# III. Model documentation and write-up

Information included in this section may be shared publicly with challenge results. You can respond to these questions in an e-mail or as an attached file. Please number your responses.

1. Who are you (mini-bio) and what do you do professionally? If you are on a team, please complete this block for each member of the team.

   I am a bioinformatic engineer with a master's degree of Biochemistry. I also interested in take part in various data science competition. Have a competition grandmaster tie at Kaggle and also have some wins at Chinese domestic competitions.

2. What motivated you to compete in this challenge?

   I am interested in geography and this challenge is friendly to amateurs.

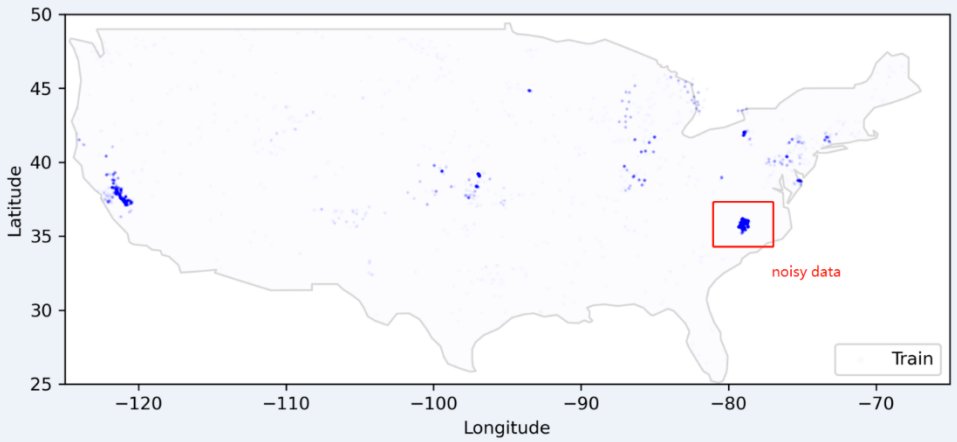
3. High level summary of your approach: what did you do and why (e.g., key features, algorithms you used, any unique or novel aspects of your solution, etc.)?

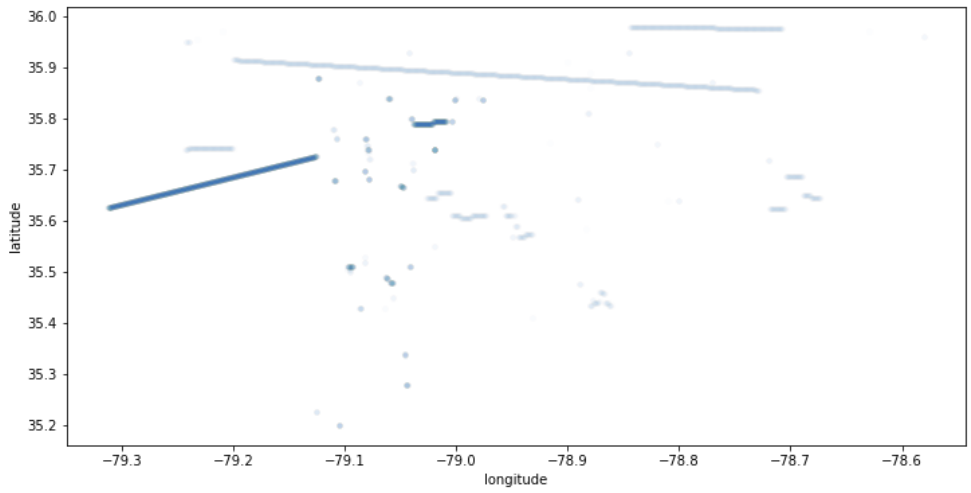
   My approach is data-driven. I select my method based on the quality and properties of the data in different regions. For the South and West regions where the data quality is somewhat low, I use a simple KNN method based on latitude, longitude. For data points in the Midwest and Northeast regions, I use a GBDT model with temperature and water region color features

4. Do you have any useful tables, charts, graphs, or visualizations from the process (e.g., summarizing model performance, testing different features, exploring the data, etc.)?

   Please refer to the EDA, and feature statistical below.

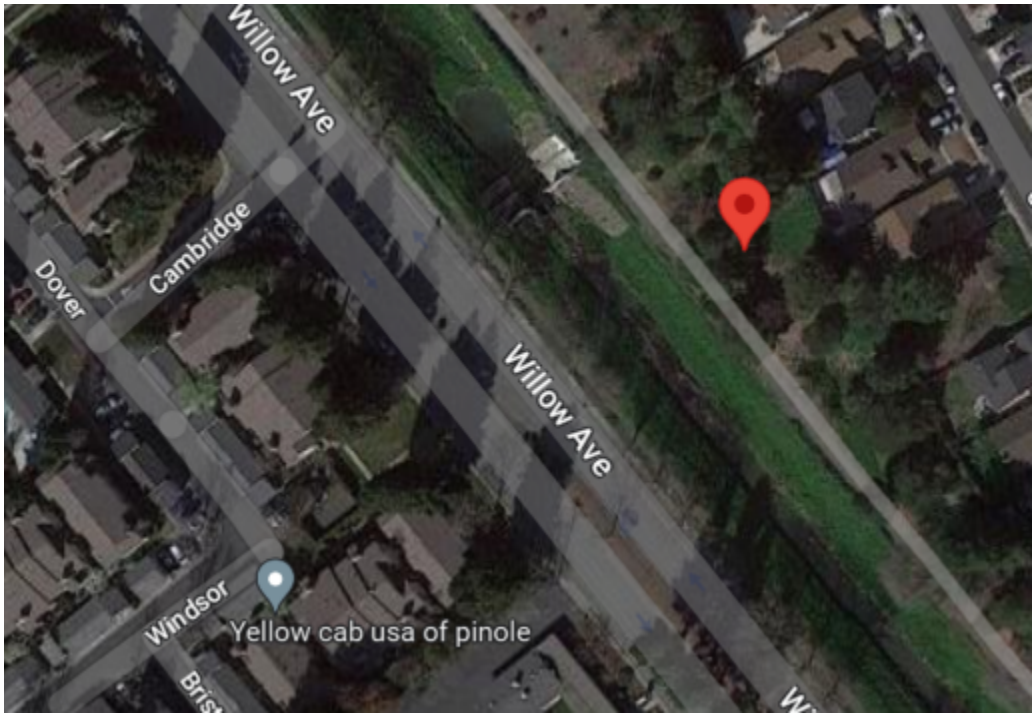
   A. Snapshot of noisy location of the South train dataset. In the figure below, there are 8589 data points in the red box, most has wrong location.



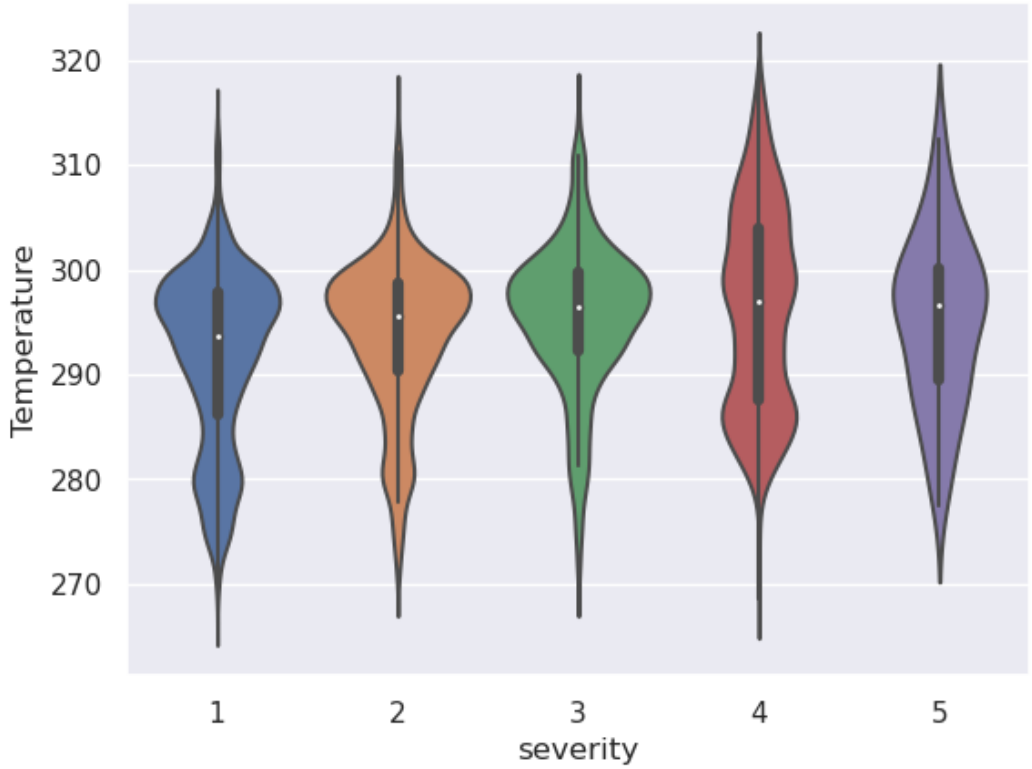




B. Typical data point at West region

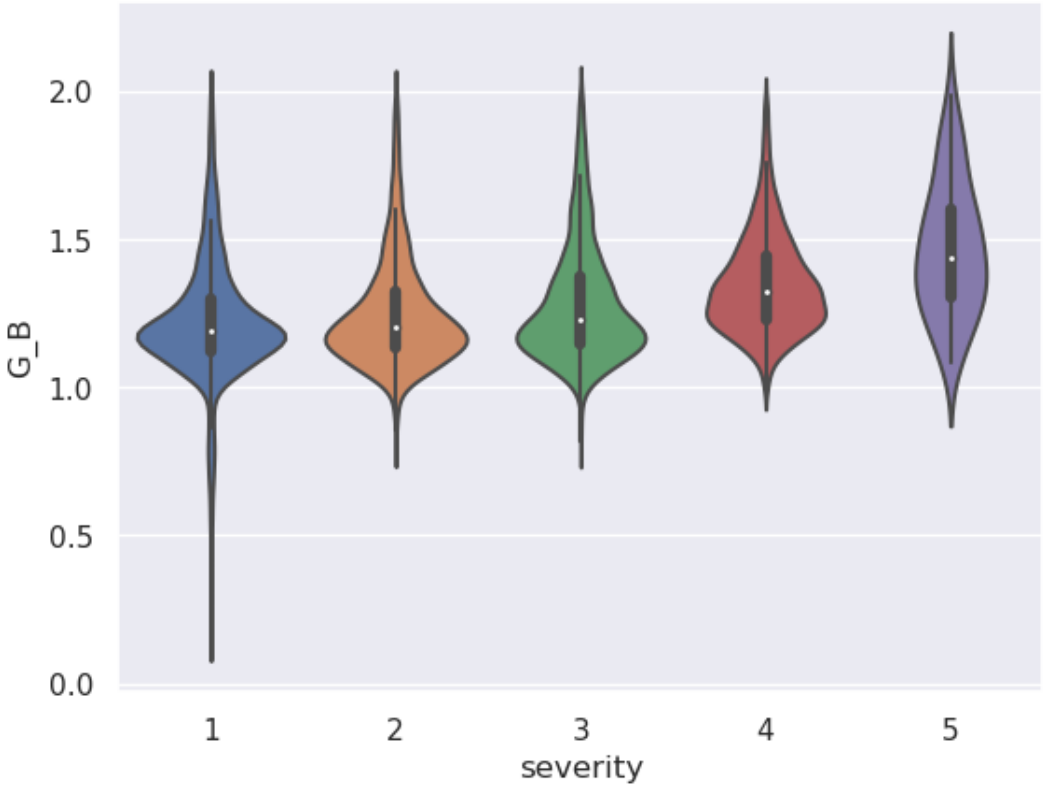
Many datapoint at west region are small river, make this region very hard to modelling with the limitation of the satellite image resolution.




C. The relationship between the temperature and the severity



D. The relationship between the watercolor and severity




E: brief ablation study of key features

| Features | CV | LB |
|---|---|---|
| Region mean | 1.10 | 0.8750 |
| KNN | 1.03 | 0.7945 |
| LightGBM + nrrr | 0.93 | 0.7580 |
| LightGBM + nrrr + water color | 0.917 | 0.7555 |

5. Copy and paste the 3 most impactful parts of your code and explain what each does and how it helped your model.

1. EDA: I calculated the water region using the CLA layer of the Sentinels image and identified noisy data points

2. For the feature engineering part, I calculated the Green/Blue and Green/Red values using a method similar to some papers..

3. As the dataset is very small, I applied heavy regularization to the LightGBM model using the parameters below..

6. Please provide the machine specs and time you used to run your model.
   - CPU (model): AMD Ryzen 5950x
   - GPU (model or N/A): RTX 3090 (not used)
   - Memory (GB): 128GB
   - OS: ubuntu server 20.04
   - Train duration: ~1h (2days to download extra dataset)
   - Inference duration: ~10min

7. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, failure modes, etc.)?

   I suggest using my model only for data in the Midwest and Southeast. The training data in the South and West regions is either low quality or too difficult to fit a good model.

8. Did you use any tools for data preparation or exploratory data analysis that aren't listed in your code submission?

   - Google Earth: to check if the downloaded satellite image is correct..

   - Jupyter Notebook: conducting the Exploratory Data Analysis.

9. What features did you select for use in your model and why? Please describe the preprocessing steps you took to develop features from raw data (e.g., filling missing values, removing outlier values, encoding categorical values, etc.), and what criteria or methods you used to select these features.

   Extra data source:

   - The NRRR data of the date of the datapoint.

   - The Sentinel image, within a 60-day data range, less than 1% cloud image.

   The key features in my model are:

   - Region and Location, these features used to select method.

- Satellite image color features: mean red, blue, green, percental 95% red, blue and green of the water region. The ratio between blue and red, blue and green value.

- The NRRR surface features.

Preprocessing steps:

a. The entire South region is considered as noisy data since most samples in this region have inaccurate location information, as shown in the figure above. It is important to note that using a powerful model with this low-quality data may result in inaccurate predictions. A simple KNN model works best in this region. Although an area-based group k-fold CV may appear to be better for training a tree model in this region, this is due to the unbalanced data label, and using a model trained with unstable data is always risky.

b. The West region is very hard to predict since most of the samples were collected from small rivers that are less than 10m wide, making data collection extremely difficult. Our Landsat data source has a resolution of 60m, while Sentinel has a resolution of 10m. This region requires a higher resolution data source for better prediction accuracy..

c. As for the rest of the regions, the water regions are relatively larger, and the location information is more accurate. Therefore, I used watercolor features and climate features for these regions to improve prediction accuracy. The watercolor features include mean red, blue, and green values, as well as the 95th percentile values of the water region. The climate features include temperature data from the NRRR dataset for the date of the data point.

10. How does your model's performance vary? Please consider model performance for various levels of cyanobacteria density, region, environmental conditions, and season. Describe under which conditions your model is the most accurate and under which conditions your model is the least accurate.

    My model's performance varies depending on different conditions:

    - The best situation is when a datapoint has both NRRR data and Sentinel data.

    - If a datapoint is located in the Northeast or Midwest regions, the performance is acceptable.

    - If a datapoint is located in the South or West regions, the performance is low.

11. How did you evaluate the performance of your model (e.g. how you approached validation, dataset splits, metrics used, etc.)?

To evaluate the performance of my model, first I use a area-based group kfold to wipe out the leakage between the neighbor dataset.

a. To cluster the data, I used K-means algorithm based on the latitude and longitude values. Data points within the same cluster were considered as a group.

b. To group neighboring data points in one-fold, I used Scikit-learn's GroupKFold function.

I used the RMSE metric to evaluate each model's performance. Since this dataset is small and noisy, I used 100 times seed average during training. By doing so, I obtained the RMSE score with 100 repeats, which allowed me to use statistical tests such as t-test to check if certain features or methods significantly improved the model's performance.

12. What simplifications could be made to run your solution faster without sacrificing significant accuracy?

    This LightGBM model is already very fast.

13. What are some other things you tried that didn't necessarily make it into the final workflow (e.g., features, preprocessing steps, etc.)? Did you test any other types of models?

    First, I tried using a CNN model with satellite images. Unfortunately, this type of model resulted in very high RMSEs. After some analysis, I suspect that the quality and resolution of the satellite images, as well as the accuracy of the positions, made it very difficult to fit the CNN model well.

    Second, I am attempting to train a Unet model to segment the water region of the Landsat image, since many data points were collected before the deployment of the Sentinel satellites. However, the resolution of the Landsat images did not meet the requirements for accurate segmentation.

14. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?

    I believe the following points would be helpful for improving the model's performance:

    - A higher resolution of the satellite images could be useful for the West region, which would allow for more accurate predictions.

    - More accurate location information would also help to improve the model's accuracy.

# Write-up rubric

Write-up submissions will be evaluated along the following dimensions for the bonus prize for best write-ups.

**Interpretability (35%):** How comprehensively does the report describe the full model pipeline from data intake to inference (e.g., feature engineering, model selection, training and validation procedures)?

**Insight (25%):** How well does the report describe the exploratory, trial-and-error process that led to the given solution (i.e., the rationale for what was tried and what was ultimately used)?

**Rigor (25%):** To what extent is the report built on sound, sophisticated quantitative analysis and a performant statistical model?

**Clarity (15%):** How understandable is the report for a broader audience (i.e., a subject matter expert without technical experience in python-based ML)?