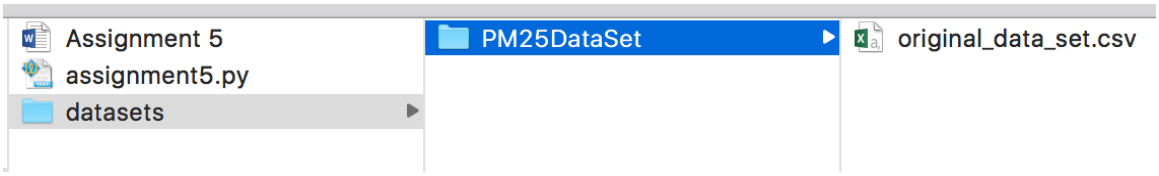
CS452 Due: 4/16/2020
Assignment 4

In this assignment, you will use machine learning to predict PM2.5 particle pollution in Beijing. The data is located at http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data. Your assignment is to develop a python script called **assignment4.py** (you can use Spyder or any other development environment of your choice) to perform the following. Be sure to use a random seed of 10 for all tasks that involve random process.

a) Programmatically download the date and save it on your local drive using the folder and file structure and names as shown below. Be sure to use a function named **fetch_data** and the call it to accomplish the task.



b) Develop a function called **load_data** that would load the data from the **original_data_Set.csv** into a Pandas data frame called **pm25DF**. Run info on the pm25DF and make sure it looks correct.

c) Observe the following:

   a. The first column is a sequence number that will not be useful for the study and it needs to be ignored.

   b. Year has very few values and hence may not be very informative so it needs to be ignored.

   c. pm2.5 is the dependent variable. The variable has missing values. Replacing the missing values for the dependent variable is not generally recommended and is a bit tricky. You will need to remove all the samples with the corresponding missing values.

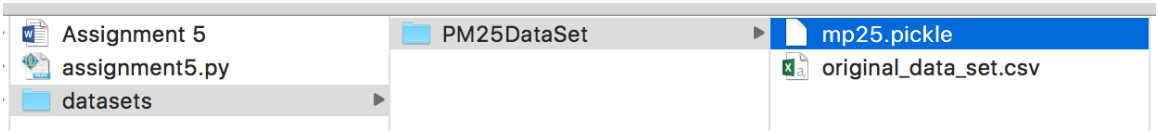
   d. cbwd is a categorical feature and needs to be processed as such.

   e. ls and lr represent the total hours of snow and rain respectively. Having a feature that is the sum of the two may add value. Be sure to add this feature to your data set.

d) Create and run a preprocessing pipeline (similar to preprocerssing_pipeline demo) to process numeric and categorical features. Make sure to remove the samples with missing pm2.5 values before you run the pipeline.

e) Separate the data set into training and test set (20%).

f) Save the preprocessed training and test set in a pickle file called **mp25.pickle** in the **PM25DataSet** directory.



g) Train the following models using 10 fold cross-validation and identify the most promising model. Use the code from experimental_modeling as starting point for this task.

   a. Linear Regression (Use default parameters)

   b. ElasticNet on 2nd degree polynomial

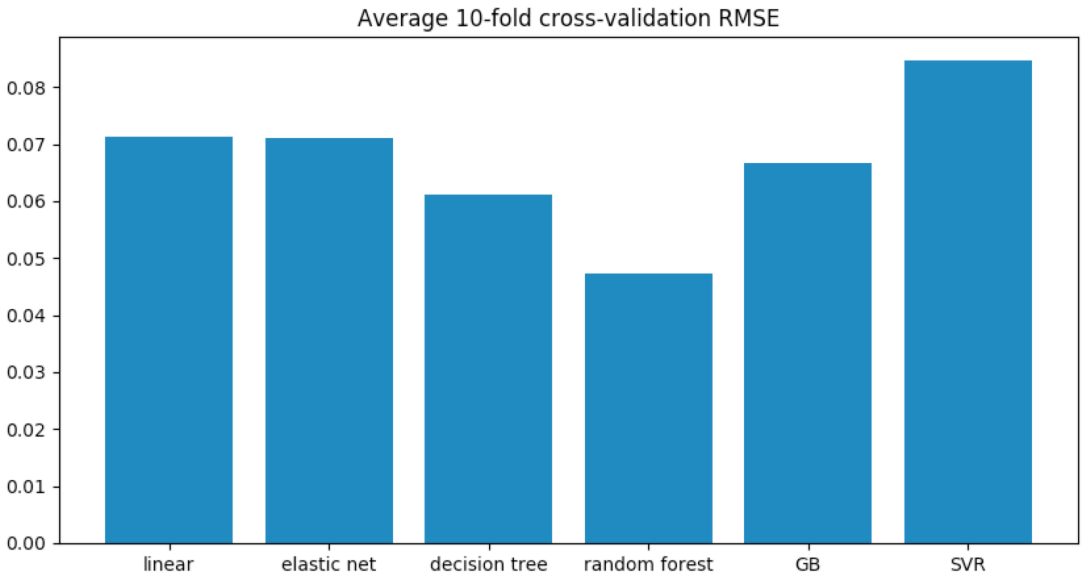
   c. Decision Tree Regression

   d. Random Forest Regression

   e. SVR

   f. Gradient Boosting

   If you have done everything right, the Average RMSE plot should look similar to the following:



h) Fine tune the best performing model using GridSearchCV on two of the numeric hyper-parameters (Note that in practice you will do this for most relevant hyper-parameters). For each of the two hyper-parameter examine the default value and two other appropriate values.

i) Find the RMSE and normalized RMSE of the best model using the best hyper-parameters selected by the grid search.

Submit a copy of your script along with the screen shot of its output.