MINI

PROJECT

GRAPHIC ERA DEEMED TO BE UNIVERSITY

4th SEMESTER

Name : Siddhant Bohra
Section : A
Class Roll No. : 56
Univ Roll No. : 2014881

CUSTOMER SEGMENTATION IN RETAIL SECTOR

Introduction :

In today's competitive world, we live in a society where the person or the company who has maximum amount of data regarding people is the most successful.

Data regarding people can comprise of many things like their name, phone number, annual income, how they like to spend, what type of things they like to buy, etc. Company's like to use this data in their favor for their success. Here comes the role of Customer Data Segmentation.

It the practice of dividing the company's customers into groups that reflects similarity among customers in each group.
Its goal is to decide how to relate to customers in each group in order to maximize the value of each customer to the business.

The process requires a though out strategy, to understand and group the customers and which data we use to do this.

In this project we have used Demographic and a bit of Behavioral Segmentation, as we have taken Age, Annual Income and Spending Score as the Segmentation Factors.

Advantages of Customer Segmentation :

1. Price Optimization
2. Enhances Competitiveness
3. Brand Awareness
4. Acquisition and Retention
5. Increase Revenue and Rate of Interest

Motivation :

For the motivation, there is no specific motivation for me to choose this topic, but since we are going to deal with databases in our future semesters so I thought learning and to get an experience of working with database will be a good option.

Also in my elective I have been learning a bit of beginners level database management, so this reason also pushed me a little bit to work on this project.

Some Details Regarding The Terminologies in the Project :

Ques : Why Machine Learning is Used?

Earlier when we used to segment data manually it was a very hard task. But now in recent years with the help of Machine Learning we can figure out the regularities within similar type of data easily and also discover the recurring pattern, which ultimately helps us to perform data segmentation easily and efficiently.

Ques : What is Unsupervised Learning?

Since in this project we need to cluster our data, we will be using Unsupervised Learning.
Unsupervised Learning is the training of a machine using information that is neither classified or labeled and allowing the algorithm to act on that data without guidance.
Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Ques : What is K-means Algorithm?

In this project we have used K-means Algorithm to cluster our data.
K-means algorithm is an iterative algorithm that divides the unlabeled data-set into 'k' different clusters in such a way that each data-set belongs to only one group that has similar properties. It the uses the method called WCSS (Within Cluster Sum of Squares) i.e., the sum of the square distances between the data points and the cluster centroids is minimum. The least variation we have within the cluster the more homogeneous data points are within the same cluster.

Ques : What is Elbow Method?

To figure out the or to choose the optimum value of 'k', in this project we are using the elbow method. In this we plot a graph between WCSS (**W**ithin **C**luster **S**um of **S**quares) and Values of K, it is called so because, when the graph is plotted, it looks like a human's elbow. The value at which the elbow starts to get stable or uniform, is the Elbow Point. And the value of 'k' for which we obtain the elbow point is the optimum value of 'k' i.e., the optimum numbers of clusters which we will use for the project.

Methodology :

In this project I've used Python Programming Language because it is easy to program in Python for database management related programming, as it consists of predefined libraries and classes which help us to deal with database management and plot graph regarding the data we have.

Firstly I've used three libraries :-

Pandas : Pandas is an inbuilt python library that is used for working with and for manipulating datasets.Pandas is a reference to "Panel Data" or "Python Data Analysis".

Matplotlib : Matplotlib is a low level graph plotting library in python that serves as a visualization utility.

Sklearn : Sklearn stands for "Scikit Learn". It is a machine learning library for python. It is basically used to train the machine with a given data-set so that it can give optimum output for the further unknown data given by user.

Then using the pandas library we are importing the database i.e., Mall_Customers.csv
Then testing some basic functionalities we look for the starting sample of database we have using head() function. We get the info about the data using the info() function. We get the number of rows and columns in the database using the shape() function. Also we can check whether our database has any null value at any point using isnull() function.

Let's begin with the segmentation process.
I've done segmentation of three types. On the basis of Spending Score and Annual Income, on the basis of Age and Annual Income, on the basis of Age and Spending Score. The process is same for all three.

So for the segmentation on the basis of Spending Score and Annual Income. First we will store the values of column of Annual Income and Spending Score in a single variable using the iloc() function of pandas library. Now since we want to plot a cluster graph for the process of segmentation, we need to perform WCSS to find the appropriate number of 'k' value which is also the number of clusters which are appropriate for the segmentation. Then for all the values of 'k' we plot a line graph between WCSS and Number of Clusters(k). Then after studying the graph we obtain an elbow point and the value of 'k' regarding that point is the appropriate number of clusters in which we can segment our data. Now for the given project, we obtain the value of 'k' = 5. So for the obtained value of 'k' we train our model using kmeans from the sklearn library, and using the function fit_predict() function. The clusters are created on the basis of WCSS algorithm, that is cluster centers are formed and the sum of square of distances from every centroid is calculated and the point belongs to group where distance from centroid is minimum. Then finally we plot the cluster graph and bar graphs for Spending Score and Annual Income. Where we can clearly see the five clusters in which data has been segmented.

Similarly, for the other two segmentation we continue with the same procedure to obtain the cluster and bar graph for them.

Acknowledgement :

For this project I would like to acknowledge the following :

1. **Edureka** : For providing the basic information about the details on Data Segmentation.

2. **Geeks For Geek** : For acting as a source for the details regarding the knowledge on Python Programming Language and its Libraries.