Seeing the Sakura Bloom

Alexandra Schoenberg

Springboard Data Science Career Track, 2022-23

The Problem

- One of Japan's most beautiful and popular attractions, for visitors and locals alike, is the blooming sakura flowers
- These flowers bloom only for a short period of time in the spring, and correctly timing a trip to see them can be tricky

Can we use ambient temperature data to create a model that will accurately predict the "best" day to see the sakura flowers in peak bloom?

Who Might Care?

Travelers from far and wide

Local families enjoying a picnic

…and anyone with an appreciation for nature or Japanese culture!

What Factors Might Impact the Bloom?

- Weather and nature are very complicated and intertwined
- The annual bloom date could possibly be affected by:
  - Temperature
  - Rainfall
  - Sunshine
  - Air quality
  - Wind patterns
  - Natural disasters - Japan is subject to earthquakes, tsunamis, and even volcanic activity
  - Etc.

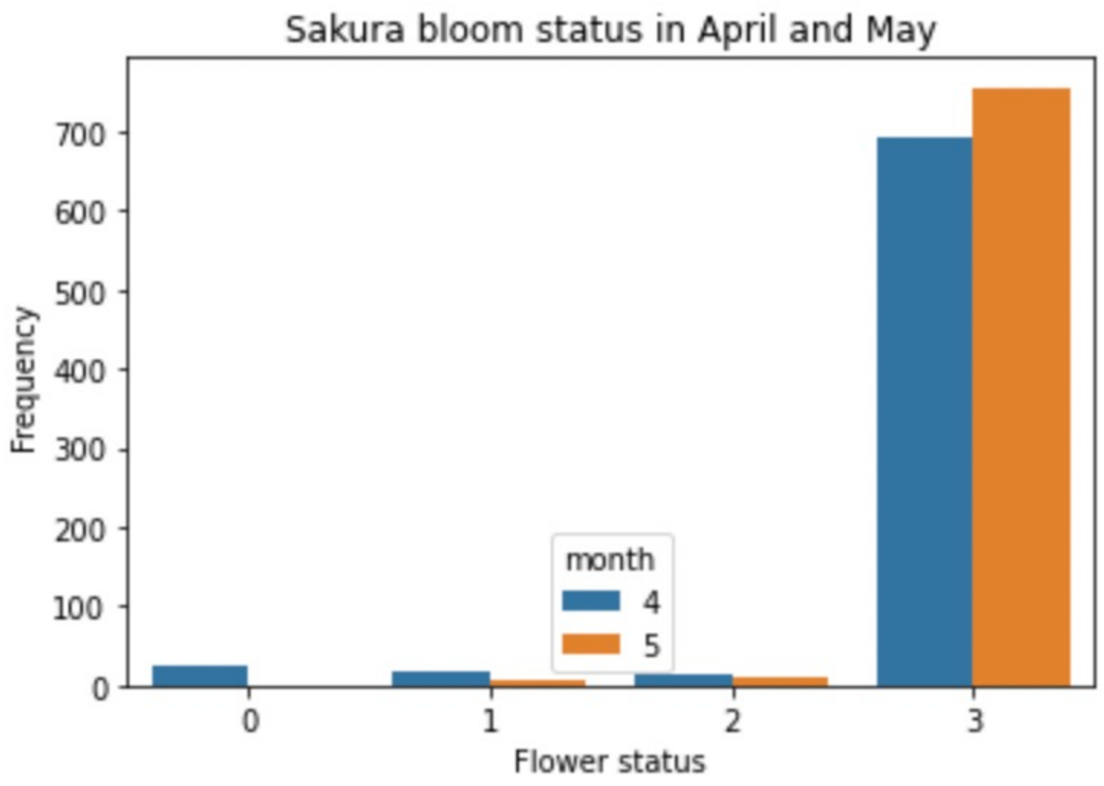
Data Information

- Ambient temperature data, sourced from the Japanese Meteorological Agency
- Sakura bloom status data, specific to Hirosaki Park in Hirosaki City, sourced from the Hirosaki City Green Association
- Data acquired for the period of Jan 1, 1997 - Dec 31, 2019
  - Initial dataset contained 3 columns (date, temperature, bloom status) and 9131 rows
  - File format: csv
  - Each record: an individual day

Data Wrangling

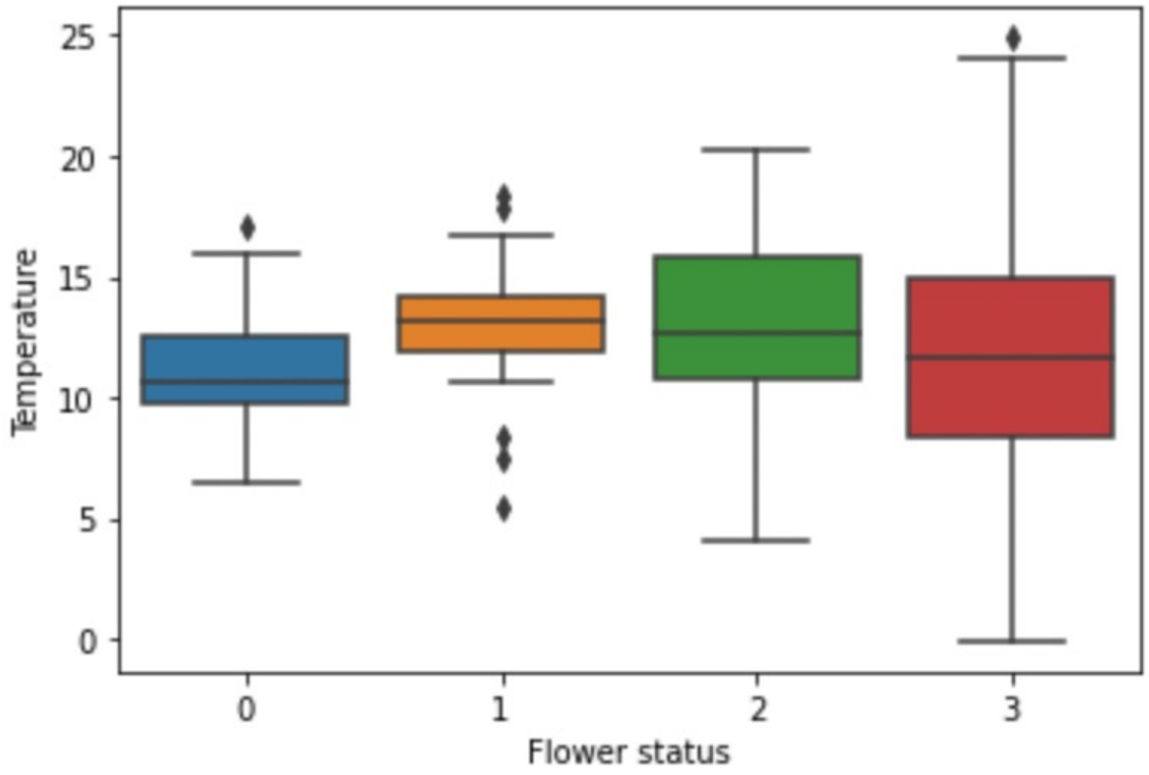
- Date values were split into 3 columns: day, month, year
- Bloom status values were labeled and imputed
    - 0: full bloom
    - 1: bloom
    - 2: scatter
    - 3: no bloom
- Historically non-blooming months were removed, leaving only April and May data
- Result was a cleaned DataFrame of 5 columns and 1525 rows

Data Exploration



The first figure indicates that, even in the blooming months, the bloom period itself is quite short.

Relationship Visualization



The second figure, a box and whisker plot, provides a visualization of the temperature ranges and medians with respect to blooming status, within the months of April and May.

Modeling

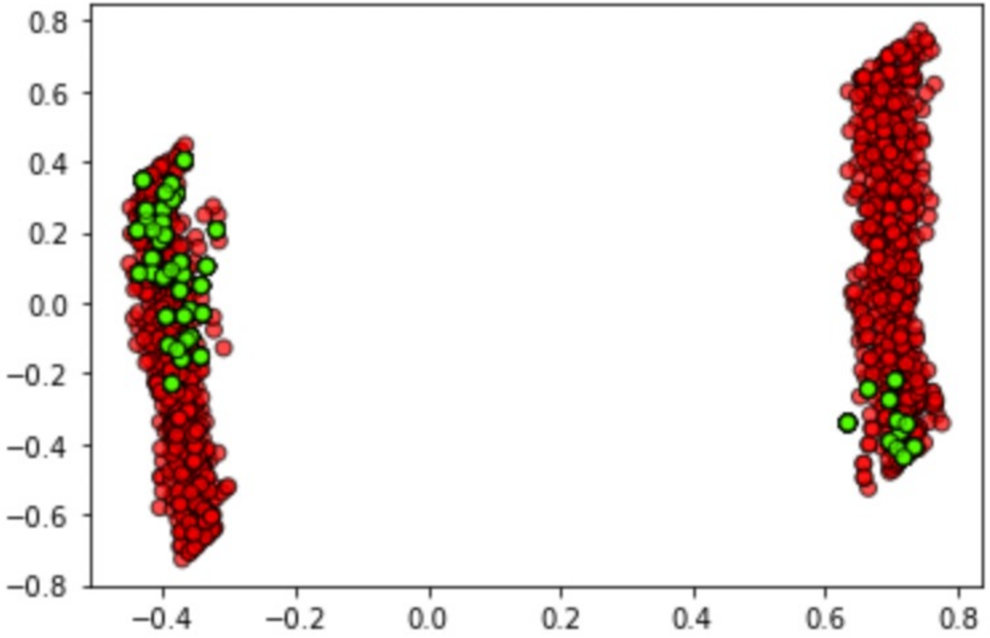
- Decision Tree Classifier model with binary categories
  - All blooming statuses (full, bloom, scatter): 1
  - Non-blooming status: 0
- Features were magnitude standardized
- Each model output included a Principal Component Analysis scatterplot and a Classification Report, providing precision, recall, F1, and accuracy scores

Modeling

- Because the data was extremely unbalanced, multiple models were created using different sampling methods:
  - imbalanced (baseline model)
  - oversampled
  - undersampled
  - oversampled with SMOTE technique
- The most successful model was the oversampled model

Model Results

- Oversampled model PCA



- Oversampled model Classification Report

	precision	recall	f1-score	support
0	0.96	0.97	0.97	363
1	0.35	0.32	0.33	19
accuracy			0.94	382
macro avg	0.66	0.64	0.65	382
weighted avg	0.93	0.94	0.94	382

Takeaways and Future Steps

- This model demonstrated that the correlation between temperature and blooming was not very strong, and that many more features and factors are likely at play in determining the blossoming timeline
- Although the model was not very successful, the project was a good lesson for me with regards to data collection and project design
- In the future, I will build upon this foundation to select more comprehensive data and asking more discerning questions