ML Hackathon 2020 May 26th

Dear All,

Welcome to TCS Hackathon 2020 May 26th.

Please read the below instructions carefully before starting the Hackathon.

1. Total Time For Hackathon is 120 minutes.
2. Upload only one zip file which should have both code and output.(Generally Jupyter Notebook files((.ipynb files)) having both)
3. Download attached dataset and save on your local machine.
4. Instructions to upload results:
5. Comments, code, output should be within ipython notebook only. No other documents required
6. Upload the file with .ipynb extension (Jupyter Notebook file) (Zip and Upload)
7. Mention any special packages need to be installed
8. Don't upload multiple files.
9. Make one ZIP file with all your .ipynb files and upload into ion (from where you downloaded your question)
10. Zip File naming convention ML_May26_CT/DT Reference number.

Any form of Plagiarism is strictly prohibited and if found, we will not consider your files for evaluation.

Problem Description:-

Churn Analysis in Telecommunication:

A customer can be called as a “churner” when he/she discontinue their subscription in a company and move their business to a competitor.

Prediction as well as prevention of customer churn brings a huge additional revenue source for every business.

Here, we use a telecom customer data set to classify the set of possible customers who are likely to churn.

In Simple terms, from set of inputs variables, need to predict particular customer is staying or discontinues service.

Please follow and perform below steps in detail in ipython notebook.

Step1:- Dataset Understanding.

- Load all the packages.

- Load Dataset.
- Dataset Analysis:-
  - Get the shape of the dataset and print it.
  - Get the column names in list and print it.
  - Describe the dataset to understand the basic statistics of the dataset.
  - Print the first three rows of the dataset
- Target Identification
  - Identify the target variable(s) and print
- Feature Identification
  - In our case by analyzing the dataset, we can understand that the columns like Phone Number might be irrelevant as they are not dependent on call usage pattern.
- Since Churn? is our target variable, we will be removing it from the feature set.
- With these assumptions we extract all the relevant columns required for our classification.

Step2:- Pre-Processing of the Data.

- Categorical Data
  - Identify the categorical variable in the data and print it
- Handling Categorical Data
  - Perform the following tasks as a part of Handling Categorical Data
    - Convert to boolean
    - One hot Encoding
- Missing Values
  - Perform the following task
    - Imputing-Missing Values
- Perform the Standardization

Step3:-Applying Classification Algorithem

- Train and Test Data
  - Split the data for training and testing(70% train,30% test)

- There are various algorithms to solve the classification problems. Code using the below algorithms.
- Decision Tree Classification:
    - random_state=seed
  - Train the model with train_data and train_label
  - Now predict the output with test_data
  - Evaluate the classifier with score from test_data and test_label

  - Print the predicted score

- SVM Classification:
  - Initialize SVM classifier with following parameters
    - kernel = linear
    - C= 0.025
    - random_state=seed

    - Train the model with train_data and train_label
    - Now predict the output with test_data
    - Evaluate the classifier with score from test_data and test_label
    - Print the predicted score
- Random Forest Classifier
    - Do the Random Forest Classifier of the Dataset using the following parameters.
        - max_depth=5
        - n_estimators=10
        - max_features=10
        - random_state=seed
    - Train the model with train_data and train_label.
    - Now predict the output with test_data.
    - Evaluate the classifier with score from test_data and test_label.

Note:- If need, please take proper assumptions in selecting above parameters(other than above mentioned values or any extra parameters) and clearly mention in comments section.

Also, if you know any other algorithm, other than above mentioned algorithms, Please feel free to use.Please mention details in comments.

Step4:-Performance Evaluation Measures.

- Generate confusion matrix.
- Calculate Precision, Recall.

Please write proper comments for every cell of code.Only Upload one zip file.

**********All The Best**********