Air Pollution PM2.5 Prediction of Guangzhou City

Presented by: Rebecca Chen

April 16, 2019

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Today's Outline

- Problem Derivation
- Dataset Description
- Prediction Models Details
- Models Comparison and Selection
- Future Work

❖ Problem Derivation

- What is PM2.5?

PM2.5 : particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair.

❖ Problem Derivation

- What is PM2.5?

PM2.5 : particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair.

- Why should we pay attention to PM2.5 pollution?

In the past few years, levels of smog have increased throughout China. Since PM2.5 is so small , it penetrates deeply into the lung, and impair lung functions.

❖ Problem Derivation

- What is PM2.5?

PM2.5 : particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair.

- Why should we pay attention to PM2.5 pollution?

In the past few years, levels of smog have increased throughout China. Since PM2.5 is so small , it penetrates deeply into the lung, and impair lung functions.

- How does this project help to ease the problem?

Accurate prediction can helpfully guide local citizens prepare for safely hangouts.

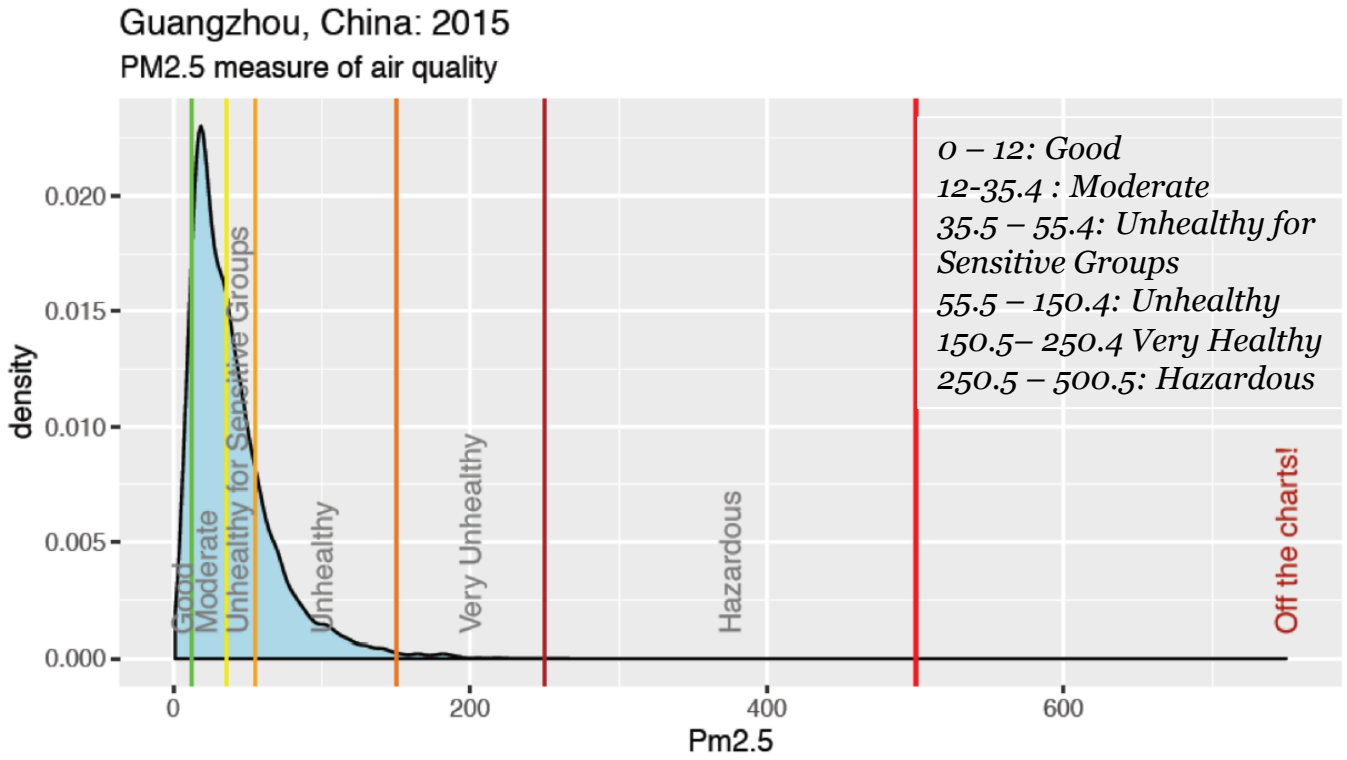
❖ Dataset Description

This dataset records 8760 observations for hourly PM 2.5 data of year 2015 in Guangzhou, China.

❖ Dataset Description - Response Variable

This dataset records 8760 observations for hourly data of year 2015 in Guangzhou , China.

How does PM2.5 look like in 2015, Guangzhou, according to the dataset?



❖ Dataset Description – Categorical Variables

hour: hour of data in this row

season: season of data in this row

cbwd: Combined wind direction

❖ Dataset Description – Numerical Variables

*DEWP: Dew Point (Celsius Degree)

TEMP: Temperature (Celsius Degree)

HUMI: Humidity (%)

PRES: Pressure (hPa)

Iws: Cumulated wind speed (m/s)

precipitation: hourly precipitation (mm)

Iprec: Cumulated precipitation (mm)

* The **dew point** is the temperature to which air must be cooled to become saturated with water vapor. When further cooled, the airborne water vapor will condense to form liquid water (dew).

❖ Model Details (1/3) – Smoothing

Fits a generalized additive model (GAM) to data with smoothing terms.

Baseline Model: No interaction

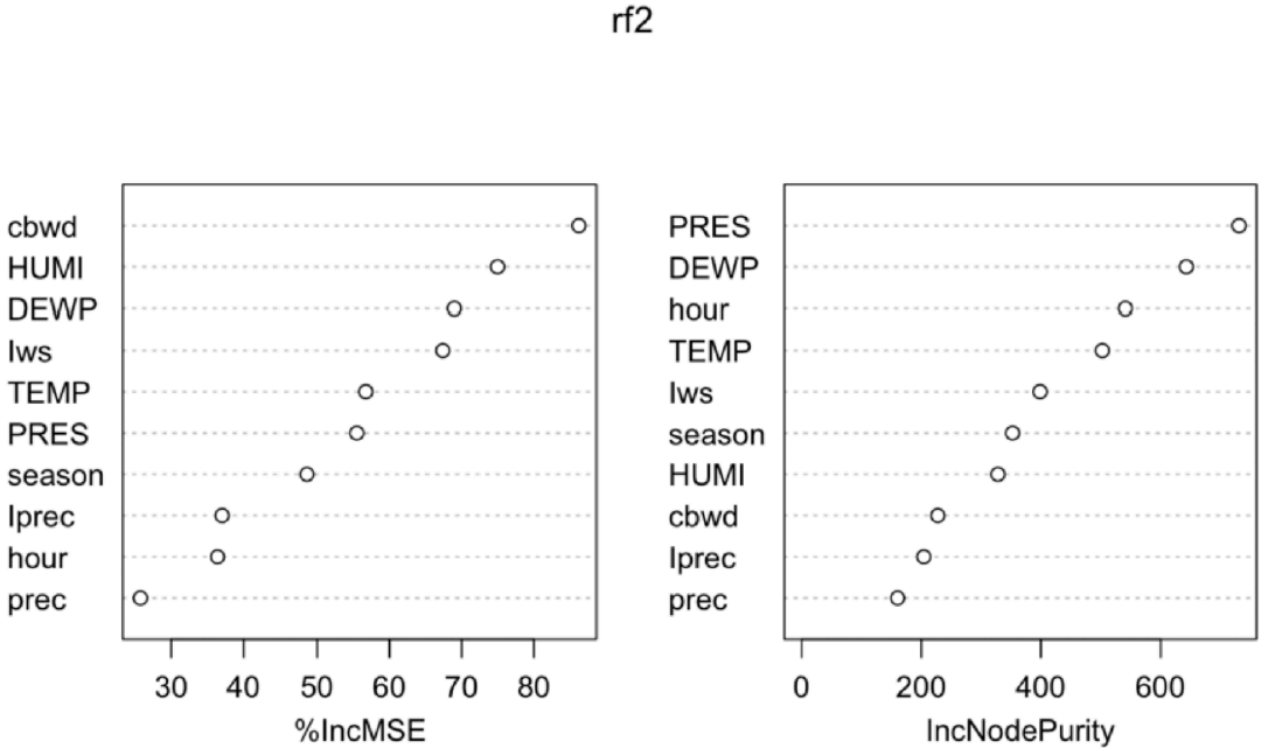
Pm2.5 ~ hour + cbwd + season + s(TEMP) + s(HUMI) +s(PRES) + s(DEWP)+ s(Iws) + s(precipitation) +s(Iprec)

Add interactions, stop at the best model with interactions of :

- DEWP, Iprec
- DEWP, TEMP
- TEMP, Iprec

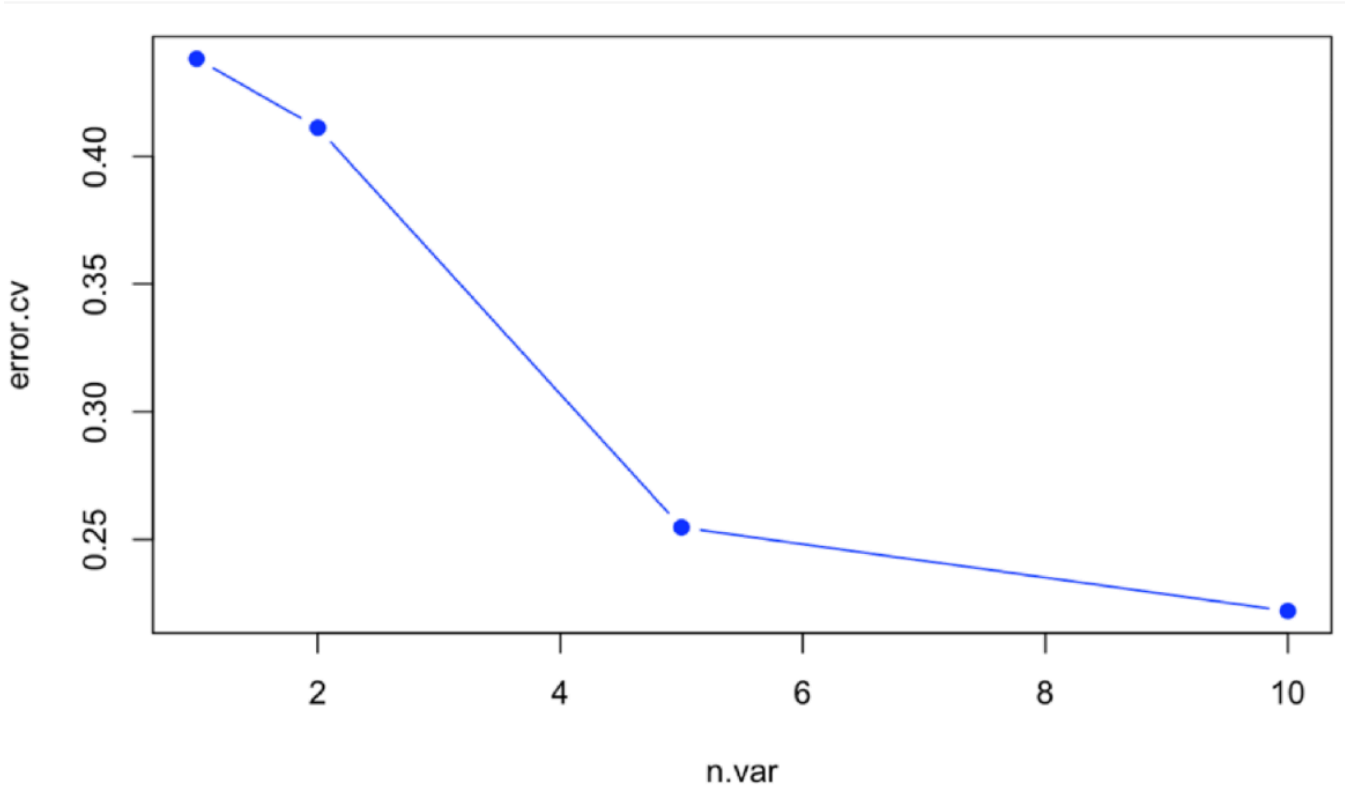
❖ Model Details (2/3) – Random Forest

Importance of Variables:



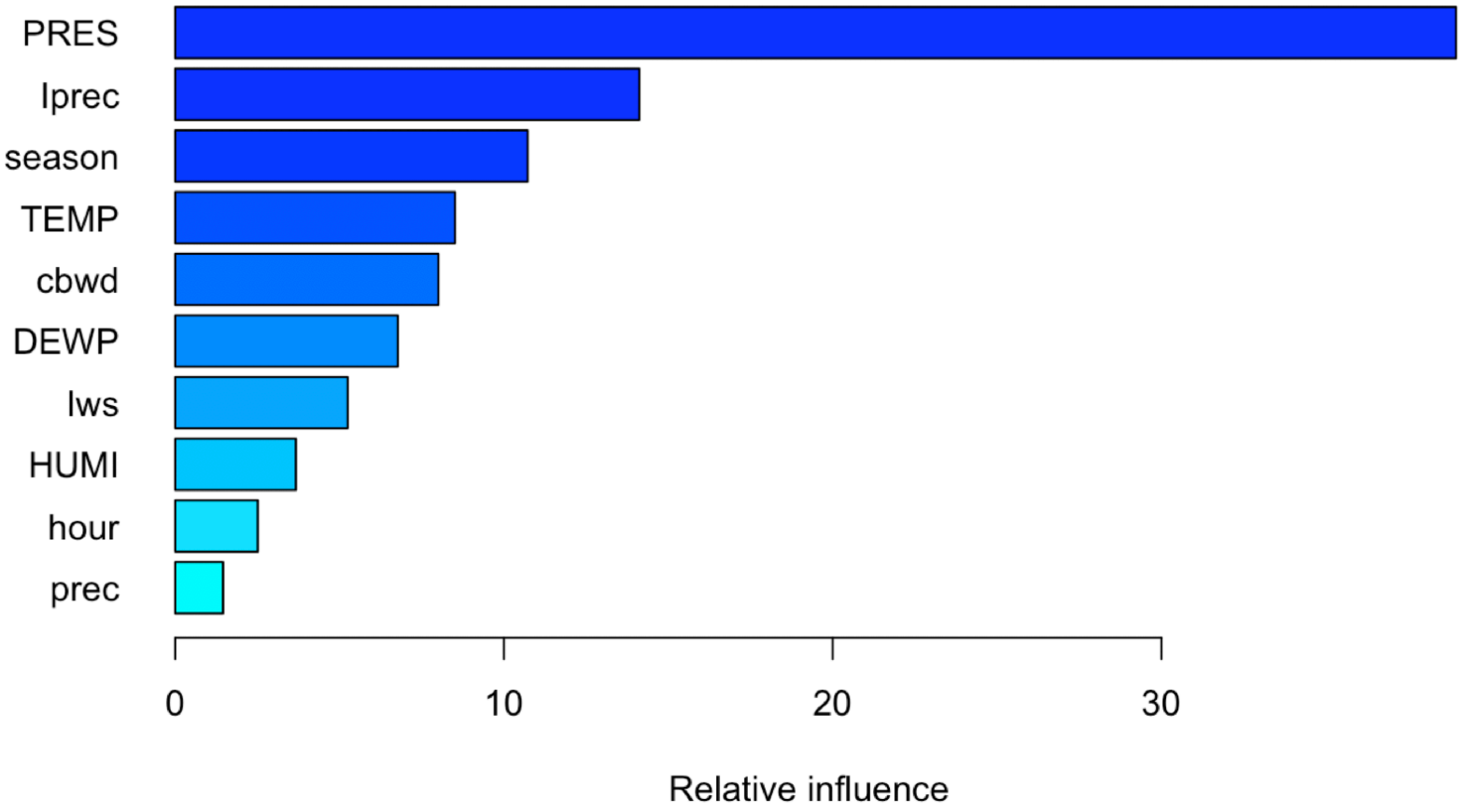
❖ Model Details (2/3) – Random Forest

Feature Selection:



❖ Model Details (3/3) – Gradient Boosting

Tuning Parameters: Shrinkage (0.8), number of trees(700)



❖ Model Selection and Comparison

Model	Interaction	5-fold CV Err	Cmpt Time (s)
GAM	No (m)	0.3247	17.986
	One (m7)	0.2924	28.191
	Two (m24)	0.2879	94.278
	Three (m25)	0.2876	194.407
Random Forest	No	0.2219	302.54
Gradient Boosting	No	0.2974	10.192

- Accuracy
- Computation Time
- Ease of Use
- Difficulty of Interpretation

Which model is better ?

❖ Model Selection and Comparison

Model	Interaction	5-fold CV Err	Cmpt Time (s)
GAM	No (m)	0.3247	17.986
	One (m7)	0.2924	28.191
	Two (m24)	0.2879	94.278
	Three (m25)	0.2876	194.407
Random Forest	No	0.2219	302.54
Gradient Boosting	No	0.2974	10.192

- Accuracy
- Computation Time
- Ease of Use
- Difficulty of Interpretation

Depends on Purposes!

❖ Future Work

- Fully considerations of interactions in smoothing parts. (factor-factor interaction , numerical – factor interaction)
- All three models indicate to remain all 10 variables, however with not quite same important variables rankings.
- More significant predictors, such as PM10, CO(Carbon monoxide), ozone.

Thanks!

Question?