Cherry Blossom Prediction Abstract

James Miller

Simplicity in model design is an often overlooked trait when it comes to predicting what appears to be complex phenomena at face value. The approach taken in this submission is using a Random Forest model with the emphasis on the process being behind engineering the training and testing data. For simplicity, daily weather data was gathered from NOAA from the Reagan National Airport website, this contained the date, average, minimum, and maximum temperature. Features added were growing degree days (GDD) loosely following the GDD model which would square the Celsius temperature if the day was above freezing, then heating units was added by the cumulative sum of the GDD. Lastly the data was engineered such that each column was the heating units on each day from day 1 to 120 while each training row corresponded to the year of the bloom. Then a random forest model was fitted on the DC weather data. Simplicity comes in where the Washington DC weather data was used for the other locations (New York, Vancouver, Liestal, Kyoto) while the actual provided bloom days for those locations were subbed in. This was done in mind working on the assumption that cherry blossom trees in DC will bloom with the same temperatures or heating units as cherry blossom trees at other locations. The number of predictors sampled for splitting at each node was chosen for each individual location's model via Out Of Bag estimation. Once the models were trained, individual predictions were made. The intervals were created by keeping all predictions from each of the 10,000 trees (arbitrarily selected) and then the upper and lower quantiles were taken at 95% and 5% respectively.