Capstone Project Instructions

Requirements

The analysis was done with python code inside jupyter notebooks instances in Amazon SageMaker. You will be able to execute the first two ipython notebooks (see below) without Amazon Web Services (AWS). However, if you wish to train the models, it will require you to set up a notebook instance inside AWS with access to a s3 bucket.

All the packages used (Pandas, Numpy, SageMaker) were loaded from mxnet_conda_p36 kernel from AWS. Thus, their version are associated with the python 3.6 . ScikitLearn were installed via pip and its version associated with it.

This folder contains several files that altogether compose the capstone project. There are three types of files: .ipynb files, .py files, .xlsx file and .pdf file.

- Ipynb files: there are 3 files of this type. Each serves a given function. They **must** to be executed in the order described below:
  - Feature_generation.ipynb should be launched **first**. It contains python code for loading the data set available in an online github repo as well as create the feature files.
  - Exploratory data analysis.ipynb should be executed **second**. It contains python code for doing exploratory data analysis. **Notice however that most of the EDA is discussed in the report.pdf not in the notebook.**
  - Training_evaluation.ipynb should be executed **third**. It contains python code for training and evaluating the machine learning models in **Amazon SageMaker**.

  * The empty folders will be filled when the notebooks are executed in this sequence.

- .py files: they are in the folders named source_sklearn. These files contain python code for lauching training jobs in AWS using the scikitlearn estimator.
  - The .py files inside /task1 folder contains the code for random forest model and kernel ridge regressor model.
  - The .py files inside /task2 folder contains code for multi-task random forest model.

- .xlsx file: it is a file inside data/ folder used in the Feature_generation.ipynb when one of the code cells are executed. The cell points to this folder.

- .pdf file: The pdf file is the report detailing the project objectives, methodology, results and conclusions.