AeroPredict

Air Quality Prediction Using Machine Learning

By Chase Woodfill and Mikayla McCormack

Why does this matter?

- Air pollution causes 7 million premature deaths globally each year (WHO).
- Long-term exposure increases the risk of heart disease, stroke, and lung conditions.
- Current air quality systems offer real-time data, but lack forecasting capabilities.
- This limits proactive decision-making for individuals and policymakers.

Our Problem:
We aim to build a machine learning model that predicts future CO levels using historical sensor and weather data—filling the gap between monitoring and meaningful early warning.

Motivation & Problem

Goal:
Forecast future carbon monoxide (CO) levels using historical air quality and weather data to support proactive public health and planning decisions.

Research Questions:

- Can we accurately predict future CO concentrations using past sensor and weather data?
- Which features (e.g., time, sensor readings, weather) are most predictive?
- How well does a Random Forest model perform in this forecasting task?

Goals & Q's

Overview:

- Source: UCI Machine Learning Repository via Kaggle
- Records: 9,357 hourly entries
- Location: Roadside in a polluted Italian city
- Timeframe: March 2004 – February 2005
- Device used: Field-deployed gas multisensor + certified analyzer

Data Includes:

- 5 metal oxide sensor readings (PT08.S1–S5)
- Reference gas concentrations: CO, NMHC, Benzene, NOx, NO_2
- Weather features: Temperature, Relative Humidity, Absolute Humidity
- Timestamp: Date and Time

The Dataset

Data Preparation:

- Loaded air quality dataset
- Combined date and time columns into a singular value
- Cleaned invalid readings (-200 → NaN)

Feature Engineering:

- Extracted time features (hour, day, month, weekday)
- Selected predictors (PT08.S1, T, RH, AH)

Modeling:

- Initially used Linear Regression
- Switched to Random Forest Regressor for better accuracy

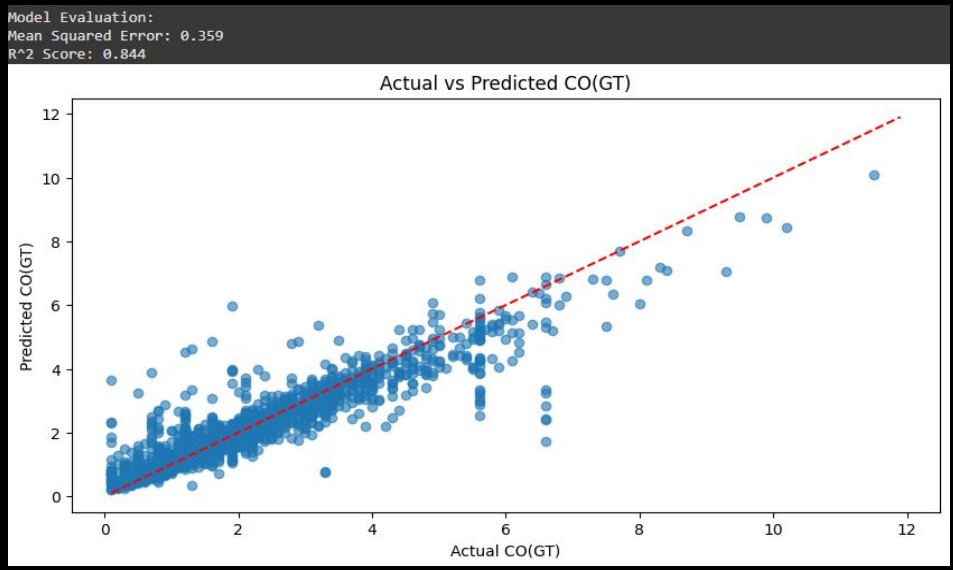
Approach

Results

This scatter plot compares actual vs. predicted CO levels.

- The red dashed line represents perfect predictions.
- Most points are close to the line, especially in the 0–6 mg/m³ range, showing strong model accuracy.
- Slight underprediction occurs at higher values due to fewer examples.

Overall, the model does a great job estimating CO levels based on historical data.



Evaluation:

- Evaluated model used Mean Squared Error and R^2 score

Linear regression:

- Mean Squared Error: 0.767
- R^2 Score: 0.667
- Random Forest Regressor:
- Mean Squared Error: 0.359
- R^2 Score: 0.844

Research Questions Results:

- **Can we accurately predict future CO concentrations using past sensor and weather data?**
  Yes – our Random Forest model achieved an R^2 score of 0.844, showing strong predictive power.
- **Which features are most predictive?**
  Time-based features (hour, month), PT08.S1(CO) sensor readings, and weather data (temperature and humidity) had the strongest influence.
- **How well does a Random Forest model perform in this forecasting task?**
  It significantly outperformed linear regression, capturing non-linear trends and interactions in the data.

Results Pt2

- **Expand to Predict Other Pollutants:** Extend the model to forecast additional harmful gases like NO_2, Benzene, and NOx, which are also tracked in the dataset. This would broaden the model's impact and make it more useful for comprehensive air quality forecasting.
- **Use Datasets from Other Regions or Time Periods:** Test and retrain the model using data from different cities, climates, or years to evaluate how well it generalizes. This could also reveal location-specific patterns and make the model more powerful in real-world deployment.
- **Deploy as a Real-Time Web App with Dashboards:** Integrate the model into a user-facing web application that displays current and predicted AQI levels, visualizations, and recommendations. This could support daily decision-making for users (e.g., whether to avoid outdoor activity).
- **Experiment with More Advanced Models:** Explore the use of XGBoost for improved accuracy and faster training, or Long Short-Term Memory networks to capture sequential patterns in time-series data. These models may better handle pollution spikes and time-dependent dynamics.

Future Work

References:

COSC426 - University of Tennessee, Knoxville

Dataset: https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set?resource=download

Github: https://github.com/mikjmcc26/AeroPredict

Questions?