Air pollution prediction based on multicollinearity

Shanika Iroshi Nanayakkara
School of Information Technology
Deakin University, Burwood, Australia
shanikan@deakin.edu.au

***Abstract*—In this study we propose prediction model to predict air pollution components which are specified as carbon monoxide, benzene and nitrous oxide. We uses multiple predictors such as temperature, absolute and relative humidity and five sensor data. We identify that these predictors are correlated among each other and we use rigid linear regression instead general linear regression method to hinder the impact of multicollinearity in this study. We clearly show the performance efficiency of proposed models using determination of coefficient, bias and variance scores for each models. We derive 95%, 82% and 68% determination of coefficient for the model of benzene, carbon monoxide and nitroos oxide respectively.**

Machine Learning, Data Mining, Regression, Prediction

I. Introduction

Air pollution is significant problem which arises due to many reasons. Among them technological evolution takes high priority. However, people need to move on with the technical world while keeping sustainable environment for the future generation. For that, almost all industrialization production processes need to concern on their emissions into the environment. For example, many industries tend to left different types of air particles into the atmosphere as a result of their production process. Though it does not seems to have quick threat on environment and the human life, it has long life impact. Consequently, people tend to find the solutions to mitigate these pollution in order to enhance the best balance between the industrialization and the environment protection. Initially, data is the major asset for everything in the world. Because, every application had data driven structure to analyse the current trends in order to predict the future impacts or enhancements. Machine learning and Artificial intelligent play vital role in the data driven applications. Therefore, in such studies present prediction and detection models based on the existing data.

In this study we reveals the high performing prediction model, using supervised learning model called regression, to predict the air pollution. For that, we use temperature values, relative and absolute humidity, and 5 sensor values as predictor variable for several months. The given target variables are compositions of carbon monoxide, benzene, and nitros oxide.

In this paper we show that the model prediction behavior on multicollinearity based predictor variables. Since this is a labeled dataset, we decide to perform supervised learning on training data to propose a model. Moreover, prepossessing helps to determine the correlations between the different features. This predictor variables indicate multicollinearity. Due to that, It is impossible to propose a linear correlation based model for this scenario. Therefore, we decide to continue our empirical results using the multi correlation based regression method called rigid regression. Our main contribution is as follows,

- First we analyse, how much of multicollinarity consists among the given predictor variable using heatmap visualization.
- Second we show how target air pollution components behave over certain period.
- Third we clearly show predicted results by using the proposed regression model.

The rest of this paper is structured as follows: Section (2) describes the preliminary studies that we conducted before proposing this model. The method of this proposed model is in Section (3). Section (4) discusses the empirical findings of the study presenting performance comparison and Section (5) involves discussion and (6) elaborate on future directions conclusions.

II. Preliminaries

Initially, regression problems anayse the correlation between the predictor variable and the response variables [2]. Most probably, regression problems define correlation between two quantitative variables. This is indicate as linear regression if it represents relationship by straight line. However, logistic regression and non-linear regression show curve shape relationship. Moreover, the challenge is selecting appropriate predictor variable for the given response variable. In this scenario we have three response variables such as carbon monoxide, benzene and notrous oxide compositions.

However, when there is more predictors available in training set it is obvious to have correlation among predictors. In such cases, we cannot use linear regression methods to derive prediction models only using two predictors. Because, Once we select one predictor variable to derive correlations between that selected predictor and the given response variable, we need to assume other predictors have no or less impact on that correlation. Depending on different factors such as experience, historical data one response may be affected by more predictors [3]. In this study we are working on same kind of scenario.

Machine learning allows machines to analyse existing data and predict on them. These predictions comes as the artificial intelligent models which can outperform on future data predictions. However, making prediction cannot be always trustworthy to have 100% accuracy. Therefore, it allows to have

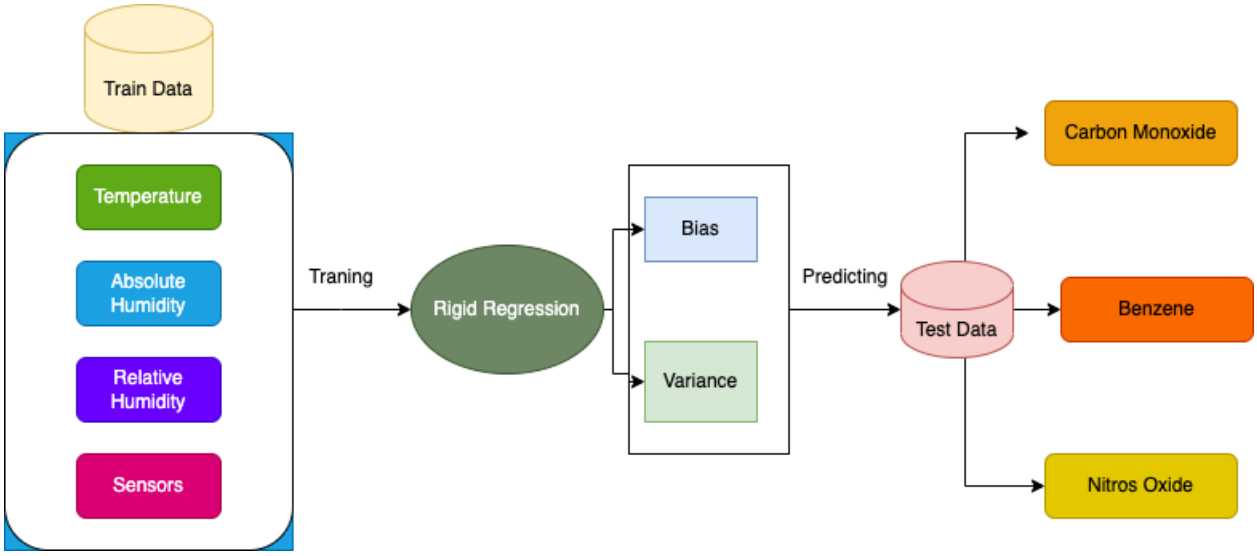

Figure 1: Framework

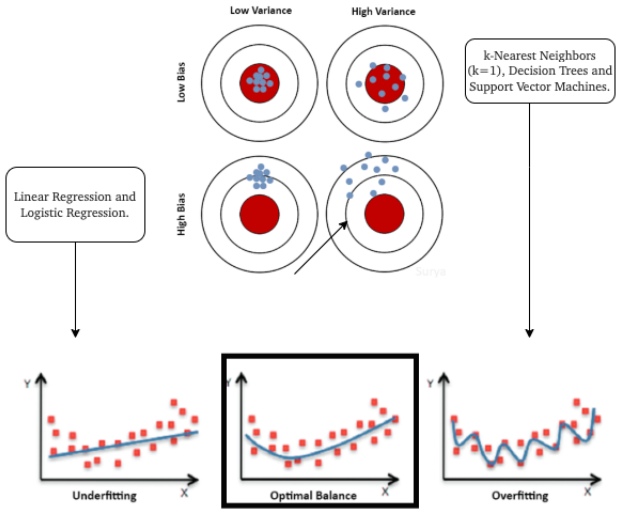

Figure 2: Bias and Variance for Prediction Accuracy [1]

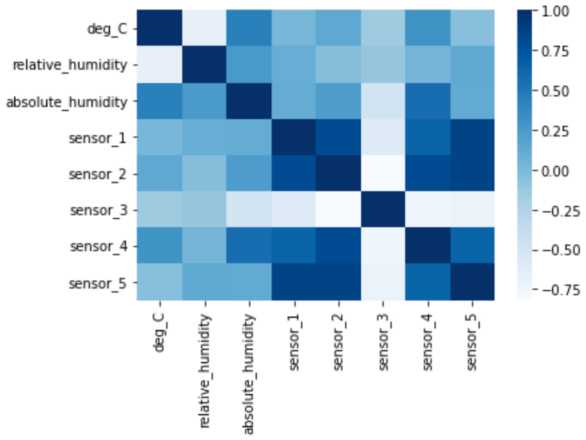

Figure 3: Fig˙heatmap

predictions errors. These errors are called Bias and Variance. We use this bias and variance values in the models to state accuracy levels of proposed models in our evaluation process. Bias is a training data error while variance is the testing data error. As in the Figure 2, we can have four types of error levels. However, low bias and low variance models are the bet fitted models. Because the both training and testing data fitted exactly on the model as in optimal balance graph. However, since it is impossible to derive non-error research work target to derive low bias, high variance models such as decision tree, support vector machines and k-nearest neighbour or high bias, low variance methods such as regression models. In this study we uses regression models and we prove that we are achieving underfittiing scenarios as illustrated in Figure 2 In linear regression, we normally derive a best fitted line to represent all training data set. However, due to the multiple core relations among predictors, we uses rigid regression in which uses penalty term to augment the loss function of the linear regression for better performance.

III. Method

A. *Prepossessing and Multicollinearity*

Prepossessing is the initial step in machine learning approaches. Because, all the models are driven on the existing training data. Therefore, training data must be clear enough having less noise. There are so many methods to clean the data such as min-max scaling, standard scaling and remove duplication. Apart from that, we need to maintain data consistency among the learning process. In this study we use, standard scaler which is available in sklearn package in python. Then convert all the data into one consistent data format. Further, we remove the duplicates and make the data more

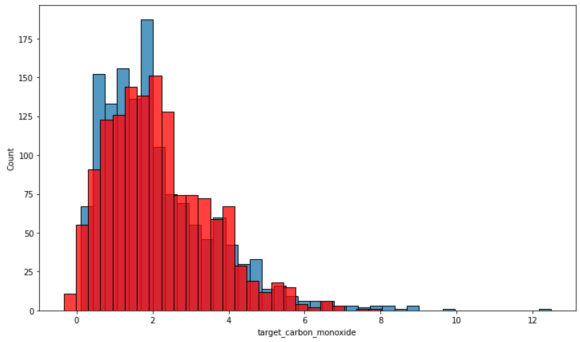

Figure 4: Prediction accuracy for carbon monoxide results [Training histogram indicates in blue color whereas testing histogram show in red color]

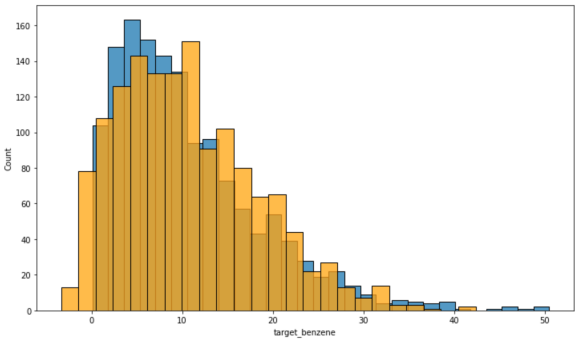

Figure 5: Prediction accuracy for benzene results [Training histogram indicates in blue color whereas testing histogram show in red color]

narrow by which allow to maintain clarity of data. We maintain only relevant data items throughout the training process by eliminating other irrelevant data such as target variables and date time data. Further, we analyse the data distribution by generating boxplot visualization as in Figure 9. This uses to identify the symmetric and skewness of target data in the training set before we develop the model.

Multicollinearity is a aspect which occurs when two or more predictor variables are correlated [2]. As a consequent, coefficients error may increase [2]. Then the each variables behave insignificantly, thought they should be significant. So that, the models are badly impacted if such model does not address this multicollinearity in an proper way. The reasons for multicollinearity are using different types of variables inaccurately, poor or null hypothesis, bad selection of a dependent variable, repetitions in variable, and a high correlation between variables and choice of dummy variables [4]. There are certain set of solutions to hinder the impact of multicollinearity in linear regression. They are obtaining more set of predictors, removing unwanted variables, deciding accurate independent variable and use rigid regression method or partial squares regression [4]. Sometimes, if all these solutions may not applicable. And then researchers decide do nothing. For example, in this study, we recognize the given training data has multicollinearity due to multi correlations among predictors [Temperature, absolute humidity, relative humidity, sensor data] as depicted in Figure 3 and Figure 7. As a solution, we decide to implement rigid regression method to derive our proposed method.

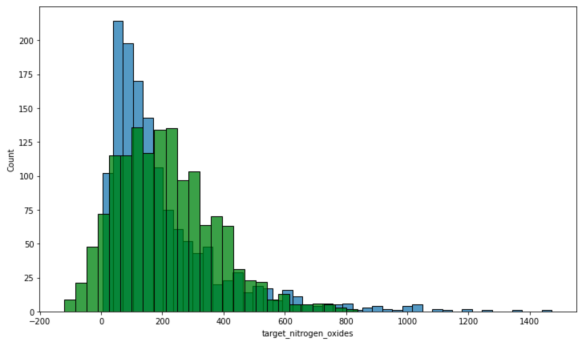

Figure 6: Prediction accuracy for nitrous oxide results. [Training histogram indicates in blue color whereas testing histogram show in red color]

Table I: Performance measurements of proposed models

	CM	Benzene	NO
Bias	0.38	3.379	12916.6
Variance	0.001	0.009	30.591
CD Training	0.82	0.95	0.68
CD Testing	0.83	0.94	0.66

CD : Coefficient Determination
CM : Carbon Monoxide
NO : Nitrous Oxide

B. *Usage of Rigid Regression*

Linear regression is most widely use learning predictive model which presents relationship as a straight line. In this approach show correlation between two variables. That means there should be one predictor for response variable/variables at a time. Moreover, the dependent variable, which is called response should be continuous and independent variable(s) (predictor variables) can be continuous or discrete. In contrast, rigid regression is a technique used to implement multicollinearity of predictor variables which are highly correlated each other [5]. Generally, rigid regression is same like linear regression unless it add penalty values to reduce the loss or error of linear regression. This error can be happen due to bias and/or variance of the variables.

C. *Efficiency Analysis*

We uses coefficient of determination to measure the performance of the proposed models. The coefficient of determination interprets how well the regression model fits testing data observations. Further, it describes how much variation can be expected in the dependent/response variable.

IV. Experiment and Analysis

In this study we propose a model to predict air pollution component behavior by using rigid regression. We use rigid

	deg_C	relative_humidity	absolute_humidity	sensor_1	sensor_2	sensor_3	sensor_4	sensor_5
deg_C	1.000000	-0.668002	0.445162	0.017513	0.133167	-0.145437	0.308202	-0.050567
relative_humidity	-0.668002	1.000000	0.249013	0.093130	-0.035152	-0.102146	0.027002	0.126466
absolute_humidity	0.445162	0.249013	1.000000	0.105977	0.236894	-0.485445	0.567376	0.124945
sensor_1	0.017513	0.093130	0.105977	1.000000	0.811898	-0.592233	0.643191	0.860849
sensor_2	0.133167	-0.035152	0.236894	0.811898	1.000000	-0.819334	0.812454	0.863464
sensor_3	-0.145437	-0.102146	-0.485445	-0.592233	-0.819334	1.000000	-0.741439	-0.706006
sensor_4	0.308202	0.027002	0.567376	0.643191	0.812454	-0.741439	1.000000	0.641120
sensor_5	-0.050567	0.126466	0.124945	0.860849	0.863464	-0.706006	0.641120	1.000000

Figure 7: Predictor Correlation Chart [Numerical representation for heatmap diagram] [1]

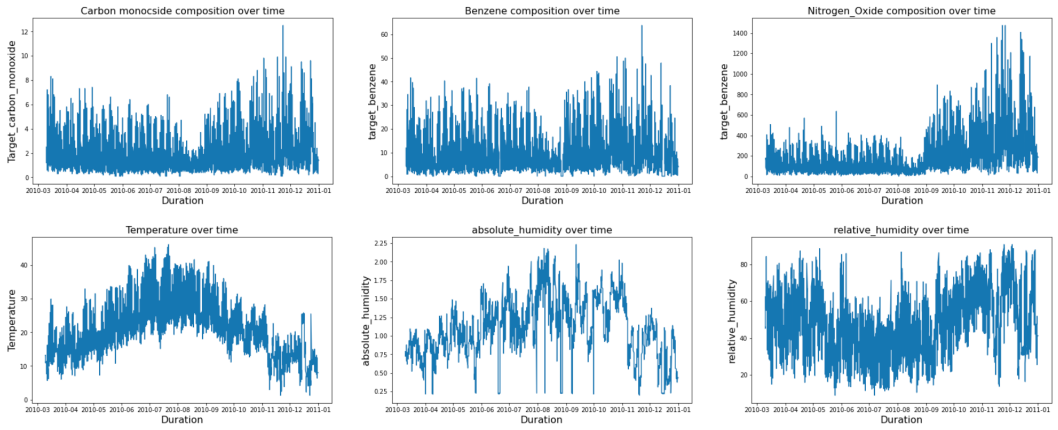

Figure 8: Impact of Air pollution and behavior of each pollution components over time. From left top corner to right top corner shows behavior of carbon monoxide, benzene and nitrous oxide respectively. From left bottom corner to right bottom corner shows how temperature, absolute humidity and relative humidity change over time.

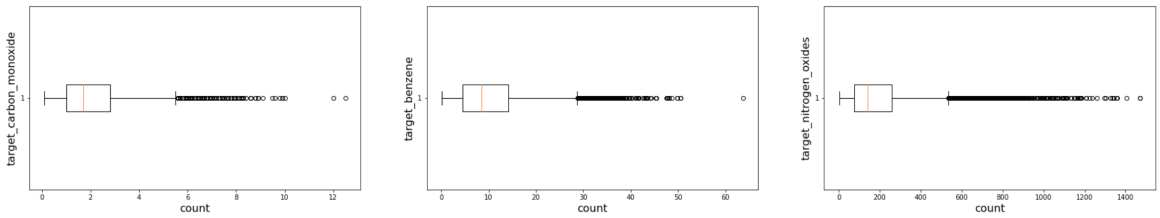

Figure 9: Boxplot illustration to summarize target data distribution

regression instead of linear regression, to hinder the multi-collinearity issue. As illustrated in Figure 8 first we demonstrate given air pollution components behavior over time. the given duration starts from march 2010 to January 2011. As in Figure 8 first row shows how each air particles behave over this time series. Based on that, we can see all three types of air shows high amount of concentration within September 2010 to January 2011. And further, if we consider temperature, it seems high variance within mid of the year. Moreover, humidity perform very fluctuation behavior over this period.

We test our proposed model using test dataset which includes same set of predictor variables. The results are given in Figures 4, 5, 6. Accordingly, test data prediction given in above figures shows that test accuracy of our model. Because, the test data histogram is properly fitted with test histogram. Further, in Table I shows how bias and variance relates with different models which are separately derived to predict compositions of carbon monoxide, benzene and nitrous oxide. Accordingly, it shows high bias and low variance which is common in all three models. This indicates underfitting behavior of data points. However, we use coefficient determination factor to show how well our proposed models fits the observed data. As mentioned in Table I, all three models has overall best performance efficiency. For example, carbon monoxide prediction model hold 0.82 (82%) coefficient determination, which indicates high efficiency of the model. And such, benzene model holds

0.95 (95%) coefficient determination, which indicates highest efficiency. Moreover, compare to the other two models, the model for nitrous oxide has less efficiency since it has 0.6 coefficient determination. This is clearly shown in Figure 6 in which fits the testing histogram with training histogram.

V. Conclusions

In this study we propose prediction model to predict air pollution components which are specified as carbon monoxide, benzene and nitrous oxide. We uses multiple predictors such as temperature, absolute and relative humidity and five sensor data. We identify that these predictors are correlated among each other. Therefore, we reveals the need of handling multicollinearity in this study. Consequently, we use rigid linear regression instead general linear regression method. We clearly show the performance efficiency of proposed models using determination of coefficient, bias and variance scores for each models. We recommend to expand this study to evaluate how well these prediction models behave on dding more noisy data. This may help to increase the prediction accuracy. Further, we suggest to implement this study to increase the amount of predictors to predict air pollution.

Acknowledgement

I would acknowledge to all the people who helped me to success this effort. Specifically I would thank for the directors and all the members in flip00 for directing me towards this projects.

References

[1] S. Gutta, "Bias and variance in simple terms," 2020. [Online]. Available: https://medium.com/analytics-vidhya/difference-between-bias-and-variance-in-machine-learning-fec71880c757

[2] G. K. Uyanık and N. Güler, "A study on multiple linear regression analysis," *Procedia-Social and Behavioral Sciences*, vol. 106, pp. 234–240, 2013.

[3] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012.

[4] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: the problem revisited," *The Review of Economic and Statistics*, pp. 92–107, 1967.

[5] J. Dong, C. Deng, R. Li, and J. Huang, "Moving low-carbon transportation in xinjiang: Evidence from stirpat and rigid regression models," *Sustainability*, vol. 9, no. 1, p. 24, 2016.