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 an associate professor at the Delft University of Technology, the Netherlands. My current research interests include turbulence modeling, numerical weather prediction, wind energy, atmospheric optics, and machine learning. Over the past fifteen years, my research has been funded by the US National Science Foundation (including an NSF-CAREER award), the US Department of Defense, the US Department of Energy, EU Horizon 2020 program, Carbon Trust (UK), TKI Wind op Zee (Netherlands), and other organizations. My research has been disseminated through more than 60 peer-reviewed journal publications.

2. What motivated you to compete in this challenge?

   Over the past 2-3 years, I have been dabbling with state-of-the-art machine learning techniques and tools. This airathon was quite attractive as I was more-or-less familiar with the scientific content. I wanted to find out how my AIML skills matched up with top competitors.

3. High level summary of your approach: what did you do and why?

   I used remote-sensing data from the Ozone Monitoring Instrument (OMI) and numerical weather prediction data from the Global Forecast System (GFS). For regression analysis, the LightGBM model is used in conjunction with an ensemble approach.

4. Do you have any useful charts, graphs, or visualizations from the process?
   None

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

   I have spent a lot of time describing my approaches in 3 separate README files. I am not repeating here for brevity.

6. Please provide the machine specs and time you used to run your model.
   - CPU (model): iMac (2020), 3,3 GHz 6-Core Intel Core i5
   - GPU (model or N/A): N/A
   - Memory (GB): 80 GB 2133 MHz DDR4
   - OS: MacOS monterey

- Train duration: ~1.5 h
- Inference duration: 25 sec

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

   If you plan to extract GFS data (tar files are included in repo), please make sure that the extraction directory is not being synched by Dropbox or other services. Each tar file contains more than 100k files; it can tremendously slow down Dropbox. I made the mistake myself.

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

   I experimented with other models (e.g., random forest, TABnet). Since their performance was much poorer than LightGBM, I did not include them.

9. How did you evaluate performance of the model other than the provided metric, if at all?

   MSE

10. What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?

    I discovered FLAML towards the end of the competition. I wish I had more time to experiment with it.

11. 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 intend to publish a peer reviewed article on this topic. I would request NASA and drivendata to release the test labels in the near-future.

    I would like to add TropoOMI data as input to see if the results can further improve.