Using Machine Learning Algorithms to Evaluate the Relationship Between Air Quality and Temperature Change

Yuxi Jin[1]

[1]Language lab

November 21, 2022

Abstract

Human activities constantly produce air pollutants, which may greatly impact climate change. Elucidating the relationship between air quality and temperature change is essential to gain a better understanding of climate change. Up until now, machine learning algorithms have been deployed to big data analysis in various fields. Here, we use the machine learning algorithms to analyze temperature and air quality data of different cities across China. Multiple linear regression and tree-based methods, including bagging, boosting and random forest, are used to train the model. With the tree-based methods, the factors highly associated with temperature change will be elucidated, which indicate their significant impact on temperature change. The results in this study demonstrate the possibility of using machine learning in atmospheric science field to predict air quality and temperature change, and how different algorithms perform regarding temperature and air quality predictions, which is informative for future air quality prediction research. The relationship between air quality and temperature change can also provide guidance to policymakers.

Using Machine Learning Algorithms to Evaluate the Relationship Between Air Quality and Temperature Change

Yuxi Jin
jinxx285@umn.edu

Introduction

Human activities constantly produce air pollutants, which may greatly impact climate change. Elucidating the relationship between air quality and temperature change is essential to gain a better understanding of climate change. Up until now, machine learning algorithms have been deployed to big data analysis in various fields. machine learning can be quite accurate when it comes to temperature predictions, both for monthly air temperature (Appelhans et al. 2015; Naing & Htike, 2015) and global temperature changes (Zheng, 2018).

Materials and methods

1. Data

1.1 Parameters

Parameters	Meaning	Unit
Temp	Atmosphere temperature at 2 meters above the ground	degree Celsius
Po	Atmosphere pressure at station level	mmHg
RH	Relative humidity at 2 meters above the ground	%
T_d	dew point at 2 meters above the ground	degree Celsius
P	Atmosphere pressure at sea level	mmHg
$PM_{2.5}$	atmospheric particulate matter with a diameter of 2.5 μm or less	$\mu g/m^3$
PM_{10}	atmospheric particulate matter with a diameter of 10 μm or less	$\mu g/m^3$
CO	carbon monoxide	ppm
NO_2	nitrogen dioxide	ppb
O_3	Ozone	ppb
SO_2	sulfur dioxide	ppb
AQI	air quality index	

1.2 Cities

Shanghai, Beijing, Guangzhou, Wuhan, Changsha, Nanning, Chengdu, Zhengzhou, Dalian, and Harbin.

1.3 Time frame

2016/1/1-2016/12/31

2. Algorithms and their features

2.1 **Multiple linear regression (MLR)** is a statistical method that uses several explanatory variables to predict the outcome of a response variable and to model the linear relationship between them.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_p x_{ip} + \varepsilon_i \ (i = 1,2, \dots n)$$

2.1.1 **Coefficient of determination (R^2)** provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

2.1.2 **Mean squared error (MSE)** measures the average squared difference between the estimated values and the actual value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

2.2 **Random Forest (RF)** operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the mean prediction (regression) of the individual trees. (See Figure 1)

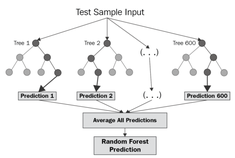
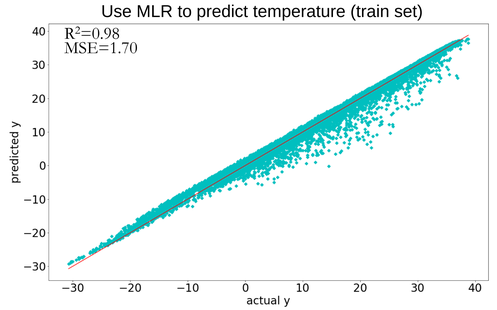
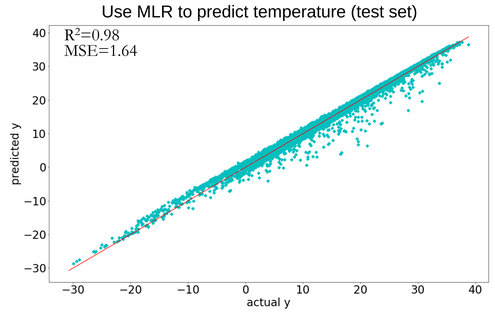


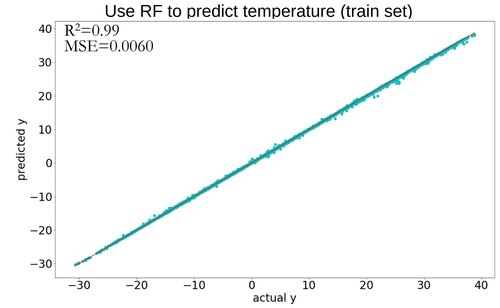
Figure 1. Random Forest Structure (Chakure 2019)

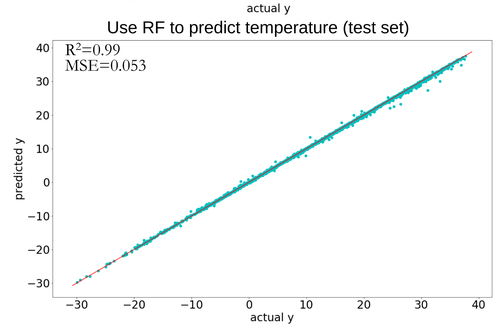
2.2.1 Variable importance represents the statistical significance of each variable in the data with respect to its affect on the generated model.

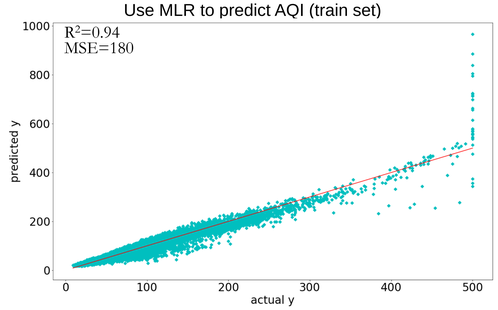
Results

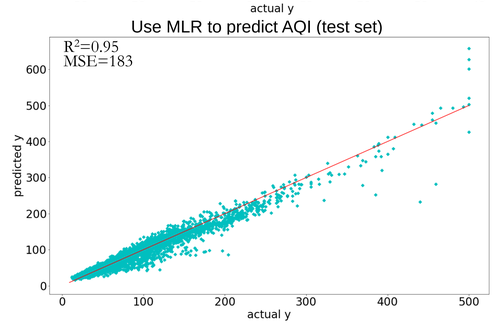


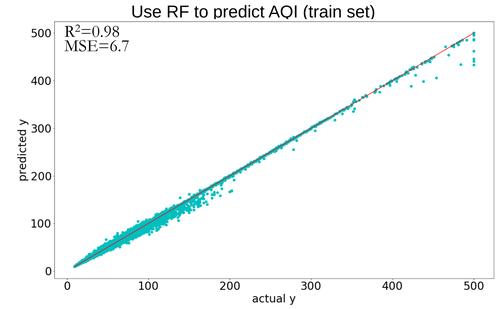


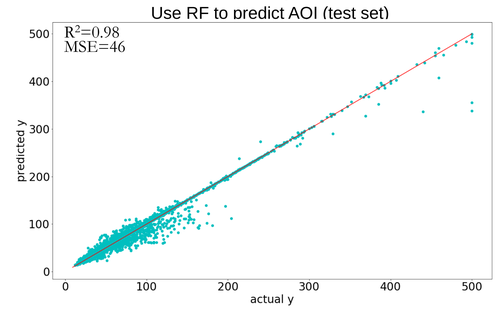




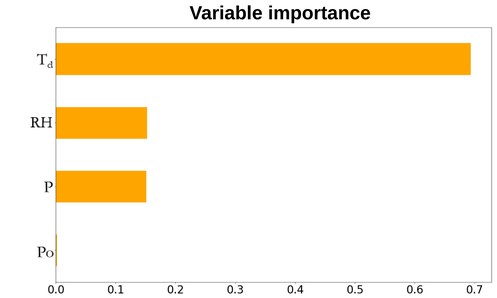




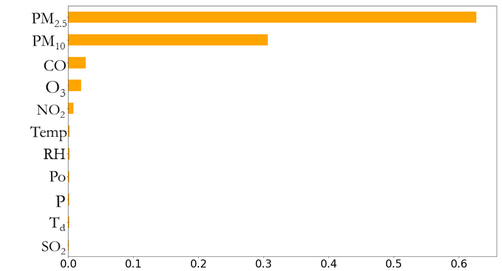




Results (continued)



When predicting temperature, **dew point (T_d)** has the most significant impact on temperature.



When predicting AQI, **$PM_{2.5}$** impacts the AQI the most.

Conclusions

- Both multiple linear regression and random forest perform well in all prediction scenarios.
- The results in this study demonstrate the possibility of using machine learning in atmospheric science field to predict air quality and temperature change.

References

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A. and Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spatial Statistics*, 14, pp.91-113.

Chakure, A 2019, *Random Forest Regression*, viewed 09 December 2019, < https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f>.

Naing, W.Y.N. and Htike, Z.Z., 2015. Forecasting of monthly temperature variations using random forests. *APRN J. Eng. Appl. Sci*, 10, pp.10109-10112.

Zheng, H., 2018. Analysis of Global Warming Using Machine Learning. Computational Water, Energy, and Environmental Engineering, 7(03), p.127.

Acknowledgments

The author gratefully acknowledge the helpful discussions with Yiming Qin from Harvard University, and Ying Liu from University of California, Los Angeles.

GET CONTACT