Cherry Blossom Prediction Journey

Cherry blossom peak blooms are hard to predict due to weather variability, regional variability, genetic variability, human impact, and lack of historical data; they also vary heavily on weather conditions including temperature, precipitation, and sunlight; these all can vary year to year, making it difficult to accurately predict bloom dates far in advance. Specific regions can be especially hard to predict if there is no long-term weather data available. Some traditional methods of predicting cherry blossom peak blooms in the past have been heavily reliant on trends within phenological observations, weather forecasting, traditional calendars, historical records, and simple linear regressions. This is seen in a variety of examples ranging from tracking when buds start to swell to the ancient Japanese tracking via lunar calendar and associating bloom with full moon in early April, to even trying to match current data with the most similar occurrences in the past expecting a repeat in history.

By analyzing these along with bloom dates, machine learning algorithms can identify correlations and make predictions based on those patterns; this can be seen already in other fields with MARS models being used to build predictive models for a wide range of applications, including finance, healthcare, and marketing to name a few. In addition, while the trend of global warming is clear and unequivocal, the rate and nature of this warming is complex and non-linear, and requires careful study and analysis to fully understand and predict.

Machine learning algorithms can process and analyze large amounts of data and identify patterns that may not be immediately visible to humans. Additionally, MLA can learn and adapt over time as new data becomes available, allowing for continuous improvement of accuracy of the predictions. They are well suited for this task of predicting peak bloom dates due to the size of the data and ability to identify complex patterns, adapt to new information; this is essential when analyzing natural phenomenon that is influenced by many factors. The MARS (Multivariate Adaptive Regression Splines) model is a powerful statistical learning algorithm that can be used for both regression and classification tasks. It is particularly useful when the relationship between the predictors and the response variable is nonlinear and/or interactive; it is designed to capture nonlinear relationships between the predictors and the response variable, which makes it particularly useful when linear models such as linear regression are not sufficient.

The method used here was the “earth” package which implements the MARS algorithm. Firstly, seven different bloom spots were analyzed; each data set contained latitude, longitude, altitude, and bloom_doy variables; as well as their respective year. This data was pre-cleaned, organized, and all had the same variables and measuring method; it was a great starting point as this provided a plentiful amount of data and a way to compare possible explanatory factors. From there, the data was all combined into one sheet in order to run the MLA algorithm and find some connection between geographic location, time, a possible uniform regression along all sites’ days until peak bloom. Once the model was created, this was then taken and applied in a closed system of each individual site and further evaluated the model within each situation.

Next, we split each site up into training and testing data for the model. The purpose of splitting data into training and testing was to evaluate how well a model can generalize to new, unseen data. By evaluating the model's performance on a separate testing set for each site, we can estimate how well

the model will perform on new, unseen data. It was important that this was planned before making the initial overall model with the MLA; the testing data was taken out of the initial resources when creating the model. This is important because if we use the same data for both training and testing, the model may overfit the training data and perform poorly on new data and then also not truly predict but rather "plug and chug" old values. Overall, the model showed promising residual charts and distribution of mean errors; a highlight of a strong R-squared value, indicating a high amount of variation explained for by the model.

This model was based on large assumptions though; a route chosen because it was unorthodox and largely unseen in prior research / brainstorming phases. This model is largely excluding temperature, emission, natural gas, or any other variables other than year and geographical coordinated alone. The thinking was that each geographical region has their own weather patterns, extremes, and terrain; in addition, they all also have their own highs and lows but share a same growing season roughly for cherry blossoms. The thinking was that focusing on raw numbers will skew the model as the four sought after sights have completely different precipitation patterns, median temperatures, and emission totals; however, their bloom dates aren't as different as expected. The final decision was to assume normal distribution within the season for all sights and make predictions by only taking into account the overall regression by year based on geographical coordinates. /Global warming is globe wide, and all sites aren't in a vacuum. With each being affected and assuming normal or nonextreme changes that happen year to year, the focal point predicted the next assumed change with complete neglect to outside variables mentioned above. The predictions don't seem to be outliers when compared to the prior data, and the RMSE wasn't incredibly worrisome either from first glance, so that stuck as the focus.

The flowering bloom dates of cherry blossom trees can vary from year to year and are influenced by a range of factors such as weather conditions and climate. According to the National Park Service, the average peak bloom date for cherry blossoms in Washington, D.C. is around April 4th. In Vancouver, the cherry blossoms typically bloom in late March or early April, depending on the weather conditions. This means that on average, the cherry blossom bloom dates in Vancouver are typically a few days to a week earlier than in Washington, D.C. However, on average, the flowering bloom dates in Vancouver, Canada are generally earlier than in Washington, D.C., United States; this was my reliance or assumption when checking or verifying the Vancouver predictions with virtually no data; with that condition satisfied was enough.

This model and MARS models did prove to have some limitations to be aware of such as overfitting. As with any modeling technique, MARS models can suffer from overfitting if they are too complex or if the number of predictors is too large relative to the number of observations. This can be especially true for higher-order MARS models that include more interaction terms. MARS models are also designed to fit additive models, which means that they assume that the effect of each predictor on the outcome is independent of the other predictors. This may not be appropriate for all types of data and may limit the flexibility of the model in some cases. Another factor which exacerbated this and can very easily skew the entire spread was its' sensitivity to outliers. MARS models can be sensitive to outliers in the data, especially if the number of observations is small, such as with one of the locations, Vancouver, which only had one observation.

This model hasn't been used plentifully throughout my research in this field, but these types of models can be a useful tool for modeling cherry blossom bloom dates and other complex environmental processes. They offer lots of promise, and they prove to have some advantages over other modeling techniques, such as their ability to handle both continuous and categorical data; it is a unique perspective and method to use when analyzing the data to even open the door into finding new ways to analyze such a delicate system.