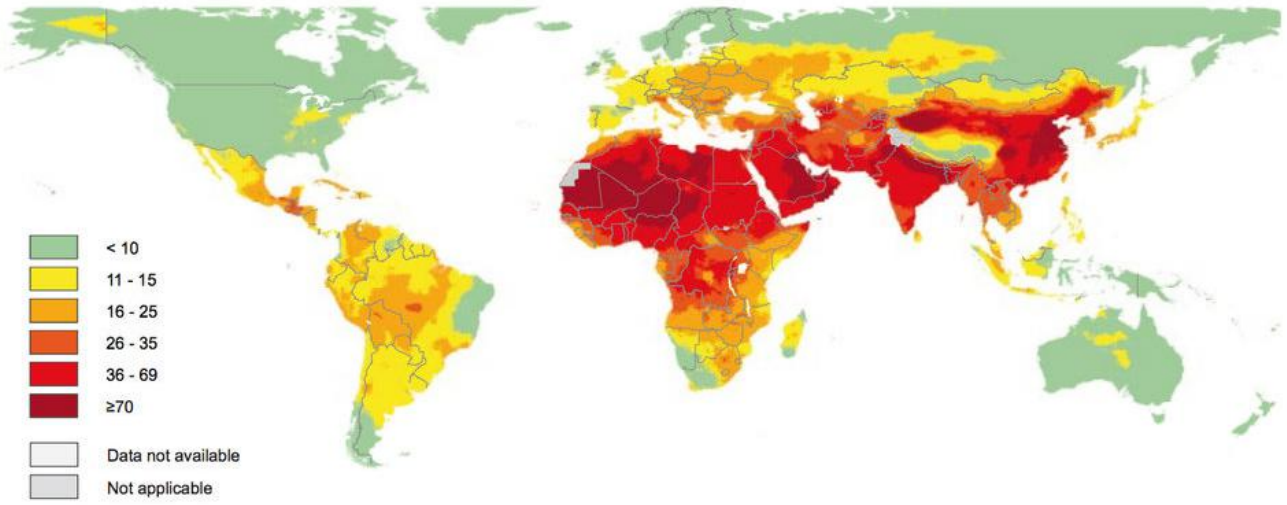
415 Data Mining Final Project-----

Prediction for Beijing PM2.5 and insights into influential factors associated with air quality

Han Zhou & Xiaoxue Xin & Bolun Xiao



- Table of Content:

1. General Data Summary and Visualization

- Data Description and overview:

We downloaded our data from the UCI online machine learning archive:https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data#

This dataset contains Beijing air quality data measured in every hour of every day from January 1, 2010 to December 31, 2014. This data set has one response and seven variables. The response being pm2.5, the pm2.5 particle concentration, and the variables are:

DEWP :Dew Point
TEMP: Temperature
PRES: Pressure
cbwd: Combined wind direction
Iws: Cumulated wind speed
Is: Cumulated hours of snow
Ir: Cumulated hours of rain

We decided to use year 2012 and 2013 data as our training data, which contains 17543 observations, and we use year 2014 data as test data, which contains 8759 observations

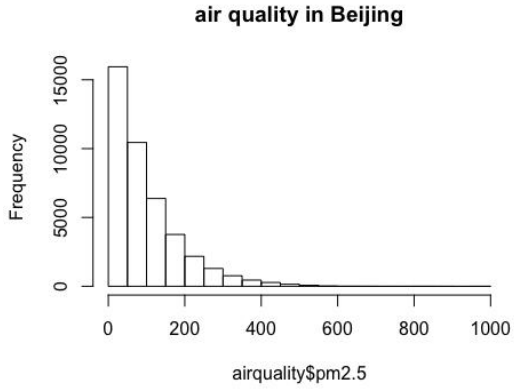
The Air Quality levels are shown in the table below:

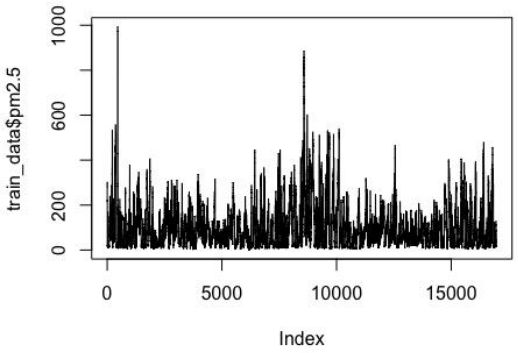
AQI Category	Index Values
Good	0 - 50
Moderate	51 - 100
Unhealthy for Sensitive Groups	101 – 150
Unhealthy	151 – 200
Very Unhealthy	201 – 300
Hazardous	301 – 400
Hazardous	401 – 500

- Data Visualization

Initially, we need to make some plots and specify the basic properties of PM 2.5 value samples. The histogram of air quality in Beijing is shown below.

From the result, we can see that the majority of air quality is fine, but there are 1759 data points whose pm2.5 is greater than 300, which is hazardous. Since our data is based on hour, there are totally two and a half month time that the air in Beijing is very bad. And the maximal value of pm2.5 is 994, which is horrible! We are really concerned with the air quality in Beijing.



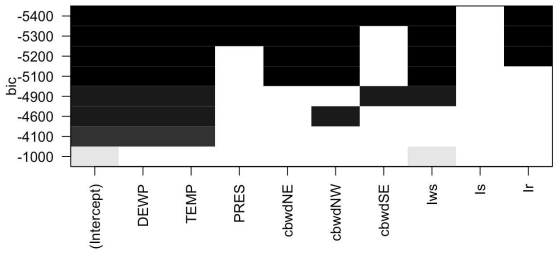


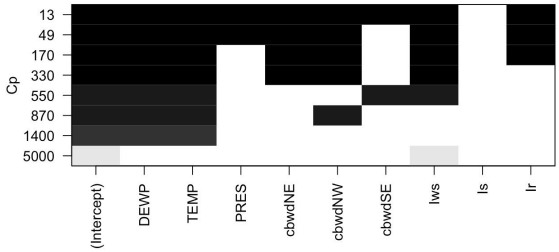
Based on our training data set, we draw a plot of pm2.5. There exists several periods of time that there are high pm2.5. The corresponding plot has been shown above.

2. Dimension Reduction and Variables Choice

- Dimension Reduction and Results

Before implementing the actual models we did variable selection using Cp and BIC





From the plots above we see that both Cp and BIC suggested a new model containing variables DEWP, TEMP, PRES, cbwd, lws, lr

3. Classification/Clustering Methods in Prediction

Initially, we try to predict the air quality by clustering methods. By introducing the Level variable, air quality has been divided into five categories. Then by Clustering and Classification Methods, we want to give accurate prediction classification for our test data and calculate the training & test errors of the predictions.

In this part, we try to LDA, QDA, Single Tree and Random Forest Method:

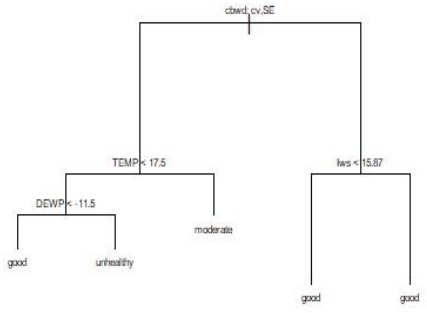
- LDA Method and QDA Method

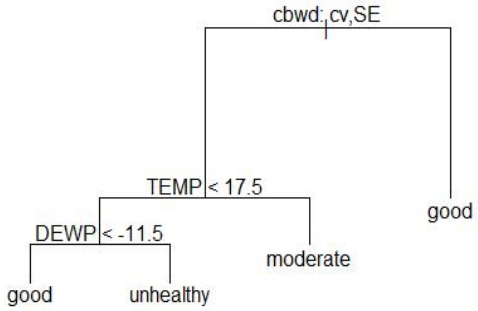
After using the LDA method on the training data and test data to make classification, the training error is 0.4973 and the test error is 0.5253. After using the QDA method, the training error is 0.5860 and the test error is 0.6035.

Although the training error and test error are relatively large in these two methods, we need to clarify that compared with QDA method, LDA should be more suitable for this data set. To be more specific, LDA has restrictive Gaussian Assumptions while QDA assumes quadratic decision boundary. Generally, QDA is more suitable for limited number of training observations since it makes some assumptions about the form of the decision boundary. Here, our training samples are quite large, thus QDA is redundant to some extent.

- Single Tree

Next, we try to use Single Tree to make classifications based on our training data and give level predictions on test data. Also, we conduct the Cross-Validation and try to simplify our decision tree. According to the CV error plot, we choose parameter lambda to be 4 since its error is the smallest with no more than 5 splits and give the corresponding two decision trees as below.





According to the output, the training error rate is 0.5389 and test error is 0.5715. It seems that the training error and test error are still quite large. Besides, for these two decision trees, the training error and test error are almost the same, which indicates that the classification accuracy can hardly be improved by simply changing parameters. Thus we try to use Random Forest Method to improve the accuracy.

- Random Forest Method

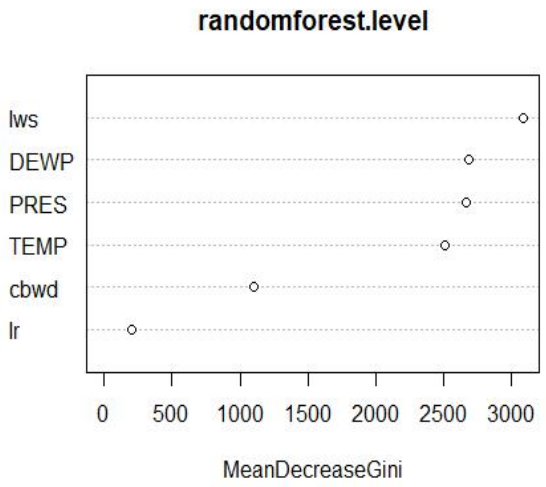
In order to use the Random Forest Method to improve the accuracy, we set number of trees to be 1000 and 5000.

For number of trees =1000:

The plot below shows the importance of variables. It seems that IWS, DEWP, PRES and TEMP are the four most important variables among all variables. According to the output, the training error is 0.3658 and the test error is 0.5684. The test error indicates that Random Forest method performs better than a single tree. Also, this training error is the

smallest compared with all other methods. The table on the right hand shows the classification error for these five groups.

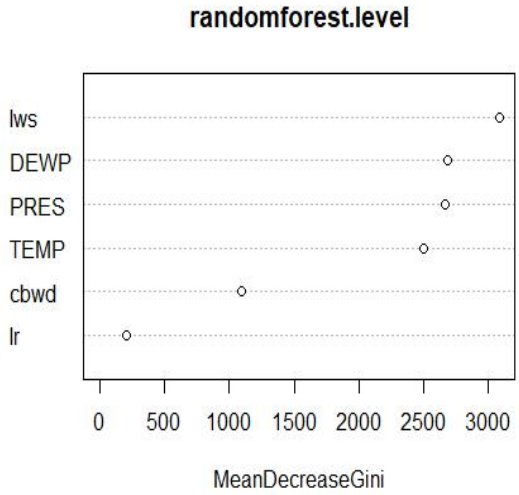
It can be seen that among these five groups, the first group has the smallest classification error, 0.1826683, which indicates that random forest methods is quite suitable for the prediction in this group.



	Classification Error
Good	**0.1826683**
Hazardous	**0.5961272**
Moderate	**0.4923077**
Unhealthy	**0.4182942**
Very Unhealthy	**0.5690608**

For number of trees =5000:

The plot below also shows the importance of variables and the conclusion is the same as the previous case. According to the output, the training error is 0.3658 and the test error is 0.5684 and it indicates that increasing the number of trees would not improve the classification accuracy. The table on the right hand shows the classification error for these five groups.



	Classification Error
Good	**0.1846945**
Hazardous	**0.6016598**
Moderate	**0.4875000**
Unhealthy	**0.4217553**
Very Unhealthy	**0.565375**

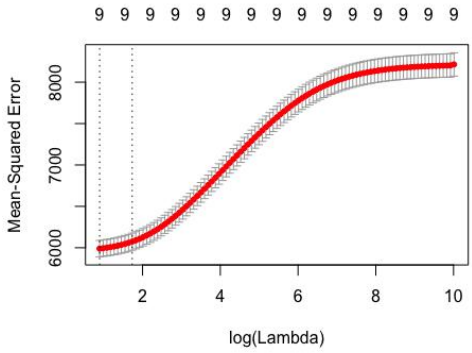
4. Regression Methods in Prediction

Next, we try to predict the air quality by regression methods. We use linear and non-linear regression to predict the specific value of pm2.5. In the case of linear regression, we use penalized regression, ridge regression and lasso.

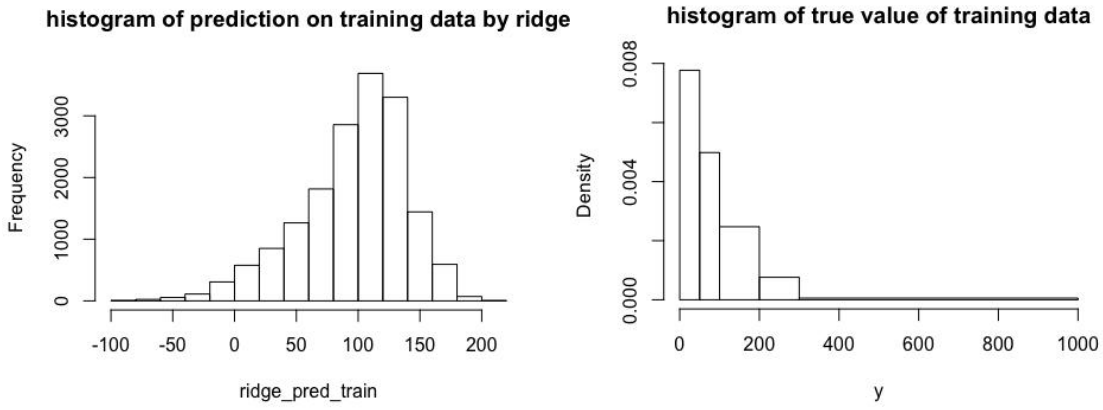
4.1 Linear Methods

- ridge regression

By using cross validation, we choose our lambda as 2.450472, which can be seen from the below plot.



Based on the specific value, we calculate the errors. The training mean squared error is 5983.056, and test error is 6644.661. Since our goal is to predict the degree of air quality, so maybe the errors will become less if we classify them into different classifications, which has tolerance toward moderate error.

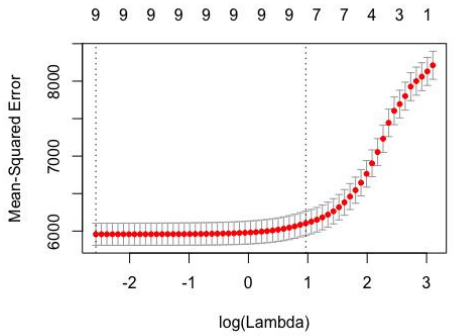


Compared with the histogram of true value, the shape of prediction is far from satisfied. And we can also see from the prediction plot that there exist some negative values,

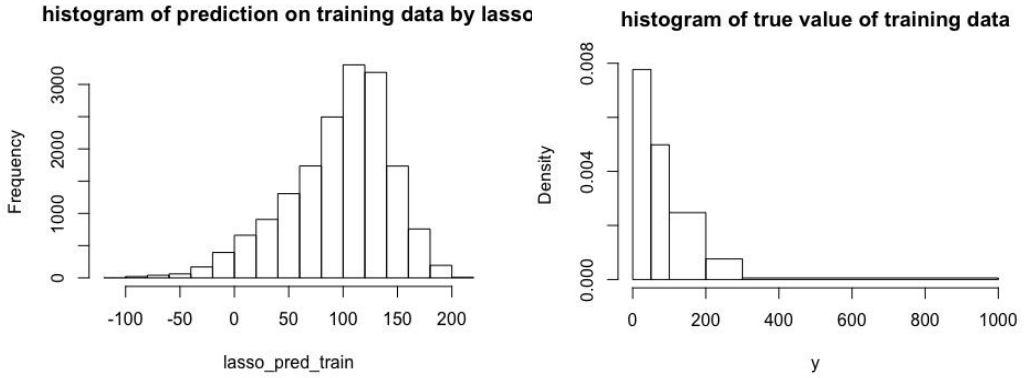
which cannot be true. The training error is 0.5980675. And the test error is 0.5856137. So, we seek other methods to see whether there exists better prediction results.

- lasso regression

Again, we use cross validation to choose our lambda, which is also shown in below plot.



The training error is 5983.056, and the test error is 6627.308, which are similar with ridge regression. Compare the two plots, the result is not satisfactory.



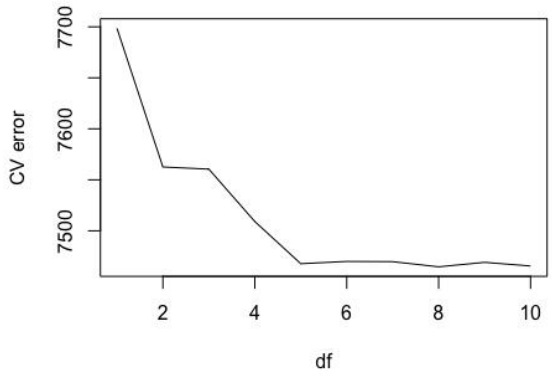
Thus, lasso and ridge may have similar result. The training error is 0.5861073, and the test error is 0.5709502.

Next, we try to use some nonlinear method to improve our result.

4.2 nonlinear regression

- natural spline

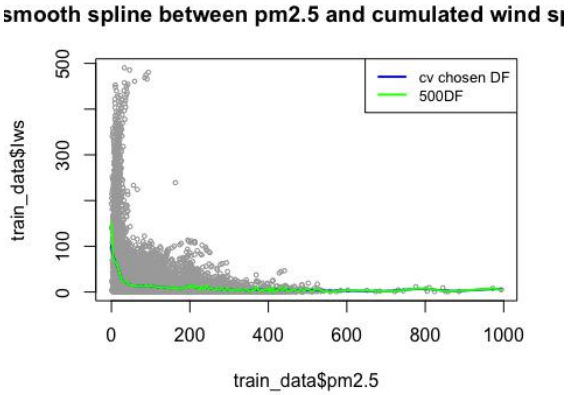
First, we use natural spline to fit our training data. And we choose the degree of freedom again by cross-validation. Thus we choose degree of freedom as 8.



The training error and test error are 0.691569 and 0.7024593, which are even bigger that the results in linear regression.

- smoothing spline

We use smoothing spine to draw a plot to see why the non-linear regression has even worse result. We can see the relationship between pm2.5 value and wind speed from the plot below.



From the plot, we can see that when the wind speed is fast, the pm2.5 is low. We can imagine that the wind blow away the harmful material. And when the speed is slow, the range of pm2.5 is big. The pm2.5 value may be influenced by the other variables. From the plot, the prediction seems not good at small values of pm2.5. When we set a big degree of freedom, i,e, more knots in our regression, it seems better at the small value part, but the result is still not desirable.

5. Summary/Conclusions

Our purpose is to build a predictive model in order to classify pm2.5 values into different air quality categories. We introduced a new categorical variable for further classification/regression analysis.

In our analysis, simple regression methods or clustering methods are not quite suitable for this data set. Random Forest Method has the best prediction performance, while its error is still over 40%. However, if we combine these two kinds of methods, we can find that when PM2.5 value is large, smoothing spline has quite good regression effect. Meanwhile, Random Forest Method has better performance on classification when the PM2.5 value is relatively smaller. Thus, if we can divide our observations into two groups in advance and predict respectively, the performance would be improved a lot.

6. Limitations/Future Work

According to the Training Error and Test error, current methods including regression and clustering ones do not perform very well on current data set.

There are two shortcomings:

- We introduced a new categorical variable quality level to the data set for classification and clustering. While in fact, the level classification regulations and standards would influence the performance of predictions models. It is possible that other level classifications regulations would be more suitable for this data set.
- Generally speaking, PM2.5's negative influence is becoming larger nowadays. It indicates that some Time Series Analysis should be introduced. Also, weather factors would influence the level of PM2.5, thus some seasonal effects should also be introduced.

7. Individual Contributions Description

Han Zhou: Data Description, Data Summary, Dimension Reduction and Presentation

Xiaoxue Xin: Linear and Nonlinear Regression Methods implement, Report Edit

Bolun Xiao: Classification and Clustering Methods implement, Report Edit