AIR POLLUTION MEASUREMENTS PREDICTION

YU LI

ABSTRACT. 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.

CONTENTS

Date: (None).
2020 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Pollution Prediction.

1. 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

2. Data Description

Before model training, data needs to be analyzed to determine the required features. Here is the statistics of training data and test data:

Table 1. Train Data Description

Elements	`Number`
$datetime$	7111
$degC$	408
$relative - humidity$	762
$absolute - humidity$	5451
$sensor1$	3882
$sensor2$	3882
$sensor3$	3882
$sensor4$	3882
$sensor5$	3882
$target - carbon - monoxide$	95
$target - benzene$	405
$target - nitrogen - oxides$	3268

Table 2. Test Data Description

Elements	`Number`
$datetime$	2247
$degC$	280
$relative - humidity$	653
$absolute - humidity$	1915
$sensor1$	1758
$sensor2$	1816
$sensor3$	1833
$sensor4$	1877
$sensor5$	2017

In order to understand the change trend of data, the data is visualized and analyzed based on the visualization results.

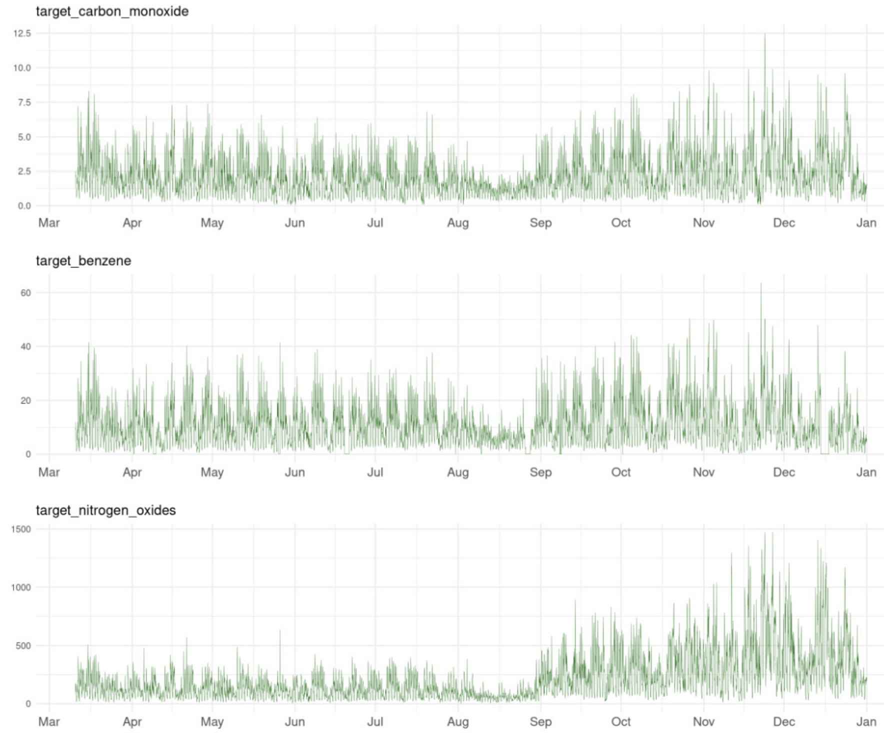

FIGURE 1. Target Overall Situation

It can be seen from the figure1 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.

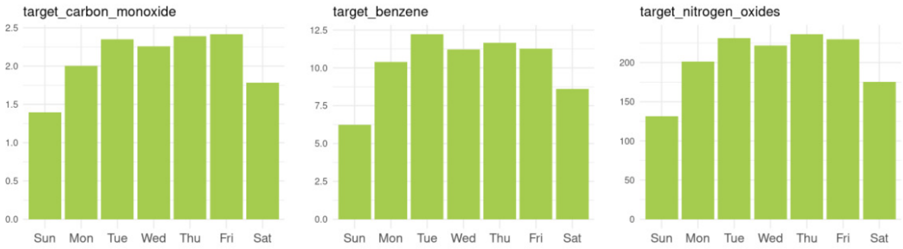

FIGURE 2. Target Weekly Situation

It can be seen from the figure2 that the content level of each pollutant at the weekend of each week will decrease significantly, so it is necessary to take whether this day is a weekend as a feature of the model.

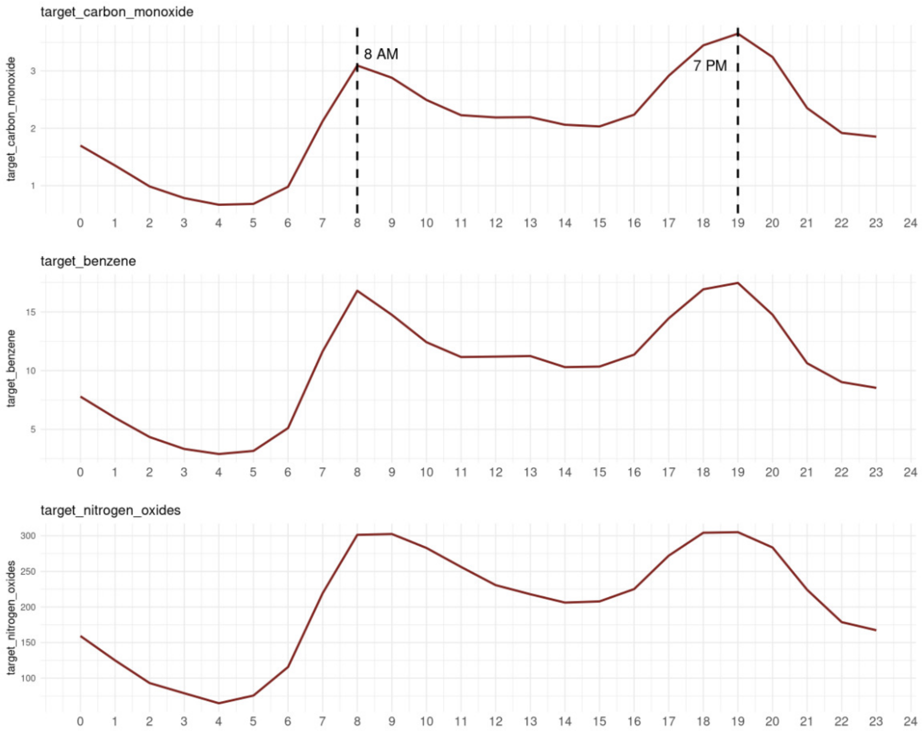

Figure 3. Target Daily Hourly Change

It can be seen from the figure3 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.

3. Feature Engineering

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

4. Model Training

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.

5. Result

- 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

(A. 1) School of Computer Science,, Nanjing University of Science and Technology, Jiangsu 246000, China

Email address, A. 1: `yli@tulip.academy`