CUSTOMER SEGMENTATION USING MACHINE LEARNING IN PYTHON

Dr. Y. Suresh, Dr. J. Senthilkumar, Dr. V. Mohanraj, Dr. Selvaraj Kesavan

1Professor/IT, Sona College of Technology, Salem, Tamilnadu, India
Email id: ysuress33@gmail.com

2Professor/IT, Sona College of Technology, Salem, Tamilnadu, India
Email id: jsenthil10@gmail.com

3Professor/IT, Sona College of Technology, Salem, Tamilnadu, India
Email id: vmohanraj06@gmail.com

4DXC Technologies, India,
Email id: selvarajkesavan@gmail.com

ABSTRACT

A client is a critical aspect in the success of any organisation. The retention of an existing client base and the development of a new client base are crucial for a company's long-term success. This necessitates an understanding of client behaviour in respect to the firm. As a result, a company seeking a competitive advantage in the market must gain a 360-degree perspective of its clients using Python, ML. Customer Segmentation is one such strategy in this approach, which aids in identifying groups of similar consumers based on their interactions with the product and then effectively implementing various marketing strategies for the suitable consumers.

Keywords: Segmentation, Python, Machine Learning, Organisation

I. INTRODUCTION:

One of the most significant uses of unsupervised learning is customer segmentation. Companies may discover different categories of clients using clustering algorithms, enabling them to target the prospective user base. We will utilise K-means clustering in this machine learning project, which is the most important approach for grouping unlabeled datasets.

Reason for Algorithm Selection:

K-means clustering is a straightforward and quick technique. It is the most often utilised strategy for client segmentation, particularly for numerical data. K-means provides a number of computational benefits, including the ability to scale effectively with huge datasets. For computation on a large dataset, hierarchical and model-based clustering approaches involve calculating a whole distance matrix, which has restricted scalability and high memory needs. K-means clustering, on the other hand, is more run-time efficient. Given these features, as well as the fact that the input dataset is big and mostly consists of numerical data, K-means proved to be an excellent option for consumer segmentation.

II. IMPLEMENTATION:

Client segmentation is the practice of dividing a customer base into many groups of people who are similar in many aspects significant to marketing, such as gender, age, hobbies, and other spending patterns.

Companies that use customer segmentation believe that each client has unique needs that need a tailored marketing strategy to satisfy. Companies want to obtain a better understanding of the customers they're after. As a result, their goal must be particular and designed to meet the needs of each and every unique consumer. Furthermore, by analysing the data acquired, businesses may acquire a better grasp of client preferences as well as the needs for identifying profitable categories. This allows them to more effectively design their marketing strategies while reducing the chance of their investment being jeopardised.
Customer segmentation is based on a number of significant differentiators that split consumers into groups that may be targeted. Data on demographics, location, economic position, and behavioural patterns are all important factors in establishing the company's approach to distinct sectors.

2.2 How to Implement Customer Segmentation?
We will do data exploration as the initial phase in our data science project. We'll import the necessary packages for this role before reading our data. Finally, we'll look through the input data to get the information we need. The next step is to specify how many clusters the k-means algorithm should construct. This is accomplished via the elbow method.

2.3 Elbow Method
The elbow approach is a heuristic approach that serves the purposes of evaluation and confirmation of cluster analysis consistency in order to discover the best number of clusters, the dataset is examined. This method is used in order to get the percentage of variation explained with respect to the number of clusters: This will be the case when we have enough clusters that the addition of another cluster does not drastically improve data modelling. When the number of clusters is plotted against the percentage of variation explained by the clusters, the early clusters will offer a lot of information (provide a lot of variation), but the marginal increase will level out at some point, resulting in a concave curve in the graph. As the elbow criteria has been set, the number of clusters to be picked is referred to as the "elbow".

2.4 K-means Algorithm
Cluster the data using the k-means method, and the number of clusters (k) we want in the final result is the first step. The first step in the cluster creation procedure is to randomly choose k objects from the dataset to be the initial cluster centres. The objects you picked are also known as cluster means, sometimes known as centroids. The objects that remain are then all given to the centroid that is the closest to it. Euclidean the closer an item is to the cluster mean, the more likely it is to be drawn from that cluster. In this stage, we have a reference to the "cluster assignment". The process is implemented to automatically adjust the mean value for each cluster in the data when the assignment is done. Once the recalculation of the centres is complete, the observations are reviewed to see whether they are now closer to a different cluster. This modification to the cluster mean is used to reallocate the objects. This is performed until the cluster assignments aren't changing anymore. Similar to the clusters detected in the previous iteration, the new clusters found in this iteration are the same.

Summing up the K-means clustering –

- The number of clusters we need to construct is specified.
- The method chooses k items from the dataset at random. The starting cluster or mean is this item. A new observation is assigned to the centroid that is closest to it. In this assignment, the Euclidean distance between the item and the centroid is taken into consideration.
- For each of the cluster's data points, a new mean value is calculated by multiplying the previous mean value by the total number of data points belonging to that cluster. The centroid of the kth cluster has a length of p and contains the means of all of the variables for the kth cluster's observations. This expression is known as the p expression and represents the number of variables.
- Within the sum of squares, iterative minimization of the total. The assignment will then cease wagging when we reach maximum iteration via iterative minimization of the total sum of the square.

Code:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
```
#Import the data into the programme.

dataset = pd.read_csv('input/Mall_Customers.csv')

dataset.head(10) #Print first ten rows of the dataset

#the dataset's total rows and columns

dataset.shape

dataset.info() # There are no missing values since all of the columns are filled with 200 items. #Missing values computation

dataset.isnull().sum()

### Choose from a variety model in the features.

#Only two characteristics (Annual Income & Spending Score) are provided, and there is no Label.

X= dataset.iloc[:, [3,4]].values

#Constructing the Model

from sklearn.cluster import KMeans

wcss=[]

###To acquire the maximum number of clusters, use a static code.

for i in range(1,11):
    kmeans = KMeans(n_clusters= i, init='k-means++', random_state=0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

#Visualizing the ELBOW approach to determine the best K value

plt.plot(range(1,11), wcss)
plt.title('The Elbow Method')
plt.xlabel('no of clusters')
plt.ylabel('wcss')
plt.ylabel('wcss')
plt.show()
# Model Build

kmeansmodel = KMeans(n_clusters= 5, init='k-means++', random_state=0)
y_kmeans= kmeansmodel.fit_predict(X)

# Imaginatively mapping all the groups
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()

### Model Interpretation

#Cluster1(Red Color) -> spending less & earning high
#cluster2(Blue Color) -> average in terms of spending & earning
#cluster3(Green Color) -> spending high & earning high [TARGET SET]
#cluster4(cyan Color) -> spending more & earning less
#Cluster5(magenta Color) -> spending less & earning less
III. APPLICATIONS:

Cluster detection may aid in the development of a customised strategy for each cluster base. By tracking consumers across months and recognising the amount of consumers migrating from one cluster to the next, clustering may also be used to determine client buying behaviour. This enables the company to better arrange its efforts for increasing income at various locations. Customer segmentation insights can be used by businesses to better target their marketing efforts to the right customers, for example, discounts and offers related to a specific shop can be sent to only those customers who regularly purchase from that shop, without bothering customers from other shops. As a result, focusing on the appropriate consumers for the appropriate transactions may help reduce marketing expenses, boost revenue, and improve customer satisfaction.

IV. CONCLUSION:

We looked through the client segmentation model in this data science assignment. Supervised learning was the process through which we created our machine learning model. We used a clustering approach called K-means clustering in particular. We evaluated and displayed the data before putting our system into action.

REFERENCES