REPORT

CUSTOMER SEGMENTATION USING K -MEANS

PREPARED BY:-SAURAV BORAH

-> ABSTRACT:

Customer segmentation:

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits. Customer segmentation enables a company to customize its relationships with the customers, as we do in our daily lives.

Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

Benefits of customer segmentation include personalization, customer retention, better ROI for marketing, and revealing new opportunities.

Unsupervised learning in machine learning:

The use of machine learning can be seen almost everywhere around us, be it Facebook recognizing you or your friends, or YouTube recommending you a video or two based on your history — Machine Learning is everywhere! However, the 'magic' of machine learning is not just limited to only these areas. Machine Learning is broadly categorized as Supervised and Unsupervised Learning.

Supervised Learning is one in which we teach the machine by providing both independent and *dependent variables*, for example, Classifying or predicting values.

Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision. In contrast to supervised learning that usually makes use of human-labeled data, unsupervised learning, also known as self-organization allows for modeling of probability densities over inputs.

Two of the main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships. Cluster analysis is a branch of machine learning that groups the data that has not been labelled, classified or categorized. Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

Some of the most common algorithms used in unsupervised learning include: (1) Clustering, (2) Anomaly detection, (3) Neural Networks, and (4) Approaches for learning latent variable models. Each approach uses several methods.

-> INTRODUCTION:

Clustering is the task of dividing the population or data points into several groups, so that the data points in the same groups are more similar to other data points in the same group than those of other groups. In simple words, the goal is to segregate groups with similar traits and assign them to clusters.

Cluster analysis can be done based on the resources in which we try to find subgroups of samples based on resources or based on samples in which we try to find subgroups of resources based on samples. We will address resource-based clustering here. Clustering is used in market segmentation; where we try to fine customers, who are similar to each other, whether in terms of behaviors or attributes, segmentation / compression of images; where we try to group similar regions, group documents based on topics, etc.

The Problem:

Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

Malls or shopping complexes are often indulged in the race to increase their customers and hence making huge profits. To achieve this task machine learning is being applied by many stores already.

It is amazing to realize the fact that how machine learning can aid in such ambitions. The shopping complexes make use of their customers' data and develop ML models to target the right ones. This not only increases sales but also makes the complexes efficient.

K-means Clustering:

Among many clustering algorithms, the K-means clustering algorithm is widely used because of its simple algorithm and fast convergence. However, the K-value of clustering needs to be given in advance and the choice of K-value directly affects the convergence result. To solve this problem, we mainly analyze four K-value selection algorithms, namely Elbow Method, Gap Statistic, give the pseudo code of the algorithm; and use the standard dataset Iris for experimental verification. K-means clustering is a method used for clustering analysis, especially in data mining and statistics. It aims to partition a set of observations into a number of clusters (k), resulting in the partitioning of the data into Voronoi cells. It can be considered a method of finding out which group a certain object really belongs to. It is used mainly in statistics and can be applied to almost any branch of study.

The approach K-means follows to solve the problem is called Expectation Maximization:

- Specify number of clusters K.

- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Stopping Criteria for K-Means Clustering:
There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

When to use Cluster Analysis?
If we are using a labeled data, we can use classification technique, whereas in case when the data is not labeled, we can cluster the data based on certain feature and try to label it on our own. So, when we use cluster analysis, we don't have labels (i.e.. data is not labeled) in the context of machine learning this is called as unsupervised learning.

Final Goal:
The goal of clustering is to maximize the similarity of observation within the cluster and maximize the dissimilarity between the clusters.

-> DATASET:

Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

- Customer ID: It is the unique ID given to a customer
- Gender: Gender of the customer
- Age: The age of the customer
- Annual Income (k$): It is the annual income of the customer
- Spending Score: It is the score (out of 100) given to a customer by the mall authorities, based on the money spent and the behavior of the customer.

-> METHODOLOGY:

Tools used for implementing and visualizing:

Various python libraries were used for completing this project. Some of them are listed below: -

• NumPy: NumPy is a library for the Python programming language, adding support for large, multi- dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

• Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

• Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

• Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

• Scikit-learn: Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbors, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

K-means:

K-means clustering algorithm is the most selected technique to cluster data. K-means is a nonhierarchical clustering and use looping to group data into K groups. The K-means clustering start the iterative process by finding the initial centroid, or central point, of each group by randomly selecting representative data from raw data to be a centroid in each K data groups. Then assign each data to the closest group by calculating the Euclidean distance between each data record to each centroid to allocate the data record to the nearest group. After that each cluster will find new centroid to replace the initial one and repeat steps of Euclidean distance computation to group data members and send each member to group of the nearest centroid. The process will stop when each group has stable centroid and members do not change their groups.

The steps of k-means algorithm can be summarized as the following:

• Specify group number and select initial centroid of each group.

• Calculate Euclidean distance for each data member and centroid to assign members to the nearest centroid.

• Calculate distance's mean of every data member and own centroid to define new centroid in each group.

• Repeat steps 2 and 3 until each group has stable centroid or same centroid.

-> IMPLEMENTATION:

The steps for implementation.

- Various libraries were imported to manipulate, visualize and modeling data.

- Certain components of the data are visualized for better insight.
- Selecting a k value for the model.
- Model implementation and obtaining the labels.
- Plotting the Clusters on as 2D graph.

CODE:

```
from google.colab import
drive
drive.mount('/content/drive')


import numpy as np
import seaborn as
sns import pandas
as pd
import matplotlib.pyplot as plt
from sklearn.cluster import
KMeans
path ='/content/drive/My
Drive/Mall_Customers.csv' df = pd.read_csv(path)
df.head()
sns.boxplot(x=df["Annual Income (k$)"])
sns.boxplot(x=df["Age"])
sns.boxplot(x=df["Spending Score (1-100)"])
genders = df.Gender.value_counts()
sns.set_style("darkgrid")
plt.figure(figsize=(10,4))
sns.barplot(x=genders.index,
y=genders.values) plt.show()
age18_25 = df.Age[(df.Age <= 25) & (df.Age
>= 18)] age26_35 = df.Age[(df.Age <= 35) &
(df.Age >= 26)] age36_45 = df.Age[(df.Age <=
45) & (df.Age >= 36)]
```

```
age46_55 = df.Age[(df.Age <= 55) & (df.Age >=
46)] age55above = df.Age[df.Age >= 56]


x = ["18-25","26-35","36-45","46-55","55+"]
y =
[len(age18_25.values),len(age26_35.values),len(age36_45.values),len(age46_55.values),len(age55above.valu es)]


plt.figure(figsize=(15,6))
sns.barplot(x=x, y=y,
palette="rocket")
plt.title("Number of Customer and
Ages") plt.xlabel("Age")
plt.ylabel("Number of
Customer") plt.show()
df.replace(to_replace ="Male", value = 0,inplace=True)
df.replace(to_replace ="Female", value =
1,inplace=True) df


from sklearn.preprocessing import
StandardScaler X = df.values[:,[3,4]]
X = np.nan_to_num(X)
Clus_dataSet = StandardScaler().fit_transform(X)
Clus_dataSet
wcss=[]
for i in range(1,11):
   kmeans = KMeans(n_clusters= i, init='k-means++', random_state=0)
   kmeans.fit(X)
   wcss.append(kmeans.inertia
_) plt.plot(range(1,11), wcss)
plt.title('The Elbow Method')
```

```
plt.xlabel('no of
clusters')
plt.ylabel('wcss')
plt.show()
kmeansmodel = KMeans(n_clusters= 5, init='k-means++', random_state=0)
y_kmeans= kmeansmodel.fit_predict(X)
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'yellow', label
= 'Centroids')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-
100)') plt.legend()
plt.show()
```

-> CONCLUSION:

K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major applications of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

Firstly, we tried to visualize the gender distribution from the given dataset. We made a bar plot to check the distribution of male and female population in the dataset. The female population clearly outweighs the male counterpart.

Then we tried to visualize the age distribution. We made a bar plot to check the distribution of number of customers in each age group. Clearly the 26–35 age group outweighs every other age group.

Finally using the K-means algorithm we tried to analyze their annual incomes and spending

scores. So, looking at the final graph we can conclude the following: -

In cluster 4 (light blue colored) we can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

In cluster 2 (blue colored) we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it's because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.

In cluster 1 (red colored) we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.

In cluster 3 (green-colored) we see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall's facilities.

In cluster 5 (pink colored) we see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall's services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

Finally, based on our machine learning technique we may deduce that to increase the profits of the mall, the mall authorities should target people belonging to cluster 5 and cluster 1 and should also maintain its standards to keep the people belonging to cluster 3 and cluster 2 happy and satisfied.