Cherry Blossom Prediction

Johnson Wei, Hanji Sun, Max Xu, Erin Ma, Yi Pan

February 28, 2023

1 Introduction

The blooming of cherry trees is an important aspect of phenological study, and researchers often analyze records of blooming dates in relation to the preceding spring and winter temperatures. This relationship between temperature and cherry blossoming is two-fold, with some studies using blooming dates to reconstruct spring temperatures in Japan dating back to the 9th century, while others focus on using observed or projected temperature to predict cherry tree blooming dates. We built multiple supervised machine learning models to predict the peak bloom dates of cherry trees in Kyoto, Japan, Washington DC, USA, Vancouver, Canada and Liestal, Switzerland.

2 Data collection

The initial data sets for our study were obtained from the Github repository of the George Mason University's Department of Statistics cherry blossom peak bloom prediction competition. These data sets included peak cherry blossom dates in Kyoto, Liestal, Washington DC, and various cities in Switzerland and South Korea. Additionally, there are data sets from the USA National Phenology Network (USA-NPN), which contained information on accumulated growing degree days, as well as average, maximum, and minimum temperatures in winter and spring, and accumulated precipitation in winter and spring for individual cherry trees located in different states across the United States. As the development of cherry blossoms is influenced by geographical location and local climate, we extracted relevant features from the USA-NPN data set and obtained climate data from the National Oceanic and Atmospheric Administration (Chamberlain 2021).

We obtained daily data on maximum temperatures, minimum temperatures, and precipitation and calculated the average values for each year. Since most cherry trees typically blossom in the spring, we considered weather data from the previous year's winter as a reasonable predictor for cherry blossom peak bloom dates. Thus, we included the average maximum temperatures, minimum temperatures, and precipitation from December of the previous year and January and February of the current year in our analysis. Moreover, The Accumulated Growing Degree Day (AGDD) data was extracted from the USA-NPN dataset.

Apart from the features obtained from the USA NPN data set, we think that global warming has caused cherry blossom dates to occur earlier in recent years. Thus, we retrieved the carbon dioxide emission data from a study published on Our World in Data.

3 Methods

3.1 eXtreme Gradient Boosting

XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm used for regression and classification tasks. It is a type of ensemble method that combines multiple decision trees to form a powerful predictive model.

In XGBoost, decision trees are created sequentially, with each new tree attempting to correct the errors of the previous tree. This process is called boosting, where multiple weak models are combined to form a strong model.

XGBoost is designed to minimize the loss function of the model by using a gradient descent algorithm. The gradient descent algorithm involves finding the direction of steepest descent to minimize the loss function. This allows XGBoost to learn from the errors of previous trees and adjust its predictions accordingly.

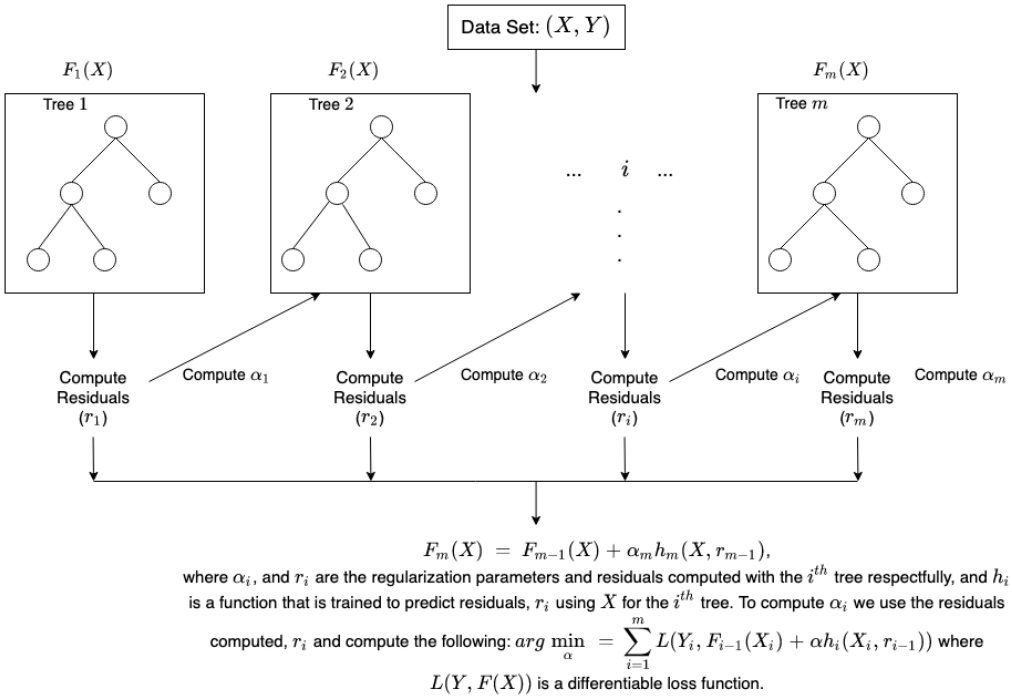

Figure 1: An illustration of the XGBoost model

3.2 Convolutional neural networks

Convolutional neural networks (CNNs) are a type of deep learning algorithm used for image recognition, natural language processing, and other tasks. CNNs are inspired by the structure and function of the human visual cortex, which processes visual information.

CNNs are particularly suited for analyzing time series data, which is essential for understanding the impact of environmental factors on cherry blossom blooms.

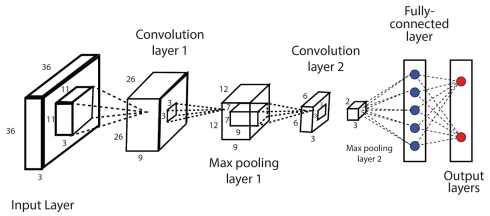

Figure 2: An illustration of the CNN model

4 Results and conclusion

We use the MSE as the metric to compare the accuracy between XGBoost and CNN models. In conclusion, the XGBoost, a decision tree model, has shown greater accuracy which has MSE=32 than the CNN model which has higher MSE=40.

However, it is not appropriate to make a general statement that XGBoost is better than CNN for predicting the cherry blossom date, as the effectiveness of a machine learning algorithm depends on the specific characteristics

of the problem and the available data. In general, CNNs are better suited for image recognition and analysis tasks where the input data has a spatial or temporal structure, while XGBoost is more appropriate for structured data where the features have a clear linear or nonlinear relationship with the target variable.

For predicting cherry blossom dates, we have used various types of input data such as historical climate data, geographic information, and other environmental factors. If we have structured data that can be represented as a table with clear features, then XGBoost might be a good choice as it can capture complex nonlinear relationships between the features and the target variable. On the other hand, if we have image or time-series data, where the input has a spatial or temporal structure, then a CNN or another type of deep learning model might be more effective.

year	kyoto	liestal	washingtondc	vancouver
0	2023	94.0	86.0	87.0
1	2024	90.0	83.0	82.0
2	2025	106.0	79.0	81.0
3	2026	86.0	92.0	89.0
4	2027	103.0	93.0	86.0
5	2028	87.0	84.0	81.0
6	2029	97.0	75.0	85.0
7	2030	96.0	86.0	94.0
8	2031	102.0	85.0	82.0
9	2032	88.0	87.0	85.0

Figure 3: Our prediction results

5 Discussion

While the precise mechanism behind cherry blossom is complex and difficult to ascertain, our study utilized suitable models that accounted for various data features. However, due to the inherent complexity of this phenological event, there are certain limitations to our models, which could be mitigated by increasing the sample size and refining the prediction of temperature for the upcoming decade. Additionally, it would be beneficial to incorporate more climate information to explore the potential relationship between cherry blossoms and global warming. Overall, although there are limitations to our study, our findings provide valuable insights into the prediction of cherry blossom peak bloom dates.

6 References

Anna Veronika Dorogush, Andrey Gulin, Vasily Ershov. 2018. “CatBoost: Gradient Boosting with Categorical Features Support.” ”https://arxiv.org/abs/1810.11363”.

Cherry Blossom Festival (U.S. National Park Service). https://www.nps.gov/subjects/cherryblossom/bloom-watch.html.

Graham, Karen. 2021. “Climate Change Likely Cause of Japan’s ‘Earliest Cherry Blossoms’.” Digital Journal, March.

Harris, Charles R., K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. “Array Programming with NumPy.” Nature 585: 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. “Lightgbm: A Highly Efficient Gradient Boosting Decision Tree.” Advances in Neural Information Processing Systems 30: 3146–54.

Mohamed Farouk Abdel Hady, Friedhelm Schwenker, and G¨unther Palm. Semi-supervised learning for regression with co-training by committee. In International Conference on Artificial Neural Networks, pages 121–130. Springer, 2009.

Uran Chung, Liz Mack, Jin I Yun, and Soo-Hyung Kim. Predicting the timing of cherry blossoms in washington, dc and mid-atlantic states in response to climate change. PloSone, 6(11):e27439, 2011.