Predicting PM2.5 levels of Beijing, China

Charles Liu, Hao Qiu, Curties Wurster, Jason Washam

4/18/2018

Abstract

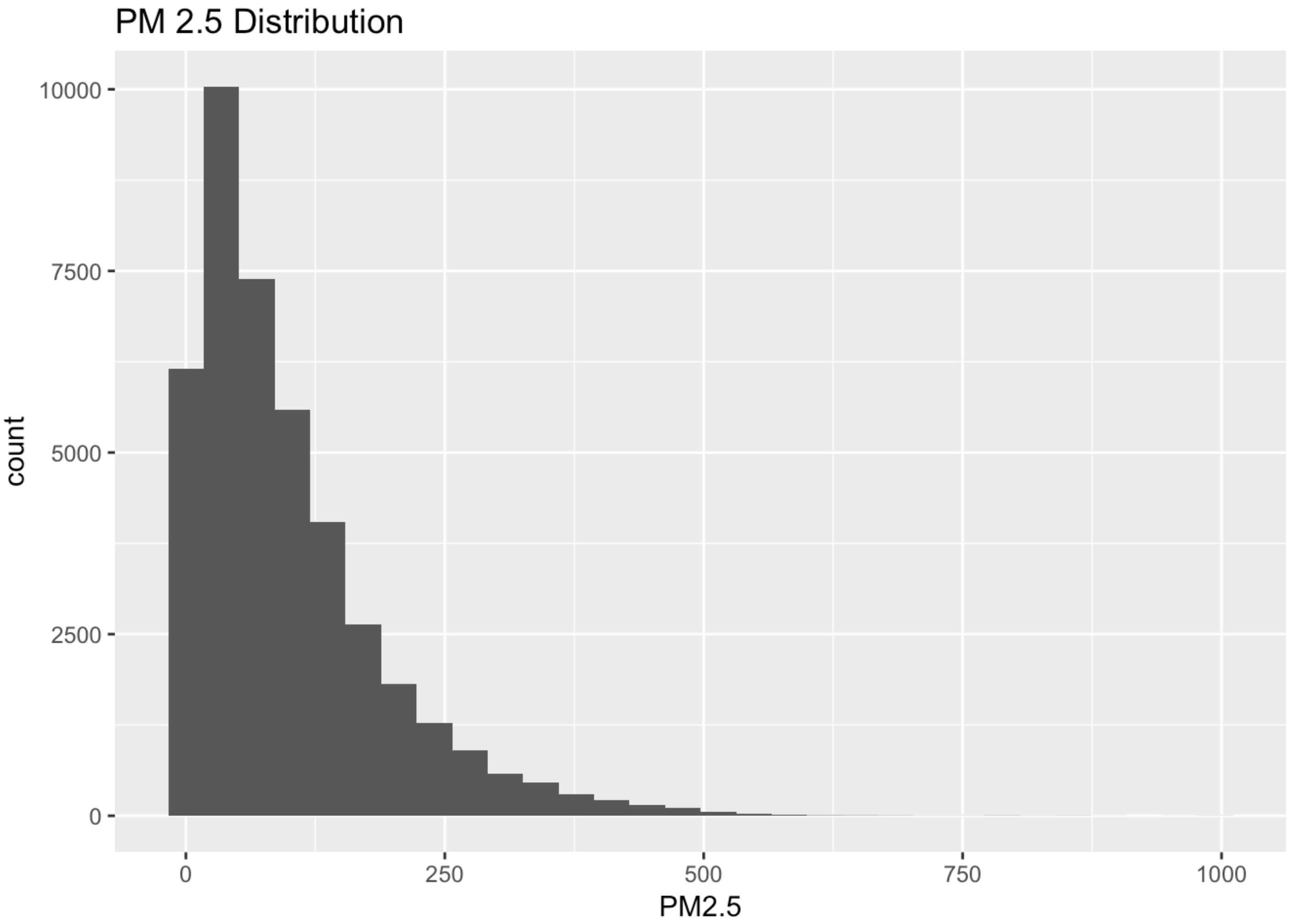
Due to the rapid growth of economics and urbanization, air quality problem catches both government and public's eyes, especially in some most populated and industrialized cities such as Beijing. Beijing, the capital of the People's Republic of China and the world's second most populous city, is known for air pollution and constantly batting against widespread health problems caused by air pollution. To help citizen of Beijing with this issue, we decided to make a statistical prediction model based on the past climate and PM2.5 data in order to give alerts to citizen of Beijing. The alerts will advise the citizens to take the necessary health caution.

Introduction

Over the past few decades, rapid urbanization and industrialization has left China in a struggle to improve and contain serious environmental issues, especially issues concerning extensive air pollution caused by PM2.5. Fine particulate matter 2.5 (PM2.5) refers to tiny particles or droplets in the air which can travel deeply into the respiratory system, reducing lung function and having the ability to cause cancer in cases of long-term exposure. The government wants to find a way to predict the PM2.5 level in the city of Beijing a day in advance in order to give alerts to the citizens to take protective measures ahead of time, decreasing the chance to inhale PM2.5. To efficiently predict the air quality of Beijing, it is crucial to understand the factors influencing pm2.5. By using significant variables provided in the data set, we aim to predict the class of pm2.5 based on the classification model and furthermore propose potential approaches to address air pollution issues.

Explanatory Data Analysis

PM 2.5 Value Distribution



Methods

Data

Data Description

Our data set contains 43,824 observations with 13 different variables. This data set concerns PM2.5 levels in Beijing from 2010 to 2014, including variables such as year, month, pm2.5 concentration(numeric), temperature, and air pressure etc. By using the selected features, we aim to predict the class of pm2.5 based on the **classification** model and furthermore propose potential approaches to address air pollution issues. Therefore, this analysis, albeit lacking in chemical analysis, is statistically meaningful.

Data Cleaning

There are 2067 NA in pm2.5 column. However since our data set contains total 43824 rows, we can remove rows where pm2.5 is NA. Also, we removed the row indexing numbers. We also created additional columns Season, and Time, which are a categorical variable indicating the month and hour of the data. Season will contain four levels, namely Spring(month 3 to 5), Summer(month 6 to 8), Autumn(month 9 to 11) and Winter(month 12 to 2). Time will have four levels, namely Morning(hour 6 to 11), Afternoon(hour 12 to 17), Evening(hour 18 to 22), Midnight(hour 23 to 24, 0 to 5). We will remove `month` , `day` and `hour` after we create Season and Time.

The details of variables can be found in appendix

Classification

We divided the response variable, PM2.5, into two classes: good and dangerous. Our good classification refers to satisfactory air conditions with little to no risk posed through air pollution. Our hazardous classification warns that outdoor activity should be completely avoided due to the fact that everyone will be affected by the air quality. We labeled PM2.5 level greater than 100 as dangerous and below 100 as good air condition. With this classification, the public will able to easily understand the result of the prediction model.

Test Train Split

Considering that we are going to perform 5-fold cross validation and calculate test accuracy to compare models, we split our data set into two parts. 80% of the data set is our training data and the other 20% is our testing data.

Model

Since we decided to use classification method, we discussed methods such as K Nearest Neighbor, Random Forest, and Linear Discriminant Analysis. In addition, we also wanted to perform cross validation in order to obtain accuracy of the model. The model accuracy and the test accuracy is measured for each model to compare and decide the best model for prediction. Using caret, we decided to use Logistic regression, K Nearest Neighbor (scaled and not scaled), Random Forest (out of bag and cross validated) and gradient boosting method for our analysis.

Result

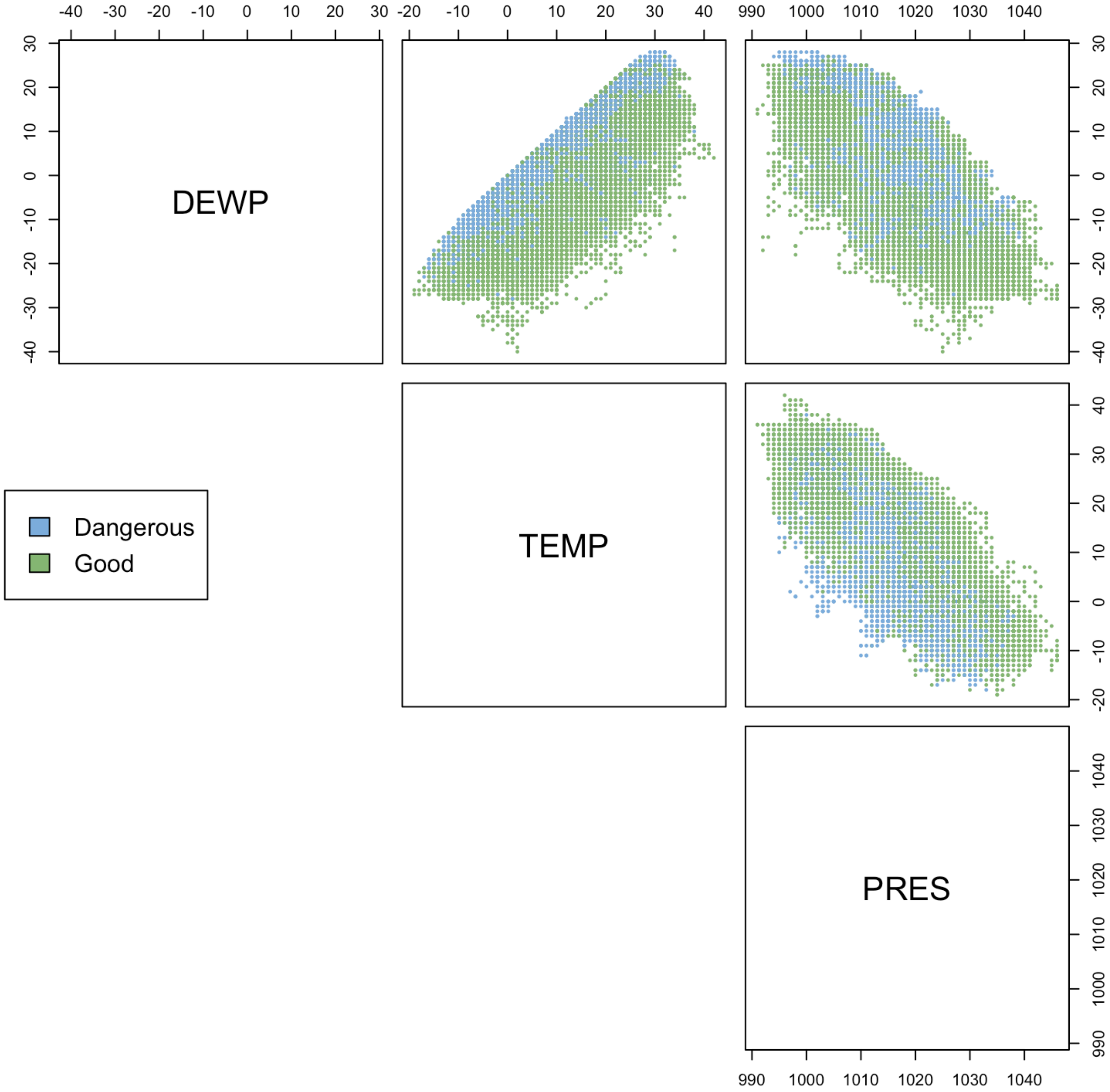
Model	Train Accuracy	Test Accuracy
logistic Regression	0.7363647	0.7399114
KNN CV	0.7521104	0.7588313
KNN CV scale	0.7746814	0.7824213
Bagging oob	0.7940190	0.7958328
Random Forest CV	0.7970725	0.8030176
Boosting, CV	0.7733342	0.7782302

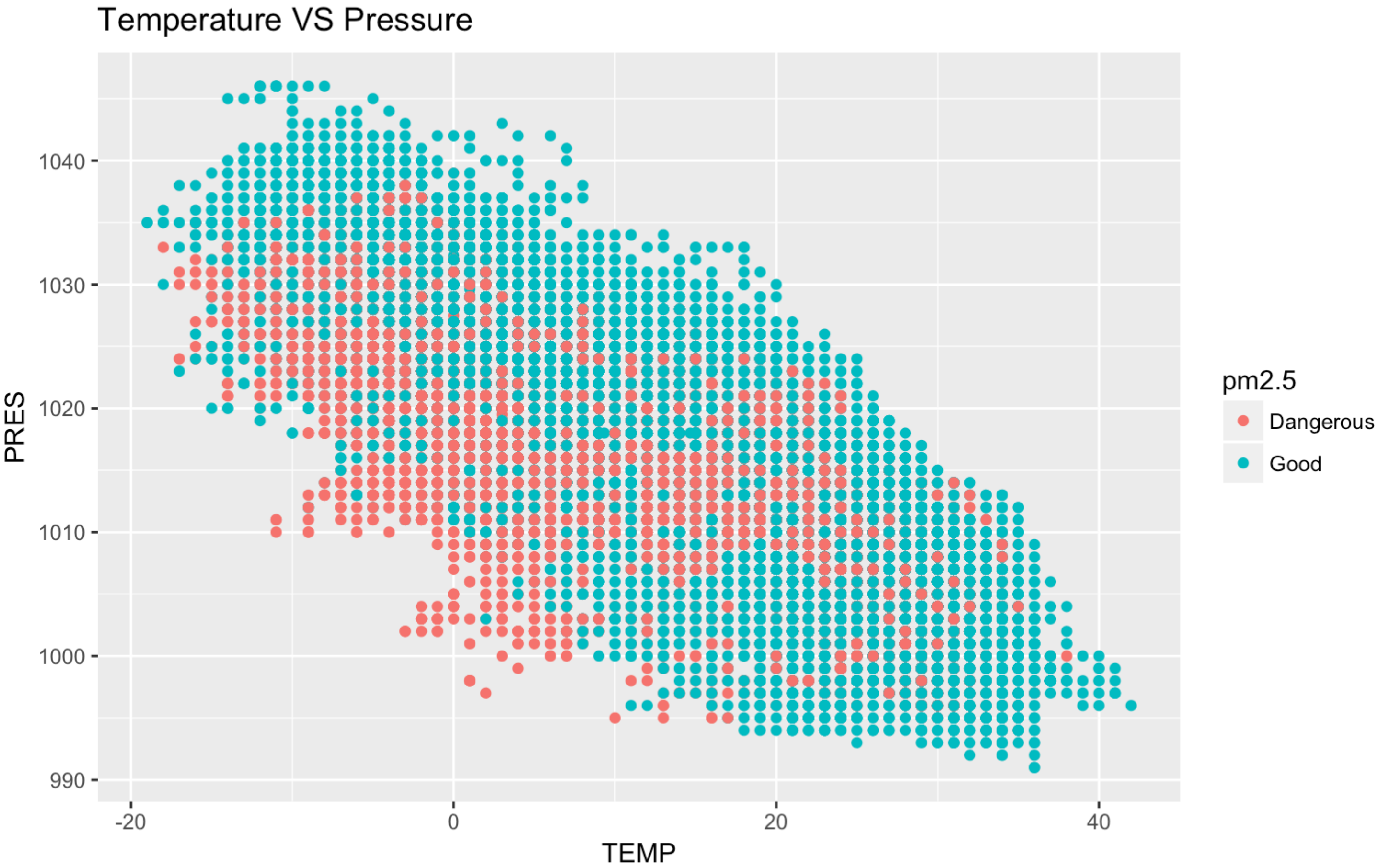
Discussion

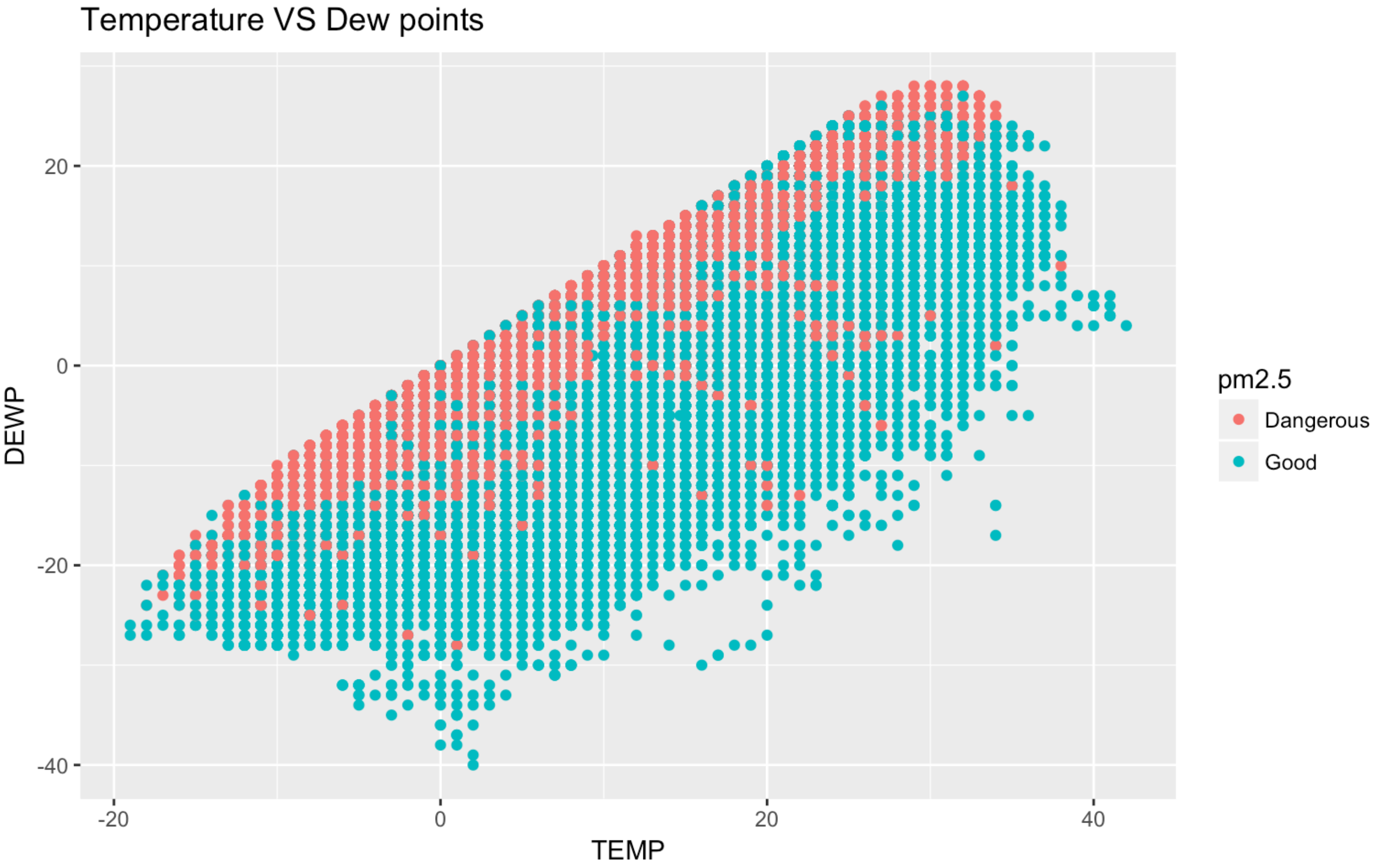
The result shows the train and test accuracy of each different cross validated models starting from simple logistic regression to generalized boosted regression. We also included the bagging model to see if ensemble method works better with our data. Overall, the cross validated random forest model with 5 randomly selected predictors (mtry = 5) had the best accuracy of predicting the PM2.5 level based on the climate data. We also check variable importance and found that dew point is the most influencing variable in predicting the PM2.5 level. In addition, the variable importance shows that the cumulated hours of snow variable is not influencing the prediction.

Appendix

Name	Data Type	Description
Year	Continuous	Year of data in this row
PM2.5	Categorical	PM2.5 Concentration (ug/m^3)
DEWP	Continuous	Dew Point (Celsius Degree)
TEMP	Continuous	Temperature (Celsius Degree)
PRES	Continuous	Pressure (hPa)
cbwd	Categorical	Combined wind direction
lws	Continuous	Cumulated wind speed (m/s)
ls	Continuous	Cumulated hours of snow
lr	Continuous	Cumulated hours of rain
season	Categorical	Season of data in this row
time	Categorical	Time of data in this row


DEWP
TEMP
PRES
Dangerous
Good
-40
-30
-20
-10
0
10
20
30
990
1000
1010
1020
1030
1040


Temperature VS Pressure
PRES
1040
1030
1020
1010
1000
990
TEMP
-20
0
20
40
pm2.5
Dangerous
Good


Temperature VS Dew points
DEWP
20
0
-20
-40
TEMP
-20
0
20
40
pm2.5
Dangerous
Good