Determining Key Factors Influencing Ozone Formation in Taichung City, Taiwan Using Machine Learning Models

Jonalyn C. Madriaga[1,2,3], Charles C.-K. Chou[1,2]

[1]Research Center for Environmental Changes, Academia Sinica, Taipei 11529, Taiwan;
[2]Earth System Science Program, Taiwan International Graduate Program, Academia Sinica, Taipei 11529, Taiwan;
[3]College of Earth Science, National Central University, Taoyuan 320, Taiwan
Corresponding email: jcmadriaga1@gmail.com; as0191576@gate.sinica.edu.tw

Supplementary Materials

EGU25-16129ECS

Methodology

1. Instrumentation

Parameter	Instruments
VOCs	PTRTOFMS Ionicon 8000
Ozone	UV Photometer, Thermo Scientific Model 49i
NOx (NO, NO_2)	Chemiluminescence Analyzer API- T200U, API-T500U

2. Machine Learning Model

GridSearchCV was applied to identify the optimal hyperparameters for both RF and XGBoost, minimizing overfitting and ensuring generalization to unseen data.

The models were run in Python 3.12.

Result: Some Preliminary Run using other set of variables

Predictors: TVOCs, NOx, RH, Temperature

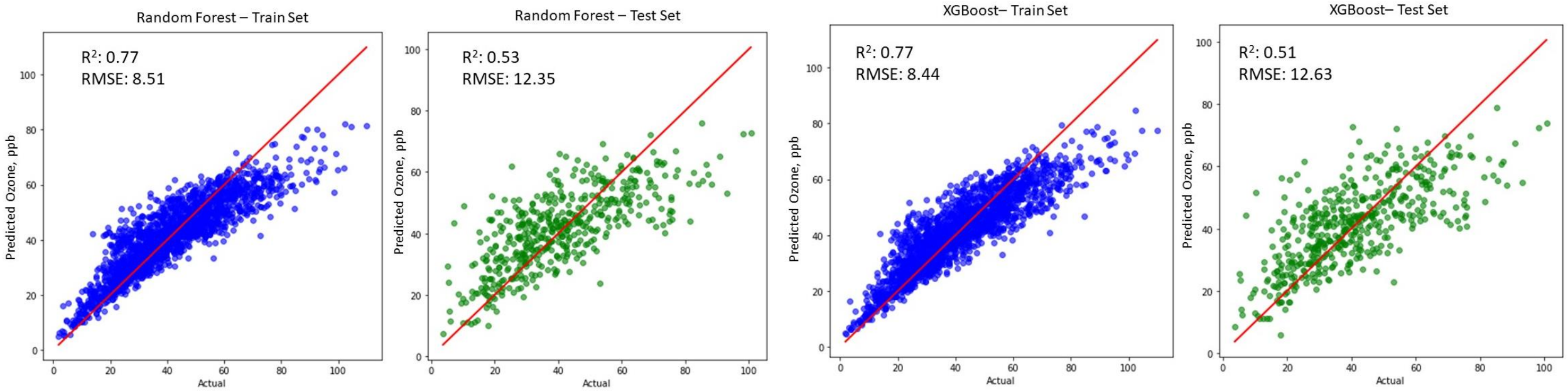

Figure S-1. Performance of Random Forest and XGBoost Best Model fitting result for both Train and Test dataset when the variables considered were only TVOCs, NOx, RH, and Temperature

Preliminary results indicate that the number of predictors significantly affects machine learning performance. When fewer predictors are used, Random Forest slightly outperforms XGBoost. A reduced number of predictors also leads to a lower R^2 score. With this specific set of predictors, the models capture only about 50% of the variance in ozone concentration.

Result: High Ozone Case

Predictors: TVOCs, NO,NO_2, RH, Temperature, Solar Radiation, Wind Speed

The data used for this analysis were limited to March 2023, the month in which the highest occurrences of elevated ozone levels were observed throughout the year.

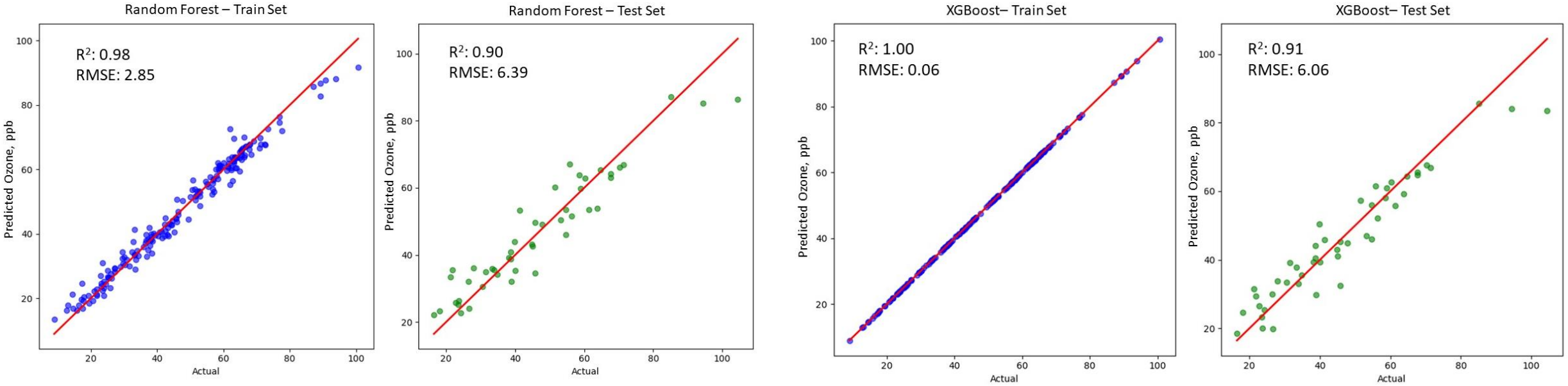

Figure S-2. Performance of Random Forest and XGBoost Best Model fitting result for both Train and Test dataset when using only March 2023 dataset and the variables considered were TVOCs, NO,NO_2, RH, Temperature, Solar Radiation, Wind Speed

Both models captured approximately 90% of the variance in ozone concentration and exhibited strong performance metrics. While the XGBoost model demonstrated slightly better results, the difference was minimal.

Result: High Ozone Case

The data used for this analysis were limited to March 2023, the month in which the highest occurrences of elevated ozone levels were observed throughout the year.

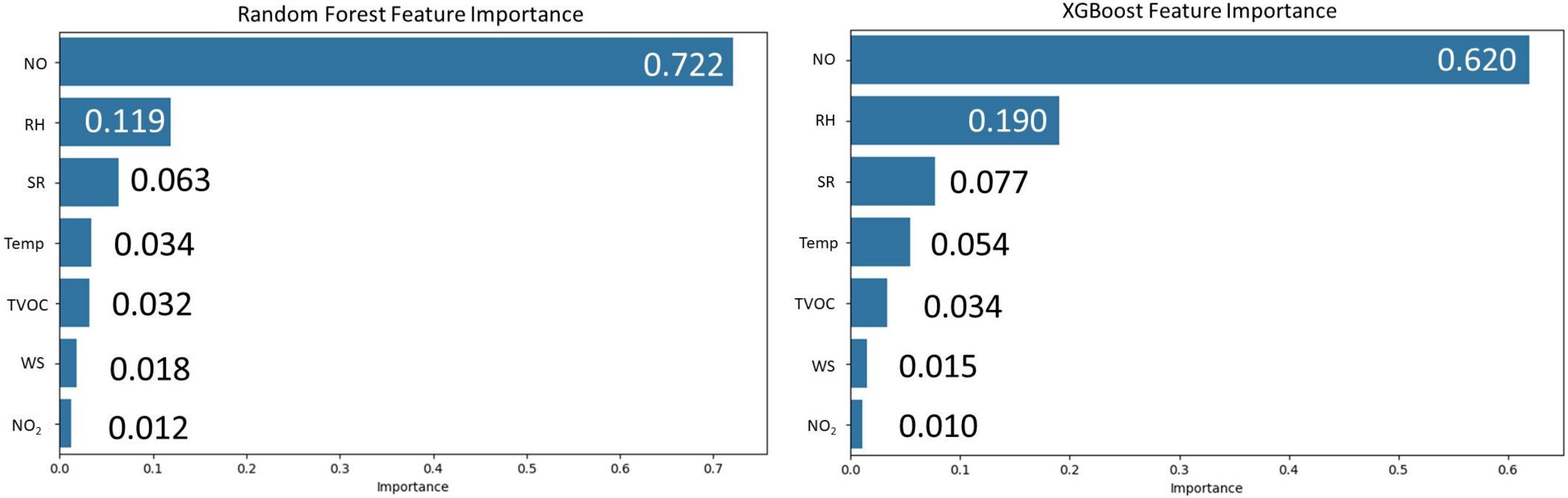

Figure S-3. Feature Importance of the Best Random Forest and XGBoost Models in Determining Variable Contributions to Ozone Predictability. The importance values are shown in descending order (from top to bottom) with the total feature importance of all features summing to 1.

During high ozone episodes, NO, relative humidity, and solar radiation were identified as the most important key contributors to ozone prediction for both ML models.