**Project Requirement and Specification**

**On**

**CUSTOMER SEGMENTATION IN RETAIL SECTORS**

**(CSE** **IV SEMESTER MINI PROJECT)**

**2020-2021**



**Submitted to: Submitted by:**

MR, ASHISH GARG MR KUSHAGRATRIVEDI

University Roll. No.: 2015079

CSE-DS- IV Sem

Session: 2020-2021

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GRAPHIC ERA DEEMED UNVERSITY, DEHRADUN**

* 1. **About Project**

As retail industry emerges there is increasing motivation for retailers to look for data or strategies that can help them segment or describe their customers in a succinct, but informative manner. This work focuses on Customer Segmentation, by integrating machine learning practices and conventional business understandings, the paths to understand customer behaviour became more intertwined and answers question like What segments or groups of customers do we have and how we can increase UI thus increasing sales?

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,we will be using unsupervised learning fot this data set. 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.To figure out 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 (Within Cluster Sum of Squares) 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.

* 1. **Requirement of Project**
     1. **Hardware Requirement**
* Processor: min 1 GHz ,recommended 1.60 GHz or more
* Ethernet connection or a wi-fi
* Hard Drive: min 32GB ,recommended 64 GB or more
* Memory(RAM):8 GB recommended
* OS:WINDOWS 10
* System type: 64 bit recommended.
  + 1. **Software Requirement**
* Python 3s version
* Jupyter Notebook
* Microsoft Excel
  1. **Modules of Project**
     1. **Load Data And Packages**
* Check out the data
  + 1. **Exploratory Data Analysis(EDA)**
    2. **Training a Model**
    3. **Segmentation of Data**
    4. **Feature Transformation/Engineering**
    5. **Modelling and predictions**
  1. **BRIEF MOTIVATION:**

I want to thank Computer Science Engineering Department who provided me this opportunity to do the mini project .I also want to thank my teachers who provided the lecture to understand all important topics related to project and taught us to write and run the program, they also took doubt sessions to clear our doubts. Completing this project has build a confidence in me and now i am able to perform more projects like these.

**REFERENCE**

1. UCI Machine Learning Repository for datasets download
2. W3schools.com for coding functionalities explanation in python
3. Google.com for some explanations

**THANK YOU!**