Cherry Peak Bloom Competition 2024

The challenge of predicting cherry tree blossoms isn't just about creating predictive models. It's also an opportunity to identify factors that can directly or indirectly impact tree growth each year. One of these factors, as mentioned on the competition's website, is climate change. That's why the data sources we used for our model involve variables like temperature, pressure, humidity, and precipitation levels to describe climate behavior.

After analyzing the data, we decided to process it daily. This decision significantly increased the number of predictor variables. So, we chose to work with a model capable of effectively handling high-dimensional data, such as Gradient-Boosted Trees. Besides being suitable for the competition's scenario, this model is robust against overfitting, which is crucial since accuracy is a goal of the competition. On a personal note, it was my first time using this model, and I found the experience very enriching.

Data Description

The data available on the competition's website unfortunately appears unbalanced. Cities like Vancouver and New York have few or no records at all. That's why we decided to use a single model for all cities. However, the external data collected only covers from January 1st, 1981, to January 27th. This led to data loss, particularly for Kyoto, which has many more records than other cities.

The variables used for this project include surface pressure (kPa), earth skin temperature (Celsius), temperature at 2 meters (Celsius), specific humidity at 2 meters (g/kg), relative humidity (%), wind speed at 10 meters (m/s), and corrected precipitation (mm/day). Both the response and predictor variables are numerical, so it makes sense to work with a regression model for this assignment.

Preprocessing and Feature Engineering

We decided to prioritize climate data for each year and location. Since cherry trees usually blossom at the beginning of spring, we chose to relate data from December 1st of each year to February 27th of the following year. The idea was to analyze how the climate behavior of the previous season affects the time it takes for cherry trees to blossom. To implement this strategy, we needed to filter and remove data for which weather metrics couldn't be collected.

The collected climate data had to be processed to display it daily in different columns. In total, approximately 620 more columns had to be added. Another important aspect regarding external data is that although there is data for the year 1981, all observations for this year were eliminated because data from 1980 couldn't be obtained.

Lastly, the most crucial step was to add the context of climate behavior to the cherry blossom data. This operation was accomplished using a left-join operation.

Model Selection and Training

For this competition, only one model was developed. However, it was necessary to run it multiple times to determine the most suitable hyperparameters. A clear example of this was determining the number of iterations or estimators needed. Based on the Deviance plot, 250 was defined as the best option for that parameter. In fact, it's the point where the Deviance values for the training data stop changing drastically.

One of the most relevant hyperparameters when defining the model was the loss function used. To calculate the point estimate for the day of the year when cherry trees will blossom, the default value "squared error" was used. On the other hand, to calculate confidence intervals, "quantiles" had to be used to determine both lower and upper values for the 95% confidence interval.

For the model training stage, it was decided to randomly distribute the data with a 70% distribution for training and 30% for testing. The reason for not using a block distribution of time series was that the predictive variables already consider the aspect of changes over time.

Evaluation Metrics

The chosen variable to evaluate the model's performance was the squared mean error (RMSE). After applying the model to the testing data, an RMSE of approximately 6.8277 days was obtained. Initially, when monthly data was used instead of daily data, the RMSE levels exceeded 9 days.

To determine the hyperparameter values that improve our model's fit to the data, we relied on deviance. A lower deviance value indicates that our model better describes the variance in the target variable. We obtained deviance values for the testing data set, ranging between 48 and 50 for the test data.

Results and Discussion

As it was mentioned before, our model obtained a RMSE value of 6.8277 days in testing data. The most relevant predictors that supported predictions made by the model are “temperature at 2 meters on February 22” and “specific humidity at 2 meters”. This is interesting because both belong to the last week of January and February.

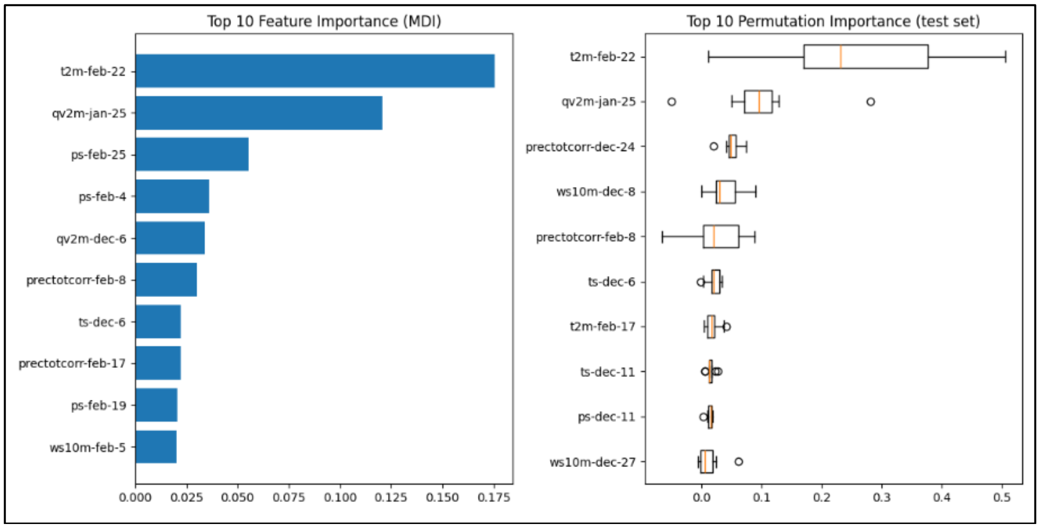


Figure 1. Top 10 more relevant variables when it comes to prediction using Gradient-Boost tree.

When comparing historical value for "Day of the Year" against "Predicted Day of the Year". We expected to see several points close to the diagonal of the plot. In fact, we were able to see that pattern in our scatter plot. Our model is not perfect, so there are some points that do not follow this pattern.

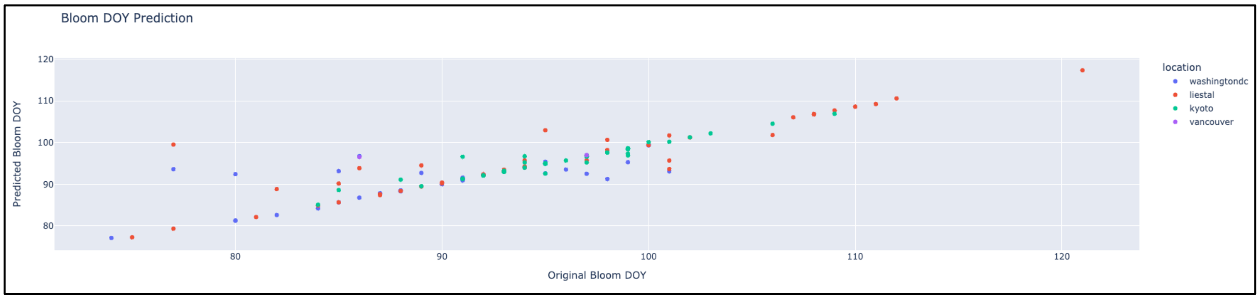
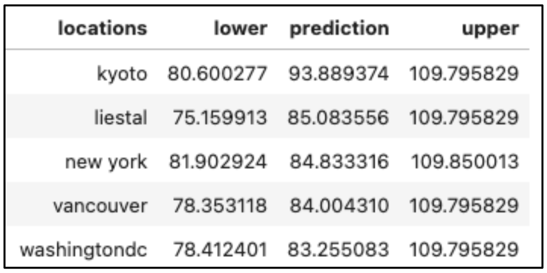


Figure 2. Day of the Year Vs Predicted Day of the Year

The table below shows the point estimates and 95% confidence interval for each location. Out of the 5 locations, Vancouver and New York are the 2 cities that did not contribute with samples values to model the relationship between climate behavior and day of bloom.



locations	lower	prediction	upper
kyoto	80.600277	93.889374	109.795829
liestal	75.159913	85.083556	109.795829
new york	81.902924	84.833316	109.850013
vancouver	78.353118	84.004310	109.795829
washingtondc	78.412401	83.255083	109.795829

Figure 3. Point Estimate and 95% Prediction Interval