Air Quality Index Prediction with Spark ML and H2O

Peng Liu, Wenjie Duan, Xuxu Pan, Jingxian Li, Min Che

Contents

DataSet 1 : Air Quality Index (AQI)

Daily AQI Dataset in United States from 1980-2019
Size: 671.8 MB
Source: https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI

- AQI: how polluted the air currently is, or how polluted it is forecast to become
- **AQI is calculated** using the measurements of average times of the concentrations of O_3, SO_2, NO_2, CO, $PM_{2.5}$, PM_{10}.
- High AQI means high level of air pollution

AQI

State Name	county Name	State Code	County Code	Date	AQI	Category	Defining Parameter	Defining Site	Number of Sites Reporting
Alabama	Baldwin	1	3	2007-01-03	55	Moderate	PM2.5	01-003-0010	1
Alabama	Baldwin	1	3	2007-01-06	23	Good	PM2.5	01-003-0010	1
Alabama	Baldwin	1	3	2007-01-09	13	Good	PM2.5	01-003-0010	1

DataSet 2-5 : Weather

Data titles	Size	Features (for all the weather parameter)
Wind	2.3 GB	‘Date Local', 'State Code', 'County Code', 'Arithmetic Mean', 'Latitude', 'Longitude','State Name', 'County Name', 'Units of Measure', 'Parameter Name'
Temperature	462.9 MB	
Barometric Pressure	133.6 MB	
RH and Dewpoint	344.1 MB	

Source: https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI
Data Fusion: join the above 5 datasets

Analytic Goal: Predict AQI

- Investigate How Weather Influence AQI
  - Weather: wind, temperature, pressure, humidity
- Investigate How Historical AQI correlated with future AQI
  - Time series analysis
- Combine the above 2 factors (weather + time)
  - Expanded features

Related Works

Previous works	Our project
Predict air quality in one city	Predict air quality in multiple counties in the United States
Very few training data (e.g 10 days data)	Data from 1980-2019
Only neural network and linear model	More algorithms included
Features: ● Air pollutants concentrations, e.g. SO2, NO2 (data leakage) ● Time series AQI data (only one previous day) ● Meteorological (weather) features	**Features:** ● Avoid features with potential data leakage ● More time series AQI data ● Meteorological (weather) features

Preprocessing Algorithms

- Joined the 5 datasets by State, County, and Date
- Imputed the missing value with median value
- Feature engineering by extracting day of week, day, month, year
- Implemented window function to get the previous 30 days AQI value

Cluster Setting & Execution Time Comparison

Name	Cluster Setting: machine specs	Cluster Setting: number of nodes	Execution Time For pre-processing (in seconds)
Wenjie Duan	m5.xlarge	3	120.12
Min Che	m5.xlarge	4	119.22
Peng Liu	m5.2xlarge	3	72.47
Xuxu Pan	m5.2xlarge	4	74.46
Jingxian Li	m5.4xlarge	3	Can not start notebook

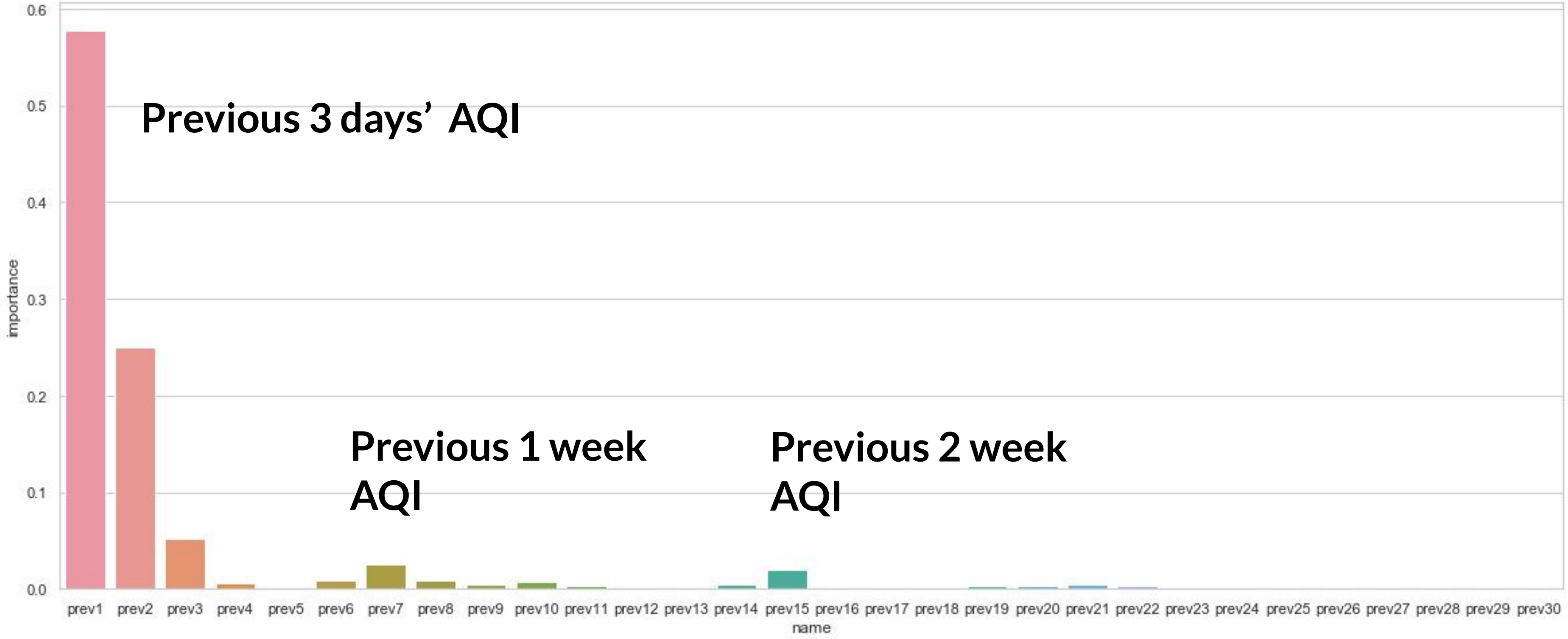
Execution time decreases, when change from xlarge to 2xlarge

Machine Learning Outcome Comparison

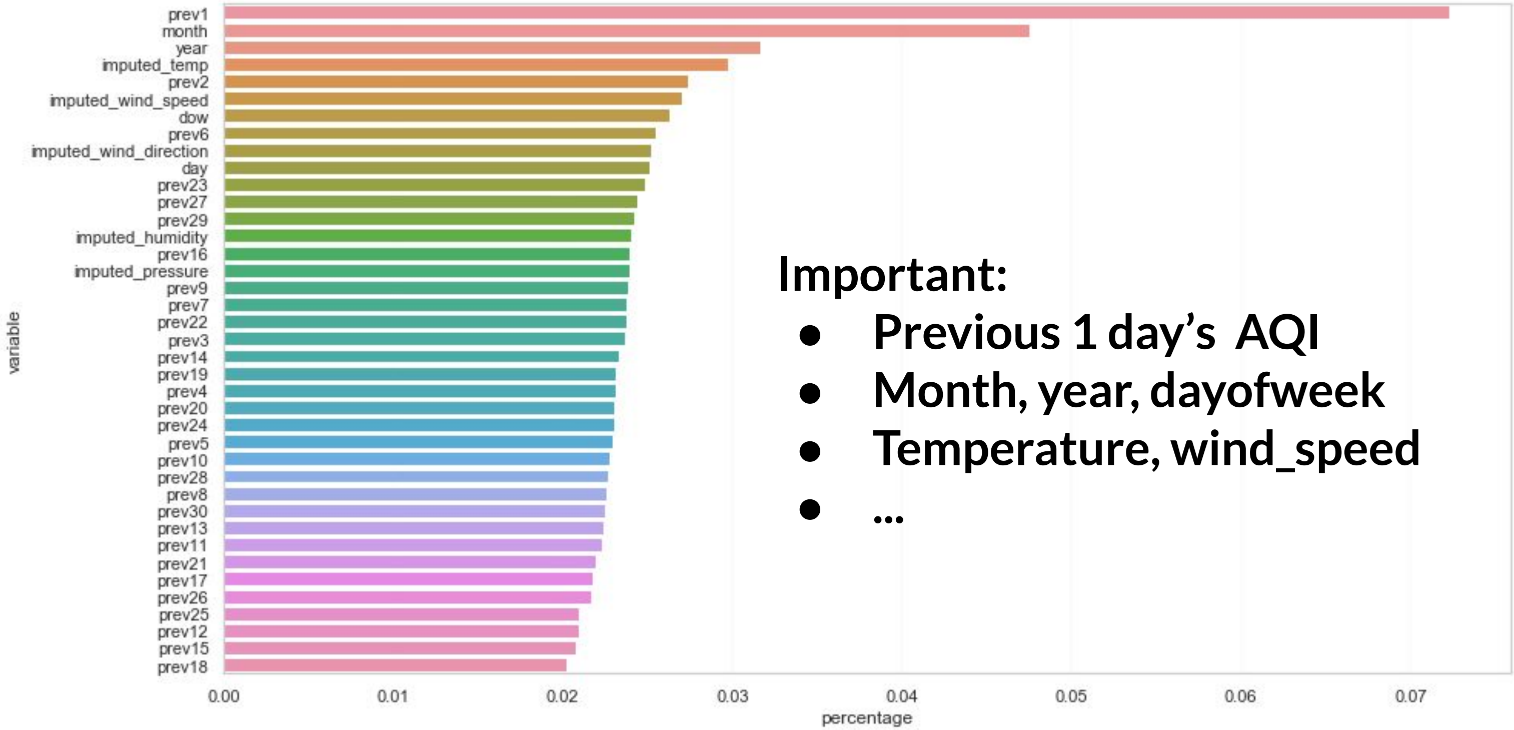
1. On Los Angeles county data

Name	Model	Features	MAE	RMSE
Wenjie Duan	SparkRandomForest	Previous 30 days AQI	24.67	32.93
Min Che	SparkRandomForest	Previous AQI + Date+Weather	25.30	33.76
Xuxu Pan	H2ODeepLearning Estimator	Previous AQI + Date+Weather	**22.05**	**29.56**

Feature Importance: previous AQIs on LA county data



Feature Importance: previous AQI + date+ weather, on LA county data



Important:

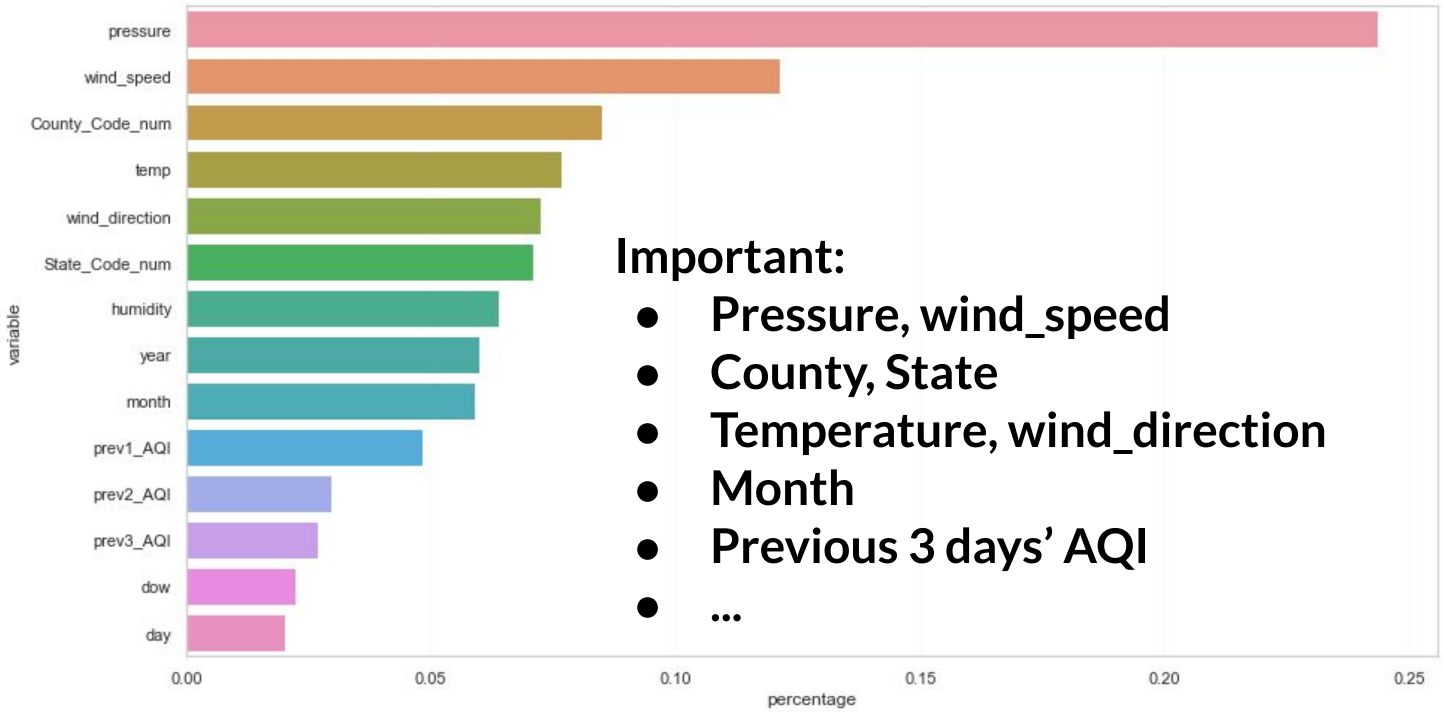
- **Previous 1 day's AQI**
- **Month, year, dayofweek**
- **Temperature, wind_speed**
- **...**

Machine Learning Outcome Comparison

2. On all counties data

Name	Model	Features	Data Range Predicted	MAE	RMSE
Peng Liu	H2ODeepLearning Estimator	Date+Weather+Previous 1 day AQI	All counties	12.3	20.7
Jingxian LI	H2ODeepLearning Estimator	Date+Weather+Previous 3 days AQI	All counties	**11.8**	**19.2**
			LA county	**12.5**	**19**

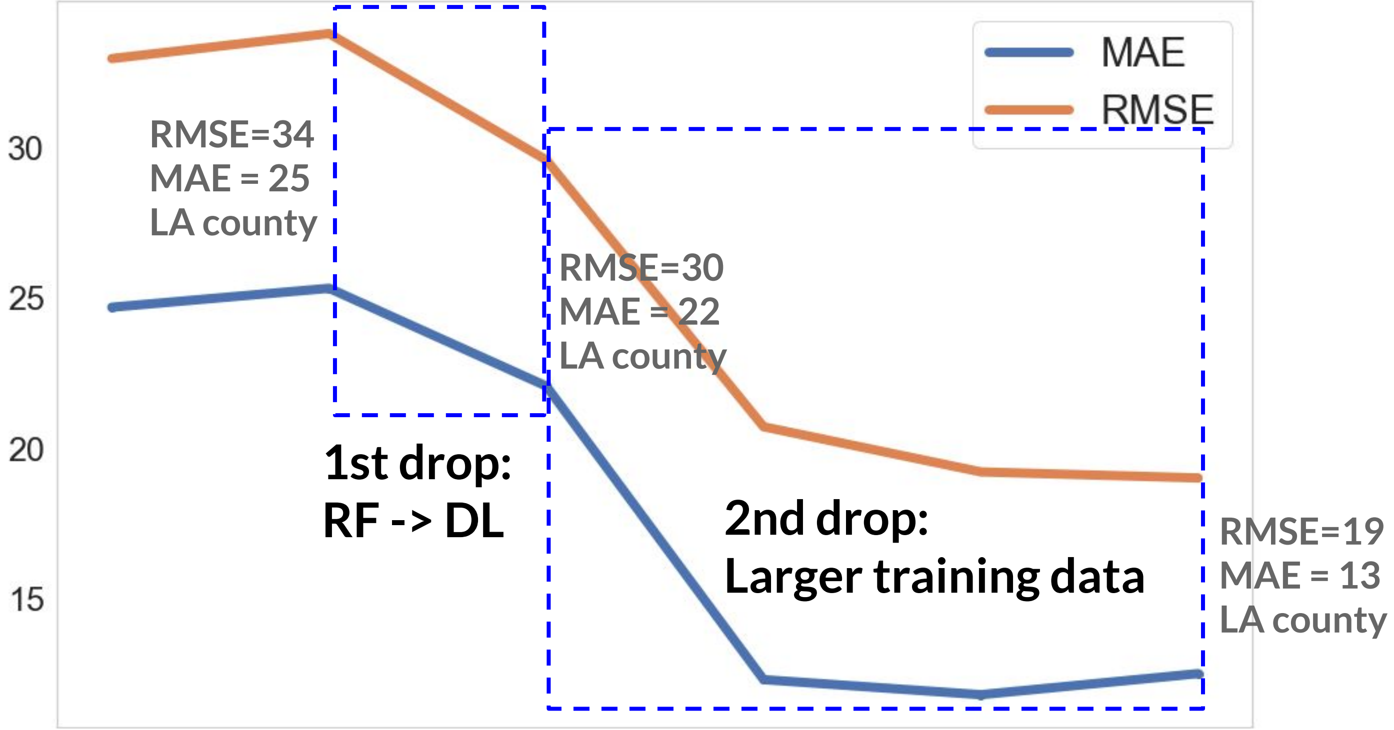
Feature Importance: previous AQI + date+ weather + County info on all counties data



Important:

- Pressure, wind_speed
- County, State
- Temperature, wind_direction
- Month
- Previous 3 days' AQI
- ...

Machine Learning Outcome Comparison



Best Model Execution Time Comparison

Name	Cluster Setting: machine specs	Cluster Setting: number of nodes	Execution Time For **best-model** (in seconds)
Wenjie Duan	m5.xlarge	3	90.21
Min Che	m5.xlarge	4	70.08
Peng Liu	m5.2xlarge	3	106.70
Xuxu Pan	m5.2xlarge	4	65.78
Jingxian Li	m5.4xlarge	3	Can not start notebook

Execution time decreases,
when change from xlarge to 2xlarge

Lesson Learned

- Deep learning models performs better than Random Forest™ models
- Models trained on larger dataset performs better than those trained on smaller dataset
- Time series model is important for air quality forecast
- Execution time decrease when cluster specs change from xlarge to 2xlarge
- However, more nodes do not guarantee higher time efficiency
- Can not open a notebook on cluster specs larger than 2xlarge ☹️

References

- Corani, G. (2005). Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecological Modelling, 185(2-4), 513-529.
- Wang, W., Xu, Z., & Weizhen Lu, J. (2003). Three improved neural network models for air quality forecasting. Engineering Computations, 20(2), 192-210.
- Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering, 2017.
- Kolehmainen, M., Martikainen, H., & Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. Atmospheric Environment, 35(5), 815-825.
- Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., ... & Osman, M. R. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. Water, Air, & Soil Pollution, 225(8), 2063.

Thank you!