ISSN:0975-9646

Air Pollution Prediction Using Machine Learning Techniques

Aananya Lakhani

Abstract: Air pollution is a big cause of concern as we have seen pollution levels increase across major South Asian cities. This not only impacts human health but also affects the economic development for any nation. In the past few years, various data collection centres have been created across different cities of India to monitor this. The data is available in both temporal and spatial domains. In this research article, we have carried out exploratory data analysis to understand the various factors that contribute to air pollution as well as find the impact of the covid-19 lockdown and consequent limited human activity, on the environment. Lastly, we have built machine learning models using support vector regressors and random forest regressors to predict the air quality index (AQI) trend.

Nomenclature:
AQI – Air Quality Index
SVC – Support Vector Classifier
RF – Random Forest
SVR – Support Vector Regressor

Keywords: Prediction, Machine Learning, Support Vector Regressor, Random Forest.

1. Introduction

With increasing economic activity around the globe, the demand for a diverse range of products and services is growing exponentially. This leads to a major impact on the environmental pollution coming from the agricultural sector, manufacturing sector, infrastructure sector as well as automobile sector [2, 6]. Major parts of all these sectors are still relying on coal and petroleum sources for energy production that is used to conduct the operations of these industries. Inspite of strict environmental regulations and awareness the scale of operations in these industries leads to emissions of various kinds of pollutants such as sulphur, sulphur dioxide, nitrogen dioxide, carbon monoxide, benzene, toluene, xylene (the triads are commonly known as BTX), particulate matter 2.5 ($PM_{2.5}$) and particulate matter 10 (PM_{10}) [3]. Air pollution has a direct impact on the health of various animals, plants and humans [6, 5, 2, 4]. Prolonged exposure to bad quality air leads to bronchial irritation, swelling of the tracheal walls or even lung cancer in the worst cases [1]. Air Quality Index (AQI) is an indicator that is used to report air quality. AQI below 10 is good, between 50 and 100 is satisfactory, between 100-200 is unhealthy for sensitive people, between 200 and 300 extremely unhealthy and beyond 300 is hazardous for normal population.

One of the most widely used modelling techniques for time series forecasting is autoregressive integrated moving average model (ARIMA) [7, 8, 9, 10]. Box and Pierce [7] have described the impact of residuals on estimations. The residual autocorrelation can be represented as a singular linear transformation which possesses the properties of a normal distribution. In his paper, Yeaw et. al., [9] demonstrated the ARIMA model's ability to predict the monthly AQI with 95% confidence level. In [10], a comparison between the ARIMA model and the exponential Holt smoothing model on 2014 AQI data from Beijing, China is done and it is shown that ARIMA is better than Holt model for better forecasting. Le and Cha proposed real time air pollution prediction using a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for temporal analysis on satellite images of Daegu city, Korea. The model demonstrated was simple in architecture but had powerful prediction capabilities.

In this paper, we will introduce the previous work that has been done in the prediction of air pollution and will explore the various statistical and machine learning techniques that have been used for this study. Thereafter in section 2, we will bring forth the theoretical machine learning technique that has been employed in this paper. Section 3 will go into exploratory data analysis of the air pollution dataset from Indian cities. Section 4 will describe the methodology and section 5, the results of air pollution prediction using advanced machine learning techniques. Lastly, the article will conclude with key learnings and findings from this research work.

2. Machine Learning Techniques

Two machine learning regressor techniques are used in this article namely, support vector regressor and random forest. These techniques are explained in this section.

2.1. Support Vector Regressor

The Support Vector Regressor is a combination of linear regression and support vector classifier. The purpose of the support vector classifier (SVC) is to create a hyperplane that creates a division across a homogenous group of datapoints. The SVC separates the training set $(x_1, x_2, x_3, \dots x_n)$ which belong to the d-dimensional space into $y_i \in \{-1, 1\}$ which denote different classes of the observation. The classes are separated by a hyperplane into a new feature space by a kernel function $K(x_i, x_j)$. The kernel function can be a linear function, radial basis function (RBF), polynomial function or a sigmoid function, dependening on the problem set [13, 14, 15].

2.2. Random Forest Regressor

The random forest classifier is a superset or an ensemble method to decision trees (DT), wherein the main drawback of DT is that they tend to overfit the training data. Random forest solves this problem as it is basically a collection of decision trees forming a forest. Randomness in the random forest comes in as the trees are different from each other and they bring randomness in the prediction thus preventing the overfitting of the dataset and resulting in improvement of overall accuracy [12]. It was proposed by Breiman in 2001 [16]. He showed prediction consistency using a simple version of random forest.

3. Exploratory Data Analysis of the Air Pollution Dataset

The dataset contains various pollutants that come from different states, cities and weather stations. The dataset is collected from 2015 until 2020. This means, that we can use this dataset to find the impact of decreased human activity on air pollution during the 2020 covid-19 lockdown. Key air pollutants captured in this dataset are PM2.5, PM10, NO, NO2, NOx, CO, SO2, O3, Benzene, Toluene, Xylene, and Air Quality Index (AQI). The data is collected on a daily basis and covers all major cities such as Delhi, Mumbai, Chennai, Ahmedabad, Bhopal, Lucknow, Bengaluru, Chandigarh, Gurugram, Guwahati, Hyderabad, Jaipur, Patna, Shillong, Kolkata, Thiruvananthapuram, etc. Before the exploratory data analysis of the dataset we have to do data engineering. This is to remove any nan or empty cells from the dataset. The empty cells and null values are removed using `*.dropna` function. Alternatively, `*.fillna` can also be used to fill the empty cell with zero. Normally, in a large dataset, empty cells are quite prominent due to data collection inefficiencies which can depend on the method of data collection. This needs to be handled by either using advanced data engineering techniques or improving upon the data collection methodology. It was found that PM10 and NH_3 has high proportion of missing values. The data is then rearranged and additional year and month columns are extracted from the date feature. Post this, the particulate matter column is created: the sum of PM2.5 and PM10. A separate Nitric metric is created which again is a sum of NO, NO_2 and NO_x pollutants in each city. The new BTX feature is also created which is the sum of Benzene, Toluene and Xylene content in various cities.

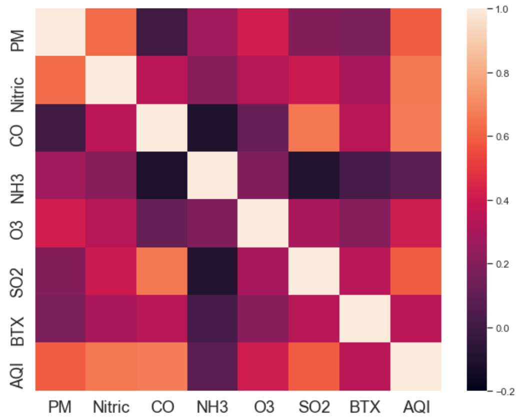

Figure 2. The correlation matrix of the pollutants and their contribution to the AQI.

Using the correlation matrix as shown in figure 2, we can see that ammonia and BTX has lowest corelation to the AQI values while nitric and carbon monoxide (CO) content have major correlation with the AQI values. From figure 3, it is evident that the overall pollution level across major cities of India decreases due to increased temperature and decreased fog. This is because cold air is dense and water condensate traps air pollutants which move slower than the warm air. The dense air traps the pollutants while warm air helps in removing the air pollutants to the outer atmosphere.

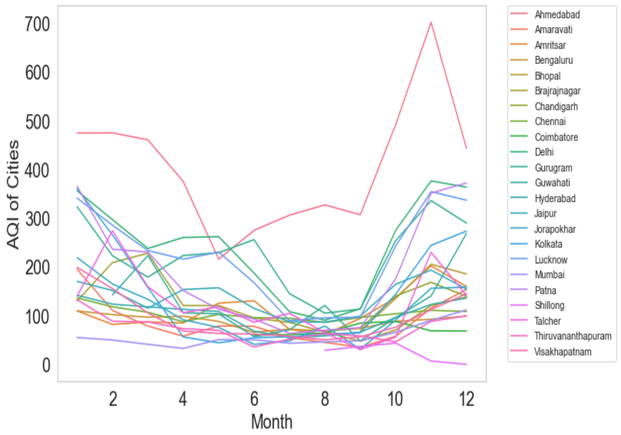

Figure 3. AQI of cities across different months in India.

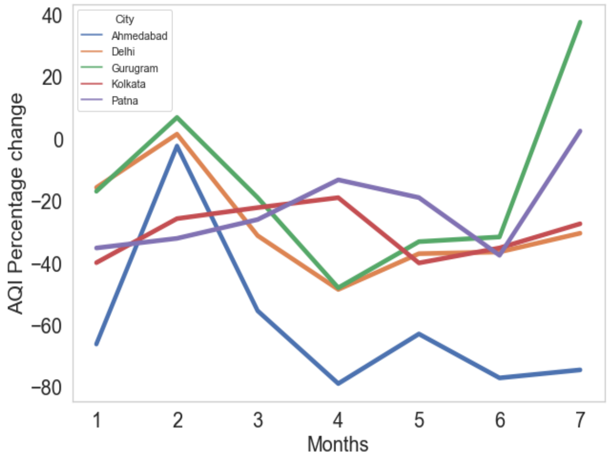

Figure 4. The percentage decrease in the AQI from 2019 to 2020 (lockdown months).

Figure 4, shows the impact of the lockdown due to Covid-19 on the AQI across 5 major cities. We can easily see that the lockdown and resultant decreased human activity significantly improves air quality and decreases AQI levels. Figure 5, shows the top pollutants and the respective top 10 cities in each pollutant metric. It can be seen that particulate matter is high in metro cities such as Delhi, Gurugram, Bhopal, Talcher, Jaipur, Kolkata, Guwahati, Chandigarh, Amritsar. Ozone concentration is high for cities like Bhopal, Jaipur, Patna, Delhi, etc. Meanwhile Delhi has the highest concentration of Nitric and Ammonia pollutants. CO, SO_2 and BTX concentration is highest in Ahmedabad this may be due to the nearby ports and the shipping industry. Overall AQI is highest for Ahmedabad followed by Delhi.

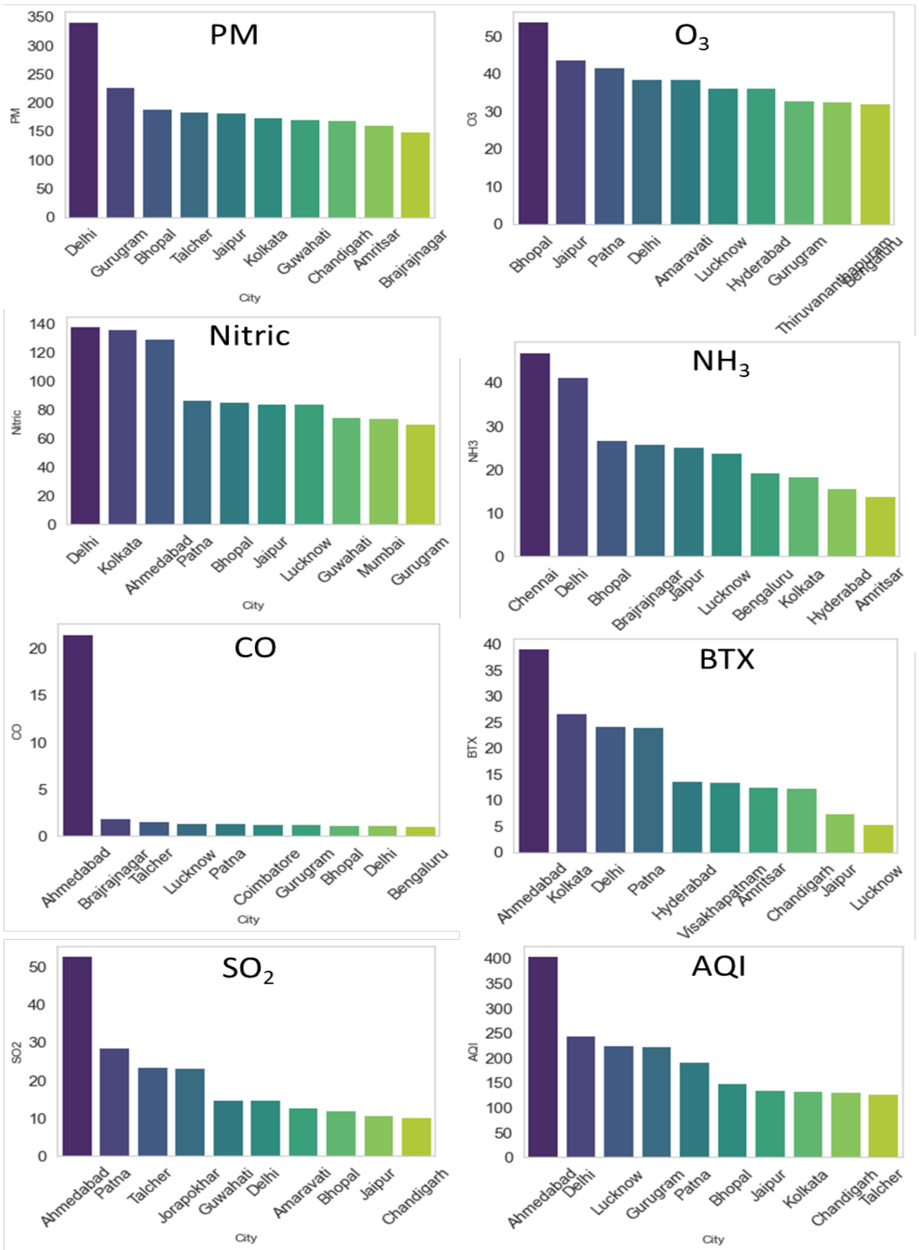

Figure 5: The above bar plots show the distribution of each pollution metric (PM, Ozone, CO, BTX, Nitric, NH_3, CO, BTX, AQI, SO_2) and the top 10 cities in those metrics.

4. Experimental Methodology

Motivated by the preceding literature, we evaluated the air pollution dataset using Random Forest and Support Vector Regressor (SVR) algorithms in our work and made predictions for the upcoming days. Lastly, we also evaluate the performance of the model.

4.1. Dataset

The data is collected from Kaggle dataset. For the prediction analysis we use the city data which is collected day wise. The dataset content and features description have already been discussed in section 3 of this paper.

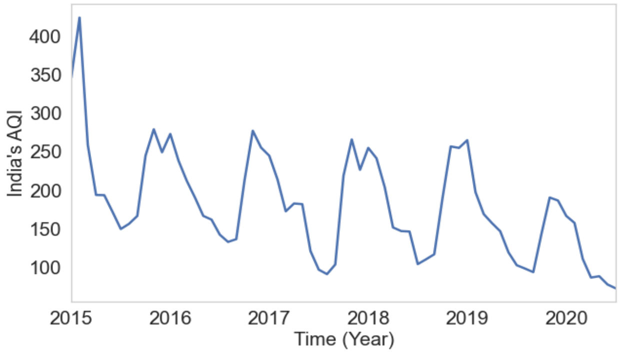

Figure 6. The above graph shows the India's AQI from the year 2015 upto year 2020.

4.2. Feature Engineering

The dataset is engineered with new AQI values for India which is basically the mean of all the AQI on a given date across various cities of India. The 'City' feature is expanded for each city into a new column to calculate the mean AQI value of India. The variation of AQI values against time is plotted in figure 6. It is not possible to feed the date feature directly into the SVR or Random Forest models as python considers it a string. Therefore, we indexed the dataset against the corresponding date and use that as an independent variable for the SVR and Random Forest model. The dependent variable is the newly created India_AQI feature in the dataframe. The data is reshaped to a 2-dimensional dataset.

For random forest regressor the data is further arranged into a stepwise function and the model is run on the stepwise X_grid dataset. For the SVR model, the training set is transformed using the 'standardscaler' method which will scale the dataset into machine optimised values for easier training and better prediction.

4.3. Parameter Tuning

The support vector classifier is imported from the sklearn library. SVR requires a kernel function which maps the lower dimensional data to higher dimensional data. The kernel in our case is known as the radial basis function. Radial basis function (RBF) $K(x, x')$ is given by the following equation.

$$K(x, x') = exp\left(-\frac{\parallel x - x' \parallel^2}{2\sigma^2}\right)$$

Where, $x - x'$ is the Euclidean distance between two featured vectors and σ is the free parameter which help us to determine how well the curve can fit around the vector points. The RBF helps in creating a hyperplane from 2D to higher dimension. The regularization parameter C is used to prevent misclassification. Smaller values lead to large margin separation of the hyperplane from the vectors while large values do the opposite. In our case, C value is chosen to be 1e4 and gamma is 10. The gamma value decides on the space influence of the vectors around the hyperplane.

The random forest regressor is an ensemble technique borrowed again from sklearn library. The trees that are built in the random forest are stored in the estimator attribute. The n_estimator was set to be 15 – the value which provided us with the best prediction. Random forest is very powerful technique and does not require depth specification as it is in itself quite deep. It is often observed that little parameter tuning would be required to get the highest efficiency which makes this technique fairly easy to implement.

5. RESULTS AND DISCUSSION

Figure 7 shows the real AQI values and the predicted values by the random forest classifier technique and the SVR model. It can be seen that the overall prediction is fairly high and predicted data is very much aligned with the real dataset. The accuracy of the SVR technique is higher compared to random forest. SVR gives an accuracy of 98 % while Random Forest gives an accuracy of 88%. Hence, for prediction of air quality index we can easily use SVR technique.

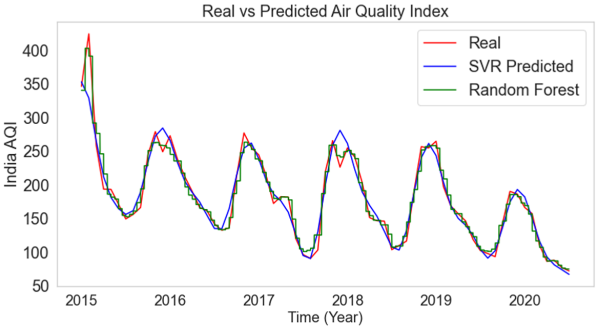

Figure 7. The prediction of a random forest regression model on the AQI of India.

6. CONCLUSIONS

In conclusion, in this article we did exploratory data analysis (EDA) and air quality index prediction on the air pollution dataset collected from different cities of India between 2015 and 2020. From the EDA it is observed that BTX and ammonia had lower correlation with AQI values. Further, air pollution decreases during the summer season and increases during winters. Delhi and Ahmedabad are worst hit cities. Indeed, the covid-19 lockdown has a significant impact in pollution reduction across cities and in some cases, pollution went down by upto -80 % (in the city of Ahmedabad). Subsequently, SVR and Random Forest models are used as a machine learning techniques to make predictions on the air quality index and both the models performed well on the training dataset with an accuracy of 98 % and 88 % respectively.

REFERENCES

1. F. Caiazzo, A. Ashok, I. A. Waitz, S. H. L. Yim, and S. R. H. Barrett, "Air pollution and early deaths in the United States. Part I: quantifying the impact of major sectors in 2005," Atmospheric Environment, vol. 79, pp. 198–208, 2013.
2. B. Holmes-gen and W. Barrett, "Clean Air Future, Health and Climate Benefits of Zero Emission Vehicles" American Lung Association, Chicago, IL, USA, 2016.
3. CERN, *Air Quality Forecasting*, CERN, Geneva, Switzerland, 2001.
4. U. A. Hvidtfeldt, M. Ketzel, M. Sørensen et al., "Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of PM2.5, PM10, and black carbon," Environmental Epidemiology, vol. 2, no. 2, 2018.
5. Y. Gonzalez, C. Carranza, M. Iniguez et al., "Inhaled air pollution particulate matter in alveolar macrophages alters local pro-inflammatory cytokine and peripheral IFN production in response to mycobacterium tuberculosis," American Journal of Respiratory and Critical Care Medicine, vol. 195, p. S29, 2017.
6. L. Pimpin, L. Retat, D. Fecht et al., "Estimating the costs of air pollution to the National Health Service and social care: an assessment and forecast up to 2035," PLoS Medicine, vol. 15, no. 7, Article ID e1002602, pp. 1–16, 2018.
7. G. E. Box and D. A. Pierce, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," Journal of the American statistical Association, vol. 65, no. 332, pp. 1509–1526, 1970.
8. C. L. Hor, S. J. Watson, and S. Majithia, "Daily load forecasting and maximum demand estimation using ARIMA and GARCH," in Proceedings of the 2006 International Conference on Probabilistic Methods Applied to Power Systems, pp. 1–6, IEEE, Stockholm, Sweden, June 2006.
9. L. Y. Siew, L. Y. Chin, P. Mah, and J. Wee, "Arima and integrated arfima models for forecasting air pollution index in shah alam , selangor," The Malaysian Journal of Analytical Science, vol. 12, no. 1, pp. 257–263, 2008.

10. J. Zhu, “Comparison of ARIMA model and exponential smoothing model on 2014 air quality index in yanqing county, Beijing, China,” Applied and Computational Mathematics, vol. 4, no. 6, p. 456, 2015.
11. V.D. Le and S. K. Cha, “Real-Time Air Pollution prediction model based on Spatiotemporal Big data”, the International Conf. on Big Data, IoT and Cloud Computing, August 2018.
12. Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. Machine Learning 63: 3–42.
13. Moula FE, Guotai C, Abedin MZ (2017) Credit default prediction modeling: an application of support vector machine. Risk Manag 19(2):158–187.
14. Pławiak P, Abdar M, Pławiak J, Makarenkov V, Acharya UR (2020) DGHNL: a new deep genetic hierarchical network of learners for prediction of credit scoring. Inf Sci 516:401–418.
15. Zhong H, Miao C, Shen Z, Feng Y (2014) Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. Neurocomputing 128:285–295
16. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.