Unsupervised Learning

Week - 1 Practice Project

Topics Covered:

- **K means Clustering**
- **Hierarchical Clustering**

Domain:

Customer Segmentation, E-commerce

Problem Statement:

XYZ.com is an e-commerce company based in Argentina. They have past order details which includes customer personal details like customer name, address, etc and order details like order quantity, sales, product code, etc.

The company wants to understand if there is any pattern among these customers and use it for making more profit. They want to analyse the sales data wrt to customers and identify high yield groups of customers.

Objective:

We will use clustering techniques to identify the segment of customers.

Customer segmentation is the activity of dividing a broad consumer or business market, normally consisting of existing and potential customers, into sub-groups of consumers (known as segments) based on some type of shared characteristics. It’s a very important activity of any e-commerce company as it can be used to identify high yield segments, those segments that are likely to be the most profitable or that have growth potential.

Data Description:

ORDERNUMBER: Order number of the product.

QUANTITYORDERED: Ordered quantity.

PRICEEACH: Price of each product.

ORDERLINENUMBER: Order line number of the product.

SALES: Sales of the product.

ORDERDATE:Order date of the product.

STATUS:Shipping status(i.e. Shipped or canceled or Resolved) (**TARGET**)

STATE: state where the product needs to be shipped

COUNTRY: Country where the product to be shipped.

DEALSIZE: Size of the product.

And so on…

STEPS:

- **Import Libraries**

1. Import necessary libraries
2. Read the provided dataset.
3. Display the top five samples of the dataframe and understand different features.

- **Data Analysis and Preparation**

4. Check the shape of the data (number of rows and column). Check the general information about the dataframe using .info() method.

5. Check the statistical summary of the dataframe and write your findings.

6. Check the percentage of missing values in each column of the data frame. Drop the missing values if there are any.

- **Univariate Analysis**

7. Check the frequency distribution of the "STATE" feature and encode its categories as mentioned below.

    a. Convert the labels of the STATUS column to 0 and 1. For Shipped assign value 1 and for all other labels (i.e. 'Cancelled',' Resolved',' On Hold',' In Process', 'Disputed') assign 0. Note we will consider everything apart from Shipped as cancel (i.e. 0).

8. Convert the "ORDERDATE" feature into datetime format.

9. Check the frequency distribution of the categorical features and write your findings. Features: - COUNTRY, PRODUCTLINE, DEALSIZE

10. Use one hot coding and convert the above categorical features into numerical data.

11. Check the distribution of the continuous features and write your findings. Features:- QUANTITYORDERED, PRICEEACH, ORDERLINENUMBER, SALES, MSRP

- **Bivariate Analysis**

12. Display how sales are varying across different months. Use an appropriate plot to show the same. Which month has the highest sales?

13. Display how sales are varying across different countries. Use an appropriate plot to show the same. Which country has the highest sales?

14. Display the correlation matrix and write your findings.

15. Display a pairplot and write your findings.

- **Data Preparation**

16. Drop unnecessary columns and check the final shape of the data

17. Standardize the data.

- **K-means Clustering**

1. Use the elbow method to find the optimal number of clusters for the K-means model.
2. Report the optimal K value and fit the K-means model.
3. Report silhouette score for the above model.
4. Add a column 'cluster' in the data giving cluster numbers corresponding to each observation.
5. Check value counts of the cluster label.
6. Perform bivariate analysis between cluster label and different features like sales, year, msrp, etc.

- **Hierarchical Clustering**

1. Generate the linkage matrix using appropriate distance metric.
2. Display dendrogram and write your findings.
3. Report the optimal value for the number of clusters and fit the agglomerative clustering model.
4. Check cophenetic correlation for the above model.
5. Write your findings and conclusions.