Modeling air quality in Mainland China

Tian Xia

5/9/2020

1.Abstract

This report is about modelling air quality in mainland China. As we all know, China is one of the worst air counties in the world. Moreover, the air quality of China became worse and worse in the recent 20 years. In this project, I have mainly two goals. The first task of this project is to use statistics way to find the interpretable relationship between the air quality index and the potential variables. The air quality index is an index for reporting daily air quality. The government always use the air quality index to show how clean or dirty the air is. Besides, I also need using the model to predict the air quality index in the future.

2.Introduction & Background

In recent years, air quality has become a gradually severe environmental problem in the world. The air pollution is destroying our health now. According to the research, nine out of ten people now breathe polluted air, which kills 7 million people every year. The health effects of air pollution are serious - one-third of deaths from stroke, lung cancer and heart disease are due to air pollution. The air quality is having an equivalent effect to that of smoking tobacco. Moreover, the economic cost of air pollution is enormous.The health impacts of air pollution are estimated to cost more than 4% of their GDP. In China, the government spent 900 billion dollars a year for the treatment of air pollution. In the United States, the figure is 600 billion dollars. I come from Chengdu, China. One of the worst air quality city in China. Therefore, I hope I can do something for my community and city to make this situation better. I hope this project can help the government and citizens to decrease the air pollutants in the air.

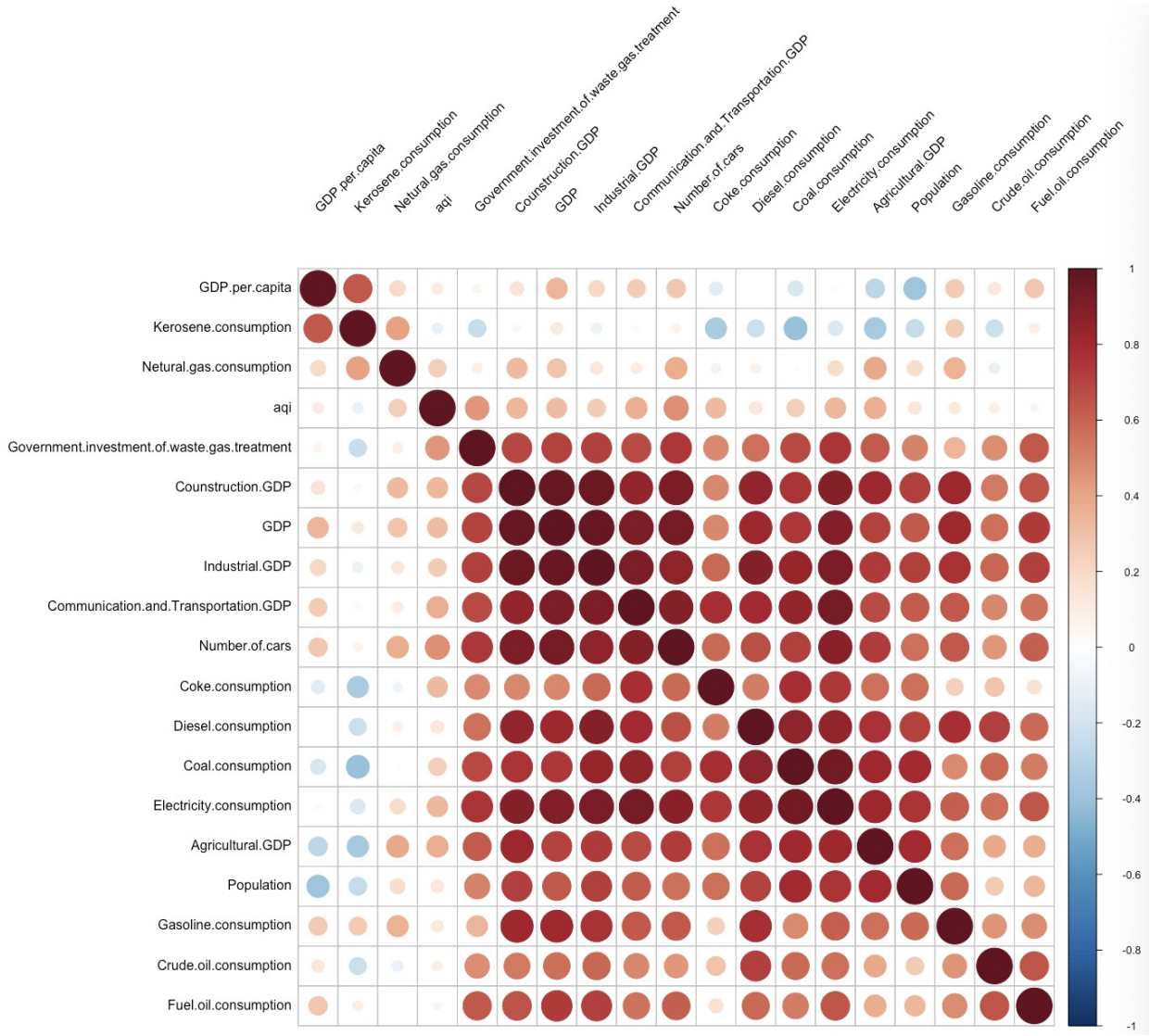
3.Methods

3.1 Data

I have two datasets for this project. The first data source is from Ming Cheng's project, which is a public project created by author Ming Cheng. The project includes every China mainland cities' AQI data for 2000-2015. The author scraped from the governer's website and calculated the AQI year average. I used the first data sources for doing cluster analysis in the first-semester project. The second dataset is the variables dataset been created by myself. The goal of this data set is to prepare the predictor variables for modelling the AQI data. All data in the second dataset is from the National Bureau of Statistics of China. This variable dataset includes 18 variables from 11 provinces with lousy air quality index. Most of the variables are: 1. Energy consumption, including coal, coke, electricity. 2. Economic indicators, including GDP, GDP per capita. By combining these two datasets, now I have a dataset with a response variable(AQI) and the predictor variables for the model.

3.2 Correlation Matrix

The correlation matrix can help us to find the relationship between variables. A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between the two variables. A correlation matrix uses to summarize data as an input into a more advanced analysis.



The result tells us some interesting findings. These 3 variables GDP, Construction GDP, Industrial GDP have a strong positive relation to each other, which means that these total 11 high air pollution provinces are may all industrial provinces. Do these findings explain the relationship between air pollution and industry factor? Moreover, since there is any relationship between the variables of the dataset. Does these relationships indicate that one or more other variables can represent some variables. In other words, there may exist a useless variable in my dataset. By removing the useless variables, the accuracy of my model may increase.

3.3 Feature Selection

The feature selection is an excellent way to find useless variables in the dataset. At the end of last semester and the beginning of this semester, I plan to use the PCA to solve this problem. However, as the result of the PCA algorithm, the variables would reconstitute one component after another. The PCA would be harmful to the interpretability of the model. Remember that, one of my goals is to find the interpretable relationship between variables and AQI. Therefore, I decided to use the feature selection, not PCA to solve the useless variables. The feature selection can find the useless variable but doesn't harm the interpretability of the model. It is easy to make the feature selection by the Boruta package; here are the results.

```
print(boruta.train)
```

```
## Boruta performed 99 iterations in 5.861714 secs.
##  15 attributes confirmed important: Agricultural.GDP,
## Communication.and.Transportation.GDP, Counstruction.GDP,
## Diesel.consumption, Electricity.consumption and 10 more;
##  2 attributes confirmed unimportant: Coal.consumption,
## Coke.consumption;
##  1 tentative attributes left: Crude.oil.consumption;
```

There has 14 variables pass the test, and one variable Coke.consumption is confirmed as an unimportant variable. Therefore, I would remove it for modelling.

3.4 Random Forest

Random forest is one of the most popular classifications and regression algorithm. Random forest's core idea is the decision trees. Most people have used a decision tree in their real lives.

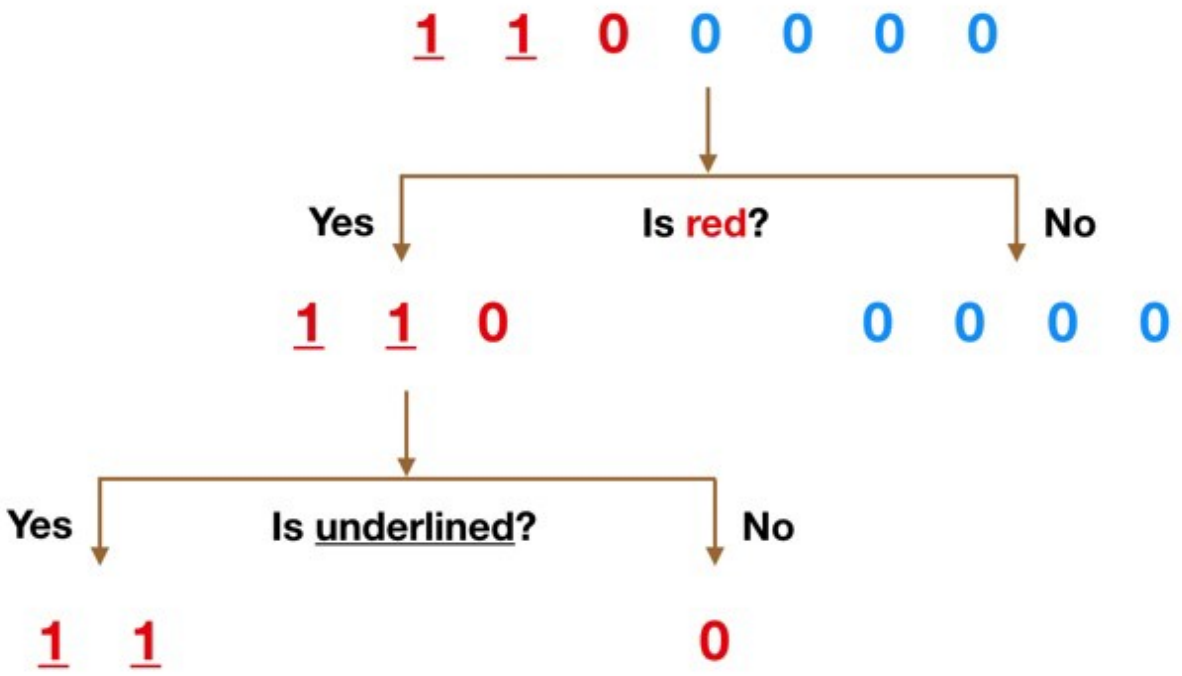


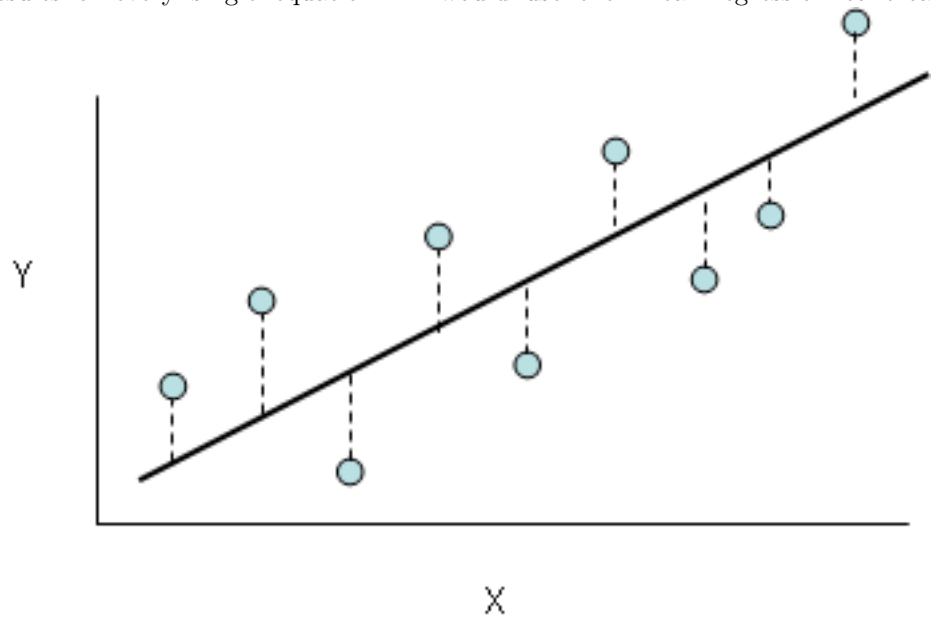
Figure 1: decision tree

This picture clearly explains the idea of the decision tree. The random forest can be seen as a combination of a lot of decision trees. Another advantage of a random forest is that random forest's model can be explained very well. We can use the important feature of random forest package to create the variables' importance plot. It shows the importance of variables from high to low.

3.5 Linear regression

Linear regression is a basic and commonly used type of predictive analysis. The basic idea of the linear regression is the least square, minimizing the sum of the squares of the residuals make in the

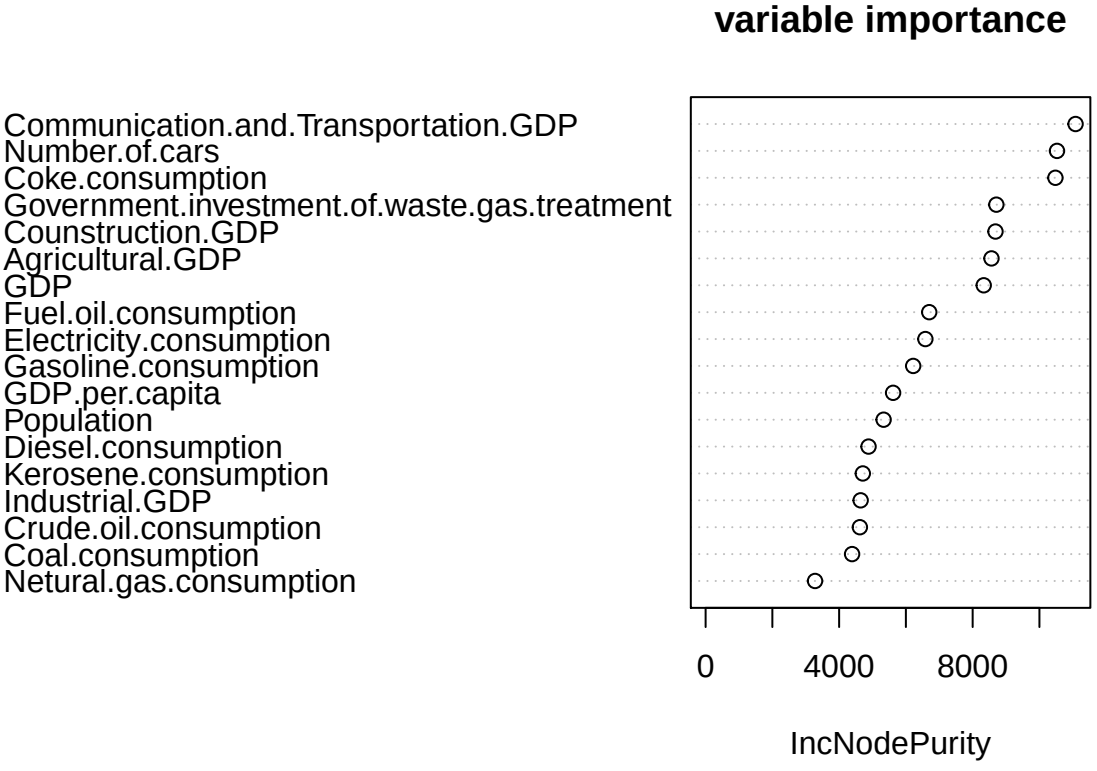
results of every single equation. I would use the linear regression to create the prediction model.



4. Result

First, I used the randomforest package to create a random forest model for my dataset. And using this model to generate the importance plot for features.

```
varImpPlot(hb_tree2, main = "variable importance")
```



The top 5

variables are Construction.GDP, Government investment of waste gas treatment, coke consumption, GDP, Communication&Transportation GDP. These 5 variables make sense to the real-life situation. The Construction would create a lot of dust and particle to the air, which is one of the most critical sources of PM2.5. The coke consumption and Communication GDP belong to the fossil-fuel category, which is the main source of CO2. The GDP is an important economic indicator, it was almost related to all other variables. The government investment variable shows the strong relationship between environment problem and China government because of China's highly concentrated political system. For example, before the 2008 Beijing Olympic Games, the air quality in Beijing was terrible. However, through a large amount of government investment, including afforestation and relocation of factories. The air quality in Beijing got significant improvement. Besides, The importance plot still has some interesting information; I'm surprised at the low ranking of factory variables. A possible reason for this situation is that Most of the selected provinces are light-industry rather than heavy-industry regions. After the importance plot, I started to create a prediction based on the random forest model. I split the dataset into a training set and test set in a ratio of 7:3. Train the model on the training set and test the prediction on the test set. Here is my result of prediction

```
table(hb_pred,testdata$aqil)
```

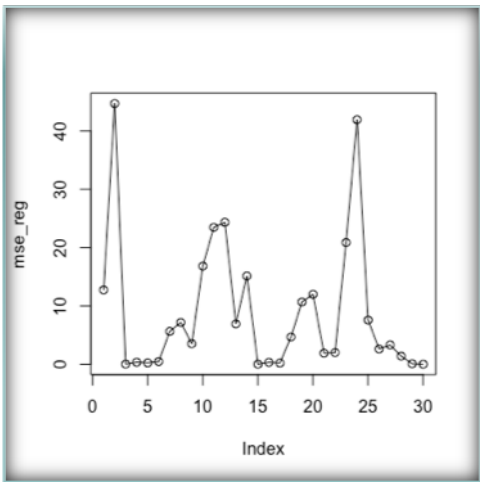
```
##
## hb_pred  2  3  4  6
##       2 26  1  0  0
##       3  0  2  4  1
##       4  0  0  0  0
```

The accuracy of the prediction is very high; the prediction only has one error. Using the random forest model to predict AQI value is very successful.

For the linear regression, I use the lm function to create the linear regression model for my dataset. And using the model to predict the test set. Here are the results.

	pred	obs
1	132.45121	152
4	115.62684	79
9	100.84259	100
52	97.24672	94
56	91.30378	94
67	73.75906	70
68	70.99999	58
76	60.32540	75
83	122.28470	112
85	96.48236	74
86	104.54197	78
99	107.97612	135
100	98.40696	84
101	95.31096	74
103	81.35027	81

Figure 2: predictiom table



We can also use the result of prediction to calculate the mse for my model. Based on these two results, most prediction result error is less than 20, which means that the accuracy of my regression model is not bad. However, the model predicts bad when the true value is too high. In the high true value cases, sometimes the error is more than 50.

5. Conclusion& Future works

In this project, my goals find the interpretable relationship between the response variable and the predictor variable. Moreover, I also want to use the model to predict the AQI value. The random forest performs both good on these two goals. It gives me a convincing importance plot and high accuracy prediction model. However, the linear regression performed not acceptable in this dataset because linear regression is terrible when there has a strong relationship among the predictor variable. Moreover, our dataset doesn't include any extremely high AQI case(AQI>150), which may lead to the model performed poorly when the true AQI value is too high. For the future works, I am interested in the Hubei air quality after the COVID-19. First, Hubei is one of the provinces in my dataset. And recently, I found a piece of news about my research on the internet. This new is about how COVID-19 decreasing the air pollutants in Hubei. Since the COVID-19, the Hubei government shut down almost all activities in Hubei, including industry and transportation. If I can found the dataset for this four-month. Then I can compare this dataset to my current works, to find more useful information.

6. Reference

https://www.displayr.com/what-is-a-correlation-matrix/

https://www.who.int/airpollution/news-and-events/how-air-pollution-is-destroying-our-health

https://www.weforum.org/agenda/2020/04/coronavirus-covid19-air-pollution-enviroment-nature-lockdown/

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

http://www.stats.gov.cn/english/