Tutorial 4: Random Forest application

In this tutorial, you will learn to apply a Random Forest for the prediction of tomorrow's maximum temperature using other temperature data (e.g. the maximum temperature data one and two days earlier, see Notebook Tutorial_4.ipynb).

The temperature data (in F) are for a station in Seattle over the year 2016. The columns are year: 2016 month: number for month of the year, day: number for day of the year, week: day of the week as a character string, temp_2: max temperature 2 days prior, temp_1: max temperature 1 day prior, average: historical average max temperature, actual: max temperature measurement, friend: a random number between 20 F below the average and 20 F above the average.

(i) Make a plot of the actual maximum temperature over the year 2016, and that of the random (friend) estimate. Is the quality of the data sufficient (check outliers, missing data points, etc.) to proceed for the problem posed?

(ii) Why is 'hot one encoding' used for the days of the week to form the feature list?

(iii) The 'baseline' is used as the 'prediction to beat'. What value of the baseline error is used in the notebook and is this reasonable for the problem posed?

(iv) Study the sklearn RandomForestRegressor class at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. What are the hyper parameters and what are their standard values?

(v) Determine the dependence on the improvement over the baseline error on the number of decision trees in the random forest.

(vi) Determine the importances for the prediction skill; why is the average temperature more important than the temperature two days ahead?