My project tasks implementations:

* Date Cleaning:

I’ve developed a python script called ‘data\_cleaning.py’ to clean the datasets from incomplete rows. The script will remove all rows missing 5 or more values also it will remove any row that’s missing one of the main features that we will use during the classification process.

* Data Transformation:

I’ve used python to aggregate the datasets to be all within the same granularity.

* Predictor script:

Most of my work is done in this section. The predictor script ‘predictor-v1.py’ consists of 3 parts as following:

1. Analyze dataset
2. Analyze dataset for specific origin and destination
3. Predict a Flight Fare

The script will prompt the user to choose an operation (from the above list) to perform and it will ask for a model type and an encoding option. The script has changed a lot throughout the project journey because at first, we encoded the categorical features using Label Encoding and the results wasn’t impressive. So, after a lot of research we found out that it’s better to encode the features using One-Hot encoding since our regression model is linear. That’s why the script has two encoding options. Also, it’s good to compare the results of each encoding side by side with the other one.

The following is a brief description of each operation (Option):

Option 1: Analyze dataset

In this part of the script, the entered dataset (must be in .csv format) will be processed and encoded according to the user input. Then, the dataset will be divided into two datasets (80% training set and 20% test set). After that, a model will be created based on the user input and it will be used to train the data and then test model score. Also, the script will display the coefficient of each feature and the overall intercept. I’m using sklearn python library to accomplish this task.

Option 2: Analyze dataset for specific origin and destination

This part of the script is similar to Option 1 excepts that this option will ask the user to enter a origin and a destination airports to analyze the data for the particular rout. Also, it will display all relevant info i.e. score, coefficients and the intercept.

Option 3: Predict a Flight Fare:

This is the interesting part of the script where the user will be asked to enter some data to predict a flight fare. In this option, the whole dataset will be used to train the regression model and the user input will be used to actually predict the flight fare. The results are interesting and accurate for most routs.

Included in the zipped file some traces for the script executions. The file name shows the parameters values enter during the execution.

Option-3-Ridge-OneHotEncoding-ABQ-BWI means the user enter the following:

* Operation = 3.
* Model = 2
* Encoding = 1
* Origin airport = ABQ
* Destination airport = BWI

Script usage:

Run the script with default parameters:

python predictor-v1.py

Run the script with debug mode to display more details:

python predictor-v1.py --debug=True