

Predictive Modelling of Urban Air Quality

Name: Soma Mani Deepak

Student ID: 24095193

Github Link: https://github.com/Manideepak788/Air_Quality_Prediction

1. Introduction:

This project's goal was to create a high-precision regression model to forecast urban carbon monoxide (CO) concentrations. We developed a strong data science pipeline using the multi-sensor chemical measurements found in the UCI Air Quality dataset. We obtained a R^2 score of roughly 0.89 by utilising sophisticated preprocessing and machine learning, offering a solid basis for an early-warning pollution monitoring system.

2. Data Understanding and Preprocessing:

2.1. Managing Missing Information

- A crucial finding from the first audit was that missing data were identified as a particular integer, -200, rather than as "null".
- Method: These were transformed into NaN values.
- Interpolation Linear interpolation was used in place of deleting records. This approach precisely maintains the trend between hourly data because air quality is a continuous time-series, guaranteeing no loss of temporal context.

2.2. Feature Engineering

- We created a number of new variables to improve the forecasting ability of the model:
 - Temporal Features: To record daily rush hours and seasonal patterns, the hour, day of the week, and month were extracted.
 - Features of Lag (CO_Lag1): The CO concentration from the preceding hour was represented by a variable that was created. The "momentum" of air pollution is explained by this.
-

3. Exploratory Data Analysis (EDA):

3.1. Correlation Analysis

We analysed the relationship between different gas sensors and the target variable (CO).

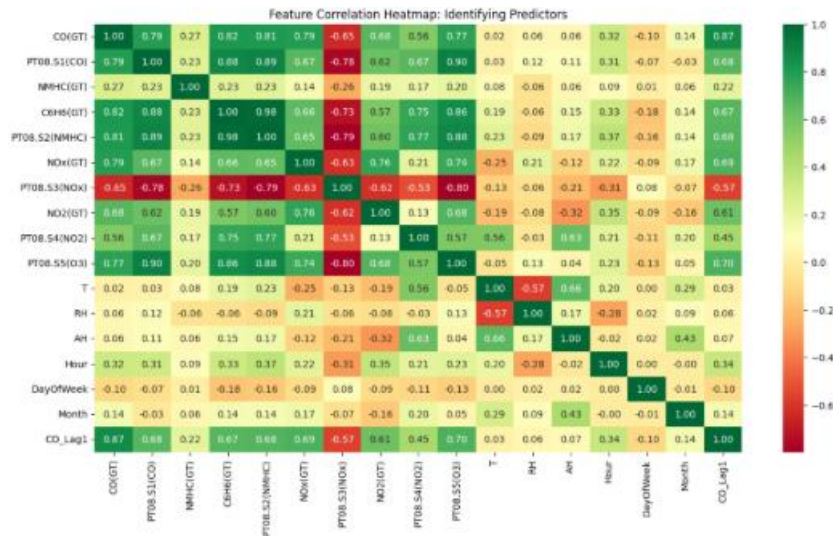


Figure 1: Heatmap showing the high correlation between Benzene ($\$C_6H_6\$$) and CO, suggesting they originate from similar combustion sources (traffic).

3.2. Target Distribution

The $\$CO\$$ distribution exhibits a "Right Skew," which means that although most hours have moderate levels, the model has to be able to forecast the frequent high-pollution spikes.

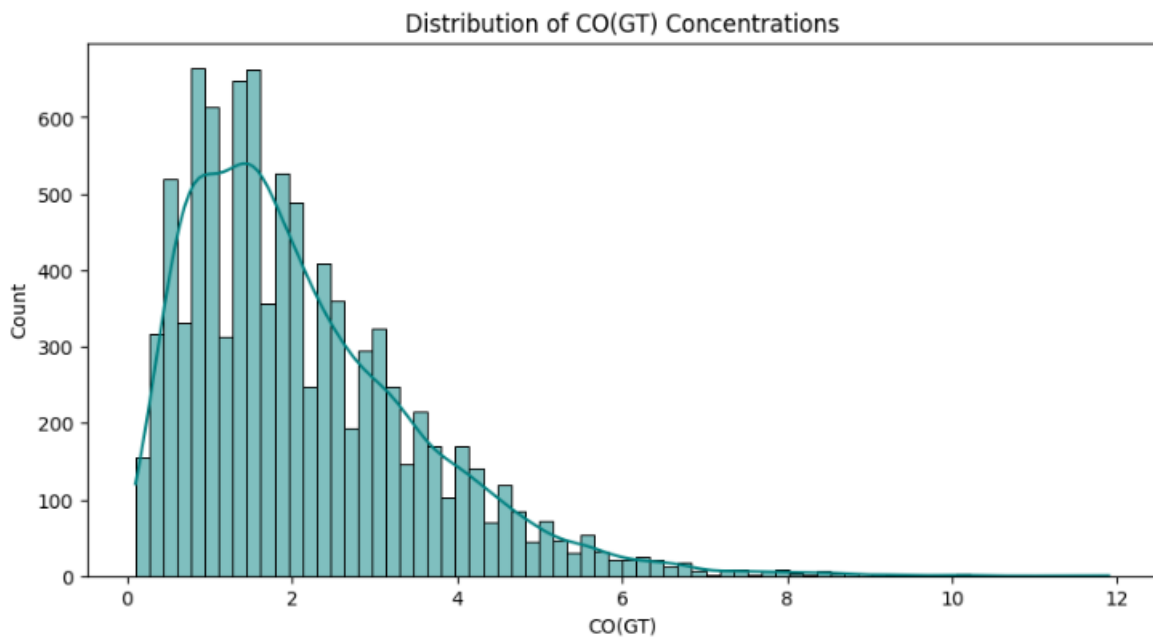


Figure 2: Frequency distribution of CO concentrations.

4. Methodology:

We benchmarked two distinct regression architectures using an 80/20 time-series split (preserving the chronological order of data):

1. **Linear Regression (Baseline):** Used to establish a performance floor.

2. **Random Forest Regressor (Champion):** An ensemble method chosen for its ability to capture non-linear interactions between humidity, temperature, and chemical sensor responses.

5. Results and Model Evaluation:

The **Random Forest** model significantly outperformed the baseline, demonstrating superior handling of the complex sensor data.

Metric	Linear Regression	Random Forest
MAE (Mean Absolute Error)	\$0.34\$	\$0.30\$
RMSE (Root Mean Squared Error)	\$0.49\$	\$0.46\$
\$R^2\$ Score	\$0.87\$	\$0.89\$

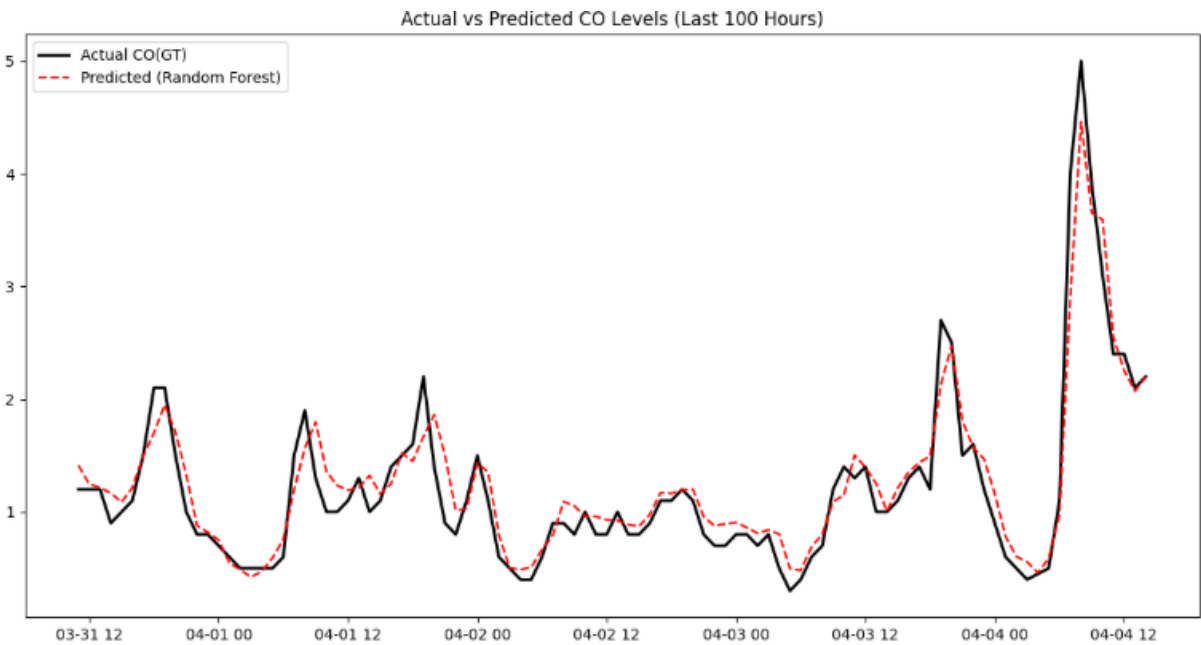


Figure 3: Time-series plot comparing actual CO levels vs. model predictions. The model accurately tracks both peaks and troughs.

6. Key Drivers of Pollution:

By analyzing the model’s internal decision-making, we identified the primary factors influencing pollution levels.

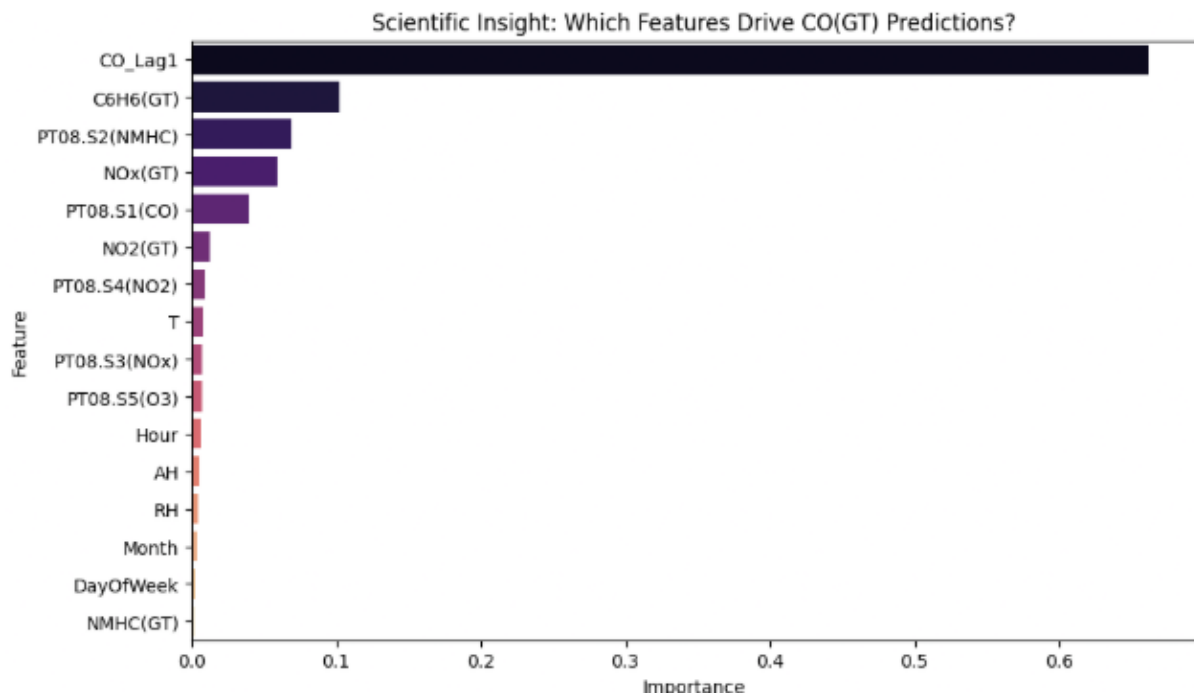


Figure 4: The top predictors for CO levels.

Scientific Insights:

- **Autocorrelation:** The `CO_Lag1` feature is the strongest predictor, meaning current air quality is heavily influenced by the previous hour.
- **Chemical Proxies:** Benzene (`C6H6`) and Hydrocarbon sensors (`PT08.S2`) are high predictors, confirming that vehicle emissions are the primary source of `CO` in this area.

7. Conclusion and Recommendations:

7.1. Conclusion

The experiment effectively showed that harmful gas concentrations may be predicted with great accuracy (89%) using inexpensive chemical sensors. We were able to capture the erratic nature of urban pollution surges by switching from a linear to a non-linear Random Forest model.

7.2. Suggestions for Stakeholders

- **Predictive Warnings:** When a surge is anticipated based on current sensor data, use the Random Forest model to send out "1-hour ahead" alerts to the public.
- **Sensor Implementation:** Benzene sensors should be given priority maintenance and calibration due to their significant role in predicting `CO`.
- **Urban Planning:** According to the time analysis, the morning and evening peaks found during the EDA phase should be the main focus of traffic control initiatives.

8. References

Dataset Reference

- Vito, S. De, Francia, G. Di., Martinotto, L., Piga, M., and Massera, E. (2008). An electronic nose, a multi-sensor tool for estimating benzene in urban pollution monitoring, was calibrated in the field. *Chemical Sensors and Actuators B*, 129(2), 750-757. 10.1016/j.snb.
- UCI Machine Learning Repository, 2007.09.060 (2008). Data Set on Air Quality. The URL is <https://archive.ics.uci.edu/ml/datasets/Air+Quality>.

Technical & Methodology References

- F. Pedregosa and associates (2011). Scikit-learn: Python Machine Learning. *Machine Learning Research Journal*, 12, 2825-2830. (Used for Random Forest implementation and linear regression).
- L. Breiman (2001). *Machine Learning*, 45(1), 5-32. Random Forests. (The Random Forest algorithm's foundational publication).
- W. McKinney (2010). *Python Data Structures for Statistical Analysis*. The 9th Python in Science Conference Proceedings. (Used to manipulate data with Pandas).

Policy & Standards Reference

- European Union (2008). Directive 2008/50/EC on cleaner air and ambient air quality for Europe. The EU's official journal. (Context for European monitoring of air quality standards).