Analysis of Air Quality Data

The objective of this assignment is to predict the hourly averaged concentrations of various air pollutants for the next 48 hours using the provided dataset.

Additionally, participants are required to validate their predictions using the Root Mean Square Error (RMSE) over the last 10 percent of the data, to ensure accurate and reliable predictions.

This assignment tests your ability to apply concepts such as data preprocessing, stationarity testing, time series modeling (ARIMA, Prophet, etc.), residual analysis, feature engineering, and multivariate forecasting.

Data Description:

The dataset consists of 9358 instances, each representing hourly averaged responses from various chemical sensors. Ground truth hourly averaged concentrations for various pollutants including CO, Non-Metanic Hydrocarbons (NMHC), Benzene, Total Nitrogen Oxides (NOx), and Nitrogen Dioxide (NO2) were provided by a co-located reference certified analyzer. Missing values are tagged with -200.

Attribute information:

0. Date (DD/MM/YYYY)
1. Time (HH.MM.SS)
2. True hourly averaged concentration CO in mg/m^3 (reference analyzer)
3. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4. True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
5. True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
6. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7. True hourly averaged NOx concentration in ppb (reference analyzer)
8. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9. True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
12. Temperature in °C
13. Relative Humidity (%)
14. Absolute Humidity (AH)

Task:

Participants are required to predict the hourly averaged concentrations of the specified pollutants for the next 48 hours based on the provided dataset. Additionally, ensure the validation scores meet the following criteria:

1. RMSE value for CO(GT) is <= 10
2. RMSE value for PT08.S1(CO) is <= 210
3. RMSE value for NMHC(GT) is <= 14
4. RMSE value for C6H6(GT) is <= 6
5. RMSE value for PT08.S2(NMHC) is <= 250
6. RMSE value for NOx(GT) is <= 190

7. RMSE value for PT08.S3(NOx) is <= 196.0619
8. RMSE value for NO2(GT) is <= 120
9. RMSE value for PT08.S4(NO2) is <= 300.7
10. RMSE value for PT08.S5(O3) is <= 400.25
11. RMSE value for Temperature is <= 12
12. RMSE value for Relative Humidity is <= 18
13. RMSE value for Absolute Humidity is <= 7

Please note that in the above-stated criterions, Units are being kept the same as in the given dataset.

Additional Requirements:

1. Model Comparison: Briefly compare at least two different modeling approaches you tried. What worked better and why?
2. Feature Importance: Highlight which features were most influential in your forecasts.
3. Residual Analysis: Analyze the residuals of your best model to check if they resemble white noise and show no autocorrelation.
4. Error Analysis: Identify periods where your model failed or underperformed. What patterns or causes can you infer?
5. Feature Engineering Insight: Describe any new features you created and their impact on the model's performance.

The file submission.xlsx gives a template for the submission of results that should be strictly followed. Also, you need to include the notebook/python files for evaluation purposes as there will be marks for data preprocessing and EDA.
Mention the above answers in the form precisely.

Links:
Dataset and sample submission file: Time Series Analysis Course Assignment - 1
Submission Form link: https://forms.office.com/r/mBYQyMQKzW