Department of Computer Science and Engineering

Faculty of Engineering

University of North Texas

Assignment 2 CSCE5300 Spring 2024

Due on or before 15th March 2024.

You need to build a Logistic Regression classifier to predict whether or not it will rain tomorrow in Australia. You must train a binary classification model.

The data set is provided. The shape of the dataset is (145460, 23).

Perform the following actions on the dataset prior to build the model.

1. Explore the dataset to find what are the categorical variables in the dataset. List categorical variable names.
2. Explore how many categorical variables have null values.
3. Get the frequency count of each categorical variable. For instance, how many discrete values in each categorical variable and how many datapoints from each distinct value.
4. Print the first five rows of the dataset.
5. What are the available columns of the dataset?
6. Drop RISK\_MM variable column.
7. Get the summary of the dataset.
8. List first five rows only for categorical variables.
9. Decompose the date field into year, month, and day fields. Then drop the original date field.
10. How many unique locations are there in the dataset?
11. Print the number of times each unique location appears in the dataset.
12. Perform One Hot Encoding for each categorical variable.
13. The RainTomorrow is the label, and all other fields are features.
14. You may use the median value (median imputation) of each field to fill null values if it is a numerical column. Use the most frequent value in case of a categorical column.
15. Normalize each numerical column (bring it to a value between 0 and 1)
16. Train your Logistic Regression model on the training dataset (70% 30% split).
17. Predict the RainTomorrow for the test set.
18. Describe the performance of your model using the confusion matrix and print TP, TN, FP, and FN. In addition, provide the accuracy and F1 score of your model.

Submission Guidelines:

1. Submit your code as a notebook file. Include comments in the file itself. Provide the answers in this file itself. You may use a text cell or keep the actual output of the program as the answer for the question. Please clearly mention the question number in each step that you perform in this file.

Useful Tips:

1. OneHotEncoder (pyspark.ml.feature.OneHotEncoder)
2. VecVectorAssembler
3. pyspark.ml.evaluation.BinaryClassificationEvaluator