Urban Air Quality Prediction Using Regression Analysis

Soubhik Mahanta
Dept. of CSE
NIT Warangal
msouhik@student.nitw.ac.in

T. Ramakrishnudu
Dept. of CSE
NIT Warangal
trk@nitw.ac.in

Rajat Raj Jha
Dept. of CSE
NIT Warangal
jrajat@student.nitw.ac.in

Niraj Tailor
Dept. of CSE
NIT Warangal
tniraj1@student.nitw.ac.in

***Abstract* - In the last several years, air pollution has risen steadily in urban environments. Cities like Gurugram, Faisalabad, Delhi, Beijing are few of the world's most polluted cities and have seen a dangerous rise in air pollution levels. Forecasting is important because of the human, ecologic and economic toll of pollution, and is a useful investment at individual and community levels. Accurate forecasting will help us plan in advance, decreasing the effects on health and the costs associated. Local weather conditions strongly affect air pollution levels. Generating deterministic models to study air pollutant behavior in environmental science research is often not very accurate because they are complex and need simulation at the molecular interaction level. Here comes machine learning to the rescue with high computing facilities to predict air pollution. This paper investigates how effective some available prediction models are in predicting the Air Quality Index(AQI) values given some input data, based on the pollution and meteorological information in New Delhi, India. We perform regression analysis on the dataset, and our results show which meteorological factors affect the AQI values more and how useful the predictive models are to help in air quality forecasting.**

Index Terms - Regression analysis, Machine Learning, Air pollution, Feature analysis, City dynamics

Introduction

A. The Air Pollution Problem

Seven of the world's most polluted cities are in India. And the severe problems in Delhi due to excessive air pollution is nothing new. Government agencies use AQI to inform the public how polluted the air currently is or it is forecast to become. Particulate matter PM2.5 is one of the major pollutants which leads to high AQIs[1]. An important factor that affects pollutant levels in the atmosphere, and hence affects AQI is the weather. This paper attempts to study this relationship between meteorological factors and AQI values and how we can apply some machine learning models to predict AQI values given weather information.

B. Regression Analysis

A reliable method of identifying which variables have an impact on a topic of interest(in this case our AQI values) is regression analysis. Which factors matter most, which factors can be ignored, and how these factors influence each other can be confidently determined by performing regression. The variable which is being predicted is called the dependent variable. In our paper, it's the AQI value. Factors that we assume to have an effect on the dependent variable are called independent variables. The various weather phenomena come under this category.

978-1-7281-1895-6/19/$31.00 ©2019 IEEE

Related Work

In this section, we will discuss mainly three broad categories of research work - air pollution control and prediction, supervised learning and feature analysis.

Air pollution control and prediction: Many approaches have applied to predictions and feature analysis of air quality. Yu Zheng[3] proposed a co-training framework with spatial and temporal classifiers. Hsun-Ping[4] used a similar Beijing dataset consisting of meteorological, traffic, POIs and human mobility to model an affinity graph and design a semi-supervised learning algorithm to recommend useful places for setting up air quality monitoring stations. İbrahim KÖK[11] talks about using LSTM recurrent neural networks for air quality forecasting. In our work, we focus on working with already existing regression models in machine learning libraries with a new dataset from New Delhi, and try to analyze which meteorological factors affect the air quality there, along with a side by side comparison about the performance of the existing models to understand their viability given proper data.

Supervised learning: Deep Air Learning [1] and Co training[3] discussed in U-Air have mostly performed semi-supervised learning because of the large amounts of unlabelled spatial and temporal data. Such is the case with most of the mentioned references. In our case, however, we remove the missing data (which is very less) or fill default values in them, and apply supervised learning models to them for our prediction.

Feature analysis: It involves an analysis and selection of features of the data (such as columns in tabular data) that are most relevant to our predictive model. DAL[1] algorithm does an association analysis to find relevant features. U-air[3] uses Decision Trees to find relevant features, while Deep Air[8] applies PCA to see which features affect the most variance of data. In our paper, we use Decision Forest, Extra Trees and Decision Trees models to show the importance of various features.

Proposed Work

1) Data

a. Dataset description

The data were obtained from two sources. One contains Air quality data and the other contains Meteorological Data.

Air quality data: We downloaded the historical data of air quality for New Delhi from the AirNow website (www.airnow.gov). Among other columns, the features of our use contained in this dataset are year, month, day, hour and AQI value for every 3 hours starting from 3 am on 1-1-2015 to 24-4-2017. It has 6700 values. There are other columns like conc, conc. units etc which are of no use to us. We just use the hour month day values to combine the AQI column as an extension to the meteorological data for our work purposes.

Meteorological data: The Delhi Weather Data was downloaded from kaggle. It contains hourly meteorological data of Delhi from 1997 to 2017. We trimmed it to every 3 hourly data from 1-1-2015 to 24-4-2017. The weather data has the columns - datetime, conds, dewpt, fog, hail, heatindexm, hum, precpm, pressurem, rain, snow, tempm, thunder, tornado, vism, wdird, windgust, windchill, wspdm. Conds describes the weather conditions like foggy, partial fog, mist, haze, light drizzle, rain, etc. dewpt gives the dew point, fog tells whether there’s fog or not, similarly for hail. hum describes the humidity, precpm the precipitation. pressurem describes the pressure. wdird denotes the direction of wind blowing, wspdm is the speed of wind flow.

Using the date-time as a reference, we combine these two datasets and add the AQI column to the weather dataset to form our weatherAndAQI dataset.

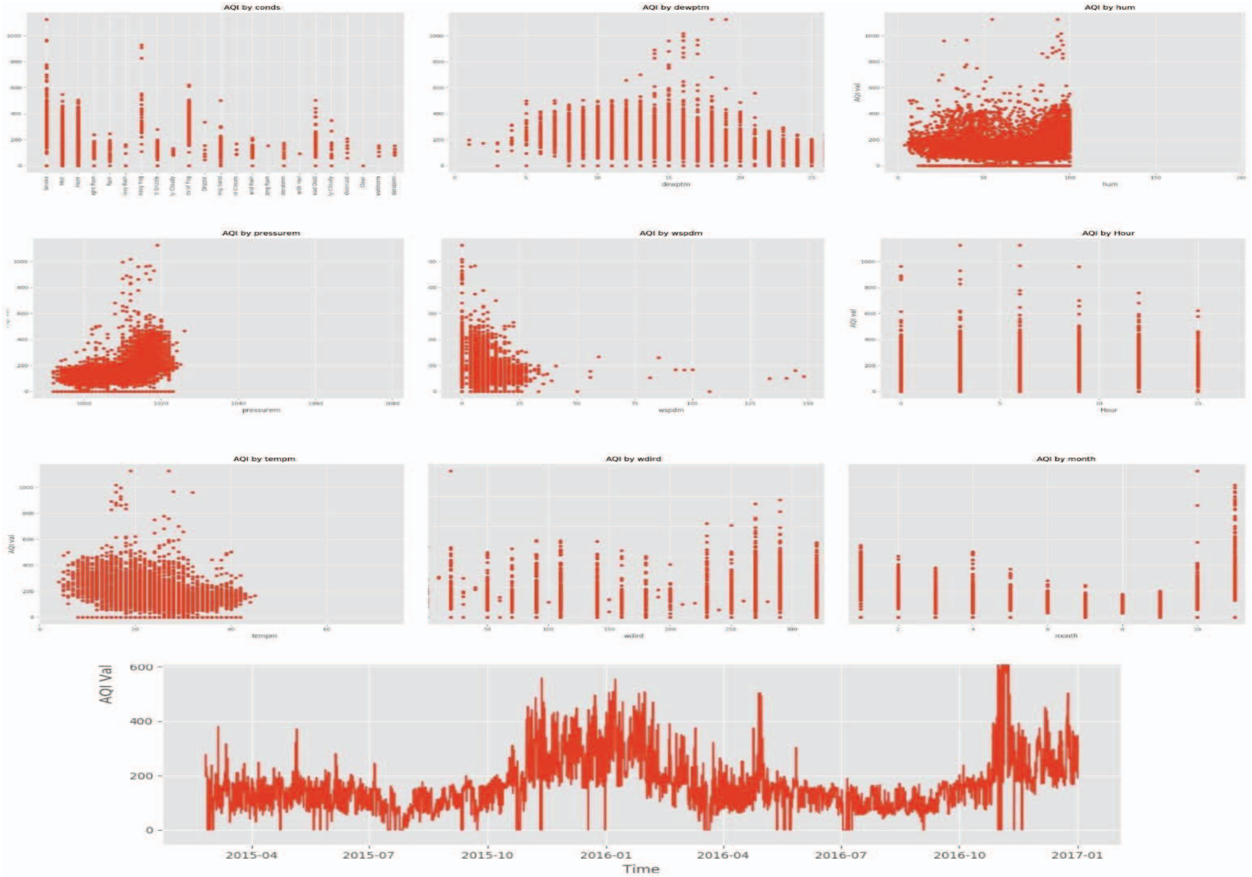


Fig 1: ***Data Visualization*** *: (from top left to right) Variation of AQI and how it varies with conds, dewptm, hum, pressurem, wspdm, hour, tempm, wdird, month, and time-series data of AQI (2015-17)*

b. Preprocessing and features

Data visualization

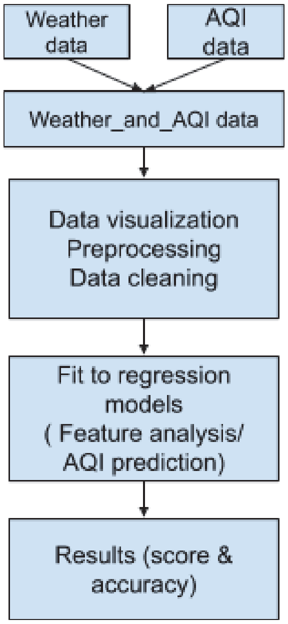
First, we drop some initial rows from the dataset because the AQI values for them were not recorded. Then we try to visualize the data and see how various features affect the AQI levels over time. As we plot the *feature vs AQI* graph, we try to find out the needed and unneeded features for our work. The plots of the features which seem to have an effect on AQI values are shown on the previous page.

Data cleaning

From the plots, we observe that datetime, conds, dewptm, hum, pressurem, tempm, wdird, wspdm, month, day, hour, AQI are our needed features. It is clear that at higher windspeed AQI is lower. Wind speed is possibly a good predictor. Similarly, AQI values are a bit higher in winter months, wind direction also matters. All other features have mostly missing values or do not significantly have any effect on the AQI value. So we drop those columns. We then create previous value features by shifting periods, so that based on previous 5 given weather and AQI data, we can predict the AQI value for the current hour. Then we take care of missing values or NaNs by dropping off those rows or filling them with the mean column values as seemed better. After doing these steps, our data is ready for fitting into the models.

2) Architecture and Prediction Models

The following figure shows the architecture of the air prediction system:



The different regression models used in the prediction system are: *Linear regression, Neural network regression, Lasso regression, ElasticNet regression, Decision Forest, Extra trees, Boosted decision tree, XGBoost, KNN, and Ridge regression*

Experiments and Results Analysis

We run on it on python 2.7 on x64 machine with 8GB RAM, 2.3Ghz intel i7 CPU, using standard plotting packages and sklearn packages. The pandas dataframe was used to implement the majority of the work.

The data was partitioned into a train set and test set in the ratio 7:3, or approx. 4400 train values and 2000 test values and each model was trained using the train set and evaluated using the test set.

Below are shows the regression plots for each of the models we have used for AQI prediction. The decision forest and Extra trees additionally have been used for feature importance analysis. The RMSE values and accuracy scores obtained are also shown in a tabular form below.

From the regression plots, we can observe that the models give pretty good results. The orange line denotes the test set points (set in ascending order) and the blue line is the corresponding predicted values in ascending order. Except for KNN and Decision tree, the blue line fits the orange line quite well, especially Linear, ElasticNet (also lasso and ridge), Extra Trees, AdaBoost and XGBoost whose prediction graph is very close to the test set graph. KNN and Decision Tree have a little less accuracy as the blue line is a bit spread out from the orange one indicating more deviation of predicted from actual values.

Table 1: Comparison of accuracy score and RMSE values for the regression models

	Accuracy Score	RMSE
Linear regression	0.84688	41.31
Neural network regression	0.82528	44.13
Lasso	0.84770	41.20
ElasticNet	0.84772	41.20
Decision Forest	0.84890	41.04
Extra trees	0.85315	40.45
Boosted Decision trees	0.83897	42.36
XGBoost	0.84562	41.48
KNN	0.69483	58.32
Ridge	0.84688	41.31

The histogram plots for RMSE and Accuracy scores are shown above.

Feature importance graph is plotted for Decision Trees, Extra Trees, and Decision Trees. The best results are shown by Extra trees predictor. It demonstrates how important the various columns of the dataset were in prediction. The Extra Trees model orders the features in decreasing order of importance as - previous AQI values, pressure, and humidity, month, temperature, conditions and hour.

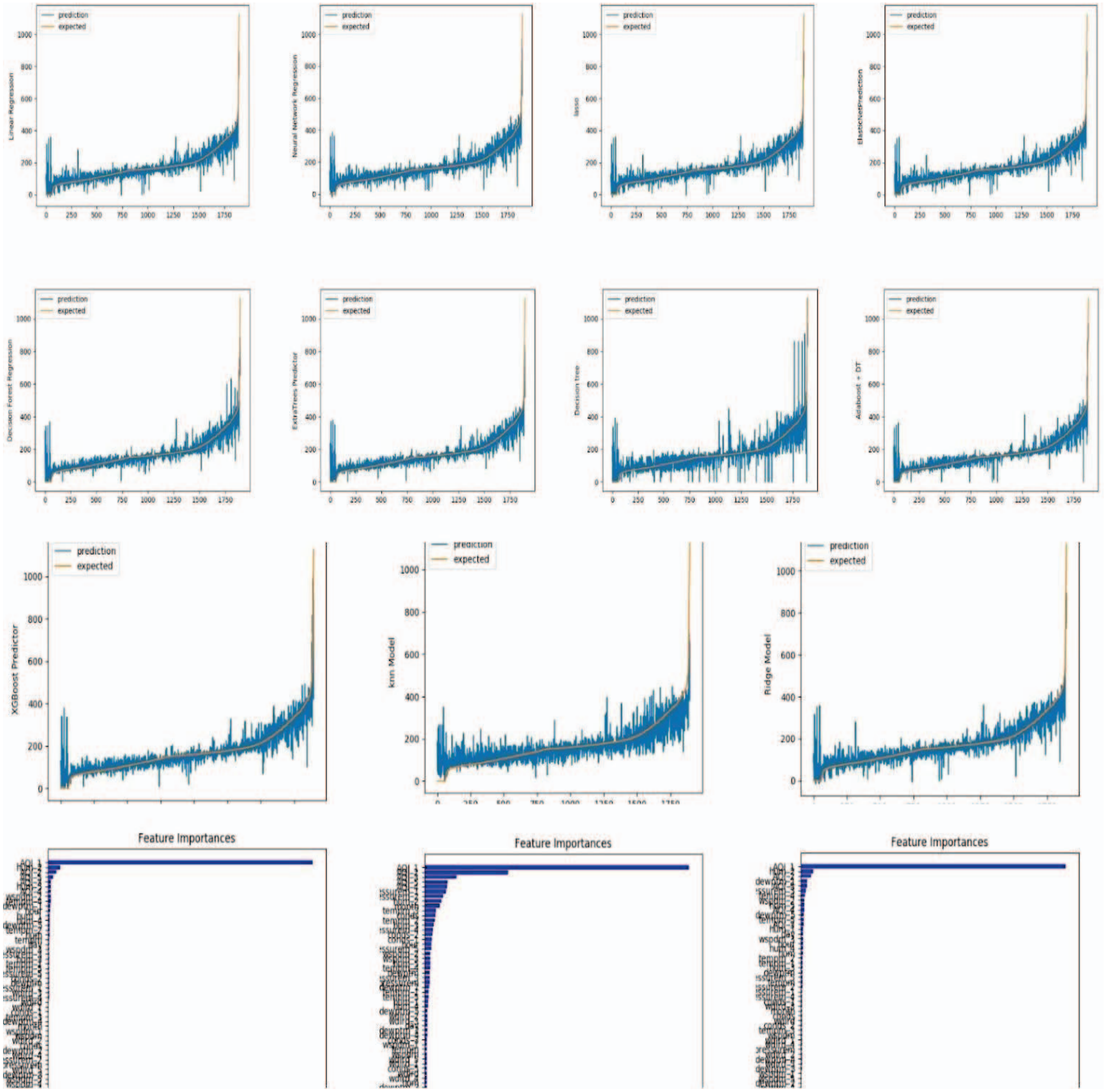
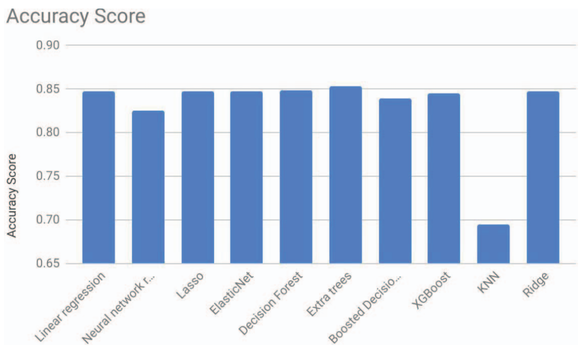
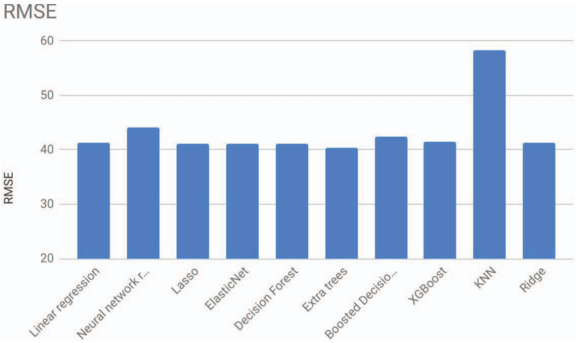


Fig 2 : ***Output plots*** *: (Top left to right) Regression plots of actual(orange) and predicted(blue) values for linear regression, neural network regression, Lasso, ElasticNet, Decision forest, extra trees, decision tree, boosted decision tree (using adaboost), XGboost, KNN, and Ridge models. The feature importance graphs as predicted by decision forest, extra trees, and decision trees are shown below*





Conclusion

In this paper, we focused on testing the usefulness of existing regression models in the sklearn library to predict the air quality index values given the past weather data. And we also tried to figure out which features are most useful for the prediction as depicted by some of those models. The results obtained show that most models achieve a decent accuracy of almost 85 %, with the highest being the Extra Trees regression model, which means that the models are quite useful as predictors.

To improve upon this accuracy, in the future, real-time and historic traffic data can be concatenated with the weather data for AQI prediction. Also, more amount of data can be used and the implementation can be done in environments like Azure ML for real-time prediction.

References

[1] QI, Zhongang; WANG, Tianchun Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. (2018). IEEE Transactions on Knowledge and Data Engineering. 30, (12), pp.2258-2297

[2] Apostolos Nicholas Refenes, Achileas Zapranis, Gavin Francis - Stock Performance modeling using neural networks: A comparative study with regression models, Neural Networks Volume 7, Issue 2, 1994, pp. 375-388

[3] Yu Zheng, Furui Liu, Hsun-Ping Hsieh - U-Air: When Urban Air Quality Inference Meets Big Data, KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1436-1444

[4] Hsun-Ping Hsieh, Shou-De Lin, Yu Zheng - Inferring Air Quality for Station Location Recommendation Based on Urban Big Data, KDD '15 Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 437-446, doi>10.1145/2783258.2783344

[5] P A Rahman, A A Panchenko and A M Safarov - Using neural networks for prediction of air pollution index in an industrial city, IOP Conf. Series: Earth and Environmental Science 87 (2017) 042016 DOI: 10.1088/1755-1315/87/4/042016

[6] Yang Zhang, Marc Bocquet: Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric environment*, Elsevier, 2012, 60, pp.632-655. ⟨10.1016/j.atmosenv.2012.06.031⟩. ⟨hal-00761344⟩

[7] Vikram Reddy, Pavan Yedavalli, Shrestha Mohanty, Udit Nakhat - Deep Air: Forecasting Air Pollution in Beijing, China

[8] Kalapanidas, Elias & Avouris, Nikolaos. (1999). Applying Machine Learning Techniques in Air Quality Prediction.

[9] Yves Rybarczyk and Rasa Zalakeviciute - Regression Models to Predict Air Pollution from Affordable Data Collections, DOI: 10.5772/intechopen.71848

[10] İbrahim KÖK Mehmet Ulvi ŞİMŞEK Suat ÖZDEMİR - A deep learning model for air quality prediction in smart cities, 2017 IEEE International Conference on Big Data (BIGDATA)

[11] Dan Wei - Predicting air pollution level in a specific city, International Journal of Engineering Trends and Technology (IJETT) – Volume 59 Issue 4 – May 2018

[12] Kunwar P.Singh, Shikha Gupta, Premanjali Rai - Identifying pollution sources and predicting urban air quality using ensemble learning methods, Atmospheric Environment, Volume 80, p. 426-437.

[13] Yifei Jiang, Kun Li, Lei Tian: a personalized mobile sensing system for indoor air quality monitoring, Proceedings of the 13th international conference on Ubiquitous computing, September 17-21, 2011, Beijing, China [doi>10.1145/2030112.2030150]