Prediction of Nitrogen Dioxide Level using Machine Learning Models

Audrey Wang

Abstract — *Air pollution has been a lingering problem to many developing countries. Overpopulated cities and unregulated industrial and vehicle emissions constantly destroy the natural environment and harm human health. Being one of the most common air pollutants, nitrogen dioxide, also known as NO2, damages our respiratory system, destroys aquatic food chains, and contributes to the formation of acid rains. In this paper, linear regression, random forest regressor and decision tree regressor from scikit learn are used to predict NO2 level in parts per million (ppm) by analyzing its correlation with other air pollutants in a specific city. The model achieved an accuracy of 69,1 percent using random forest and a 66 percent accuracy using decision tree. The accuracy of the model can increase with a more detailed dataset, but further scientific research and discovery are required to actually predict acid rain based on just NO2, since the amount of NO2 that will cause acid rain is uncountable and many more variables go into the formation of acid rain.*

I. INTRODUCTION

As we release more pollution into the atmosphere as factory and vehicle byproducts, acid rains are formed and are constantly damaging the environment and harming wildlives under and above the ocean. If the occurrence of acid rains can't be limited, the phenomenon of this man-made disaster will threaten the health and safety of billions. By predicting the potential development of acid rain with nitrogen dioxide (NO2) levels, one of the major components of acid rain, we could prevent it by limiting the number of specific gasses produced. For this project, I will be using machine learning models to predict the NO2 Mean in parts per million (PPM) of a given area based on the level of other pollutants in the air. The dataset and features recorded are primarily composed of numerical data, so the result will also be presented as a number. Any NO2 levels that are greater than 20 ppm will be considered dangerous to life and health, especially the respiratory and immune system. Since it's a supervised regression problem, I will use linear regression, random forest regressor, and decision tree regressor to achieve the goal.

I was able to find one research essay that also developed around the theme of air pollution, titled: “Detection and Prediction of Air Pollution using Machine Learning Models” by Aditya C R. Both of our projects used meteorological factors to predict a specific pollution index by employing supervised learning algorithms. The difference is that they simplified the result into a classification problem due to the uncertainty of the data, predicting the PM2.5 is either high or low. I remained at finding a single numerical prediction for the NO2 level. The limitation with classification is that it’s hard to re-utilize the categorical result since it’s not a number. However, simplifying the problem increases the model’s accuracy and lowers the impact of deviations, allowing the model to be much more reliable.

II. METHODOLOGY

Dataset — Since the relation between air pollutants and acid rain is comparatively a rare topic, I couldn't find a proper dataset for my project other than the one Inspirit AI provided. This dataset named US Pollution 2000-2021 from Kaggle provided information on the daily pollution levels in US cities. It predominantly focused on recording the four types of air pollutants: Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Carbon Monoxide (CO), and Ground-level Ozone (O3). The data contains more than 69 thousand entries and 24 columns of numerical and language data. I filtered out the immaterial features to tune the dataset, and was left with six remainings, which include the SO2 Mean, the CO Mean, the O3 Mean, the State, and the Year for independent variable X and NO2 mean for dependent variable Y. Both the state and year features have noteworthy influences on the NO2 Mean, which optimistically will improve the accuracy in predicting future NO2 means.

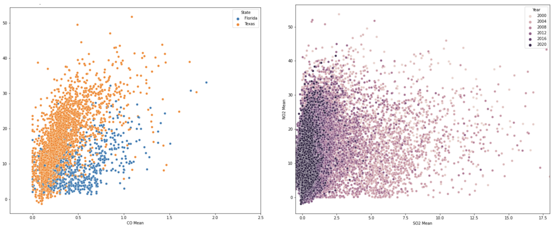
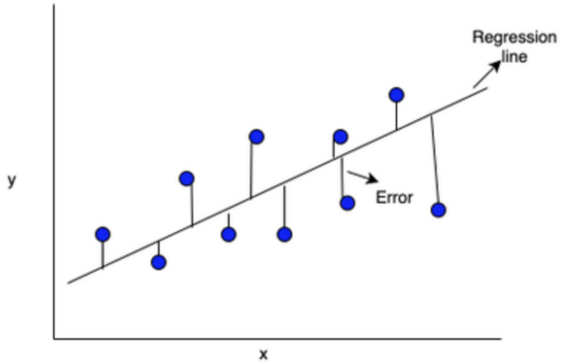


Figure on the left shows the impact of the state feature, figure on the right shows the impact of the year feature.

Note that since the state names are language data, they were enumerated and replaced by numbers for convenience. Additionally, I subtracted the year of each entry by 2000 to amplify the feature's impact on predictions. And finally, for the data splitting, I put 30 percent into the testing set with a random state value of 30.

To predict numerical results on a continuous scale (gas level in parts per million), I applied three supervised learning regression models from sklearn to the pollution dataset. Linear regression, as the first and the simplest model, fits a linear equation to the dataset, representing the correlation between the NO2 level and the number of other gasses in the air. To elucidate the linear fitting process, the model will formulate a line with the least error average with all the data points and make predictions based on that regression line.



[8]Figure above shows how the linear regression line is formed based on the error.

Then much different from linear regression, the decision tree model predicts values by categorizing the data's features. Tuples with different values fall into separated tree branches; the leaf nodes will represent individual labels. The pollution data will follow approximate pathways to find a predicted level of NO2 by classifying its features using this model. And last but not least, the random forest regression creates a prediction based on the average of multiple decision tree models' results, providing a higher accuracy than normal decision tree regressions.

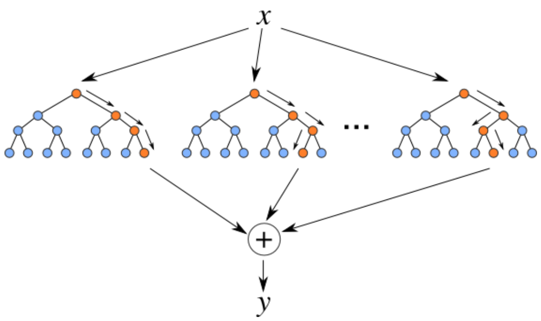
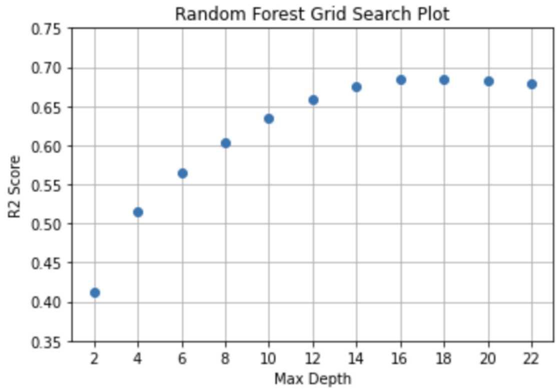


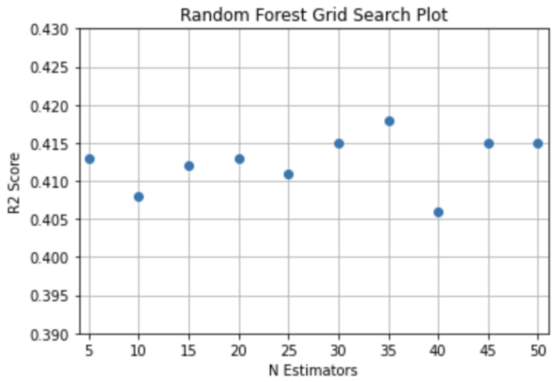
Figure above shows an example model illustration of random forest regression.

During the first round of the training and testing process, all three models ran under the default hyperparameter setting. Epochs were also not applied to the training set since the regression models use the whole dataset all the time. I fitted the 70 percent training data to the model and used the model to predict a new set of y values (NO2 levels) with the 30 percent of X testing data. I imported the R squared measure from sklearn metrics to calculate the accuracy by comparing the predicted y values with the actual y testing values. R2 scales from 0 to 1 and measures how accurate the regression model can predict from 0 to 100 percent. I applied the same process to all three models.

III. RESULTS

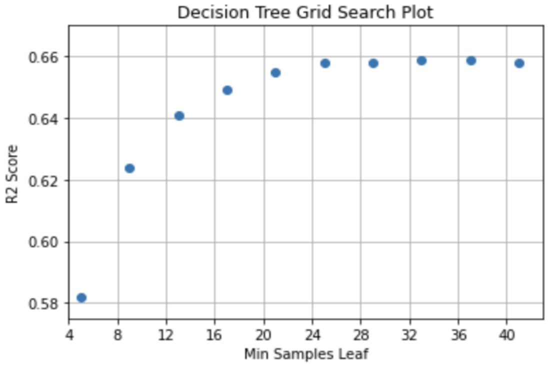
Since the vanilla linear regression does not have any hyperparameters, the final R squared value of the model will be 0.491, a nearly 50 percent accuracy. To tune the random forest regressor, I modified two hyperparameters that take integer inputs, which are the depth/split of the tree and the number of trees in the forest. As shown in the graph below, the model has the highest R2 value of 0.685 when max depth equals to 16. Imprecisely, I tested for the best n estimators value without building off from the best max depth value, and found out the model with a 35 n estimators value has the highest R2 value of 0.418.

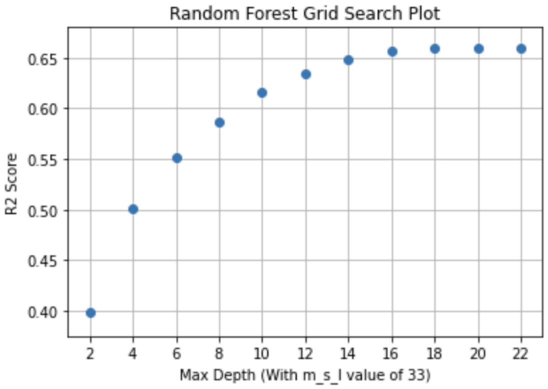




Combining the two optimal tuning hyperparameters, the random forest regression model ended up with a R2 value of 0.691, approximately a 70 percent accuracy (highest score among all models).

I applied the same tuning process to the decision tree regressor, but with the max depth variable and the hyperparameter that describes the minimum number of samples required to be a leaf node. With a minimum sample leaf value of 33 and overlaying with max depth of 16, the decision tree regression model can achieve a R2 value of 0.660.





IV. CONCLUSION

Although pollution is always taken into account while developing manufacturing and scientific projects, it won't be the factor that can stop corporations from earning money. Through all the years that pollution has been highlighted and discussed, only a handful of measurements were taken because the majority still hasn't felt threatened by the consequences of harming the environment. However, human-made disasters such as acid rain, greenhouse gasses, and ozone depletion are warning signs that have shown short-term impacts on the population. The goal is to take preventive, instead of reactive, measures while regulating air pollution. Using the random forest regressor, the model can predict the level of NO2 with a 69 percent accuracy to forecast the possibility of acid rain. Since so many other environmental factors contribute to the formation of acid rain, this model couldn't get a higher R2 score with its limited number of features. Furthermore, we need more advanced research on the correlation between the air pollutants' levels and the pH level in the air to truly predict acid rain based on the amount of NO2 in the air. I believe this model, with the proper, detailed dataset, could provide advantageous assistance for the Environmental Protection Agency (EPA) to regulate air pollution emissions by predicting the level of any gas.

ACKNOWLEDGMENTS

I would like to thank Inspirit AI and my mentor Matthew Radzihovsky for giving me this precious opportunity and expert guidance to complete an AI research project on a topic that I'm extremely inquisitive about from start to finish.

REFERENCES

[1] Blog.arcadia.com. 2022. 15 Key Facts and Statistics About Acid Rain. [online] Available at: <https://blog.arcadia.com/15-key-facts-and-statistics-about-acid-rain/> .

[2] Nj.gov. 2022. [online] Available at: <https://nj.gov/health/eoh/rtkweb/documents/fs/1376.pdf>.

[3] Clarity.io. 2022. Air Quality Measurements Series: Nitrogen Dioxide. [online] Available at: <https://www.clarity.io/blog/air-quality-measurements-series-nitrogen-dioxide>.

[4] GazDetect. 2022. Nitrogen dioxide detector & sensors, NO2 - GazDetect. [online] Available at: <https://en.gazdetect.com/gas-information/no2-nitrogen-dioxide-detector/>.

[5] Let's Talk Science. 2022. What is Acid Rain?. [online] Available at: <https://letstalkscience.ca/educational-resources/stem-in-context/what-acid-rain>.

[6] C R, A., Deshmukh, C., D K, N., Gandhi, P. and astu, V., 2018. Detection and Prediction of Air Pollution using Machine Learning Models. International Journal of Engineering Trends and Technology, 59(4), pp.204-207.

[7] US EPA. 2022. Overview of Greenhouse Gases | US EPA. [online] Available at: <https://www.epa.gov/ghgemissions/overview-greenhouse-gases> .

[8] Cloudera. 2022. Understanding Linear Regression. [online] Available at: <https://community.cloudera.com/t5/Community-Articles/Understanding-Linear-Regression/ta-p/281391>.

[9] Scikit-learn.org. 2022. scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation. [online] Available at: <https://scikit-learn.org/stable/index.html> [Accessed 27 August 2022].