Using Machine Learning to Combat Air Pollution by Forecasting Tropospheric Ozone Levels

Juarez, Eliana (School: V. Sue Cleveland High School)

Ground-level ozone is a photochemical smog and secondary pollutant that is harmful to urban populations by increasing the risk of heart and lung disease and harming agricultural crops. It is particularly severe in the developing world, where it is a leading cause of premature deaths. To warn populations of hazardous ozone levels, I developed a code to analyze and compare different machine learning algorithms to reliably predict the ozone concentration 24 hours in advance. This project first used hourly records of five weather variables and 12 air pollutant variables over the course of one year in Delhi, India. To create the best model, this project tuned, trained, and tested seven machine learning algorithms and compared their predictive abilities using cross-validation. Among the seven models, R-squared values varied from 0.39 to 0.61, with XGBoost, Random Forest, and K-Nearest Neighbors Regression ranking highest. When re-trained by separate seasons across five years, predictive capabilities of all models were significantly higher, with a maximum R-squared of 0.754 during winter, outperforming models created in previous studies on ozone forecasts. When tested, the three best performing models could reliably predict ozone concentrations 24 hours in advance, where 50% of the predictions had a percentage error of less than 10%. These results show that weather and pollutant data have sufficient predictive power for 24-hour ozone warnings, and that machine learning can greatly improve upon simpler forecasting methods. Thus, advanced data monitoring and computing can improve safety for people worldwide.

Awards Won:
Raytheon Technologies Corporation: Each winning project will receive $1,000.