Supervised Learning and Ensemble Techniques

Week 2 Practice Project

Topics Covered:

- **Naive Bayes, K-Nearest Neighbors,Support Vector Machine**

Domain:

Marketing

Objective:

Predicting customer's response to a particular product with the help of a variety of classification models like Naive Bayes, K-Nearest Neighbors, Support Vector Machine on the marketing campaign data.

Problem Statement:

An E-com company has recently run a marketing campaign around its customers. The company has collected various attributes regarding its customers like Education, Marital_Status, Country,& various data regarding the amount of money they spend on buying products for the company's website. Now, For the launch of a new product, the companies want to know whether the customers will respond to that product or not. As data scientists, we have to predict if a customer will respond to a product or not.

Data Description:

The dataset can be found here

Domain: Marketing

Feature Details:

ID: Customer's unique identifier

Year_Birth: Customer's birth year

Education: Customer's education level

Marital_Status: Customer's marital status

Income: Customer's yearly household income

Kidhome: Number of children in customer's household

Teenhome: Number of teenagers in customer's household

Dt_Customer: Date of customer's enrollment with the company

Recency: Number of days since customer's last purchase

MntWines: Amount spent on wine in the last 2 years

MntFruits: Amount spent on fruits in the last 2 years

MntMeatProducts: Amount spent on meat in the last 2 years

MntFishProducts: Amount spent on fish in the last 2 years

Response: Response to the product (Target)

and so on..

The complete feature details can be found in the above mentioned link.

- **Data Loading and Exploration.**
    1. Import necessary libraries.
    2. Display a sample of five rows of the data frame.
    3. Check the shape of the data (number of rows and column). Check the general information about the dataframe using .info() method.
    4. Check the percentage of missing values in each column of the data frame. Drop the missing values if there are any.
    5. Check if there are any duplicate rows.
    6. Remove the extra spaces in the ‘Income’ column name.
    7. Check the dtype of values in column 'Income'. Convert the values in the 'Income' column to numeric format.
    8. Check the basic statistics of the data-frame using describe() method.
    9. Write a function which will take the data frame as input and will plot a bar plot which represents the percentage of distribution of each label of ‘Country’ column.

10. Plot a bar plot which represents all the qualifications of the customers country-wise. What is the percentage of graduates in Spain?

11. Plot a percentage segment graph between the ‘Marital_Status’, and ‘Education’ of customers.

12. Plot a percentage segment graph between the ‘‘Education” and ‘Country’ of customers.

- **Model Building and Evaluation.**

1. Plot a count-plot of the target variable.

2. Drop 'ID', 'Year_Birth', 'Dt_Customer', 'Country', 'Education', 'Marital_Status' columns.

3. Store the target column (i.e.Response) in the y variable and the rest of the columns in the X variable.

4. Split the dataset into two parts (i.e. 70% train and 30% test) using random_state=42.

5. Train a Naive Bayes model and plot the below metrics.
    - Plot the error plots for the naive bayes models on the given alpha values (i.e. alpha=[0.00001,0.0001,0.001,0.1,1,10,100,1000]). Choose the best alpha.
    - Train the model on the best alpha value. Print the accuracy score, classification report, roc_auc curve for both the train and test set.

6. Train a K-Nearest Neighbors classification model and plot the below metrics.
    - Plot the error plots for the K-nearest neighbors models on the given K values(i.e. k_value=[3,5,7,9,11,13,15]). Find the best K .
    - Train the model on the best K value. Print the accuracy score, classification report, roc_auc curve for both the train and test set.

7. Train a Support Vector classification model and plot the below metrics.
    - Plot the error plots for the support vector classification model on the given C values(i.e. c_value=[0.01,0.1, 1, 10, 100, 1000]). Find the best C.
    - Train the model on the best C value. Print the accuracy score, classification report, roc_auc curve for both the train and test set.