.
The sihoutte score is **0.68 (approx)**.
Davies Bouldin Score:**0.40 (approx)**.

4. Kmeans with PCA Model

From the table, taking 5 clusters with 2 principles components gives the optimal clusters considering high silhouette score and low Davies Bouldin score.

Table 2.0 PCA with different clusters

In the last page pto

Fig 8.0 clusters with Kmeans and PCA

The sihoutte score is **0.55 (approx)**.
Davies Bouldin Score: **0.58 (approx)**.

Conclusion

Discovering all of the various groups that contribute to a more meaningful customer base allows you to get inside customer's heads and give them exactly what they want, increasing participation and profits. According to the findings, younger individuals aged 20 to 40 are more likely to use the product(s)/service(s) than older persons. The company should target ads to this demographic since they will have a higher turnover and conversion rate. Female clients are somewhat more expensive than male customers, and they spend more even if their yearly income is less than $50,000. This might be a needy group in severe need of the product(s). This is an area where the company should collect additional data for future study.
Below is the table in which all the models score is displayed.

S.NO	CLUSTERING MODEL	Silhouette score	Davis-Boulton score
1.	KMeans clustering model	0.44406692	0.82225964
2.	DBSCAN clustering model	0.20473300	2.23593606
3.	Agglomerative Clustering (Dendrograms & PCA) model	0.68651389	0.40911039
4.	PCA with KMeans clustering model	0.552626	0.584301

Table 1.2

These are the algorithms that were utilised, and the scores were calculated; the best one is Agglomerative Clustering (Dendrograms & PCA) model one with the best sihoutte score. As a result, this method was judged to be the best of all algorithms.

References

[1] R. M. S. V. Vaidisha Mehta, "A Survey on Customer Segmentation using Machine Learning Algorithms to Find Prospective Clients," *2021 9th International Conference on Reliability, Infocom Technologies and Optimization,* vol. 1, p. 4, 4 sep 2021.

[2] M. A. Camiller, Market Segmentation, Targeting and Positioning, Cham, Switzerland.: Springer,, 2017, p. chapter 4.

[3] U. J. S. M. S. M. a. E. K¨. K. Windler, ""“Identifying the right solution customers: A managerial methodology,” Industrial Marketing Management,," Vols. vol. 60,, p. pp. 173 –186, , 2017..

[4] R. Thakur and L. Workman, ""“Customer portfolio management (cpm) for improved customer relationship management (crm): Are your customers platinum, gold, silver, or bronze?"," *Journal of Business Research,,* Vols. vol. 69,, no. 10, pp. pp. 4095 – 4102,, 2016.

[5] W. Smith, "Product Differentiation and Market Segmentation as Alternative Marketing Strategies.," *Journal of marketing, ,* Vols. (1),, no. 21, pp. pp. 3-8., 1956. .

[6] S. B. ,. V. S. T. C. Tushar Kansal, "Customer Segmentation using K-means Clustering”,," *IEEE,* no. 1, p. 4, 2018.

[7] M. N. M. “, "Demographic Strategy of Market Segmentation”,," *INDIAN JOURNAL OF APPLIED RESEARCH ,* Vols. Volume : 6,, no. 5, p. 6, 2016.

[8] Y. H. Z. W. K. L. X. H. a. W. W. “. Hui Liu, "Personality or Value: A Comparative Study of Psychographic Segmentation Based on an Online Review Enhanced Recommender System”,," *MDPI, ,* 2019.

[9] S. G. “, "The basis of market segmentation: a critical review of literature”,," *European Journal of Business and Management,* vol. vol 3 , 2011.

[10] W. H. Susilo, " “An Impact of Behavioral Segmentation to Increase Consumer Loyalty: Empirical Study In Higher Education Of Postgraduate Institutions At Jakarta”,," *5th International Conference on Leadership, Technology, Innovation and Business Management,,* 2016.

	Model	Pca	Params	Centers	Silhou-ette	Davies Bouldin
2	Kmeans	2	{'n_clusters': 5, 'init': 'k-means++', 'random...	[[41.55103875105344, 1.870875408052184], [-44....	0.552626	0.584301
3	Kmeans	2	{'n_clusters': 6, 'init': 'k-means++', 'random...	[[54.806617718362595, 18.992161509063617], [-4...	0.534448	0.663913
1	Kmeans	2	{'n_clusters': 4, 'init': 'k-means++', 'random...	[[-13.167019014458756, -4.623026887457996], [4...	0.499151	0.671398
7	Kmeans	False	{'n_clusters': 6, 'init': 'k-means++', 'random...	[[56.155555555555544, 53.37777777777778, 49.08...	0.452055	0.747522
0	Kmeans	2	{'n_clusters': 3, 'init': 'k-means++', 'random...	[[41.55103875105344, 1.870875408052184], [-10....	0.451053	0.731013

Table 2.o