COLLEGE CODE: 5127

APPLIED DATA SCIENCE

CUSTOMER SEGMENTATION USING DATASCIENCE

Group-5

BATCH MEMBERS:

1. **A.VIGNESH**
2. **E.SANJAY**
3. **S.ARAVINDHAN**
4. **U.SANTHOSH**

INTRODUCTION:

Mall Customers Segmentation Project is based on customers of mall where we predict which customers has earn more money and spending high by using customer behavior and purchasing data.

Customer segmentation involves implementing data science methods to divide the customer base into smaller groups based on certain characteristics. It assists marketing managers in better understanding their customers' preferences and presenting them with better-targeted advertisements.

ABOUT THE DATA:

Where did we get the dataset?

Kaggle:

The dataset provided on Kaggle,

https://www.kaggle.com/datasets/akramw4/mall-customers

offers a valuable resource for our project aimed at customer segmentation using datascience.

Dataset Details:

The data (mall customers) contains the following information:

Mall customers – data set about customers who purchased in mall;

Customer Id – customer's those who had purchased ;

Gender - Customer's gender is mentioned where those customer is male or female;

Age - The age of the customer is mentioned;

Annual income- Cluster 1 — This cluster represents the customer_data having a high annual income as well as a high annual spend. Cluster 2 — This cluster denotes a high annual income and low yearly spend. Cluster 3 — This cluster denotes the customer_data with low annual income as well as low yearly spend of income.;

Spending scores - 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.

BEGINNING WITH THE PROJECT

To begin building a project data about customers in Mall, we first need to load the dataset.

We have a dataset file in a common format like CSV, here are the steps to load the dataset:

1.Importing the required Libraries(data.csv):

In this step, we import the necessary Python libraries and modules to work with our data and perform various data processing and machine learning tasks.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
```

2.Importing the data set(read data set; create matrix):

This step involves loading our dataset into memory. We use libraries like pandas to read data from a CSV file or other formats.

Importing dataset:

https://www.kaggle.com/datasets/akramw4/mall-customers

Print first 5 data's(head):

```
customer_data.head()
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Getting some informations about the dataset:

```
customer_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              200 non-null    int64 
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64 
 3   Annual Income (k$)      200 non-null    int64 
 4   Spending Score (1-100)  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Checking for missing values:

```
customer_data.isnull().sum()
CustomerID 0
Gender 0
Age 0
Annual Income (k$) 0
Spending Score (1-100) 0
dtype: int64
```

Splitting the data into train and test:

TRAIN- xtrain , ytrain

TEST- xtest , ytest

Once we have pre-processed the data into a format that's ready to be used by the model, we need to split up the data into train and test sets. This is because the machine learning algorithm will use the data in the training set to learn what it needs to know.

It will then make a prediction about the data in the test set, using what it has learned. We can then compare this prediction against the actual target variables in the test set in order to see how accurate the model is.

We will do the train/test split in proportions. The larger portion of the data split will be the train set and the smaller portion will be the test set.

This will help to ensure that we are using enough data to accurately train the model. In general, we carry out the train-test split with an 80:20 ratio(also 70:30 ratio)

Feature Scaling:

Feature scaling is a method in which we scale the data into an accurate and scalable size for the purpose of increasing accuracy and reducing error.

It basically prevents the large variance of data points to be used in the algorithm and allows us to achieve better results.

```
from sklearn.preprocessing import StandardScaler
```

Conclusion:

Using market segmentation, companies are able to identify their target audiences and personalize marketing campaigns more effectively.

This is why market segmentation is key to staying competitive. It allows you to understand your customers, anticipate their needs, and seize growth opportunities.