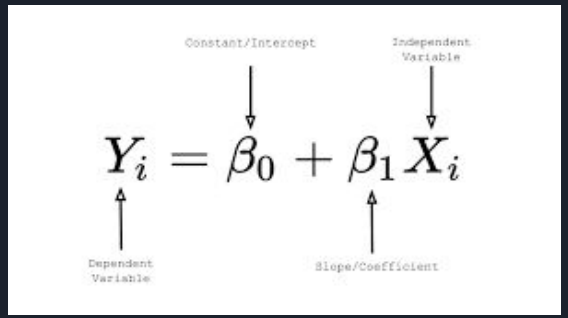
Predictive Modeling of Atmospheric Carbon Monoxide Concentrations

By: Deergh Gandhi and Shayaan Niazi

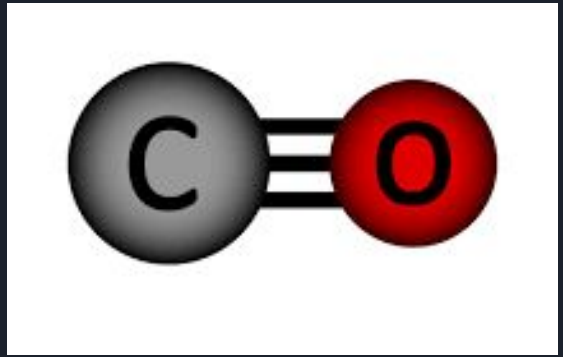
Introduction

- For our aim, we set out to use Machine Learning to predict the concentrations of Carbon Monoxide
- The Linear Regression model was chosen as our primary tool for this prediction.
- Potential to revolutionize real-time air quality monitoring, but it also has significant benefits for urban planning, ensuring healthier and more sustainable cities."

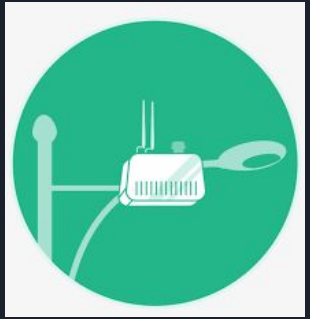


Project Description

- Central Objective: Predict the invisible yet pervasive threat - CO concentrations.
- Characteristics of CO: It's silent, being both odorless and colorless, but has major health implications.
- Using a real-world dataset allows us to bridge the gap between theoretical predictions and actionable insights.



Dataset Overview

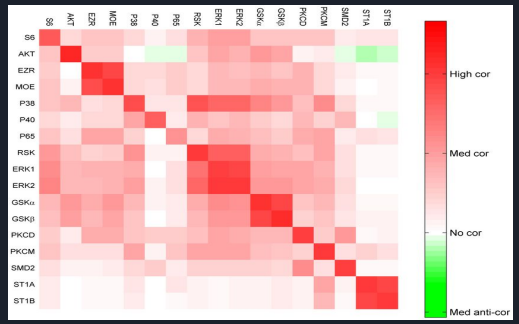


- Data sourced from specialized 'Air Quality Chemical Multi Sensor Devices’
- Strategically positioned sensors in a region in Italy known for pollution
- Data spans from March 2004 to April 2005, capturing seasonal variations
- Comprehensive data attributes, including CO, NMHC, Benzene, and others

Methodology

Methodology - Part 1

- Methodology began with thorough data cleaning. Addressed anomalies, removed redundant columns, and handled missing data.
- We ventured into exploratory data analysis
- Visualized data distributions to gauge individual variable characteristics.
- Utilized correlation matrices to uncover relationships between variables.



Methodology - Part 2

- Transitioning from EDA, we arrived at the Model Selection phase. We opted for Linear Regression.
- Why Linear Regression? Its simplicity and effectiveness made it apt for the nature of our dataset.
- To ensure our model's efficacy and guard against overfitting, we divided our data. 80% was allocated for training and the remaining 20% for testing.

Results

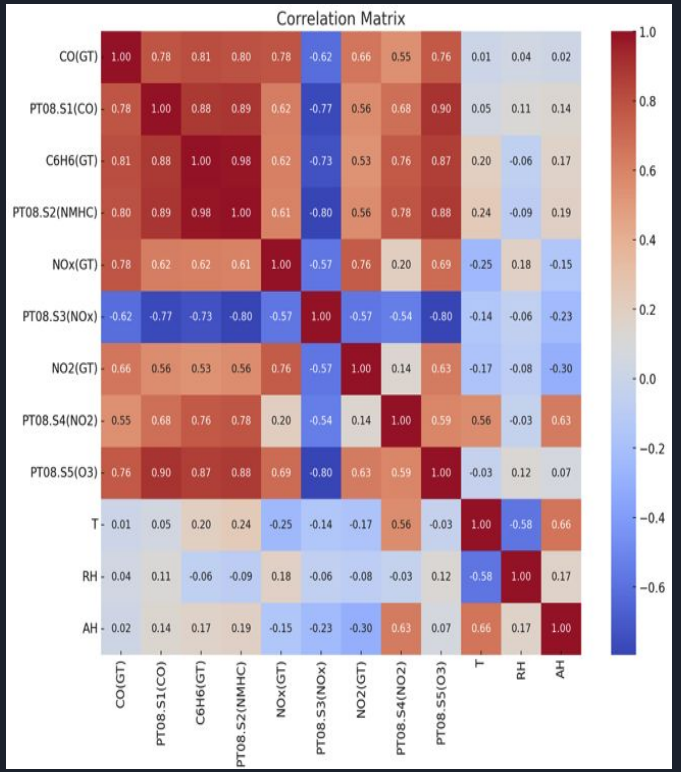
Results - Model Metrics

Our model's performance can be gauged using key metrics like RMSE and MAE.

RMSE of approximately 0.57 suggests our model's predictions deviate from actual values by about this magnitude.

MAE, being around 0.38, indicates the average absolute difference between predicted and actual values.

An encouraging sign was the model's consistent performance across both training and testing datasets.



Results - Insights

Diving deeper into the model's insights, certain features particularly influenced our predictions.

NOx(GT) and PT08.S1(CO) were paramount in determining CO concentrations.

In analyzing the residuals, we found them to be majorly centered around zero, which is a positive indication of our model's unbiased errors.

Conclusion

Work Distribution

Deergh:

- Powerpoint slides
- Presentation script

Shayaan:

- Written Report

The rest of the project including working with the data and code functions were equally distributed between us