- **DataCleaning.py** in Data Cleaning file used the original train and test file from DrivenData to generate cleaned data
- There are some missing longitude and latitude values of certain pumps. From data exploitation, the longitude and latitude of each site is highly related to the outcome of the pumps. Since the locations of pumps can be estimated based on some of the given features, the longitude and latitude of pumps were estimated to fill the missing locations in the first place
- The whole training process contains two stages, the first stage is composed of 6 models (2 XGB, 2 Extreme Forest, 2 Random Forest), the second is composed of 2 models (1 XGB and 1 Logistic). **Stacking-1-Stage.py** and **Stacking-2-Stage.py** are the codes related to the 2 stages. **Blending.py** is the code for the final blending stage. For each model in each stage, there is a corresponding code that can be used for parameter tuning in the **Parameter Tuning** file