DSE 6220 Course Project

The purpose of this project is to showcase all of the skills you have learned throughout the course. The objective is to create an ML model with SparkML from any data set of your choice. You will be graded on your entire ML pipeline and the organization of your project. There are a few requirements that I will list below. If you do not have any data that you can use you can always download data from the UC Irvine Machine Learning library (link below).

Your project must be submitted via a ZIP file which will contain all of the artifacts that you used to create your model. The artifacts must include a modular Pythonic codebase, requirements.txt, and a one page writeup of your findings.

<https://archive.ics.uci.edu/>

Requirements

1. Data Engineering
   1. At least two features that you have engineered from the data
      1. This can be a combo of features or from an external source
      2. Example: binning a continuous feature into categorical features
   2. Normalization
      1. If your ML model requires your numerical data to be normalized
   3. One-hot-encoding
      1. If you have categorical features
2. Train-test-validation split
   1. Create a train, test, validation split
   2. Hold out 10% of your observations into a validation set
3. Experiments on multiple models
   1. Do not just turn in a random forest and call it a day, try multiple models and select the model that best fits the data without overfitting
4. MLFLow integration (week 8 lab)
   1. Integrate your model into MLFlow and store a screenshot of your tracking UI in your folder
5. Module requirements
   1. Your folder structure should be setup as follows:

<NAME>\_dse6220project

-- \_\_init\_\_.py file if you are writing modular code

--Data (folder containing your original data and each split as defined in part2)

--Experiments (folder containing notebooks you used to experiment)

--requirements.txt (google how to get this)

--main.py

1) a python file that contains your final ‘production model with MLFLow integration

2) Make predictions on your validation set with your model from MLFlow