GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

PM 2.5 particle matter prediction

A Report for the Evaluation 3 of Project 2

Submitted by

Utkarsh bharal

(1613105131 / 16SCSE105071)

in partial fulfilment for the award of the degree of

Bachelor of Technology

IN

Computer Science and Engineering with Specialization of

Cloud Computing and Virtualization

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr Manoj kumar , Assistant Professor

APRIL / MAY- 2020

Table of Contents

Abstract

This project sequentially applies a set of Data Science techniques to gain insights from the PM 2.5 Dataset of an area. Data Science analysis of this data will benefit the research and community.

Many metropolitan cities in India is highly populated, like New Delhi, reported at 28,980,000 residents as of 2018, whose air quality is currently poor due to high levels of urban smog generated by neighbouring states. The primary motivation for the further study of air quality in India comes mainly from the impact of this smog on human health and concern for the residents of New Delhi who are regularly exposed to these air quality conditions. This proposal plans to focus on the monitoring of smog-related particles, PM2.5, or particulate matter measuring 2.5 microns in diameter or smaller, as a main contributor of urban smog in the area, where, although various data sets exist in the form of single point, ground based sensing monitors, a more comprehensive view of the geographic distribution of air quality could be determined through the application of remote sensing techniques like satellite monitoring. This data would be desirable both in order to increase awareness of air quality hazards on human health as well as to analyze the ground based sensor data already existing.

Introduction

With the industry development and human activities, more and more contaminants are introduced into the environment. The major form of the pollution is air pollution, water pollution, plastic pollution etc. These pollutions have a negative effect on the human health, causing serious disease and the death of human.

The air pollution in modern cities is a severe problem, which significantly affects human's life and health. PM 2.5 is a measurement a type of particulates or aerosol with a scale size less than 2.5 micrometers, which usually suspends in the atmosphere. The majority of this aerosol consists of some chemicals such as organics, sulphate, amine, nitrate, black carbon and so on. The cause of PM 2.5 is very complex, such as protosomatic emission or production, and secondary emission or production. The protosomatic production includes but not limited to vehicles emissions, power plants emissions or even natural fires. The secondary production mainly comes from varieties of chemical reactions between different chemicals in the atmosphere, whose process is usually very complicated to investigate. Besides, the physical condition of atmosphere is also an important factor to affect PM 2.5, such as the temperature, pressure, humidity, wind direction, wind speed, amount of rain, and amount of snow.

In this project, we are trying to predict the PM2.5 level using some machine learning algorithms. We will be creating some new feature using old variables to increase the accuracy of our model.

Problem Statement:

Predicting the PM2.5 level using factors like Dew, Temperature, Pressure, Combined Wind Direction, Cumulated Wind Speed (m/s), Cumulated hours of snow, Cumulated hours of rain.

Data Description and Preparation:

The data, which we got, had 13 variables and 43824 observations. The data had the following variables:

1. No. (row number)
2. Year (year of the data collected)
3. Month (month of the data collected)

4. Day (day of the data collected)
5. Hour (hour of data collected)
6. PM2.5 (PM2.5 concentration in ug/m^3)
7. DEWP (dew point)
8. TEMP (temperature)
9. PRES (Pressure)
10. Cbwd (cumulative wind direction)
11. Iws (cumulative wind speed)
12. Is (cumulative hours of snow)
13. Ir (cumulative hours of rain)

PM2.5 is the target variable.

No.

No. is of integer data type, it just provides a number to each record.

year

Year is of type integer, it tells the year in which data is collected. Different years in the data set are 2010, 2011, 2012, 2013, 2014.

month

Month is of integer data type, it tells the month in which data is collected.

day

Day is of integer data type, it tells the day in which data is collected.

hour

Hour is of integer data type, it tells the hour in which data is collected.

PM2.5

It is the target variable. It is of type integer. There are 2067 observations missing in PM2.5. Missing value contribute to 4.71% of the data. Since missing value are less than 5% so these missing values are removed. Mean of PM2.5 is 98.61 and minimum value is 0 and maximum value is 994.

DEWP

It represents the dew points. Data type is integer. There are no values missing in this variable. Distribution of the data is normal and there are no outliers present. Mean is 1.87. Minimum value present is -40 and maximum is 28.

TEMP

It represents the temperature. Data type is numeric. There are no missing values present. Distribution of the data is normal and there are no outliers. Mean is 12.45. Lowest value is -19 and highest is 42.

PRES

It represents the pressure. Data type is numeric. There are no missing values present. Distribution of the data is normal and there are no outliers. Mean is 1016 hPa. Lowest value is -991 and highest is 1046.

cbwd

This variables gives the cumulative wind direction. No missing values are present. Variable is factor. Different levels of this variables which were given are: “cv”, “NE”, “NW”, ”SE”. Since “cv” has no meaning to it and by researching we got to know that it should be “SW” so “cv” is replaced by “SW”.

Iws

This variable represent the cumulative wind speed. Data type is numeric. No values are missing. Distribution of data is rights skewed. Mean of the data is 23.87, minimum value is 0.45 and maximum value is 565.49.

From this it is clear that there are some outliers present. Using the boxplot we can see that there are some outliers for sure. 99% of data is explained by values till 268.47 and thus values above 268.47 are capped i.e. values above 268.47 are replaced by 268.47.

Is

Represents the cumulative amount of snow. Data type is integer. No missing values are present. Distribution of data is right skewed.

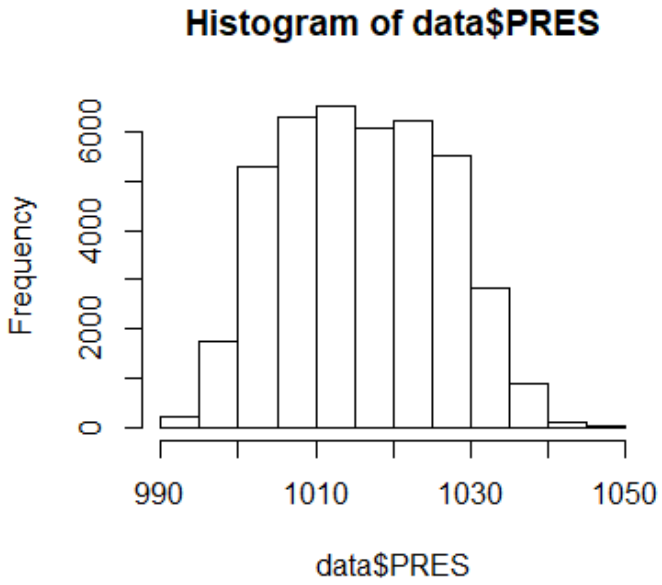
Ir

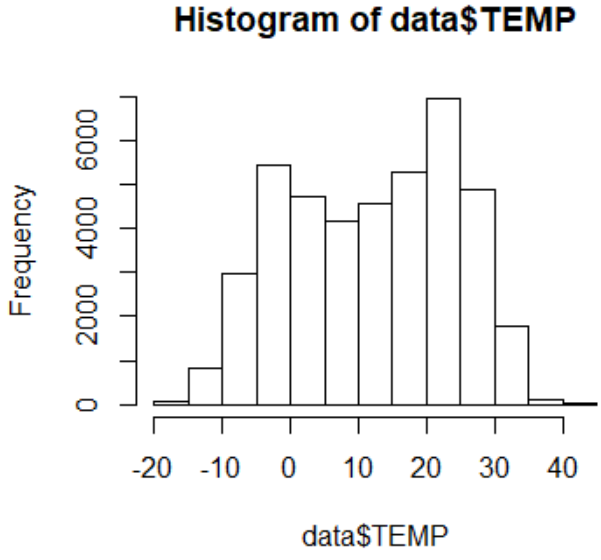
Represents the cumulative amount of rain. Data type is integer. No missing values are present. Distribution of data is right skewed.

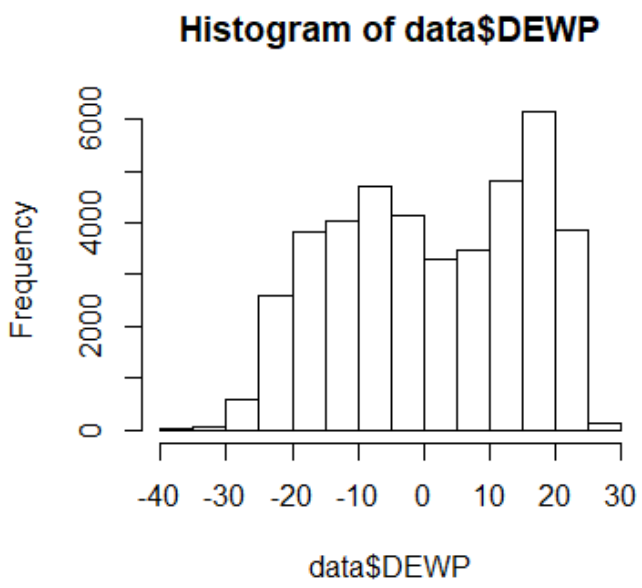
Exploratory Data Analysis

Distribution of data

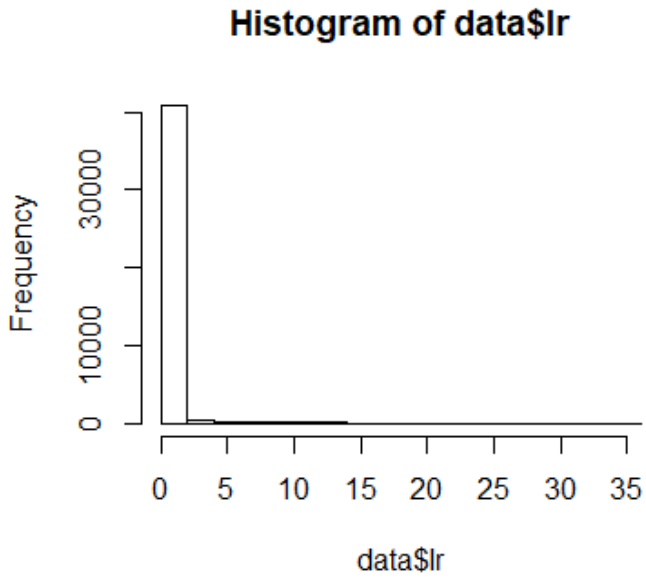
DEWP, TEMP, PRES are distributed normally as shown in the plots below.

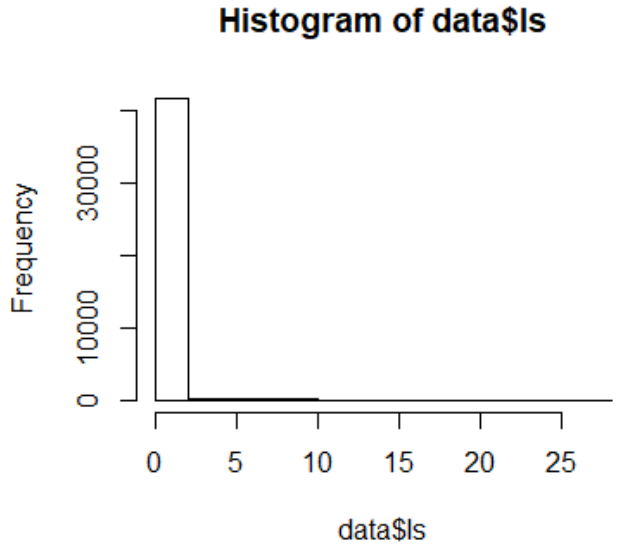


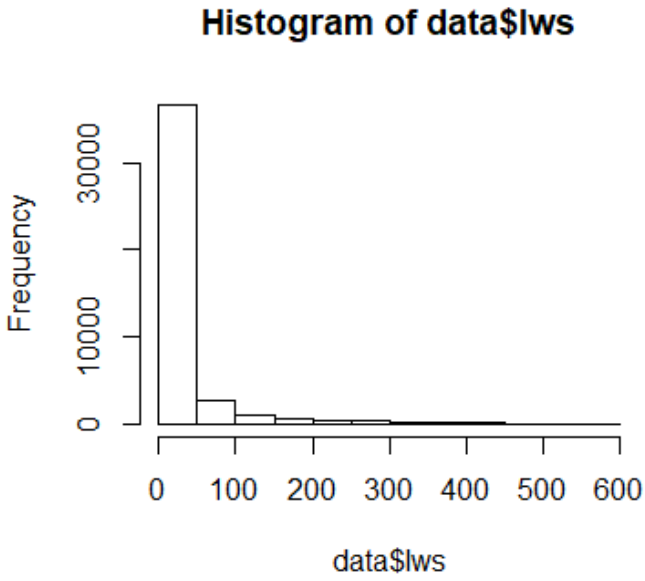




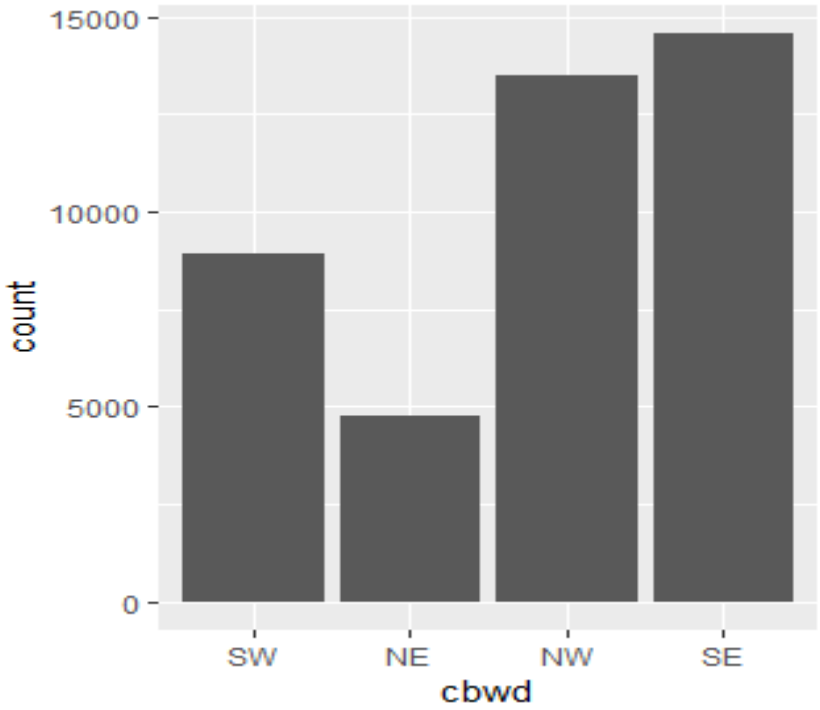
Data distribution of Iws, Is, Ir is right skewed. Plots are shown below:




Histogram of data$ls
Frequency
0
10000
30000
0
5
10
15
20
25
data$ls

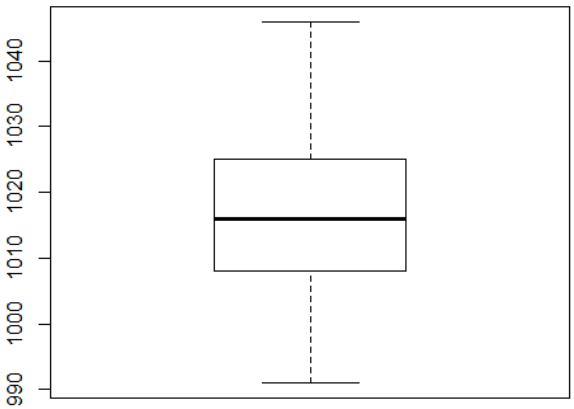

Histogram of data$lws
Frequency
0
10000
30000
0
100
200
300
400
500
600
data$lws

Distribution PM2.5 over different wind direction is show below:

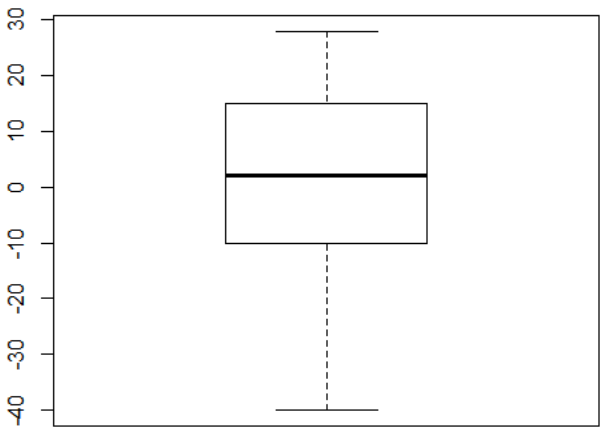


Boxplots to check the outliers

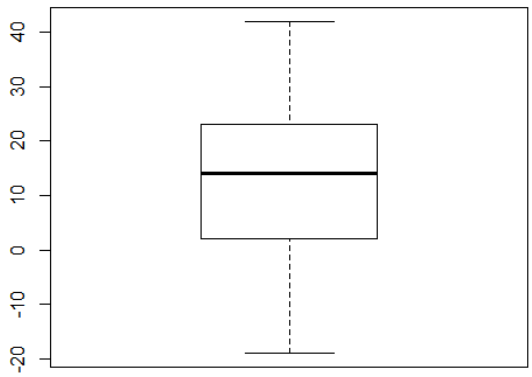
There are no outliers present in PRES, TEMP, and DEWP as shown in the plots below:



Boxplot for PRES

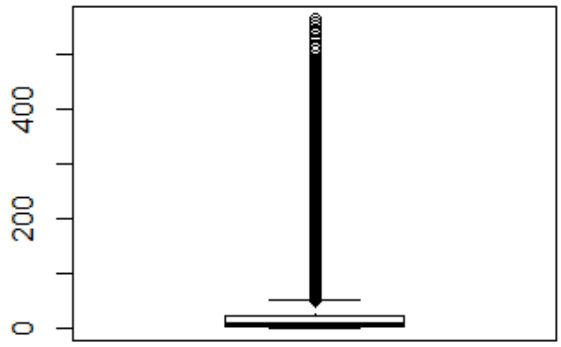


Boxplot for DEW



Boxplot for TEMP

There are some outliers present in Iws i.e. cumulative wind direction as show in the boxplot below.



The outliers are removed by capping them by the values, which represents the 99% of the data**.**

Correlation between the variables

The image below clearly shows the correlation between different variables.

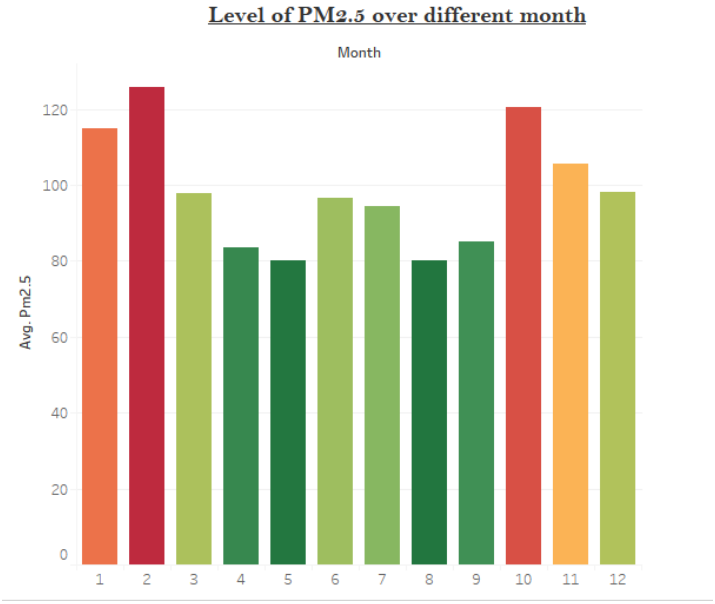
	month	day	hour	pm2.5	DEWP	TEMP	PRES	Iws	Is	Ir
month	1.0000000000	0.0069009497	-0.0005427315	-0.02406878	0.23449198	0.17213525	-0.06631703	0.014663576	-0.062883254	0.0388739475
day	0.0069009497	1.0000000000	0.0003268606	0.08278849	0.03353677	0.02287140	-0.01049679	-0.004943655	-0.037448800	-0.0001019233
hour	-0.0005427315	0.0003268606	1.0000000000	-0.02311644	-0.02178383	0.14944294	-0.04183130	0.058865349	-0.002454728	-0.0087407540
pm2.5	-0.0240687836	0.0827884927	-0.0231164430	1.00000000	0.17142327	-0.09053400	-0.04728231	-0.247784449	0.019265576	-0.0513687055
DEWP	0.2344919827	0.0335367692	-0.0217838316	0.17142327	1.00000000	0.82382123	-0.77772212	-0.293105921	-0.034925232	0.1253407561
TEMP	0.1721352533	0.0228713977	0.1494429386	-0.09053400	0.82382123	1.00000000	-0.82690281	-0.149612519	-0.094784798	0.0495444536
PRES	-0.0663170285	-0.0104967856	-0.0418312957	-0.04728231	-0.77772212	-0.82690281	1.00000000	0.178871492	0.070537123	-0.0805322089
Iws	0.0146635756	-0.0049436552	0.0588653488	-0.24778445	-0.29310592	-0.14961252	0.17887149	1.000000000	0.022630317	-0.0091569394
Is	-0.0628832535	-0.0374488001	-0.0024547276	0.01926558	-0.03492523	-0.09478480	0.07053712	0.022630317	1.000000000	-0.0097638617
Ir	0.0388739475	-0.0001019233	-0.0087407540	-0.05136871	0.12534076	0.04954445	-0.08053221	-0.009156939	-0.009763862	1.0000000000

DEWP & TEMP, TEMP & PRES, PRES & DEWP are highly correlated to each other.

Maximum correlation of PM2.5 is with DEWP and Iws and this gives an idea that at least one these variable will be useful in predicting PM2.5.

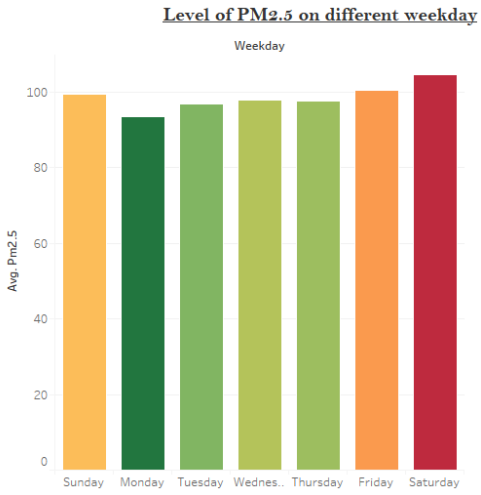
Relation between PM2.5 and other variables

PM2.5 value in different months of a year is shown in the graph below:



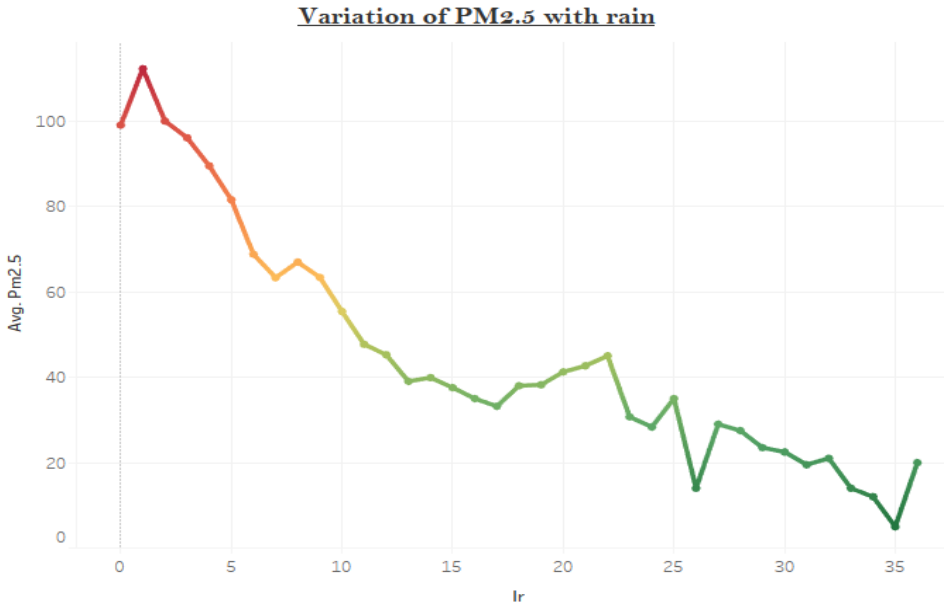
From this graphs it can be seen that level of PM2.5 is high in winters i.e. from October to March. In summers level of PM2.5 is comparatively low.

PM2.5 in different days of a week is show in the graph below:



In this graph we can see that level of PM2.5 is high on weekends and low on rest of weekdays.

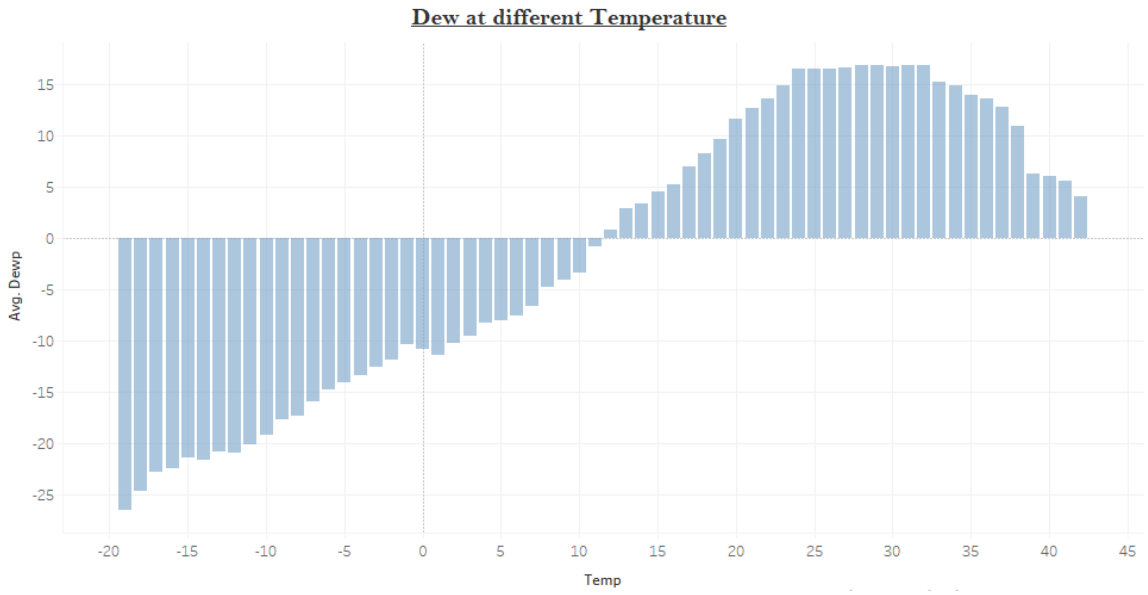
Variation PM2.5 with rain.



From this graph we can infer that PM2.5 level decreases as there is more and more rain.

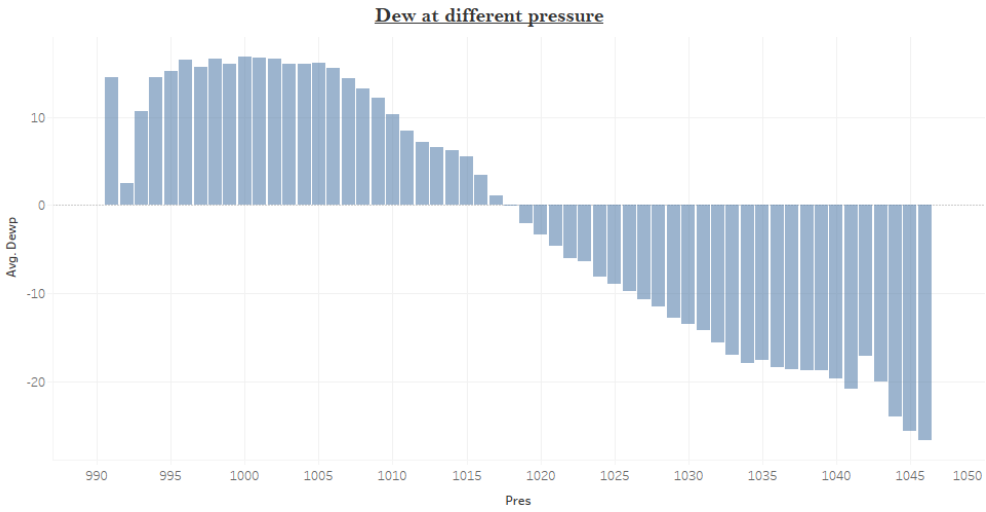
Relation between other variables

Variation of Dew with Temperature



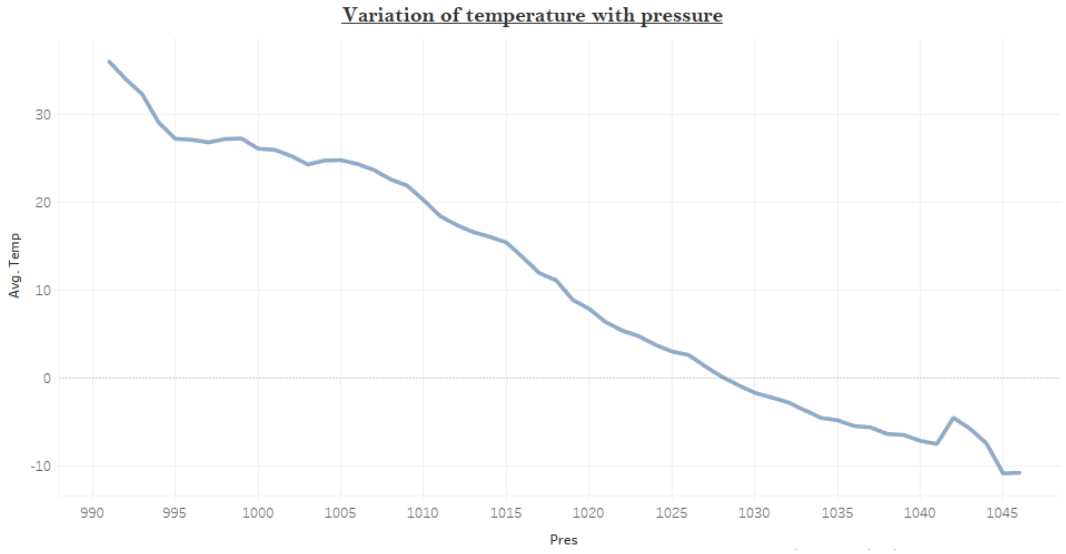
Dew is linearly related with temperature. As the temperature increases dew also increases but it reaches a maximum value and then it decreases little bit.

Variation of Dew with Pressure



Dew decreases with increase in pressure.

Variation of temperature with pressure



From the graph it can be seen that as the pressure is increased then temperature gets decreased.

Feature Engineering

The variables, which are given in the data, are never enough to predict the target with a good accuracy, to improve the accuracy some new variables have to be created using the variables, which are provided.

The new variables created are:

- first_quater_day
- second_quater_day
- third_quarter_day
- fourth_quater_day
- first_three_months
- four_to_six_month
- seven_to_nine_month
- ten_to_twelve_month
- Week1
- Week2
- Week3
- Week4
- Week5
- wind_speed_0to50
- dew_neg20_to_zero
- dew_zero_and_above

Modelling

To predict the PM2.5 three different models were created and all of them were giving different accuracy. The data was divided into train and test data with train containing 80% of the data and test containing 20% of the data. The train and test ratio was kept same in all different algorithms.

The models used are:

- Linear Regression
- Decision tree
- Random forest

Linear Regression: There were two different function used for linear regression.

1. Lm()– This was the first function to create a regression model. It was giving an accuracy of 27% which is poor. Adjusted R square value was around 34.
2. Earth()- This was other function which was used to create the regression model. It increased the accuracy to 35% with adjusted R square value of 38.

Decision tree: This was the other model that was used. The accuracy this model gave was 33% which was actually less than the Earth().

Random Forest: Random forest was the final model made. This gave the best accuracy of 63% which is acceptable.

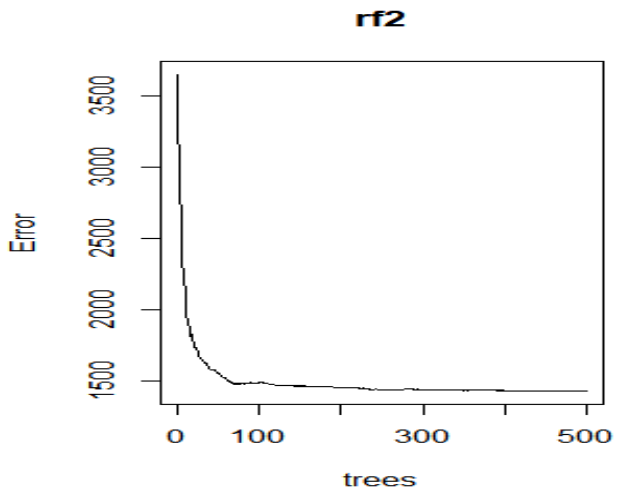
A random forest is a data construct applied to machine learning that develops large numbers of random decision trees analyzing sets of variables. This type of algorithm helps to enhance the ways that technologies analyze complex data.

The first model that was made using random forest gave an accuracy of 59%. The model was able to explain 82% of the variation.

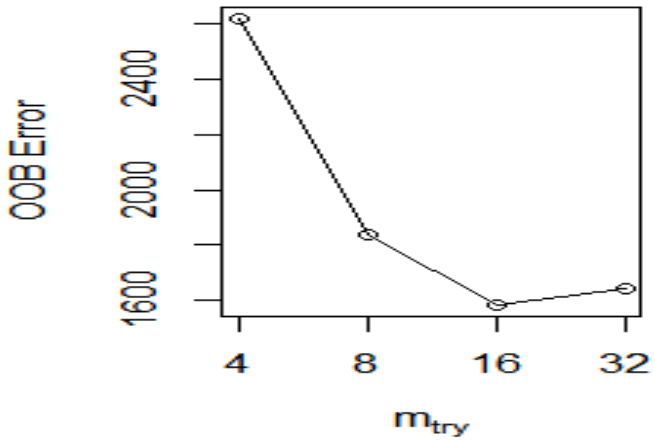
The ntree i.e. total number of trees made by the model were taken to be by default as 500 and mtry i.e. number of variables randomly selected at each split were taken to be 5.

OOB error: Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests utilizing bootstrap aggregating (bagging) to sub-sample data samples used for training.

While plotting OOB error against ntree it was clearly seen that OOB error gets constant after 250 trees. The plot is shown below:



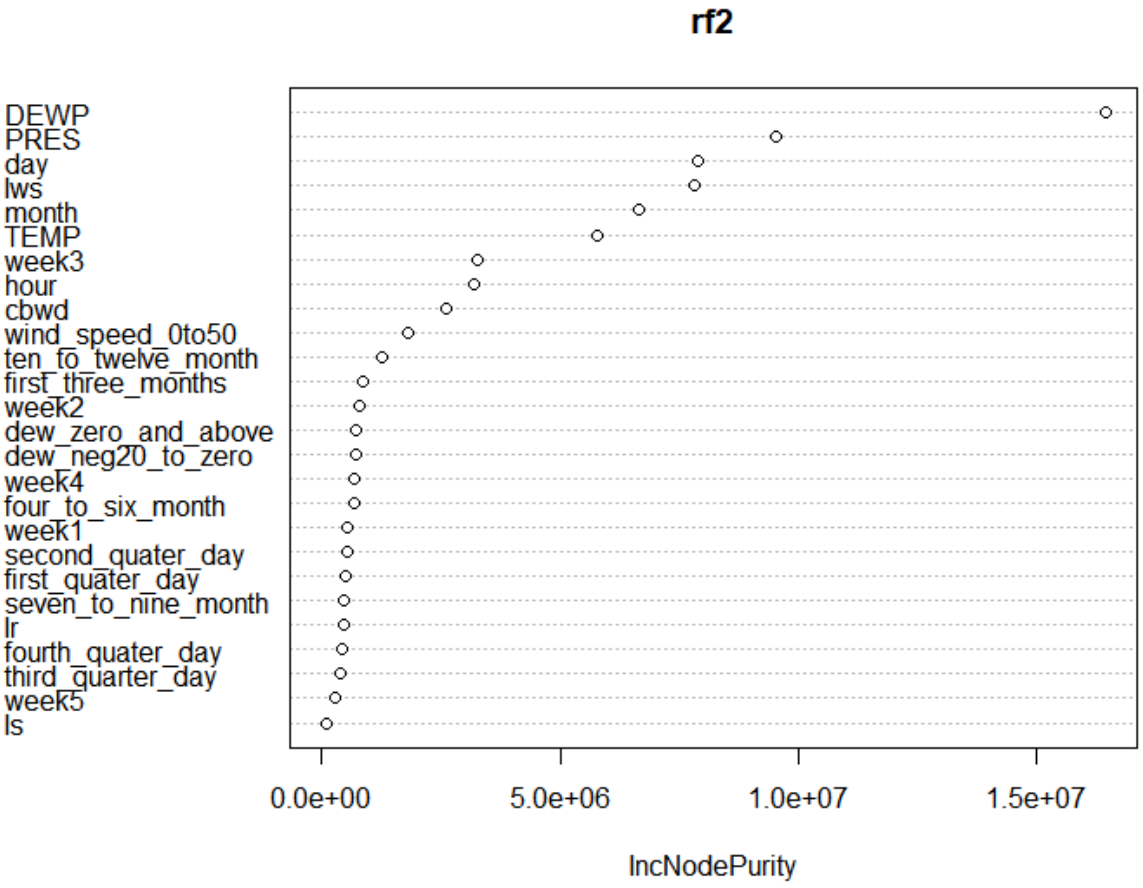
Tuning the random forest: For tuning ntreetry were taken to be 250 as error gets constant after these many trees. Improvement value after each tree was taken as 0.05. After tuning following graph was made:



From this it can be inferred that the error keeps on dropping till mtry value is 16 and it shoots up after that. This means that while creating the random forest model we have to keep mtry as 16.

After tuning the second random forest model is made with ntree as 250 and mtry as 16. This model gave an accuracy of 63% and was able to explain 85% of variance in data.

Importance of different variable in the model is shown below in graph:



From this graph it can be clearly seen that DEWP, PRES, day, are very important variables in this model. The least important variables are week5 and Is.

Conclusion

In this project, we have developed efficient machine learning methods for air pollutant prediction. We have formulated the problem as regularized and employed advanced optimization algorithms for solving different formulations. We have focused on alleviating model complexity by increasing the number of features and on improving the performance by using a structured machine learning algorithms.

Our results show that the proposed Random Forest achieves much better performance than the other two model formulations and that the regularization by enforcing prediction models for several consecutive features to be close can also boost the performance of predictions. We have also shown that advanced optimization techniques are important for improving the convergence of optimization and that they speed up the training process for the data. For future work, we will further consider the commonalities between different meteorology factors and combine them with the machine learning, which may provide a further boosting for the prediction.

After doing the project we can conclude that the area from which data is collected under high level of PM2.5 which is harmful for people living there especially old people and small kids. Level of PM2.5 is high in winter i.e. from the October to March. Level of PM2.5 is low in summers i.e. from April to September. In a week the most polluted days are Friday, Saturday, and Sunday.

The model that is created predicts the value of PM2.5 with an accuracy of 63%.

References

1. India Population 2018; *http://indiapopulation2018.in/population-of-delhi-2018.html*

2. On the source contribution to Beijing PM2.5 concentrations, Nadezda Zikova, Yungang Wang; *https://www.researchgate.net/publication/299459463_On_the_source_contribution_to_Beijing_PM25_concentrations*

3. Pm 2.5 prediction, *https://www.kaggle.com/c/pm25-prediction/data*