AIR QUALITY
Time-Series Forecasting

JOH Minho (54900228)
KU Dayeon (55824711)
LEE Yewon (55308685)

TABLE OF CONTENTS

01 BACKGROUND & PROBLEM FORMULATION

02 DATA ANALYSIS

03 METHODS & JUSTIFICATION

04 EXPERIMENTAL RESULTS

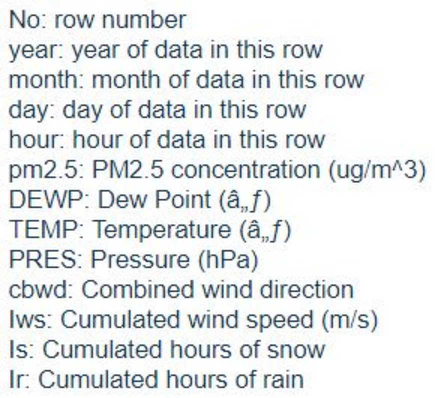
05 DISCUSSION & EVALUATION

Background

UCI

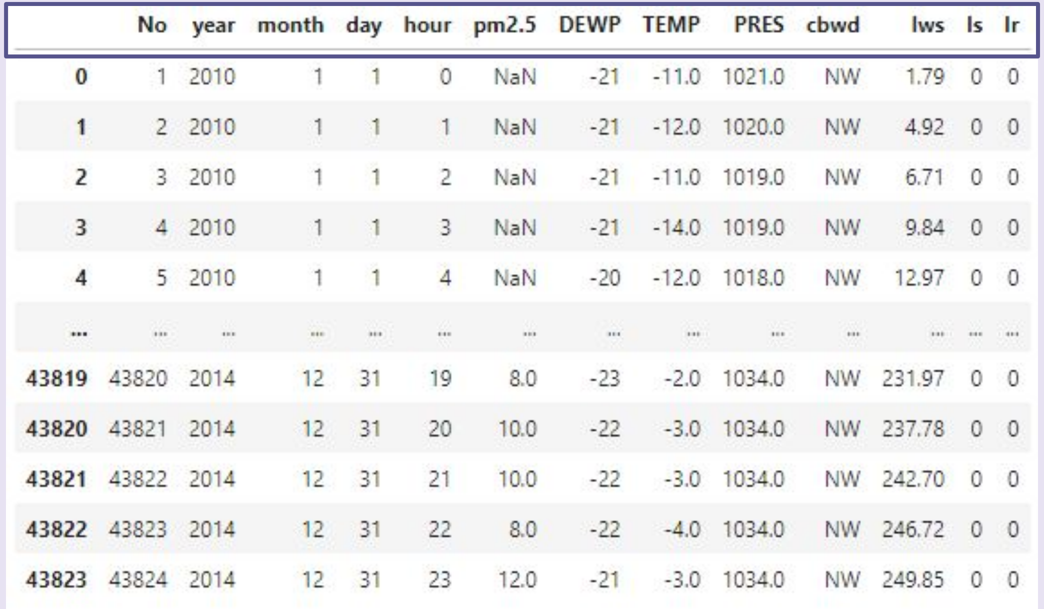
Machine Learning Repository

Attribute Information



No: row number
year: year of data in this row
month: month of data in this row
day: day of data in this row
hour: hour of data in this row
pm2.5: PM2.5 concentration (ug/m^3)
DEWP: Dew Point (â„ƒ)
TEMP: Temperature (â„ƒ)
PRES: Pressure (hPa)
cbwd: Combined wind direction
Iws: Cumulated wind speed (m/s)
Is: Cumulated hours of snow
Ir: Cumulated hours of rain

PM2.5 DATASET



	No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
0	1	2010	1	1	0	NaN	-21	-11.0	1021.0	NW	1.79	0	0
1	2	2010	1	1	1	NaN	-21	-12.0	1020.0	NW	4.92	0	0
2	3	2010	1	1	2	NaN	-21	-11.0	1019.0	NW	6.71	0	0
3	4	2010	1	1	3	NaN	-21	-14.0	1019.0	NW	9.84	0	0
4	5	2010	1	1	4	NaN	-20	-12.0	1018.0	NW	12.97	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
43819	43820	2014	12	31	19	8.0	-23	-2.0	1034.0	NW	231.97	0	0
43820	43821	2014	12	31	20	10.0	-22	-3.0	1034.0	NW	237.78	0	0
43821	43822	2014	12	31	21	10.0	-22	-3.0	1034.0	NW	242.70	0	0
43822	43823	2014	12	31	22	8.0	-22	-4.0	1034.0	NW	246.72	0	0
43823	43824	2014	12	31	23	12.0	-21	-3.0	1034.0	NW	249.85	0	0

43,824 rows x 13 columns

Problem Formulation

- Air pollution becoming severe problem in China
  - Reduce visibility
  - Cause air to appear hazy
- Predict and forecast PM2.5 values
  - Time series data and variables of Year, Month, Day, and Hour
  - Other target related variables: DEWP, TEMP,

    PRES, cbwd, lws, ls, lr
- Use RMSE to evaluate our model
  - Measures average magnitude of the error between prediction and true values

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

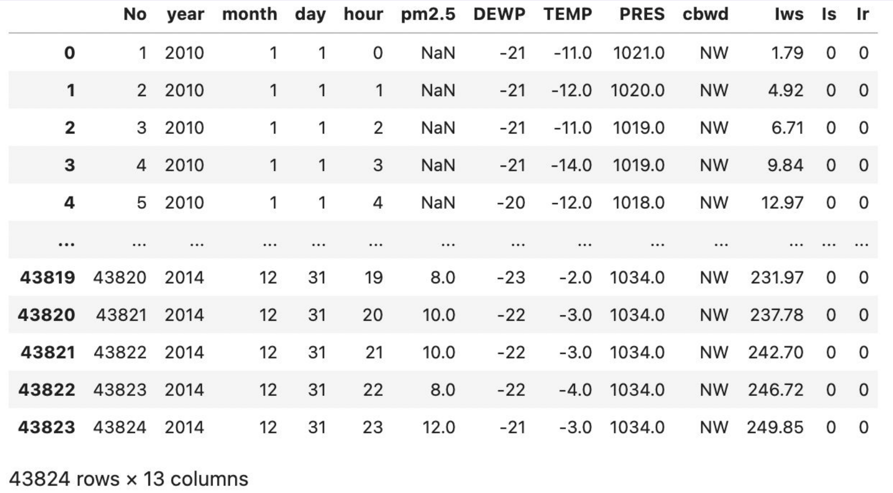
DATA ANALYSIS & PREPARATION

1) DROP NaN

2) Categorical → Numeric

3) to_datetime

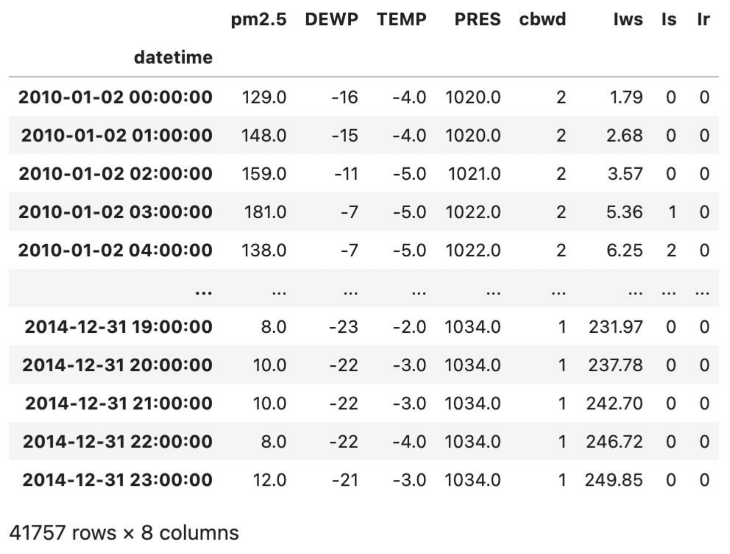
4) train test split



	No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
0	1	2010	1	1	0	NaN	-21	-11.0	1021.0	NW	1.79	0	0
1	2	2010	1	1	1	NaN	-21	-12.0	1020.0	NW	4.92	0	0
2	3	2010	1	1	2	NaN	-21	-11.0	1019.0	NW	6.71	0	0
3	4	2010	1	1	3	NaN	-21	-14.0	1019.0	NW	9.84	0	0
4	5	2010	1	1	4	NaN	-20	-12.0	1018.0	NW	12.97	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
43819	43820	2014	12	31	19	8.0	-23	-2.0	1034.0	NW	231.97	0	0
43820	43821	2014	12	31	20	10.0	-22	-3.0	1034.0	NW	237.78	0	0
43821	43822	2014	12	31	21	10.0	-22	-3.0	1034.0	NW	242.70	0	0
43822	43823	2014	12	31	22	8.0	-22	-4.0	1034.0	NW	246.72	0	0
43823	43824	2014	12	31	23	12.0	-21	-3.0	1034.0	NW	249.85	0	0

43824 rows × 13 columns

Before



	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
datetime								
2010-01-02 00:00:00	129.0	-16	-4.0	1020.0	2	1.79	0	0
2010-01-02 01:00:00	148.0	-15	-4.0	1020.0	2	2.68	0	0
2010-01-02 02:00:00	159.0	-11	-5.0	1021.0	2	3.57	0	0
2010-01-02 03:00:00	181.0	-7	-5.0	1022.0	2	5.36	1	0
2010-01-02 04:00:00	138.0	-7	-5.0	1022.0	2	6.25	2	0
...	...	...	...	...	...	...	...	...
2014-12-31 19:00:00	8.0	-23	-2.0	1034.0	1	231.97	0	0
2014-12-31 20:00:00	10.0	-22	-3.0	1034.0	1	237.78	0	0
2014-12-31 21:00:00	10.0	-22	-3.0	1034.0	1	242.70	0	0
2014-12-31 22:00:00	8.0	-22	-4.0	1034.0	1	246.72	0	0
2014-12-31 23:00:00	12.0	-21	-3.0	1034.0	1	249.85	0	0

41757 rows × 8 columns

After

METHODS & JUSTIFICATION

ARIMA

Time-series model

SARIMAX

Time-series model

DECISION TREE & RANDOM FOREST

Regression model

VAR & PROPHET

Time-series model

RESULTS - RMSE

1

ARIMA

RMSE value
102.28

2

PROPHET

RMSE value
68.81

3

DECISION TREE

RMSE value
102.64

4

SARIMAX

RMSE value
111.05

5

VAR

RMSE value
161.17

6

RANDOM FOREST

RMSE value
88.55

Normalized RMSE = (RMSE)/(Max - Min)

1
ARIMA

Normalized RMSE 1.107

2
PROPHET

Normalized RMSE 0.745

3
DECISION TREE

Normalized RMSE 1.111

4
SARIMAX

Normalized RMSE 1.202

5
VAR

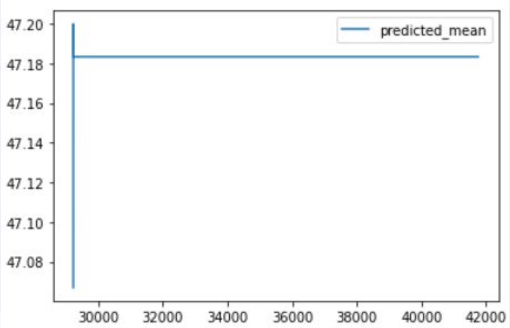
Normalized RMSE 1.747

6
RANDOM FOREST

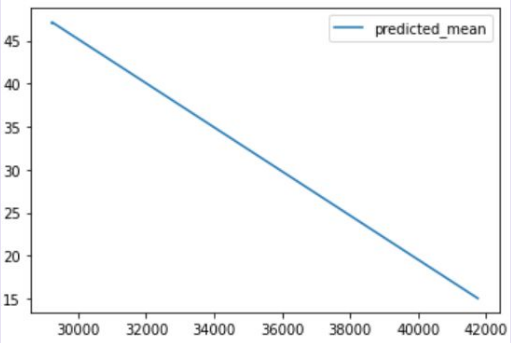
Normalized RMSE 0.960

Prediction Plot

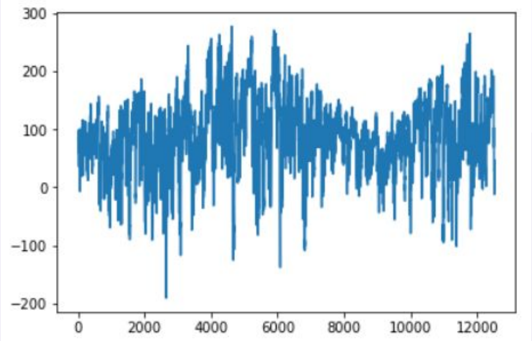
ARIMA



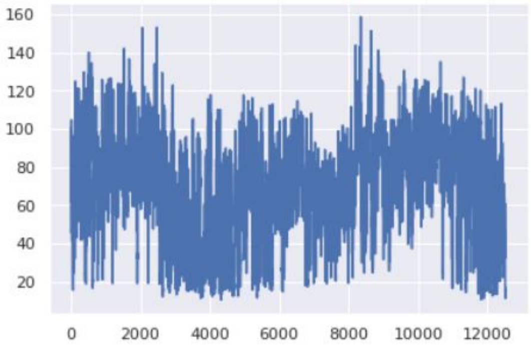
SARIMAX



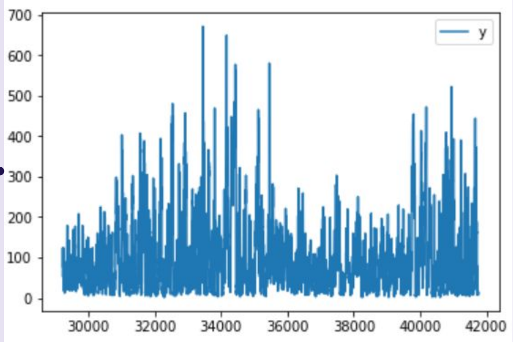
PROPHET



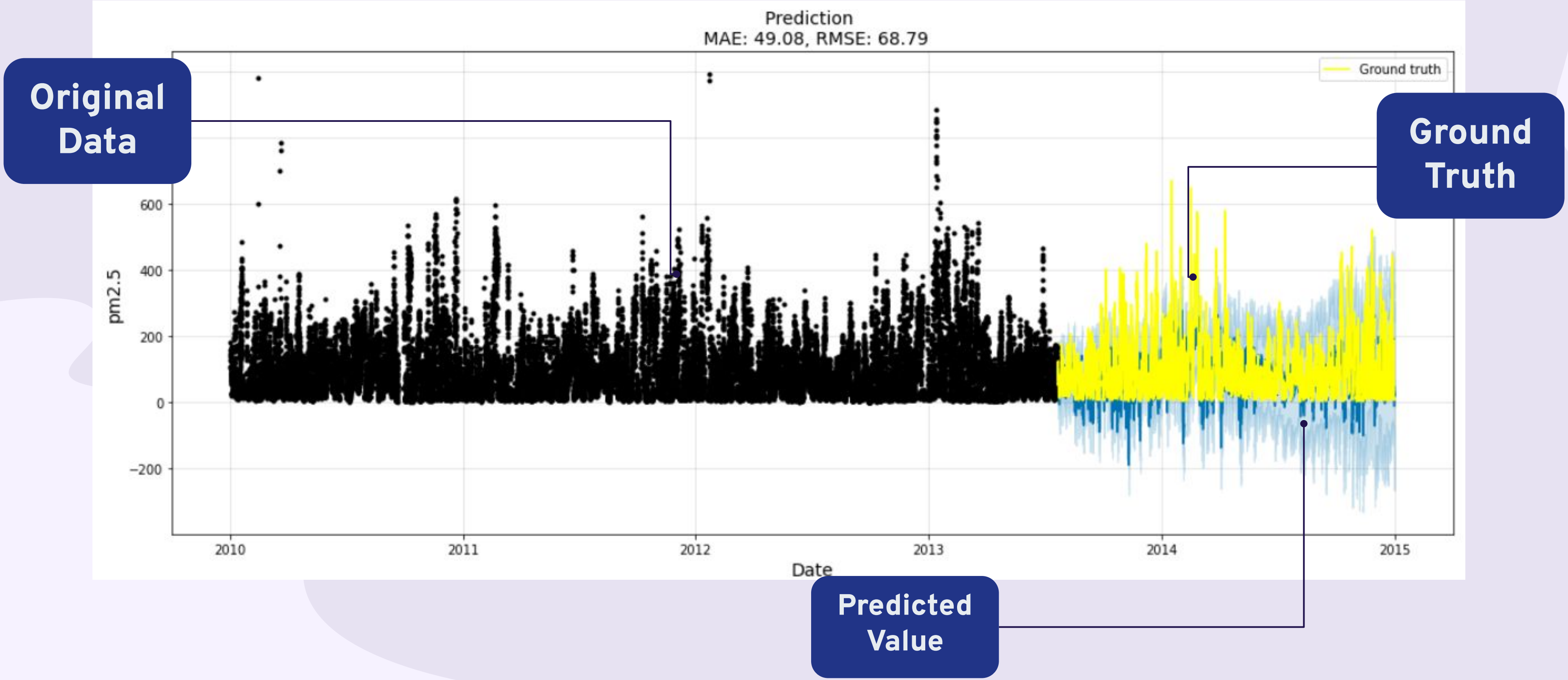
Random Forest



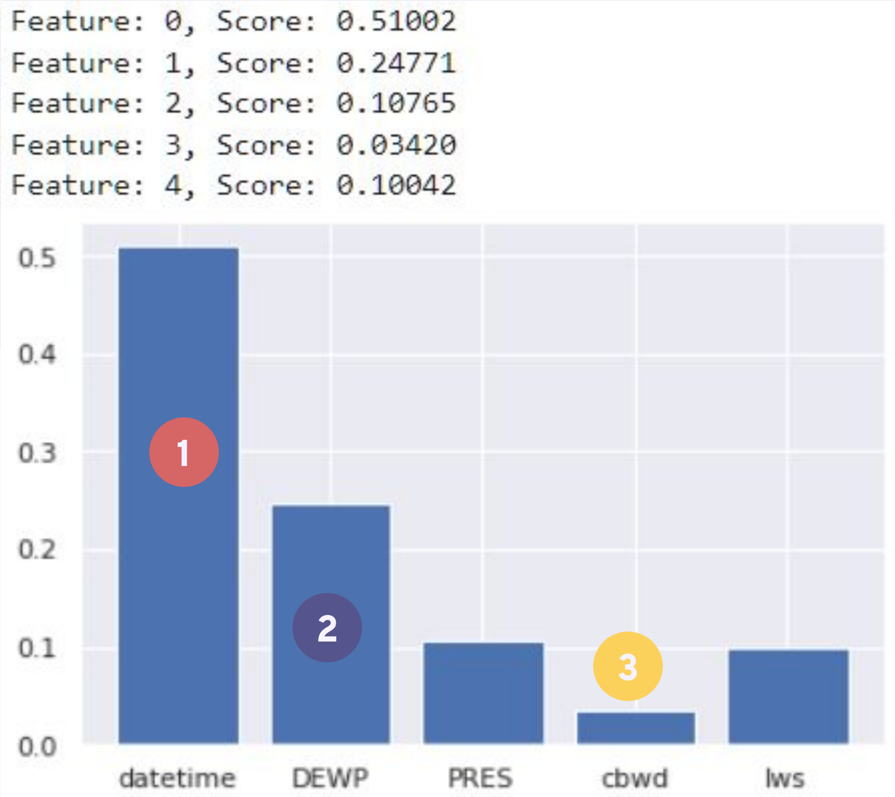
Actual Value



Prediction Plot - Prophet



Feature Importance



```
Feature: 0, Score: 0.51002
Feature: 1, Score: 0.24771
Feature: 2, Score: 0.10765
Feature: 3, Score: 0.03420
Feature: 4, Score: 0.10042
```

1 **Most Important Feature**
Datetime ⇒ 0.51002

2 **Second Most Important Feature**
DEWP ⇒ 0.24771

3 **Least Important Feature**
cbwd ⇒ 0.03420

Discussion & Evaluation

Limitation on different area

- Focused and created based on data from Beijing

Attribute values might appear different in other places

- Result in totally different PM 2.5 values

Make improvement on models

- Learn other data from numerous different areas with different pollution levels and climates