Problem statement:

One of the main indicators of air quality is PM2.5 (particulate matter) which can be generated by traffic exhaust, industry, burning of fossil fuels and many other sources.In addition to that the particles are so small that they are invisible to the naked eye and when inhaled do not just collect in the lungs and cause respiratory disease but can also enter the bloodstream and contribute to heart disease and stroke.The main objective therefore is to forecast PM2.5level at per hour per day.

Data Preparation and Understanding

The data provided spans from 2019 to November 2021.Since this is a time series assignment,we need to respect the time component in preparation our train and test datasets and also in building our model

In Machine Learning we need to have two sets of data.One to build our model(Training dataset) and the other version to evaluate our model(Test dataset).Therefore ,I split the provided dataset into 2 sets based on time.My Training dataset spans from 2019 t0 December 2020 and test dataset spans from Jan 2021 to November 2021.This is implemented in Notebook1, 1-0-lawrence-preparing_target

Preparing Target Variable(pm_2_5)

Since we are required to forecast pm_2_5 per hour per day,and in the original dataset provided there are multiple records with the same hour.I prepare my target variable by grouping by site and timestamp then taking the average within that particular hour such that we have one observation per every unique hour per site.(This is implemented in Notebook1, 1-0-lawrence-preparing_target)

Exploratory Data Understanding(EDA)

In 2.0-EDA-Lawrence-Moruye I dig deep into the dataset in order to get a more understanding of my data.There are various interesting trends brought out by the visualizations.Please refer to 2.0-EDA-Lawrence-Moruye notebook

Modelling

In this notebook 3.0-modelling-Lawrence, I've done some data cleaning(to remove constant features,duplicate columns),feature Engineering and building a Machine Learning model to do the forecasting.

In order to validate my model correctly ,I build my own custom validation strategy.From my earlier prepared Training dataset,I also prepare a validation dataset by splitting the earlier prepared train data into two halves.The first half is for training and the second is for validation.

I then fit a boosting algorithm(light gbm) to my prepared dataset.

Evaluation

I used RMSE as an evaluation metric.I evaluated the my trained model on test dataset and attained a RMSE of 0.8282413986444274

How to improve the performance of the model.

Here is a list of ideas which I think can be incorporated into the baseline work to improve the performance.

1.Do More feature engineering
2.Do feature selection
3.Trying different validation strategies(TimeSeries split/or KFold)
4.Trying different models,from different hypothesis classes
5.Stacking,boosting or blending.
6.More data cleaning

Resources:

1.https://www.epa.gov/air-trends/particulate-matter-pm25-trends
2.https://community.wmo.int/activity-areas/gaw/science-for-services/gafis