Air Pollution Measurements Prediction

Yu Li[1]

[1] Nanjing University of Science and Technology, China

Introduction

In this competition you are predicting the values of air pollution measurements over time, based on basic weather information (temperature and humidity) and the input values of 5 sensors. The three target values to you to predict are:

target-carbon-monoxide

target-benzene

target-nitrogen-oxides

Data Description

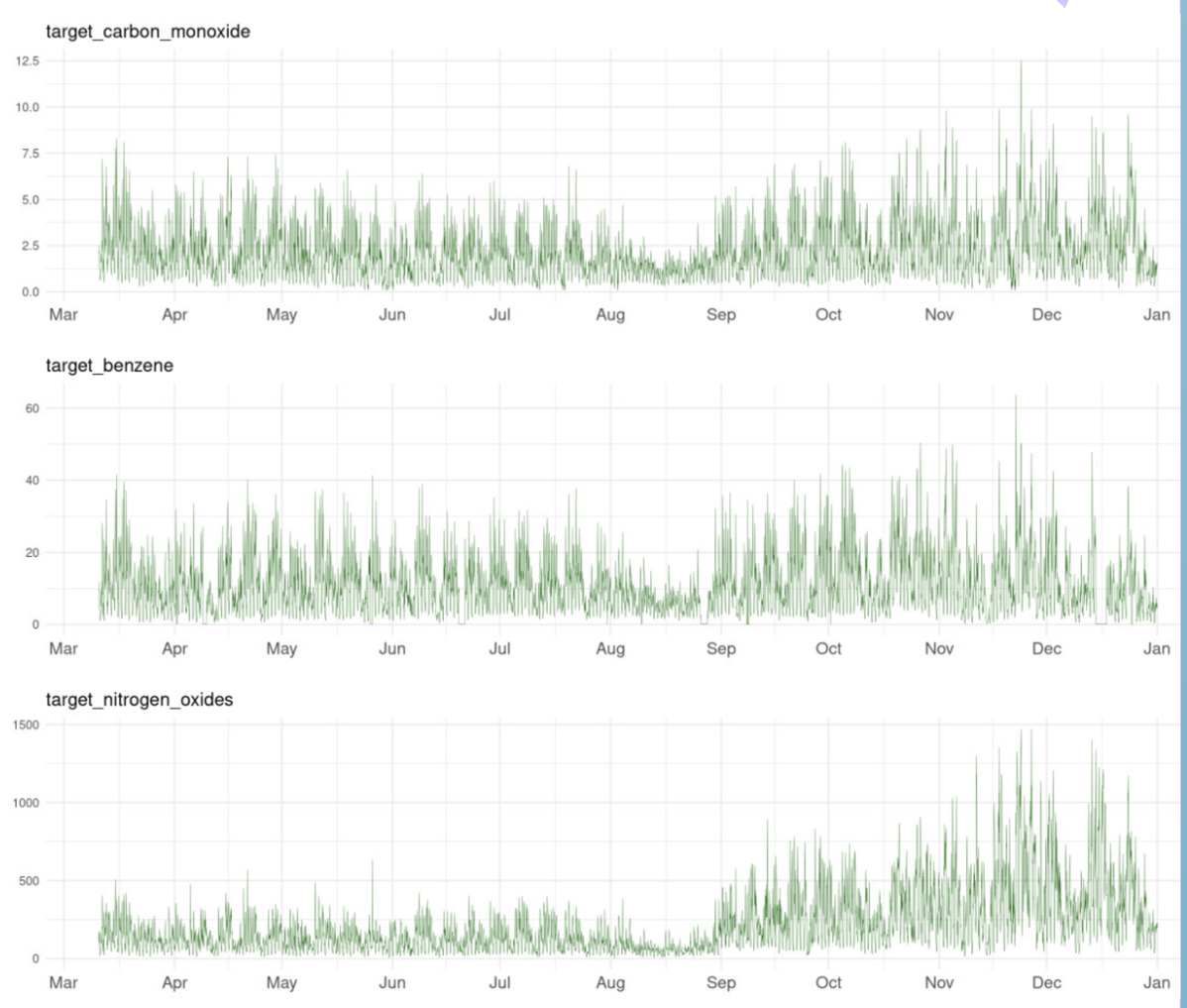
Train:

Elements	Number
datetime	7111
degC	408
relativehumidity	762
absolutehumidity	5451
*sensor*1	3882
*sensor*2	3882
*sensor*3	3882
*sensor*4	3882
*sensor*5	3882
targetcarbonmonoxide	95
targetbenzene	405
targetnitrogenoxides	3268

Test:

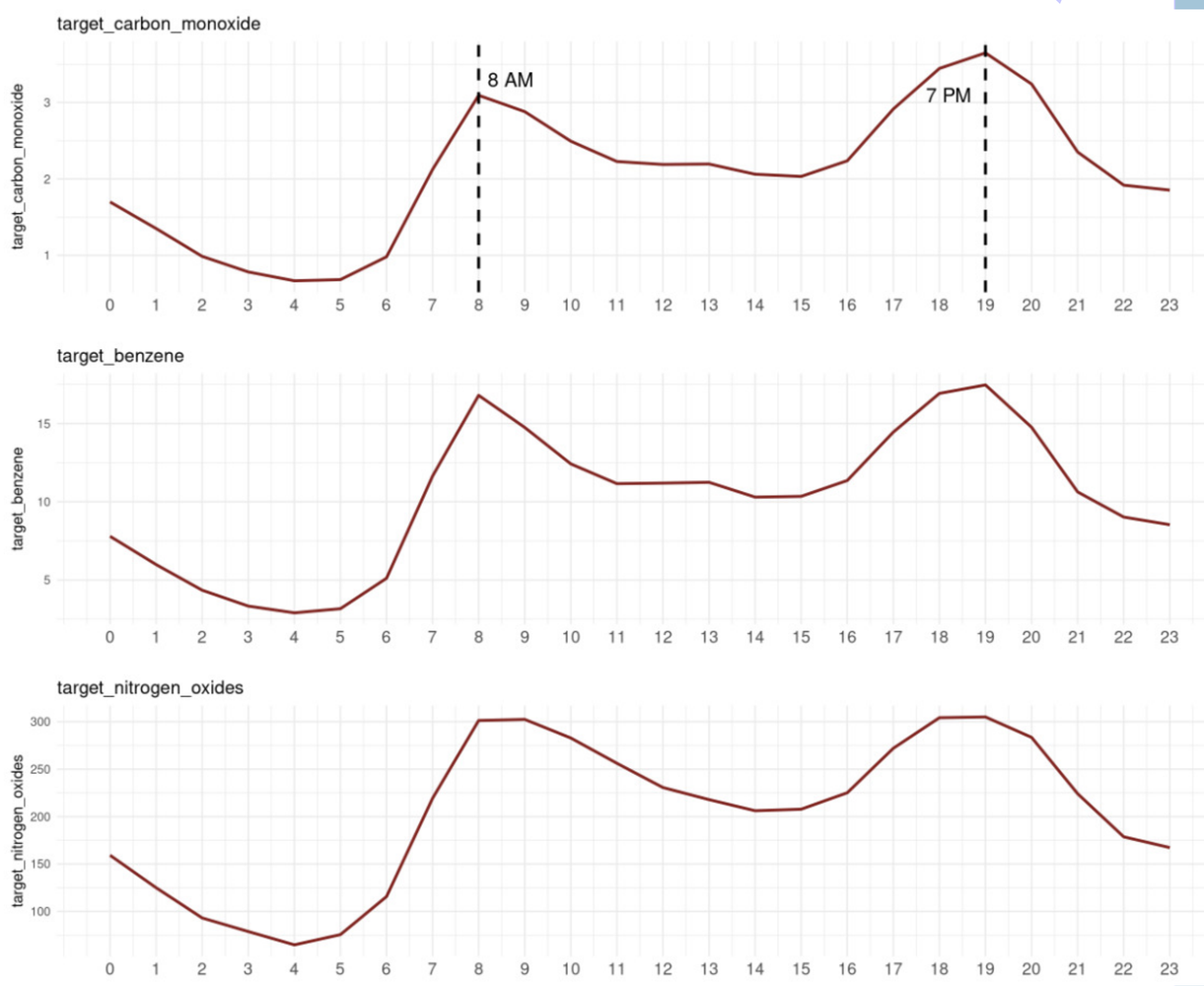
Elements	Number
datetime	2247
degC	280
relativehumidity	653
absolutehumidity	1915
*sensor*1	1758
*sensor*2	1816
*sensor*3	1833
*sensor*4	1877
*sensor*5	2017

Data Visualization



Overall Situation It can be seen from the figure that the values of the three target pollutants in August each year will be lower, gradually rising from September, and significantly higher than the level before August, so it is necessary to take the month as a feature of the model.

Data Visualization



Daily Situation It can be seen from the figure that the level of each pollutant is the lowest at 5:00 a.m. every day, and then gradually rises to 8:00 a.m. to reach the first peak, and then gradually falls to 4:00 p.m., and then rises to 7:00 p.m. to reach the second peak, and then continues to decline, so it is necessary to take time as a feature of the model.

Feature and Model

Features According to the analysis of training data, the following features are used for model training: absolute-humidity, deg-C, relative-humidity, sensor1-5, month, week, is-weekend, hour

LGBMRegressor Data fitting using LGBMRegressor, the algorithm is easy to use. It only needs to put the set features and three prediction targets into the model for training, but there is no parameter optimization, which has a certain impact on the training results.

Result

Evaluation Use RMSLE(Root Mean Squared Logarithmic Error) to evaluate the results

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(\hat{y}_i + 1) - log(y_i + 1))^2}$$

Private Score :0.33979

Public Score :0.387

TULIP Team for Universal Learning and Intelligent Processing

The 11th International Conference on Knowledge Science, Engineering and Management (KSEM 2018), 17-19/08/2018, Changchun, China