Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/20/20231089

Prediction of air pollution caused mortality exploiting machine learning models

Jianfei Li

College of Computer Science and Technology, Ocean University of China, Qingdao, Shandong, 266100, China

lijianfei@stu.ouc.edu.cn

Abstract. Air pollution is one of the most serious problems facing all mankind. It can not only cause various diseases and even lead to premature death of human beings. Especially, it could cause more harm to children, the elderly, and people who have been ill. Air pollution mainly occurs in developing countries with low income. Therefore, in order to better deal with the harm caused by air pollution, people need to predict the air pollution deaths. Based on the data set of air pollution deaths from 1990 to 2017, this paper investigated the performance of four machine learning models in predicting air pollution deaths. After the evaluation and comparison of four indicators, linear regression model and Ridge regression model are selected for further prediction considering their superior performance. This study finds that the overall trend of air pollution deaths worldwide is declining and will fall to zero by 2059 without other interference.

Keywords: air pollution, machine learning, ridge regression, linear regression.

1. Introduction

Air pollution is one of the most serious problems facing all mankind. According to the investigation and statistics of the World Health Organization (WHO), about 7 million people die from various diseases caused by air pollution every year. Air pollution is caused by the release of various harmful pollutants into the air, including ozone, sulfur dioxide, nitrogen oxides and particulate matter [1]. These pollutants can cause a range of cardiovascular, neurological and respiratory diseases, thus causing great harm to human health. In some cases, exposure to air pollution can even lead to premature death [2]. The highest levels of air pollution are generally found in middle-income and low-income developing countries, where air pollution deaths are concentrated. Vulnerable groups such as children, the elderly and people with pre-existing health conditions are the most affected because of the low immunity [3].

With the improvement of people's behavior and awareness of air environmental protection, there is more demand for quantitative prediction of air pollution deaths, and an effective prediction of air pollution deaths would quantify the death due to air pollution to provide help to inform policy decisions aimed at reducing air pollution levels and protecting public health [4]. In addition, comparing the prediction of indoor and outdoor air pollution deaths can also help to study the relationship between different air pollution and some diseases [5].

DOI: 10.54254/2755-2721/20/20231089

In recent years, with the continuous development and progress of machine learning, it can provide humans with a lot of highly accurate prediction models. Therefore, based on the data set of air pollution deaths worldwide from 1990 to 2017, this study will analyze and compare the data through data visualization to find the correlation. In this study, linear regression, Ridge regression, Lasso regression and polynomial regression are used to predict the number of future air pollution deaths, and MAE, RMSE, Explained variance, and R square are used to evaluate and compare the above four machine learning models. Finally, the appropriate model is selected for further prediction, and the most accurate algorithm is used to determine whether a country suffers a large number of deaths due to pollution, and to effectively predict the number of deaths likely to be caused by air pollution in the future. Accurate projections of air pollution deaths in each region will allow regions with different levels of pollution to adjust policies in a more timely manner and implement measures to protect the atmosphere. In addition, this study can also provide referable information, models and conclusions for the prediction of air pollution deaths and related research in the future.

2. Method

This section covers the details of the experiment including the dataset, the machine learning model, and the evaluation method. This experiment environment and tools are based on python3, Jupyter Notebook, matplotlib, pandas, sklearn, etc.

2.1. Dataset

The dataset in this study came from "death due to air pollution" in Kaggle [6]. Table 1 shows the meanings of the different keys in the dataset.

Table 1. Explanation of features in the dataset.

Year	It represents the years (from 1990 to 2017).
Entity	It represents the name of the country or region.
Code	It represents the abbreviation of the name of the country or region.
Total air pollution	It represents the total deaths of air pollution.
Outdoor particulate matter	It represents the outdoor deaths of air pollution.
Indoor air pollution	It represents the indoor deaths of air pollution.
Outdoor ozone pollution	It represents the deaths of ozone pollution.

The dataset collected indoor air pollution deaths, outdoor air pollution deaths, ozone pollution deaths and total air pollution deaths for 231 different countries from 1990 to 2017, and all four pollution deaths are measured in units per 100,000.

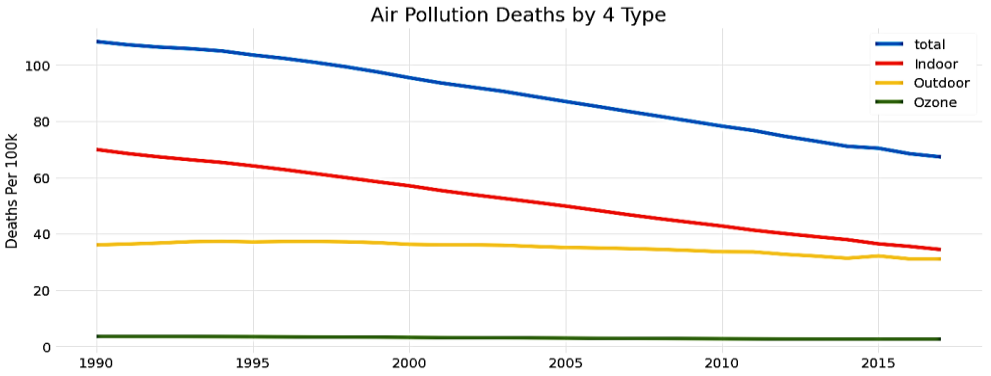

Figure 1. Average trend in air pollution deaths for 4 types.

DOI: 10.54254/2755-2721/20/20231089

As shown in Figure 1, outdoor air pollution deaths and ozone pollution deaths remained basically unchanged from 1990 to 2017, but indoor air pollution deaths declined significantly and accounted for about 60% of total air pollution deaths. Since the influence of outdoor air pollution and ozone air pollution is negligible, and the downward trend of total air pollution deaths paralleled the downward trend of indoor air pollution deaths, this article selects the total air pollution deaths as input when predicting the future number of air pollution deaths.

2.2. Machine learning models

2.2.1 Linear regression. The least square method is the core of linear regression. It compares and analyzes the performance of linear regression models by fitting the relationship between independent variables and dependent variables in the original data [7]. According to the number of independent variables, linear regression models can be divided into unary simple regression (including only one independent variable) and multivariate complex regression (including two or more independent variables). Since there is only one independent variable (year) in the experiment in this article, only simple regression will be introduced in detail.

In a problem with the dependent variable y and the independent variable $x_1, x_2, \dots, x_n$, the change in y is caused by two aspects. The first aspect is the function $f(x_1, x_2, \dots, x_n)$ made up of $x_1, x_2, \dots, x_n$; The other part is composed of many other factors (such as random factors) that are not taken into account in the independent variables. These factors can be collectively referred to as random error, which is represented by b. The calculation formula is as follows:

$$y = f(x_1, x_2, \dots, x_n) + b \tag{1}$$

In formula (1), the function $f(x_1, x_2, \dots, x_n)$ is called the regression function of y pairs $x_1, x_2, \dots, x_n$.

In the field of machine learning, linear regression model is denoted as:

$$y = \sum_{i=1}^{n} w_i x_i + b = w^T x + b \tag{2}$$

In formula (2), y is the prediction function, x is the feature input, b is the bias quantity and w is the model parameter, w can be calculated by the method of minimum loss function through iterative training.

2.2.2 Polynomial regression. When the original data is not linear, the performance of the linear regression model is often not as good as expected, so it is necessary to introduce the nonlinear regression model. There are many strategies for nonlinear regression, one of which is to transform nonlinear regression into polynomial regression. In order to extract nonlinear changes in data more accurately, polynomial regression model will be realized by increasing the degree of freedom of the model (introducing features of higher powers) [8].

The main idea of polynomial regression is to fit the polynomial regression equation through the historical data and use the polynomial regression equation to predict the new data. The equation of polynomial regression is as follows:

$$h_\theta(x) = \theta_0 x^0 + \theta_1 x^1 + \theta_2 x^2 + \theta_3 x^3 + \cdots + \theta_n x^n \tag{3}$$

In formula (3), x is the feature input and θ is the model parameter.

The essence of the polynomial regression model is the same as that of the linear regression model. By constructing the optimal function with the minimum deviation of regression model, and introducing the gradient descent method and the least square method to fit the relevant parameters

DOI: 10.54254/2755-2721/20/20231089

more accurately, the most suitable weight θ of the independent variable characteristics can be solved, so as to obtain a polynomial regression model with strong data fitting performance.

2.2.3 Ridge regression. Overfitting is one of the most important problems encountered in the process of model optimization. Ridge regression model can well solve the overfitting problem encountered in linear regression model. In the process of minimizing the loss function, users pay too much attention to the reduction of the loss value on the training set and ignore the generalization ability of the model. Adding regular term is a good way to solve the overfitting problem. In order to limit excessive w in linear regression, prevent model fitting, so the Ridge regression model is different from the linear regression model by adding a L2 regularization formula with a weighted penalty term [9]. The formula for implementing Ridge regression is:

$$l = \frac{1}{2m}\sum_{i=1}^{m}(y^{(i)} - \sum_j \theta_j x_j^{(i)})^2 + \frac{\lambda}{2}\sum_j \theta_j^2 \quad (4)$$

In formula (4), x is the feature input and w is the model parameter. Like linear regression, the least square method can be used to solve the parameters of the ridge regression model. The solution of the characteristic weight value θ is obtained by setting the reciprocal of the derivative of θ to be 0. λ is a parameter. When λ is greater than 0, if λ is increased, the deviation of Lasso regression models will be more and more, while the variance will be smaller and smaller. Therefore, it is necessary to determine a suitable λ to balance the deviation and variance of model species, so as to obtain the best model effect.

2.2.4 Lasso regression. In order to limit excessive w in linear regression, prevent model fitting, so the Lasso regression model is different from the linear regression model by adding a L1 regularization formula with feature selection. The purpose of using the L1 regularization formula is to find the features that have a significant impact on the results of the regression model, because the special selection inherent in the L1 regularization formula will exclude the invalid redundant features whose coefficients are compressed to 0 [10]. The formula for implementing Lasso regression is:

$$l = \frac{1}{2m}\sum_{i=1}^{m}(y^{(i)} - \sum_j \theta_j x_j^{(i)})^2 + \lambda\sum_j|\theta_j| \quad (5)$$

In formula (5), x is the feature input, θ is the model parameter and $\boldsymbol{\lambda}$ is a parameter.

2.3. Evaluation methods

2.3.1 Mean absolute error (MAE). MAE is used to evaluate the performance of regression model fitting data. When the performance of regression model is better, the error is smaller, and the MAE value will be smaller. MAE is the most easily understood measure of regression error, because there is no need to consider that positive and negative residuals cancel each other out. In the process of calculating the residuals of each feature point data, each residuals will take its absolute value, and the average value of these residuals is the evaluation value of MAE. Although MAE is easy to understand and implement, its results cannot be directly used to judge the model. Only through comparison, users can evaluate the performance of the regression model.

2.3.2 Root mean squared error (RMSE). The performance of the regression model is inversely proportional to the value of MSE (the worse the performance, the greater the value of MSE), because the result of MSE is the average of the sum of the squares of error of the predicted and true values of the model. RMSE, also known as standard error, is equivalent to the error between the actual and predicted values of the model further accurate on the basis of MSE, Because MSE becomes RMSE by taking the square root in order to maintain the same dimension as the target variable. RMSE

DOI: 10.54254/2755-2721/20/20231089

corresponds to the L2 norm and MAE to the L1 norm. The higher the number, the more the calculation is related to the larger values and ignores the smaller ones, so this is why RMSE are more sensitive to outliers. RMSE can be used as a standard to evaluate the accuracy of this measurement process.

2.3.3 Explained variance. The upper limit of Explained variance is 1 and the lower limit is 0. When the result is close to 0, it indicates that the performance of the regression model is poor (the degree of explanation is weak). On the contrary, when the result is close to 1, it indicates that the performance of the regression model is very good (the change of the dependent variable can be well explained by the characteristics of the dependent variable). Explained variance does not mean that variance is explained, it simply means that one or more variables can be used to predict things more accurately than before. The formula for implementing Explained variance is:

$$Explainde\ variance = 1 - \frac{var(y-\hat{y})}{var(y)} \tag{6}$$

In Formula (6), y represents the true value, $\hat{y}$ represents the predicted value and var represents variance.

2.3.4 R square (R^2). The upper and lower limits of R square are 1 and 0 respectively. The model performance and the explanatory degree of variables are proportional to the value of R squared (the model performance is the best when the value of R squared is 1; When R squared is 0, the performance of the model reaches the worst. Because R squared can measure the extent to which the dependent variable is changed by its own variable, the proportion of multiple regression equation explained can be estimated by the variation of the dependent variable y. The equation of R square is as follows:

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i-\frac{1}{n}\sum_{i=1}^{n}y_i\right)^2} \tag{7}$$

In Formula (7), $\hat{y}_i$ represents the observed value and y_i represents the real value in the model.

3. Result

As shown in Table 2, this article adopts four different machine learning models (respectively linear regression, Ridge regression, Lasso regression and polynomial regression), and uses four evaluation indicators, namely R square, Explained variance, MAE and RMSE, to evaluate and compare the ability of the models to predict the number of air pollution deaths. Lasso regression performed the worst, and although it performed the best in explained variance, it performed far worse than the other three machine learning models in the other three metrics. Polynomial regression performed so-so, doing well in explained variance, but so-so in the other three measures. Two machine learning models, linear regression and Ridge regression, performed best in the four evaluation indicators.

Table 2. Evaluation results of four machine learning models.

	R square	Explained variance	RMSE	MAE
Ridge regression	0.7158	0.9634	1.0490	0.9792
Linear regression	0.7264	0.9631	1.0292	0.9573
Polynomial Regression	0.6874	0.9636	1.1002	1.0341
Lasso regression	0.5475	0.9677	1.3237	1.2755

Figure 2 is a visualization of air pollution deaths over the next 50 years using the best-performing linear regression and Ridge regression. The predictions of the two machine learning models are

DOI: 10.54254/2755-2721/20/20231089

essentially the same, so much so that the prediction curves basically overlap each other. Air pollution deaths will continue to decline over the next 50 years and reach zero by 2059.

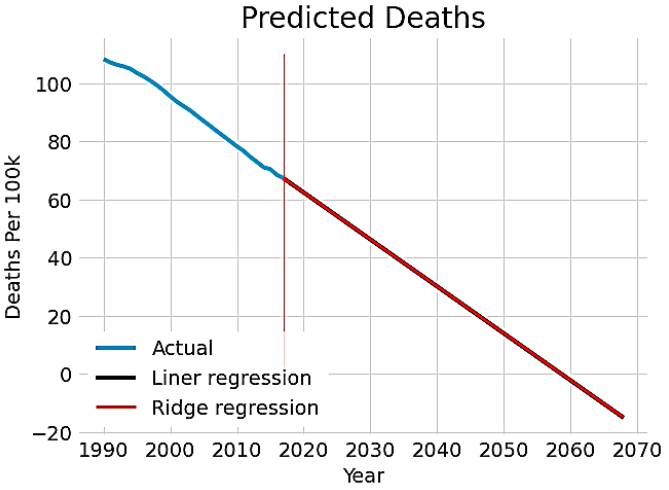

Figure 2. Prediction of air pollution deaths.

4. Discussion

The reason why the performance of polynomial regression is not as good as that of linear regression is that the original data has been close to the linear relationship, and it is not effective to fit the data with curves of higher powers. The reason why Lasso regression is not as good as Ridge regression is that the properties of L2 regularization and L1 regularization are different [11].

The number of air pollution deaths around the world is decreasing, partly because economic development has provided better medical care and partly because people are becoming more environmentally conscious. However, air pollution deaths are increasing in Libya. In Figure 3, the number of air pollution deaths in Libya has been increasing since around 2007, most likely because the economy is dominated by the oil industry.

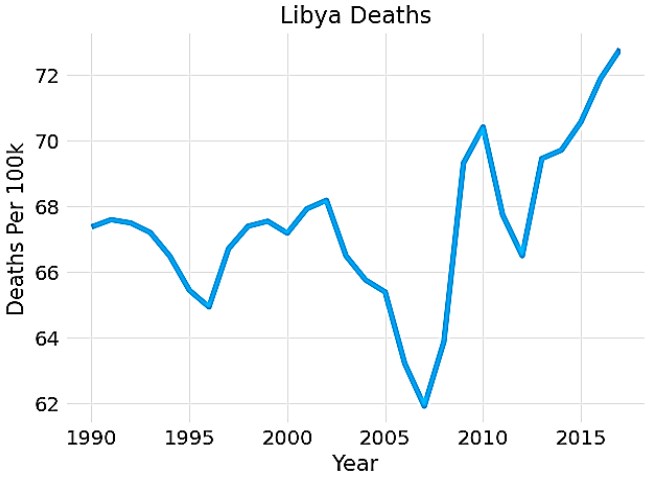

Figure 3. Air pollution deaths in Libya.

Although linear regression and Ridge regression do well in predicting total air pollution deaths, they do not do well in predicting national air pollution deaths for some regions. In Figure 4, the linear regression model is poor at fitting data for air pollution deaths in Southern Sub-Saharan Africa

DOI: 10.54254/2755-2721/20/20231089

because the original data are not linearly correlated. So the analysis can be improved by training various machine learning models for each country and region, and adopting the best model for each situation, rather than just relying on linear regression and Ridge regression. The advantage of this is that each country can be provided with more specific and accurate information, so that it can formulate policies and medical programs tailored to its own national situation.

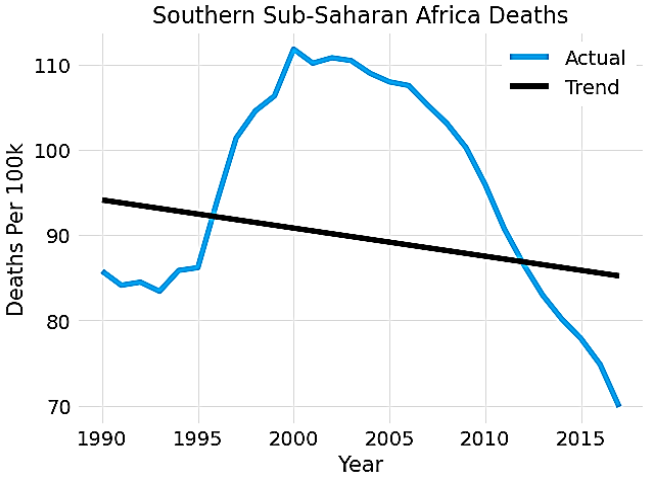

Figure 4. Real values and model fitting of deaths in Southern Sub-Saharan Africa.

In Table 3, both linear and Ridge regressions predict that air pollution deaths will decline to zero by 2059. This is a questionable prediction because it does not take into account more external factors (such as economic and scientific development, certain natural disasters, etc.). In the future, more factors can be considered on the basis of this study, so as to provide more specific and rich information for the protection of air environment

Table 3. Comparison of air pollution deaths predicted by two models.

Year	Linear regression (Per 100,000)	Ridge regression (Per 100,000)
2017	67.2522	67.2641
2018	65.6371	65.6499
2019	64.0220	64.0356
2020	62.4068	62.4214
2021	60.7917	60.8072
2022	59.1766	59.1930
2023	57.5615	57.5788
	……	
2058	1.0328	1.0809
2059	-0.5823	-0.5333

5. Conclusion

This study finds that air pollution deaths have been declining in all countries except Libya, and will drop to zero by 2059, which can be explained by the fact that First of all, deaths caused by air pollution mainly occur in middle-income and low-income developing countries. Rapid economic development and advances in science and technology have made it possible for more and more low-income and middle-income countries to use greener and healthier energy and materials, and for people to enjoy more reliable medical assistance. Secondly, Green and sustainable development has been an important concept to maintain the healthy development of the earth, coupled with people's

DOI: 10.54254/2755-2721/20/20231089

increasing behavior and awareness of air quality protection, have effectively limited the generation of air pollution. Finally, the increase in air pollution deaths in Libya could be attributed to the country's oil-led economy and its long history of war.

Finally, the linear regression model and Ridge regression model used in this study perform well in predicting the pollution deaths of the Space Agency in most countries, but not so well in some countries. This is because the raw data for a small number of countries are not linear, which results in the poor fitting performance of the models used in this study. Therefore, in the future, the most appropriate machine learning model can be designed for each country on the basis of the above, so that more accurate prediction can be achieved. At the same time, the decline in air pollution deaths is certain, but the study's finding that the total deaths of air pollution will drop to zero by 2059 should be taken with a grain of salt because it does not take into account many external factors. Therefore, in the future, a machine learning model including more factors (economy, technology and population, etc.) can be built on the basis of the above, so as to obtain a richer and more specific prediction.

References

[1] Glencross, D. A., Ho, T. R., Camina, N., Hawrylowicz, C. M., & Pfeffer, P. E. (2020). Air pollution and its effects on the immune system. Free Radical Biology and Medicine, 151, 56-68.

[2] Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. The lancet, 360(9341), 1233-1242.

[3] Air pollution, URL: https://www.who.int/health-topics/air-pollution#tab=tab_1, 2023/05/26

[4] Hales, S., & Howden-Chapman, P. (2007). Effects of air pollution on health. BMJ: British Medical Journal, 335(7615), 314–315.

[5] Christensen, G. M., Marcus, M., Vanker, A., Eick, S. M., Malcolm-Smith, S., et, al. (2023). Joint Effects of Indoor Air Pollution and Maternal Psychosocial Factors During Pregnancy on Trajectories of Early Childhood Psychopathology. medRxiv, 2023.04.07.23288289v1, 1-19.

[6] Akshat, G. (2020), Death due to air pollution, URL: https://www.kaggle.com/datasets/akshat0giri/death-due-to-air-pollution-19902017

[7] Zhaokui, H. (2022), Solving Linear Regression Equation by Moment Estimation, Studies in College Mathematics, 25(4), 41-43.

[8] Shi, X. (2023). Analyzing hospitality leader–follower dyads with polynomial regression: a critical reflection. International Journal of Contemporary Hospitality Management. 1-10.

[9] Schreiber-Gregory, D. N. (2018). Ridge Regression and multicollinearity: An in-depth review. Model Assisted Statistics and Applications, 13(4), 359-365.

[10] Ranstam, J., & Cook, J. A. (2018). LASSO regression. Journal of British Surgery, 105(10), 1348-1348.

[11] Zhou, Z., Yang, X., Ji, J., Wang, Y., & Zhu, Z. (2023). Classifying fabric defects with evolving Inception v3 by improved L2, 1-norm regularized extreme learning machine. Textile Research Journal, 93(3-4), 936-956.