Word Count: 797

With cherry blossom trees blooming earlier than ever in the past decade due to climate change, forecasting their peak bloom remains a formidable challenge. The National Parks Service underscores the difficulty of this task, stating that predicting peak bloom is nearly impossible more than 10 days in advance due to the trees' sensitivity to weather conditions.

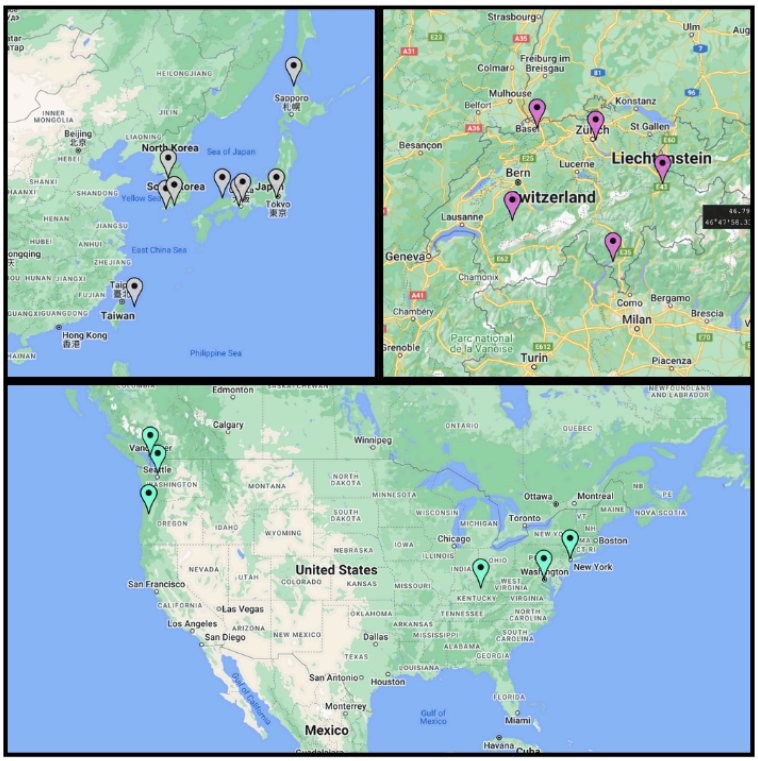
Given this inherent challenge, our team set our sights on creating a model that allows us to predict peak bloom in as little as two months in advance. For this task, we utilized over 245 unique features of the previous year's weather data to predict peak bloom. We successfully trained a Gradient Boosting Regressor (GBR) model which allows us to predict the peak bloom date for any location using only the previous year's weather data.

Notably, our approach utilizes “representative locations”: programmatically-selected locations that represent the inherent geographical diversity in the dataset. These locations allow us to lessen the time required to source data, while minimizing any resulting negative impacts on the model’s performance.
Our model was then used to predict the peak bloom date intervals for: NYC (USA), Washington, D.C., Vancouver (Canada), Kyoto (Japan), and Liestal-Weideli (Switzerland).

The data provided by the event organizers will form the foundation of our training dataset. Some files – like Switzerland and Japan – are already in our ideal format. However, the USA-NPN dataset needed some additional formatting before it was usable. In addition to the re-formatting of this file, we also needed to isolate the data for NYC. Therefore, two dataframes were created; one for NYC and another for all other locations in the US.

Amongst the supplied datasets, there were data for numerous locations from all across Asia, Europe, and North America. While it would be ideal to have extensive background data from all locations, given the time constraints of this competition we instead took a different approach. In order to emulate the geographical diversity of the dataset in as few locations as possible, we identified so-called "Representative Locations" for which we acquired extensive historical data.

The Representative Locations were selected using a KMeans clustering algorithm, which clustered all available locations based on their latitude and longitude. Once clustered, we designated the location with the most number of observations as that cluster's Representative Location. From this, we were able to preserve the geographical diversity of the dataset in as few locations as possible. For consistency's sake, we kept a constant k-value of k=5, giving us five Representative Locations for each dataset. We did this partly to ensure that no singular area was overrepresented in the supplemental data.



Disparate sources for historical weather data cause a lot of issues when trying to create a cohesive dataset for multiple locations across multiple jurisdictions. Our search for a single data source brought us to the publicly-accessible database called Visual Crossing (no affiliation) whose data spans the world-over and has ample coverage for our use-case. In total, we gathered data for the above-mentioned Representative Locations, as well as the five locations included in the final prediction.

When exploring the data offered by Visual Crossing, we identified candidate features based on related domain knowledge to this problem space. We know that cherry blossoms require specific conditions in order to bloom, and so we sought to quantify those conditions through the data we sourced.

Through this domain knowledge, we identified the following as areas which most-greatly affected peak bloom:

- Temperature
- Dew point
- Humidity levels
- Precipitation
- Wind conditions
- Atmospheric pressure
- Cloud cover / Sunlight

We initially reviewed four candidate models for this prediction task: SVR, Linear Regression, Random Forest Regression, and Gradient Boost Regression. Following an analysis of each model's performance, it was decided that we would use a Gradient Boost Regression (GBR) model. Using Cumulative Feature Importance with a threshold of `0.90`, we conducted our feature extraction and narrowed our feature space from 240+ features down to the 34 most-important features. Finally, we used Grid Search to perform our HPO, evaluating approximately 18,000 unique combinations of hyper-parameters based on their R2 values.

~~~

Despite the National Parks Service's assertion of the near-impossibility of accurate predictions more than 10 days in advance, our team endeavored to push the boundaries by creating a model capable of forecasting peak bloom up to two months ahead. Through meticulous analysis and experimentation, we developed a Gradient Boosting Regressor (GBR) model that not only enables us to predict peak bloom dates with impressive accuracy but also offers the flexibility to adapt to different locations worldwide.

A key innovation in our approach lies in the concept of "Representative Locations," strategically chosen to encapsulate the geographical diversity of the dataset while streamlining data sourcing efforts. Leveraging advanced clustering techniques, we identified these locations, allowing us to maintain model performance while minimizing the burden of data acquisition.
~~~