Dataset: https://archive.ics.uci.edu/dataset/360/air+quality

GROUP 24-003

474 Milestone 06

Project Title: *How The City Decides When You Can Breathe* 0. Prediction Goal(s)

0.Prediction goal(s):

1. Our Prediction Goal: Our primary goal is to forecast the hourly CO concentrations in the air using predictive modeling

2. Why it matters:

Carbon monoxide is a harmful pollutant, especially for people with respiratory conditions. By creating an accurate hourly forecast, we can:

- Allow residents to plan safer times to be outdoors
- Influence policymakers to identify and respond to peak pollution times
- Improve public health intervention decision-making

3. Identifying Regression/ Classificiation:

This is a regression problem, because the target variable CO(GT) is a continuous (measured in mg/ m^3).

Dataset Overview: This project uses the Air Quality UCI Dataset, which included 9,357 observations with 14 predictor variables, both numeric and temporal features. 18% of the values were missing, primarily in NMHC(GT), which we dropped due to the high proportion of missing data. The response variable, CO(GT), represents the true carbon monoxide concentration measured in mg/ m^3 and is continuous. The distribution of CO(GT) is highly right-skewed, with many readings being very low and occasional peaks during pollution events.

1. Modeling Approach

a. Model Choice

We selected Multiple Linear Regression to be our baseline model because:

- It is simple, easy to interpret, and computationally efficient
- It provides a good baseline for our future advanced models
- It is appropriate for our continuous target variables

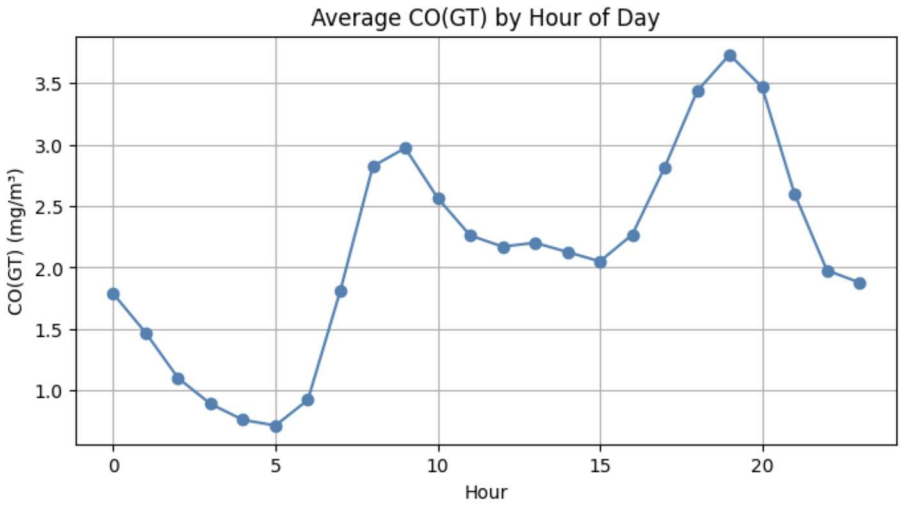
b. Implementation & Feature Selection

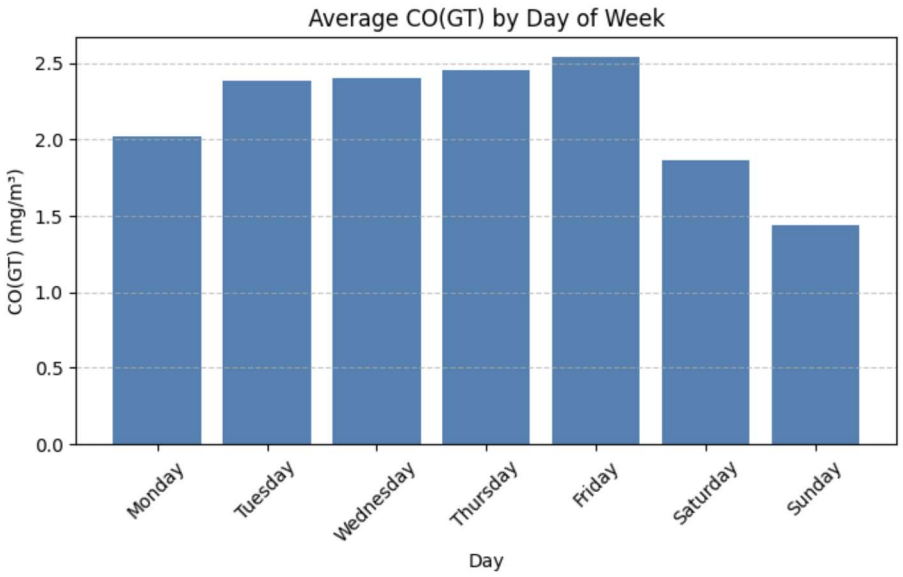
Data Cleaning:

Prior to modeling, we cleaned the dataset by removing faulty readings like (-200), dropped NMHC(GT), and imputing missing values where appropriate.

c. Feature Engineering:

We created three new features, all temporal in nature, taking the datetime and converting it to hour of day, day of week, and month of year. We call these features “Hour”, “Month” and “DayOfWeek”. The reason we did this is fairly standard, to capture relationships not seen in the raw date feature which is not exceptionally useful on its own. These three features actually became some of our strongest ones with exceptionally high correlations to the target of carbon monoxide. Of course, this makes sense, as the carbon monoxide concentration is the majority from automobiles. Therefore, the concentration will be higher when there are more cars on the road. The graphs below show the correlation between the features hour and day and the target of carbon monoxide concentration.





Stepwise Feature Selection + Model Implementation:

We utilized a stepwise feature selection approach (bidirectional) using AIC as the selection criterion to figure out which predictors contribute the most when explaining variance in response to the response variable CO(GT). Throughout the process, it iteratively added/removed predictors in order to find the best subset for the linear regression model.

d. Cross-Validation Results:

The initial feature pool included:

- **Pollutant & Sensor variables:** PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3), C6H6(GT), NOx(GT), NO2(GT)
- **Environmental variables:** T, RH, AH
- **Temporal features:** Hour, DayOfWeek, Month

Result of Stepwise Selection:

- **Retained:** PT08.S1(CO), PT08.S2, PT08.S3, C6H6(GT), NOx(GT), NO2(GT), T, RH, Hour, DayOfWeek, Month
- **Removed:** NMHC(GT), MonthName (high missingness or redundancy)

Model Evaluation: k-fold Cross-Validation:

- We evaluated the baseline multiple linear regression model using 10-fold cross-validation, a standard technique that helps reduce overfitting risk and gives a more reliable error estimate.

Evaluation Metrics:

- **RMSE (Root Mean Squared Error):** Average prediction error in mg/ m^3
- **MAE (Mean Absolute Error):** More robust to outliers
- **R^2 (Coefficient of Determination):** Proportion of variance that can be explained

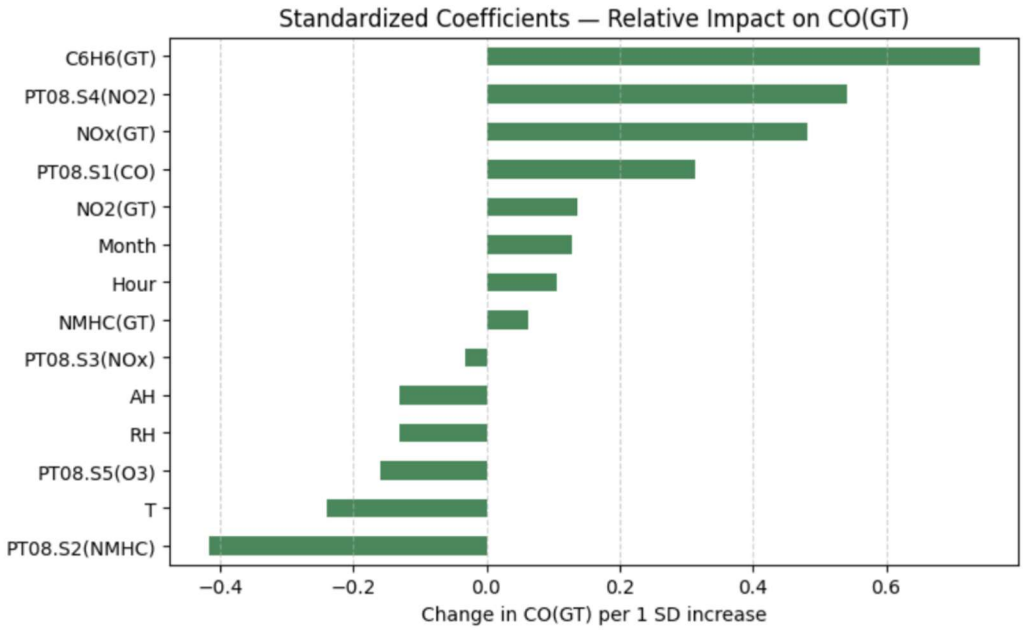
Metric	Mean CV Score	Std. Dev.
R^2	0.896	0.021
RMSE	0.258 mg/m^3	0.034
MAE	0.194 mg/m^3	0.026

- The strong R^2 confirms that the selected predictors capture a large proportion of the variation in hourly R^2 concentration.
- The low RMSE and MAE showcase that the model creates accurate forecasts even under cross-validation

3. **Interpretation & Next Steps**
    - Summarize baseline model performance results (e.g., cross-validation scores).
    - Reflect on possible improvements (e.g., feature engineering, hyperparameter tuning).

```
R² on full data: 0.896

Model Coefficients (sorted):
C6H6(GT)          0.101818
Month             0.036275
Hour              0.015237
NO2(GT)           0.002916
NOx(GT)           0.002301
PT08.S4(NO2)      0.001580
PT08.S1(CO)       0.001468
NMHC(GT)          0.000887
PT08.S3(NOx)     -0.000125
PT08.S5(O3)      -0.000404
PT08.S2(NMHC)    -0.001605
RH               -0.007685
T                -0.027581
AH               -0.333205
dtype: float64
```



As we can see from this visualization, there are plenty of strong predictors, but some such as benzene (C6H6) and Nitrogen Oxide (NO2) in the positive direction are especially strong. On the other hand, Nitrogen Dioxide and Temperature are strong predictors in the negative direction. Indicating a lower carbon monoxide concentration when the temperature is

higher. This initial model had an R^2 of .896, which is a very strong start, but it should be noted that R^2 is not everything in the scope of this problem. We are not simply trying to predict exactly how much carbon monoxide is in the atmosphere, but instead we are trying to find solutions to make the streets safer to walk in for millions of pedestrians. Therefore, predictors that are actionable and causal can help us make the city safer. As automobiles and peak commute times are the main causal drivers of carbon monoxide concentration, we should focus on cars that are more efficient or implement more green spaces in urban areas.

In the future, we will do additional feature engineering to try and judge the interaction between some of our features, such as potentially separating weekends versus weekdays, as there tend to be differences in trends on those days. Furthermore, it is important to focus more on examining which of these predictors are truly causal. For example, benzene is the strongest correlated predictor, which indicates automobiles are a factor, but simply lowering the benzene concentration would not necessarily lower carbon monoxide concentration unless gas cars went with it. We need to work to figure out which of these drivers can be implemented throughout the world to make our cities safer for pedestrians at all times of the day.

https://colab.research.google.com/drive/1GcytZTHgkFVx7wiA5k5aRVlGNkMaU_h8?usp=sharing

---

1b. More Complex Model Implementation & Tuning

1. **Model Choice**

   We selected a Random Forest Regressor for our more complex model. Unlike our linear regression, which assumed a linear relationship between our predictor and target variables, Random Forest can capture non-linear relationships and other interaction effects between temporal, pollutant and lagged features. With the air quality dynamics being very complex with things like rush-hour peaks, lagged pollutant backup, and sensor interactions, this makes Random Forest a strong candidate for improving our forecast.

2. **Hyperparameter Tuning with Cross-Validation**

We performed hyperparameter tuning using GridSearchCV with a time series split cross-validation strategy in order to respect the temporal structure of the data. The parameter grid included:

- n-estimators: 50

- max_depth: [8,None]
- min_sample_leaf: [2]

This configuration allows us to model while balancing bias and variance effectively while also handling temporal lag features.

3. Model Selection & Final Comparison

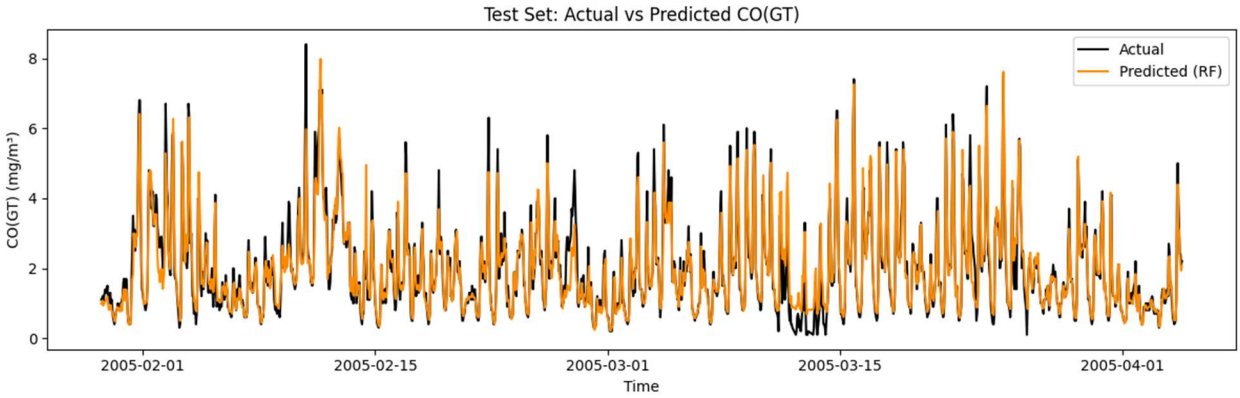
We evaluated the tuned Random Forest Model tuned on a held-out test set and compared it to our initial Multiple Linear Regression.

Model	R^2	RMSE (mg/m^3)	MAE (mg/m^3)
Linear Regression (Base)	0.896	0.258	0.194
Random Forest (Tuned)	0.913	0.397	0.259

- R^2 increased from 0.896 to 0.913, which indicates that the model captures more variance in CO concentrations
- Additionally, RMSE and MAE are higher for RF, which reflects that while the model tracks trends well, it is very sensitive to sharp pollution spikes.

4. Visualization

Below is our time series plot of actual CO(GT) readings to the random Forest Prediction. The model follows daily trends and peak hours closely, capturing the temporal structure better than our baseline model.



5. Interpretation

As thought by the improvement in R^2, the Random Forest better leverages non-linear relationships in lagged and temporal features compared to the linear regression, accurately predicting readings during most of the period. However, at some sharp peaks, the model underestimates the sharp CO peaks, like mid-February and late March. This helps explain the higher RMSE despite a better R^2. This model accurately captures overall trends but isn't good at predicting the magnitude of sudden spikes. Temporal and lagging features clearly helped improve the model with peak timings, even though the magnitude might not always be exact.

6. **Next Steps:**
    - Add additional lag and rolling features to better capture short term pollutant spikes
    - Experiment with other high performing gradient boosting models like XGBoost, LightGBM to better predict peaks
    - Incorporate other external covariates like weather and traffic volume if it's available
    - Explore quantile regression forests or probabilistic models for uncertainty intervals

Vis
https://colab.research.google.com/drive/1GcytZTHgkFVx7wiA5k5aRVIGNkMaU_h8?usp=sharing

Random Forest

https://colab.research.google.com/drive/1YJ7kUqE9kpQKN-2vt-SuXDEa1zxyy5Lo#scrollTo=rseS9FxX-MhT