Abstract

Extensive research conducted by the Environmental Protection Agency (EPA) has shown that airRecently, it has been discovered that tropospheric ozone has a detrimental effect on human health and climate, so it's of paramount importance to measure its concentration in air to take mitigation measures. Nevertheless, there are no reliable and fast instruments to measure ozone concentration on site. In order to address this issue, a machine learning model is proposed. The AQ-bench dataset was employed to predict average ozone concentrations in the world. First, feature scaling and encoding was performed. Then, in contrast to the benchmark, new features were generated to take into account relevant chemical reactions and geographic conditions that influence ozone formation. Next, we implemented linear regression, regularized linear regression, random forest, gradient boosting regressor, neural network, support vector regression and voting regression models to determine which one fits the data more accurately. The correlation coefficients of our linear regression, random forest and neural network were 59.4%, 59.3% and 60% respectively, while the benchmark reported values of 53.7%, 58.3% and 59.8%. Experiments demonstrate that the voting regression model is the best model, obtaining a correlation coefficient of 66.3%. This is the highest value amongst our proposed models and higher than any model reported on the benchmark.

1. Introduction

Pollutants such as volatile organic compounds, nitrogen oxides, tropospheric ozone and particulate matter have increased the amount of lung and heart disease in the population, as well as other health problems in vulnerable age groups [1]. From these, tropospheric ozone is particularly difficult to monitor in air quality research and is arguably the most detrimental to human health and ecosystems because it acts as a greenhouse gas and affects atmospheric circulation, cloud formation and precipitation levels [2, 3]. Ozone (O3) is a toxic greenhouse gas. While stratospheric ozone protects life on the planet's surface from ultraviolet radiation, tropospheric ozone is detrimental to human health, vegetation and climate. Additionally, it is responsible for one million premature deaths around the globe and has a harmful impact on crop productivity [3]. Tropospheric ozone concentration is challenging to predict using theoretical models because it lacks direct emission sources and is instead created through a number of chemical reaction chains involving a wide range of precursors [2]. To address this challenge, in this project we develop a Machine Learning framework to predict ozone concentrations around the world. These predictions will help authorities (such as the EPA) pinpoint which factors are most responsible for tropospheric ozone production, devise measures to control this pollutant, and implement risk prevention measures to protect the population in cases where the predictions surpass the established threshold.

2. Data

The AQ-bench dataset has been gathered from 5577 meta stations scattered over the world, compiled from TOAR database. The purpose was to determine the resulting ozone metric given all the environmental influences.

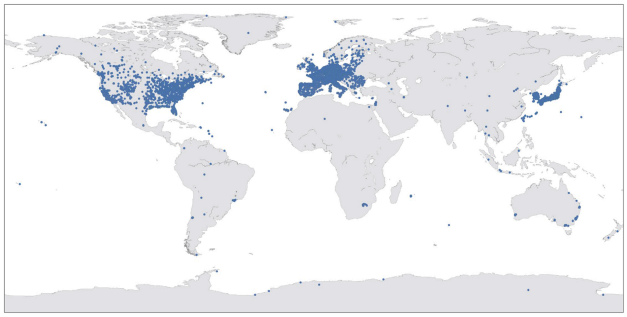


Figure 1: Meta Stations all over the world

The data has been gathered from 2010 to 2014, in a span of 5 years for aggregated air quality. These weren't influenced by short term weather and emission forcings due to the elongated period of gathering the data. Many environmental influences are on short timescales, however, this dataset will be used to predict long-term air quality conditions at these sites, thus recording data for a long time is necessary. Variety of metadata with diverse information about environmental influences has been stored throughout the day. These metadata are sometimes not directly related to tropospheric ozone; however, these can act as influencing factors to those factors which have a direct influence on tropospheric ozone. For this project, aim has been to find the average tropospheric ozone. The dataset has a total of 38 features which are relevant to our study. Some data is missing, which would have been essential for this study. This data will be generated using empirical models from the data that was collected. SOme data sets needs to be scaled as the numeric features need to have appropriate weights. Similarly one-hot encoding is needed on this dataset for categorical dataset. For the models developed by AQ Bench [2] to predict average ozone, benchmark dataset has been used.

3. Related work

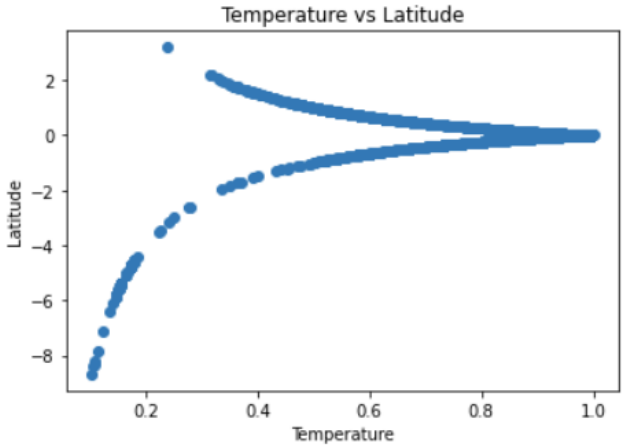
In recent years, machine learning has achieved remarkable success in areas such as pattern, image and speech recognition by usage of increasing computing power, innovative algorithms and high data availability[2].This has aroused the interest of environmental scientists in

exploring the application of machine learning and data-driven methods in their fields. The strength to be exploited is the ability of machine learning algorithms to find complex relationships in large multivariate, inhomogeneous dataset.We are focusing more on identifying the features and their correlation for prediction of tropospheric ozone concentration.

4. Methods

The AQ bench dataset was best suited for the problem as it contained the data from various regions across the globe at different times and made the data time independent. The cleaned dataset[4] used here described the geological features for the meta-data station, namely its altitude, its relative altitude, its location (co-ordinates), and the country the data is from. It also gave a detailed description of the flora around the meta-data station. The amount of vegetation and the type of it were also mentioned in it. Another facet of this dataset was the population information that it provided. All these factors played an essential role in estimating the tropospheric ozone concentration, but we can still improve as these are not a true representation of how the reaction takes place. It is as good as saying that one student studies in the school which is at X location and has Y kinds of food available in the canteen, there are Z teachers. All these things while true are not indicative how the student will perform in the exam. There has to be more information to be unearthed, such as how much time the student would spend with the teacher studying, the canteen food is to the student's preference, the location where the student lives is close by to the school so that there is lesser loss in the time in the commute.

Similar is the case with our problem as well. The factors affecting or creating the tropospheric ozone are mentioned but their interaction is not clear. Thus, we did feature engineering and created the features of how factors are interacting and producing tropospheric ozone. For this purpose we look into the chemistry of the ozone formation, where it is formed from the VOC (volatile organic compounds) and NO_x in the presence of sunlight. The reaction is triggered by the temperature as well, the higher the temperature the more the ozone is produced. The dataset being used did not give the sunlight intensity and the temperature at the meta-data station. Thus our approach was to model the temperature and sunlight such that they improve the quality of the data and actually mimic the physical nature of the respective features. Now since the data was time independent, we were restrained to use relative temperatures, that is the temperature is more in the regions near the equator and less in the regions near the poles, but this decrease cannot be a linear decrease as it is not how temperature decreases in the environment, it so for that we did a sigmoid of the negative log values of the latitude. This was done to get a much more accurate decrease in the temperature. Also the sunlight was modeled in the similar manner, which was the exponential with the exponent as the cos of the absolute value of the latitude. Another major feature that we added was the longitude. Longitude is a circular feature and was dropped in the paper, but we considered the longitude as the cosine of the value. This alone increases the R2 score of the models.



=====

Figure 2: Temperature profile plotted against the latitude

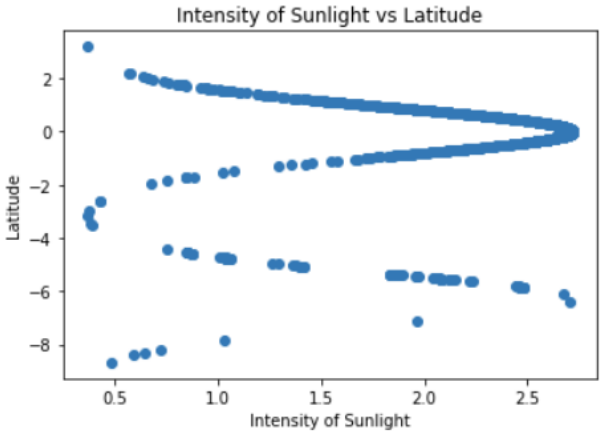


Figure 3: Intensity of Sunlight profile plotted against the latitude

Once we had our missing ingredients we implemented features which would depict the conditions of the reaction. The this was done by taking into account the reason of VOC emissions, which were human interventions, intense sunlight, temperatures higher than what they are used to be in and the amount of NO_X present around it. The amount of NO_X emitted was also modeled from the population and the night light in the area. Another major component is the amount of water that is always present as it will decide the

type of vegetation that will be inherent to the region. This was a feature which we considered.
Then new features were created which were a combination of the existing features which would give a better depiction of how the reaction takes place. This methodology was then tested which is elaborated in the subsequent sections.
The experiments were done in such a manner that we implemented the standard models that were mentioned in the paper and then moved onto hyperparameter tuning and then ensemble methods. The ensemble methods proved to be better than the standard models.

5. Experiments

The data procured from AQ Bench data set has also been used in several ML models to predict the atmospheric values in future. In their study, they have worked out predictions of all atmospheric parameters using the models; however, for our project, we aimed to predict average tropospheric ozone concentration values in the atmosphere. Thus, from their study, we focused on the prediction of average tropospheric ozone values only. The models used benchmark data rather than using cleaned up data or without any modifications. In our study, as discussed in methods, we have tried to make the data more susceptible to give better predictions using the ML models. There are 3 models for this dataset available in publications. The models used for their study are Linear Regression, Random Forest and Neural network.
The predictions based on these models with benchmark data are not too reliable. For evaluation R2 has been considered, based on that value, the accuracy percentage of predicting for each model with this dataset can be determined. Before any feature engineering, or using the benchmark dataset, the Linear regression model has the least prediction accuracy of 53.7%. The next best model used was the Neural network with a prediction accuracy of 58.3%. And the best model studied by the publication is of random forest with an accuracy of 59.8%. Some models are performing better than the others for predicting average tropospheric ozone values. We tried the feature engineering discussed above to try to make the predictions better. We started by using the same models to compare the cleaned data.
Figure 2 compares the results that were discussed in publication with the results we obtained. It was observed there was some improvement in prediction accuracies for the three models when data was put through one hot encoding, data scaling and introducing new features. The best improvement was observed in linear regression, the prediction accuracy went from 53.7% to 59.4%. For random forest there was very low improvement, 59.8% to 60%. The best model after adding features and data scaling was Neural Network, with a prediction accuracy of 61.1%.

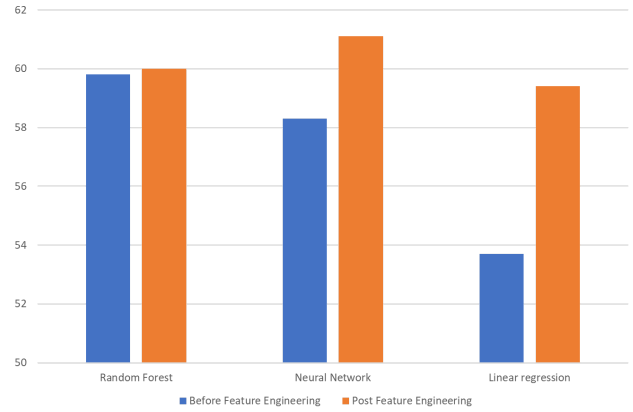


Figure 4: Model Comparison with different data sets

To have better predictions from this model, we tried to implement PCA. However, that did not yield good results. A lot of data points were getting dropped as compared to what was anticipated. Moreover, for each country, a new feature was introduced as a part of one-hot-encoding. Thus, we decided to not implement PCA for this dataset as there was a possibility of underfitting.
We saw improvement in results; however, it was still not good enough. We tried to make some improvements in predictions by implementing some other Machine Learning models as well, as it was possible that using some other methods might yield better results. Before implementing other models, some hyperparameters of the Random Forest method were changed. The result showed an improvement, initially, estimators were taken to be 100, and that gave a prediction accuracy of 59.8%. When the estimators were set to 300, the prediction accuracy went up slightly to 60.0%
Three other models have been implemented with this dataset. It was observed that the new models that were applied were better models with higher accuracy. The three models that were implemented other than Linear regression, Random forest, and Neural network were Support Vector Regression (SVR), Gradient Boosting, and Voting regression. The prediction accuracy for SVR was observed to be 62.7%. The next best model for this dataset is gradient boosting, with a prediction accuracy of 65.0%. The best model however, is voting regressor, this had a prediction accuracy of 66.3%. Figure 3 shows how the 6 models compare with each other.

Figure 5: Comparison of new models

The best model so far was the voting regression model. The visual representation of this model's predictions and errors are shown in the heatmap figures below. Figure 4 shows the ground truth for this dataset, Figure 5 shows the predictions from this dataset and voting regression model, and lastly Figure 6 shows the errors in the model's predictions.

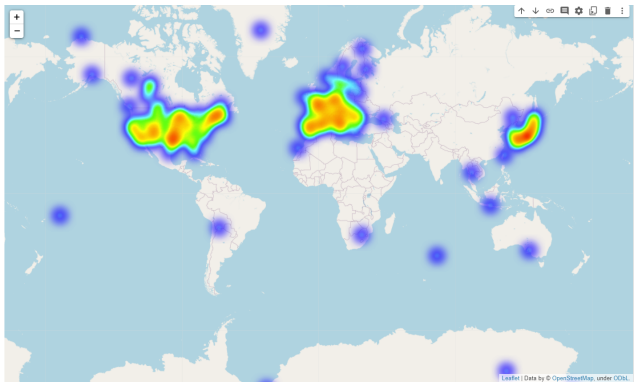


Figure 4: Ground Truth

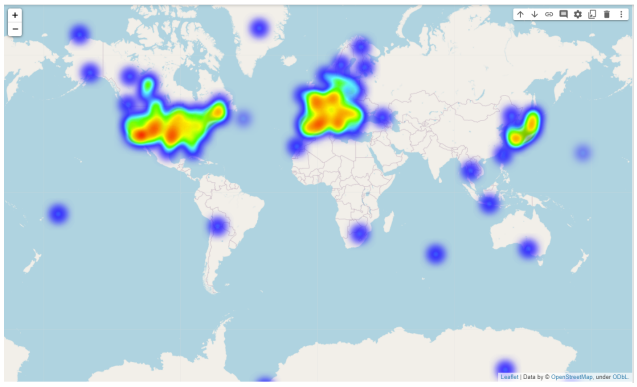


Figure 5: Predicted

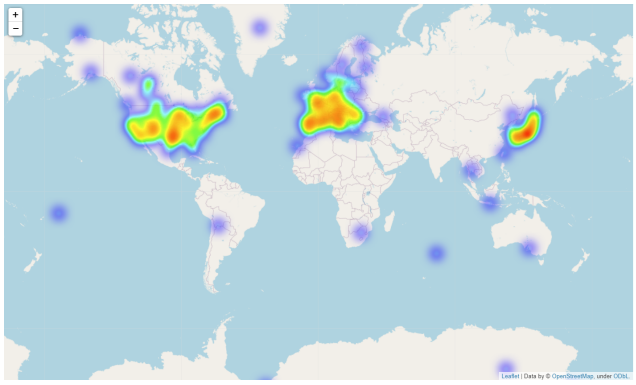


Figure 6: Error plot

6. Conclusions and Future Work

The report explores the various methods we employed to predict the tropospheric O_3 concentration from the data that were collected from meta stations situated at different locations around the world. Six different models are fitted to the dataset and a comparative study is conducted with the AQ dataset, as the benchmark dataset measuring 5577 air quality metrics. The report explores the methodology to increase the accuracy of the models from the base models mentioned in the paper. In addition, the method of feature engineering based on the chemical reactions and physical properties are explored. The addition of temperature as a new feature proved to be the most important feature as it relates the chemical reaction encouraging the forward reactions favoring the ozone concentrations. Post feature engineering the accuracy of 61.1 % for neural networks was achieved as compared to 58.3% for the benchmark model. Among all the models deployed for the study, Linear Regression model showed the most increase in the accuracy of 59.4% as compared to 53.7 % was observed. For the Random forest model a slight increase of 0.2 accuracy is reported post exploring the feature engineering. Furthermore, three additional models: (i) Support Vector Regressor, (ii) Gradient Boosting and (iii) Voting Regression are explored in the report. Of all the models deployed for this study, the Voting Regression with R^2 score of 66.3 % is achieved.

Regarding the application of feature engineering, an extension for the near future is the identification and incorporation of new features into our models for improved accuracy. The models are developed using 171 features (after one hot encoding) that can be added and different models can be employed for a robust solution. Furthermore, this report considered the study of average ozone concentration, which doesn't account for the specific concentration during day time and during artificial light conditions at night. One can study the effects of

anthropogenic activities during a 24 hour period to make robust data-driven decisions.

References

[1] United States Environmental Protection Agency, “Research on Health Effects from Air Pollution,” US EPA, Sept. 29, 2022. https://bit.ly/3y1wlgy.

[2] C. Betancourt, T. Stomberg, R. Roscher, M. G. Schultz, and S. Stadtler, “Aq-bench: A benchmark dataset for machine learning on Global Air Quality Metrics,” Earth System Science Data, vol. 13, no. 6, pp. 3013–3033, 2021. https://essd.copernicus.org/articles/13/3013/2021/.

[3] Climate and Clean Air Coalition, “Tropospheric Ozone,” UN Environment Programme, Sept. 29, 2022. https://bit.ly/3dTSMO8.

[4] AQ-Bench: a benchmark dataset for machine learning on global air quality metrics

[5] https://scikit-learn.org/stable/modules/ensemble.html