TITLE OF THIS PAPER

AUTHOR 1, GANG LI, AND AUTHOR 3

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.

Contents

Date: (None).
2020 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

Formula for Introduction

GLi:
A good paper introduction is fairly formulaic. If you follow a simple set of rules, you can write a very good introduction. The following outline can be varied. For example, you can use two paragraphs instead of one, or you can place more emphasis on one aspect of the intro than another. But in all cases, all of the points below need to be covered in an introduction, and in most papers, you don't need to cover anything more in an 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.

Motivation

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.

What is the specific problem considered in this paper?

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,

Contribution

- 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.

Most of the other studies try to reveal prediction models without considering multicollinarity. However, the importane of our method is addressing multicollinarity in high level. 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.

A roadmap for the rest of the paper

and [**?**] or **?**].

GLi:
Do all of your tenses match up in a paragraph?

2. Preliminaries

Initially, regression problems anayse the correlation between the predictor variable and the response variables [**?**]. Most probably, regression problems define correlation

A note with no line back to the text.

GLi:
This is comment from Gang.

QWu:
Response from QW

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 [**?**]. 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 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 **??**, 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 **??** 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.

GLi: Gang Li has worked up to here.

3. Method

3.1. **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 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 **??**. This uses to identify the symmetric and skewness of target data in the training set before we develop the model.

QWu: Qiong Wu has worked up to here.

Table 1. Precision Comparison on Event Detection Methods

	OR Event Detection	AC Event Detection	TC Event Detection
precision	0.83	0.69	0.46
recall	0.68	0.48	0.36
F-score	0.747	0.57	0.4

Multicollinearity is a aspect which occurs when two or more predictor variables are correlated [**?**]. As a consequent, coefficients error may increase [**?**]. 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 [**?**]. 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 [**?**]. 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 **??** and Figure **??**. As a solution, we decide to implement rigid regression method to derive our proposed method.

4. Experiment and Analysis

5. Conclusions

The quick brown fox jumps over the lazy dog. Jackdaws love my big Sphinx of Quartz. Pack my box with five dozen liquor jugs. The five boxing wizards jump quickly. Sympathizing would fix Quaker objectives.

Acknowledgement

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

The authors would like to thank ...

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy

(A. 2) School of Information Technology, Deakin University, Geelong, Australia
Email address, A. 2: gang.li@deakin.edu.au

(A. 3) School of Information Technology, Deakin University, 221 Burwood Highway, Vic 3125, Australia
Email address, A. 3: xxx@deakin.edu.au