Advances in Economics and Business Management (AEBM)
p-ISSN: 2394-1545; e-ISSN: 2394-1553; Volume 10, Issue 1; January-March, 2023, pp. 28-33
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Machine learning-Based Customer Segmentation

Sandeep Rajani[1] and Dr. Birendra Goswami[2]

[1]Research Scholar, Sai Nath University, Ranchi
[2]Professor, Sai Nath University, Ranchi
E-mail: [1]s_rajani_in@yahoo.com, [2]bg.ranchi@gmail.com

Abstract—*Every firm wants to make a lot of money and stay ahead of its rivals. Understanding business consumers is essential if a firm wants to achieve significant revenue and a strong competitive position. However, because the firm's revenue is entirely dependent on its customers, effective consumer analysis within the company helps it grow. AI has the potential to give businesses the ability to strategically place the appropriate products in front of the right customers at the right time. This task benefits from machine learning. The most significant unsupervised learning difficulty, clustering, in particular, can produce categories that group like individuals. Clusters are what these categories fall under. Using [Kaggle repository] we were able to get data about Customers with membership cards might be identified by their Customer ID, age, gender, annual income, and spending score. This final score is based on information about customer behavior and purchases. In contrast to traditional market analytics, which frequently fail when the client base is very vast, big data concepts and machine learning have facilitated a larger acceptance of automated customer segmentation methodologies. This is accomplished in this study using the k-means clustering technique. If any like to sell some new products that have just hit the market, then for each of the items, he has to focus on a particular kind of customer.*

Keywords: *machine learning, customer segment, k-Mean algorithm, sklearn.*

1. INTRODUCTION

A battle fatigue is happening right now. Numerous companies and brands are competing to hear from and understand their customers before their competitors. Although businesses are accustomed to responding to client contacts, events, and behaviour after the fact or in real time, it is becoming evident that this is inadequate. There must be action made in order to keep consumer happiness at its highest level.

Customer behavior analysis is crucial for businesses. It aids businesses in better comprehending the requirements and preferences of their customers. The company can then provide the client a better service or a product that meets their needs. Information systems can be used to acquire data on things and customers. As business competition intensifies and historical data becomes more readily available, data mining methods are increasingly employed to reveal crucial and strategic information that may be hidden within an organization's data. [2]

AI can assist in modifying marketing strategies and providing individualized customer experiences by anticipating how consumer behavior may alter current business models. AI/ML data analytics systems that can project measures like customer loyalty, affinity, predicted transaction value, and purchase likelihood could do this.

This article aims to utilize data mining techniques to identify customer segments within a business enterprise. A group of business customers is referred to as a customer segment when they all belong to the same customer base and share the same market characteristics.[3] I politely suggest effective use of CRM data mining for effectively forecasting consumer segment. CRM enables businesses to effectively gather, store, and analyze consumer-related data, as well as to make it accessible to all corporate business personnel. [1] Analytical CRM, or ACRM, is utilized for customer analysis in CRM. It involves the examination of stored customer data using machine learning techniques to uncover intriguing patterns among customers. In ACRM, consumer analysis is carried out using ML techniques. In order to solve the customer analysis problem more effectively, the effort seeks to determine the optimal machine learning algorithm. In this study, the customer dataset utilized was obtained from the [Kaggle repository]. Various machine learning algorithms were employed in the experimental procedure, and the effectiveness of each was assessed using a range of validity scores.

2. LITERATURE SURVEY

2.1 Customer Classification

As businesses strive to expand their customer bases, the competitive nature of the industry has increased due to the need to fulfil the needs and desires of their patrons. [4] Meeting the wants and needs of each individual customer can be a challenging task as they often have varying demands, preferences, demographics, sizes, and other characteristics. As a result, it is not an effective business strategy to treat all customers in the same manner. In order to tackle this problem, the strategy of customer segmentation has been embraced, which involves grouping customers into smaller clusters based on common traits or actions. This enables a more precise approach to fulfilling their requirements. [5]

2. 2 Large scale Data repository

In recent times, there has been a significant increase in Big Data exploration, which pertains to an extensive volume of structured and unstructured information that cannot be analyzed using conventional methods and algorithms. Organizations collect vast amounts of data regarding their clients, vendors, and operational activities. In addition, millions of interconnected sensors, such as those found in vehicles and mobile phones, provide data on activities like sensing, manufacturing, and dispatches.[6] Gathering data is a crucial aspect of research in various academic fields, such as the natural and social sciences, humanities, and business. The primary aim of collecting data is to acquire trustworthy information that facilitates precise analysis and well-informed decision-making. The dataset used in this study was obtained from the [Kaggle repository] and contains information on mall customers, including ID, gender, annual income (in thousands of dollars), and spending score.

2.3 Clustering data using K-means

Clustering is the technique of grouping data together into a dataset based on their similarities or shared characteristics. There are several algorithms that can be used to cluster datasets, depending on the specific criteria being used.[7] As there is no universal clustering algorithm, it is crucial to select the appropriate clustering strategies based on the specific needs and objectives of the analysis. Using a Jupyter notebook, exploratory data analysis is carried out in this paper. In addition, the algorithmic hypotheticals are justified. Customers are segmented using K- means, one of the prevalent classification algorithms involves categorizing each data point into a pre-defined cluster using the K-means algorithm., to identify hidden patterns in the data that can aid in decision-making by producing graphs and the client parts. Also, it's demonstrated how to determine which order a new consumer falls into.

3. METHODOLOGY

The methodology for this study begins with obtaining the customer dataset, followed by data processing to eliminate any missing or noisy data. Several machine learning classifiers, including KNN, are then used to perform customer analysis. The resulting outcomes are measured using various performance criteria such as precision, recall, sensitivity, and specificity. The model that achieves the best performance is ultimately selected. Multiple machine learning algorithms are used to conduct customer analysis.

3.1 Data accumulation

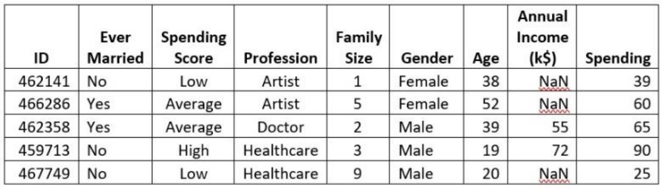
This is a data medication phase. The aim is to improve the performance of clustering algorithms by updating all data points at a standardized rate.[8] The dataset comprises information about customers from a shopping mall, including their client ID, age, annual income (in thousands of dollars), gender and spending score. We read the introductory data stored in the "*shop.csv"* file into a Data Frame using pandas.

Clients = pd.read_csv("shop.csv")

We can see that we've ID, Gender, Age, Annual Income expressed as price x1000, and the spending score as we anticipated.

Clients.head( )

Table 1: Clients.csv (first 5 rows)

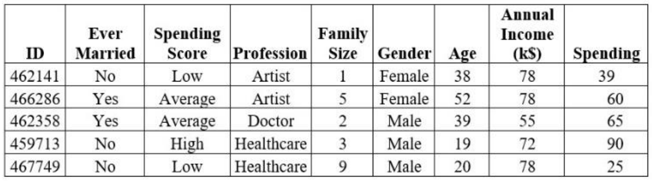


ID	Ever Married	Spending Score	Profession	Family Size	Gender	Age	Annual Income (k$)	Spending
462141	No	Low	Artist	1	Female	38	NaN	39
466286	Yes	Average	Artist	5	Female	52	NaN	60
462358	Yes	Average	Doctor	2	Male	39	55	65
459713	No	High	Healthcare	3	Male	19	72	90
467749	No	Low	Healthcare	9	Male	20	NaN	25

First, we look to see if the dataset has any missing values. Missing values cannot be handled by the K-means method. We found that some data is missing which will be filled with mode value. By using the following statement for col in *Clients.columns:*

Clients[col].fillna(Clients[col].mode()[0], inplace=True)
Clients.isnull().sum()
Clients.head() finally the data becomes

Table 2: Clients.csv (first 5 rows)



ID	Ever Married	Spending Score	Profession	Family Size	Gender	Age	Annual Income (k$)	Spending
462141	No	Low	Artist	1	Female	38	78	39
466286	Yes	Average	Artist	5	Female	52	78	60
462358	Yes	Average	Doctor	2	Male	39	55	65
459713	No	High	Healthcare	3	Male	19	72	90
467749	No	Low	Healthcare	9	Male	20	78	25

Additionally, we can look for duplicate rows. Thankfully there are no redundancies. Finally, we examine the DataFrame's representation of each variable. Direct manipulation of categorical variables is not possible. Distances are the basis of K-means. It depends on the kind of category variables how those variables are converted.

3.2 Methods of customer classification

There are numerous partitioning techniques, each with a unique severity level, data needs, and goal. The absence of sufficient coverage results in the exclusion of studies that describe artificial neural networks, particle identification, and advanced forms of ensemble.

We can start monitoring how the variables are distributed. Let's define two functions right now. The first one will get the variables' descriptive statistics. We can graph the variable distribution with the aid of the second one.

We'll receive the descriptive statistics. We will obtain the counts in each category if the variable is not numeric.

spending = Clients["Spending Score"]

statistics(spending)

Table 3: Spending Scored descriptive statistics

Spending Score	
Low	1600
Average	617
High	381

graph_pie(spending)

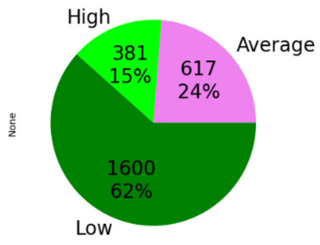


Fig. 1: Graph representing percentage of spendings

Then, we'll evaluate the Age.
age = Clients["Age"]
statistics(age)

Table 3: Mean, standard deviation, median, and variance for the Age descriptive statistics

	Mean	Standard Deviation	Median	Variance
Variable				
Age	43.621632	16.937246	41.0	286.87031

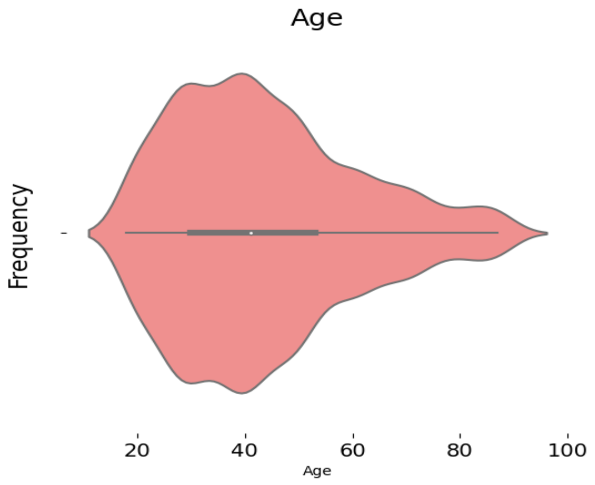


Fig. 2: Violin graph representing histogram of Age

And at the, we'll explore **Annual Income** variable.

inc = Clients["Annual Income (k$)"]
statistics(inc)

Table 4: Mean, standard deviation, median, and variance for the Annual Income descriptive statistics

	Mean	Standard Deviation	Median	Variance
Variable				
Annual Income (k$)	76.660893	8.610068	78.0	74.133274

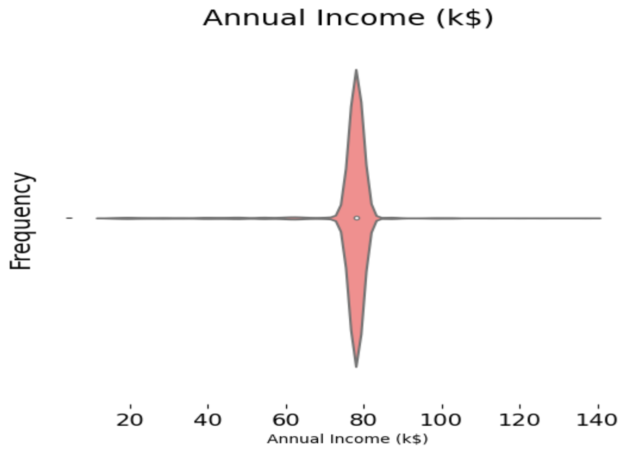


Fig. 3: Violin graph representing histogram of Annual Income($)

Gender = Clients["Gender"]

statistics(Gender)

Table 5: Gender descriptive statistics

	Gender
Female	1191
Male	1407

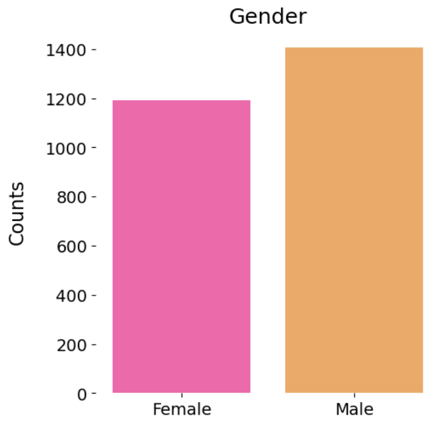


Fig. 4: Graph representing histogram of Gender

We'll also examine the relationship between the numerical parameters. We'll employ the pairplot seaborn function to achieve that goal. We are looking to see if there is a gender difference. To obtain different colors for points belonging to females or males, the hue parameter will be adjusted.

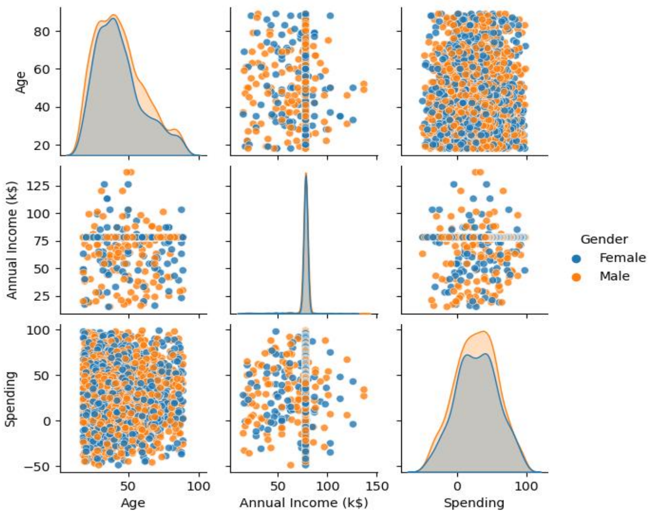


Fig. 5: Graph representing gender difference using pair plot seaborn function

Our dataset's variables follow a normal distribution. The differences are rather close to one another. With the exception of age, which has a lower variance than the other factors. Principal Component Analysis (PCA) can be used to determine which dimensions best maximize the variance of the involved features after confirming that we can use k-means. The categorized variable will be converted into two binary variables for the same (0 & 1)

Table 6*: Categorial variable gender converted to 0 & 1 for present of absence*

	Age	Annual Income (k$)	Spending	Male	Female
0	38	78.0	39	1	0
1	52	78.0	60	1	0
2	39	55.0	65	0	1
3	19	72.0	90	0	1
4	20	78.0	25	0	1

In order to apply Principal Component Analysis , we are going to use the function from sklearn module.

```
print(pca.components_)
```

```
[ [3.39028045e-03  -9.00378776e-04  -9.99993807e-01  -2.02365073e-04  2.02365073e-04]
 [ 9.99214467e-01 -3.94785342e-02   3.42330420e-03  -2.96710452e-04  2.96710452e-04] ]
```

```
print(pca.explained_variance_)
```

```
[919.843159 287.3063123]
```

Vectors are defined by these seemingly abstract numbers. The components determine the vector's direction, and the explained variance determines the vector's squared length.

A biplot, a sort of scatter plot, can be used to illustrate this. The primary component score for each point serves as a representation of that point. It also aids in the discovery of connections between the original variables and the major components.

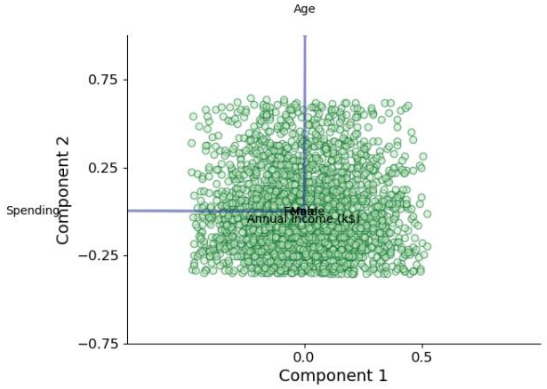


Fig. 6: Biplot representing score regarding the principal components.

We can see that the two most crucial factors are Annual Income and Spending Score.

3.3 K-Means confront

The K-means clustering algorithm is frequently employed to obtain an understanding of patterns and variations present in a database.[9] In marketing, it’s frequently used to make customer groups and understand the relationship among these groups. The K-means clustering algorithm typically uses Euclidean distance to determine the similarity or dissimilarity between two data points. To begin, we will set the number of clusters we want to use. There are several methods to determine the optimal number of clusters, such as the elbow and silhouette methods. The elbow method involves examining the total within-cluster sum of squares (WSS), which we aim to minimize. We will execute the K-means algorithm for a variety of k values, in our case, k = 5, and calculate the total WSS for each k. We will then plot the WSS versus the number of clusters and identify the elbow point, which is considered the appropriate number of clusters. Additionally, we will use the hue parameter to assign different colors to the points belonging to males and females.

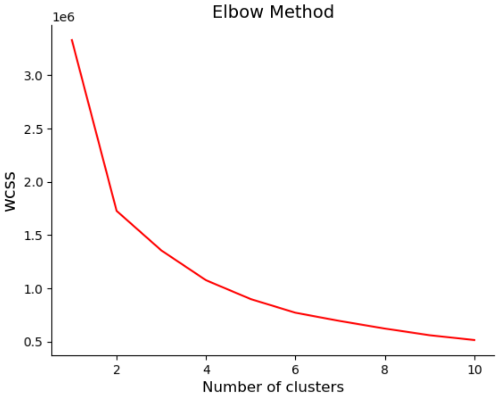


Fig. 7: Intra-cluster variation Elbow Method

3.4 Clustering Centroids

In order to identify the random state, we use the K-means technique, assuming that there are 5 clusters at this time. We then execute the procedure 10 times with different centroid seeds. Our clusters appear to be:

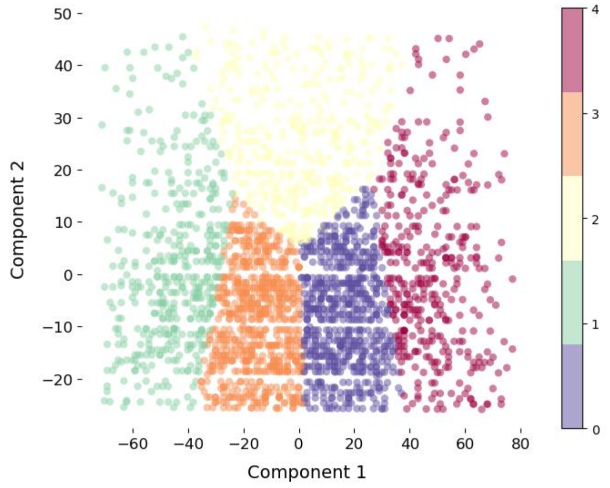
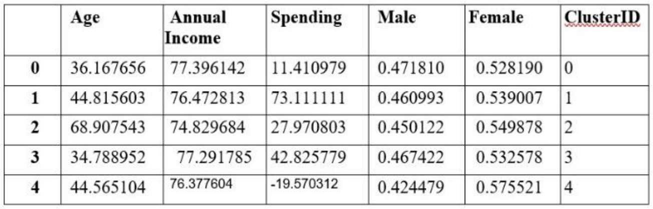


Fig. 8: Domains grouped into 5 clusters

Table 7: Centroids



	Age	Annual Income	Spending	Male	Female	ClusterID
0	36.167656	77.396142	11.410979	0.471810	0.528190	0
1	44.815603	76.472813	73.111111	0.460993	0.539007	1
2	68.907543	74.829684	27.970803	0.450122	0.549878	2
3	34.788952	77.291785	42.825779	0.467422	0.532578	3
4	44.565104	76.377604	-19.570312	0.424479	0.575521	4

The score for Annual Income and Spending appears to be the most crucial factor. Segment 0 includes people with low incomes who also have similar spending patterns. Segment 1: People with high incomes and large expenditures. Customers in sector 2 who earn in the middle range yet maintain the same level of spending. Then, in sector 4, we have clients whose income is very high but who also make the greatest purchases. People who earn little money yet spend a lot make up section 5's final group.

Imagine that tomorrow we've a new member. And we want to know which section that person belongs. We can prognosticate this by using Kmeans predict feature as: consider that the age,annual income, spending and gender(1 for male,0 for female) is 73,88,74,1,0

newClient = np.array([[43, 76, 56, 0, 1]])

new_client = kmeans.predict(newClient)

print(f"The new customer belongs to segment {new_client[0]}")

Prediction will be that: The new customer belongs to segment 1

4. CONCLUSION

The internal clustering validation method was chosen for this study, rather than the external clustering validation method which requires external data, such as labels. This is because our dataset was unbalanced. If done correctly, customer segmentation can benefit a company.

Analyzing exploratory data is done in this Jupyter notebook. In addition, the algorithmic assumptions are verified. Customers are segmented using K-means, which generates a graph representing the customer segments. Furthermore, the method of determining the category of a new customer is also demonstrated.

In this paper, we demonstrate unsupervised learning in action and produce recommendations for a possible client using data from the real world. Today, a lot of businesses amass a ton of data on their clients and customers, and they are keenly interested in discovering the significant connections that are concealed in their clientele. Knowing this information can help a business create future goods and services that best meet the desires or requirements of its clients. With that knowledge, we can provide recommendations in this section for further potential customers.

REFERENCES

[1] Dalla Pozza, I., Goetz, O., & Sahut, J. M. (2018). Implementation effects in the relationship between CRM and its performance. Journal of Business Research, 89, 391–403. https://doi.org/10.1016/j.jbusres.2018.02.004

[2] Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities. S.l: Packt printing is limited

[3] 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 Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1

[4] 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 Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1

[5] T.Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. Anu Manufactured cluster analysis using a new algorithm from structured and unstructured data. International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2

[6] McKinsey Global Institute. Big data. The next frontier is creativity, competition and productivity. 2011. Accessed at: www.mckinsey.com/mgi on July 14, 2015.

[7] 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

[8] A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3

[9] Vishish R. Patel1 and Rupa G. Mehta. MpImpact for External Removal and Standard Procedures for JCSI International International Science Issues Issues, Vol. 8, Appeals 5, No 2, September 2011 ISSN (Online): 1694-0814