School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech (CSE Semester VI):
Machine Learning (CSE 523)

Project Submission #2: Linear Regression

Submission Deadline: March 04, 2020 (11:59 PM)

- **Group No.: 21**
- **Project Area:** Environment and Climate Change
- **Project Title:** Air quality prediction using machine learning algorithms
- **Name of the group members :**
  1. Vishal Saha (AU1741004)
  2. Rushil Shah (AU1741009)
  3. Muskan Matwani (AU1741027)
  4. Mohit Vaswani (AU1741039)

Submissions:

1. **URL links :** click here
2. **Implementation code:** 'code.pdf' contains implmenetation code and that file is attached in the drive folder above. Kindly see it. Drive also contains .ipynb file.
3. **Inference:**
   Ques 1. **What have we done and why is it important?**

   Our topic is 'Air quality prediction using machine learning algorithms' and we have selected an Indian Dataset consisting various pollutant data of all the states and their location and year information. The pollutants available in the data are: SO2, NO2, RSPM, SPM and PM2.5. From the given pollutants, pollutant index of each of the pollutant was calculated by EPA method and AQI was derived from the pollutant indexes by taking maximum of the calculated values.

   1. **Data cleaning and preprocessing**
   Then, Data was cleaned and pre-processed so as to make the training process more robust. It is vital as unnecessary columns/features do not provide any information and consumes more time in the training process. Pollutant columns having NaN values were replaced by their column mean value to avoid problems in further calculation.

   2. **Exploratory Data Analysis**
   Then, exploratory data analysis was done which is an approach to analyzing data sets to

summarize their main characteristics, often with visual methods. It is useful to find the distribution and correlation of various pollutants with each other and the target variable (AQI). It is also useful to analyze the location and time where pollutant's quality is worst in order to highlight and take useful and early measures for the same. We take a broader look at patterns, trends, outliers, unexpected results and so on in the dataset, using visual and quantitative methods to get a sense of the story it tells.

3. **Linear Regression - MLE, MAP, Polynomial regression**
Then, Linear Regression using MLE and MAP method is performed. Polynomial regression is implemented on the multiple independent variables to analyze the testing and training error with different degree of polynomial. It is important to analyze the relationship between the degree of polynomial and training and testing error so that we can better find the degree that can best predict and Regularization is also implemented in order to study the effect of change in the regularization parameter and the testing error.

4. **Different Regression techniques**
Then, Several different Regression techniques such as Linear Regression, Decision tree regression, Gradient Boosting Regression and Random forest regression were also implemented on the dataset in order to predict AQI. Their RMSE is further observed and analyzed to differentiate among their accuracies.

Ques 2. **How have we implemented?**

Data cleaning and preprocessing
Our dataset consists of the following features - stn_code, $sampling_date$, state, location, agency, type, so2, no2, rspm, spm, $location_monitoring_station$, $pm2_5$, $date$

1. We at first removed the unnecessary columns from the dataset such as - $stn_code, agency, sampling_date, loc$

2. Further, we found the percentage of missing values in each of the column currently present in the dataset.

3. Since we do not know the distribution of data, we removed the outliers from each of the pollutant columns as outliers have a major effect on the mean of the feature data.
4. Null values in the pollutant's feature column were replaced by the mean of the column.
5. Calculation of AQI and pollutant index of each of the pollutants :

The **pollutants/Independent variables** are: NO2, SO2, RSPM (Respirable suspended particulate matter), and SPM (Suspended particulate matter).
The **target/dependent** variable : AQI (Air quality index)

Given all the pollutants concentration in the dataset, we find pollutant index by the following formula:

$$AQI_P = AQI_{min} + \frac{P_{Obs} - P_{Min}}{(P_{Max} - P_{Min})}(AQI_{Max} - AQI_{Min})$$

where, P_{Obs} = observed 24-hour average concentration in g/m3
P_{Max} = maximum concentration of AQI color category that contains P_{Obs}
P_{Min} = minimum concentration of AQI color category that contains P_{Obs}
AQI_{Max} = maximum AQI value for color category that corresponds to P_{Obs}
AQI_{Min} = minimum AQI value for color category that corresponds to P_{Obs}

After calculating pollutant indexes of each of the pollutant, we find AQI by taking maximum of all the pollutant indexes of the pollutants :

$$AQI = max(AQI_{NO2}, AQI_{SO2}, AQI_{RSPM}, AQI_{SPM})$$

We have got one column 'AQI' in our dataset. Now, further analysis would be done on the same.

Exploratory Data Analysis

1. Analysis of pollutants with respect to year:
Year is extracted from the date column and null values are ignored while taking the year into account. The dataset was grouped by on year and AQI is plotted with respect to each year present in the rows of the dataset.

2. Analysis of pollutants with respect to state.
We first grouped the dataset on state and calculated mean of the pollutants with respect to each state. Further, for each pollutant, bar graph is plotted with x axis as states of India and y axis as that pollutant concentration found in that state.

3. Analysis of pollutants with respect to type of the pollutant.
We first grouped the dataset on type feature and then for each pollutant, bar graph is plotted with x axis as various type of the pollutants.

4. Analysis of relationship of pollutants with the target variable AQI.
Each pollutant such as SO2, NO2, RSPM and SPM is related to the AQI variable. Hence, we plotted seperate graphs of AQI with respect to each of the pollutants to study their correlarion and distribution.

Linear Regression - MLE, MAP, and Polynomial regression
Linear regression can be applied to our model in order to predict the AQI given the independent variables (pollutants) such as NO2, SO2, RSPM, and SPM.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

where x_1 = concentration of pollutant SO_2
x_2 = concentration of pollutant NO_2
x_3 = concentration of pollutant $RSPM$

x_4 = concentration of pollutant SPM

We start with Maximum likelihood equation of the parameters θ. In order to find that, we find θ_{ML} that maximizes the likelihood function given below:

$$P(Y|X,\theta) = \prod_{n=1}^{N} P(y_n|x_n,\theta)$$

Applying this, we get

$$\theta_{ML} = (X^T X)^{-1} X^T y$$

where X belongs to R^{NXD} and y belongs to R^N
Here, in our case, D = 4 (4 pollutants) and N = 435000 (total number of rows in the dataset)
Hence, θ belongs to $R^D = R^4$

We divided the dataset into 80% training set and 20% testing set. Then, at first took θ without θ_0 (without intercept) and then applied the formula (using X_{train}) to calculate θ_{ML}. And then predicted y_pred by multiplying X_{test} with θ_{ML}.

We performed the same above analysis considering the intercept term in θ and found out that

$$RMSE_{intercept} < RMSE_{withoutintercept}$$

Further, we performed the polynomial regression:

$$y = \phi(x)^T \theta$$

$\phi(x) = \begin{bmatrix} x^0 & x^1 & x^2 ... & x^K \end{bmatrix}^T$.
$\phi(x)$ is a nonlinear feature transformation of the input example x.

$$\Phi = \begin{bmatrix} \phi_0(x1) & \phi_1(x1) & .... & \phi_{K-1}(x1) \\ \phi_0(x2) & \phi_1(x2) & .... & \phi_{K-1}(x2) \\ ... & & & \\ \phi_0(xN) & \phi_1(xN) & .... & \phi_{K-1}(xN) \end{bmatrix}$$

On applying the Φ matrix, we get
for each example, y and $\mathbf{x}$

$$y = \theta_0 + \theta_1(x_1 + x_2 + x_3 + x_4) + \theta_2(x_1^2 + x_2^2 + x_3^2 + x_4^2) + ...$$

and θ_{ML} becomes

$$\theta_{\mathbf{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

in this case, Φ is of the order $R^{NX(k+1)}$
and example $\mathbf{x}$ is a 4X1 vector (as D=4)

We calculate θ_{ML} from the above equation by taking X_{train} into consideration and then, predict y_{pred} values using X_{test} and θ_{ML}.
We calculate RMSE errors of training and testing set on a range of degree of polynomials and observer that as degree of polynomial increases, the RMSE error for testing increases whereas, RMSE error for training decreases, leading to overfit the model.

Further, we perform regularization on the model and plot the RMSE error with respect to change in the hyperparameter value - regularization parameter. We observe, as the hyperparameter increases, the RMSE error decreases. For that, θ_{ML} becomes

$$\theta_{\mathbf{ML}} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$$

where λ = hyperparameter - regularization parameter to help prevent overfitting of the model.

Different Regression techniques

We perform the following regression techniques on our model:
1. **Linear Regression** : It constitutes of applying the parameters in a linear fashion and best predict the model when there is a linear relationship between independent and dependent variables.

2. **Gradient Boosting Regression** : It includes an effective solution that can be utilized for classification as well as for regression problems. The generalization of boosting to a random differentiable loss function is called as the GRBT.

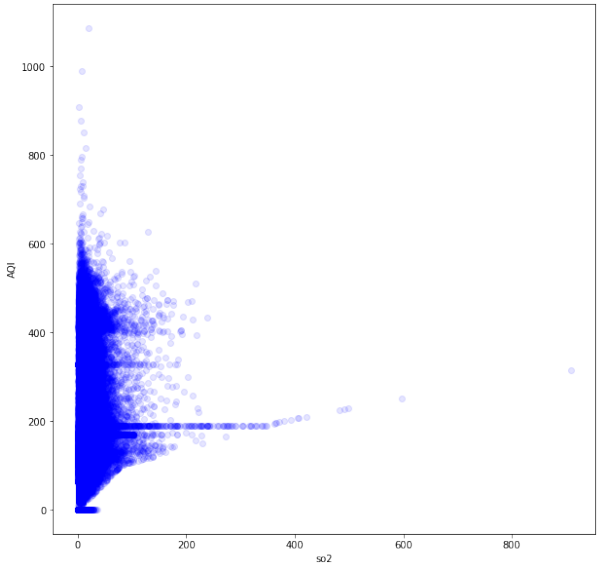
3. **Random Forest Regression** : The random forest ensures that every tree in the ensemble is generated from a sample with replacement from the training set.

4. **Decision Tree Regression** : The main objective of the Decision trees is to produce a predictive model for the values of the outcome variable using simple decision rules that have been derived from the features of the dataset.

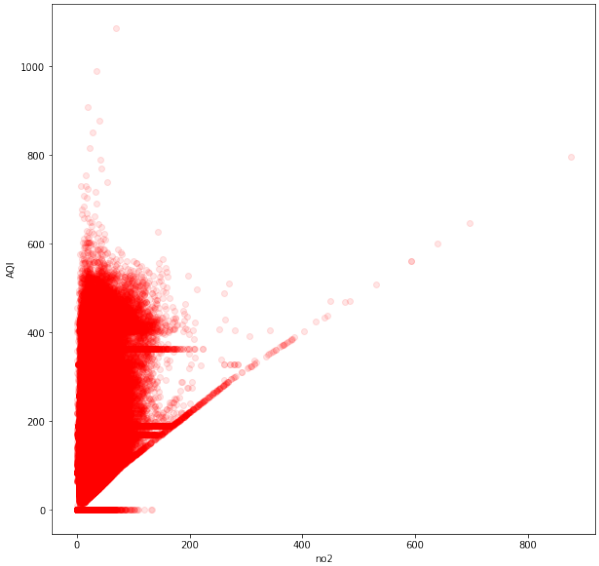
We at first, divide the dataset into 80% training and 20% testing set and call the regression functions to train the data and predict on the testing set. For evaluation, we calculate the RMSE error between predictions and actual test set values for each of the regression techniques and observe the best possible regression technique to fit our model by checking the minimum RMSE amongst all the values.

Ques 3: **Results and Inferences DISTRIBUTION OF THE POLLUTANTS WITH RESPECT TO THE TARGET VARIABLE, AIR QUALITY INDEX (AQI)**

- **AQI V/S SO2**

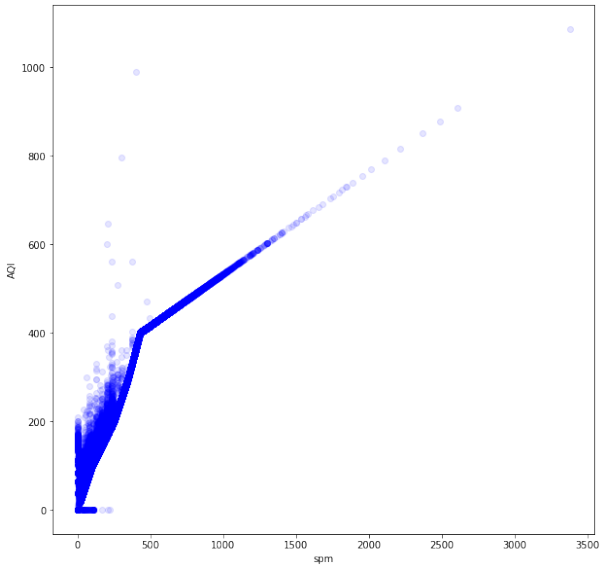


AQI V/S NO2

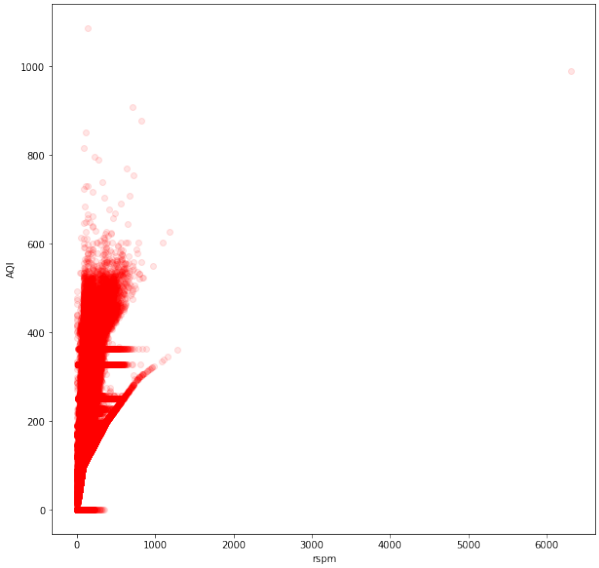


The above graph shows the correlation and dependence of the pollutants SO2 pollutant and NO2 pollutant on the target variable AQI. The left graph is plotting AQI v/s SO2 curve and right graph is plotting AQI v/s NO2 curve. We can infer that both of the graphs do not share total linear dependence on the target variable.

- **AQI V/S RSPM**



AQI V/S SPM



The above graph shows the correlation and dependence of the pollutants RSPM pollutant and SPM pollutant on the target variable AQI. The left graph is plotting AQI v/s SO2 curve and right graph is plotting AQI v/s NO2 curve. We can infer that both of the graphs do not share total linear dependence on the target variable.

ANALYSIS OF POLLUTANT CONCENTRATION WITH RESPECT TO TYPE

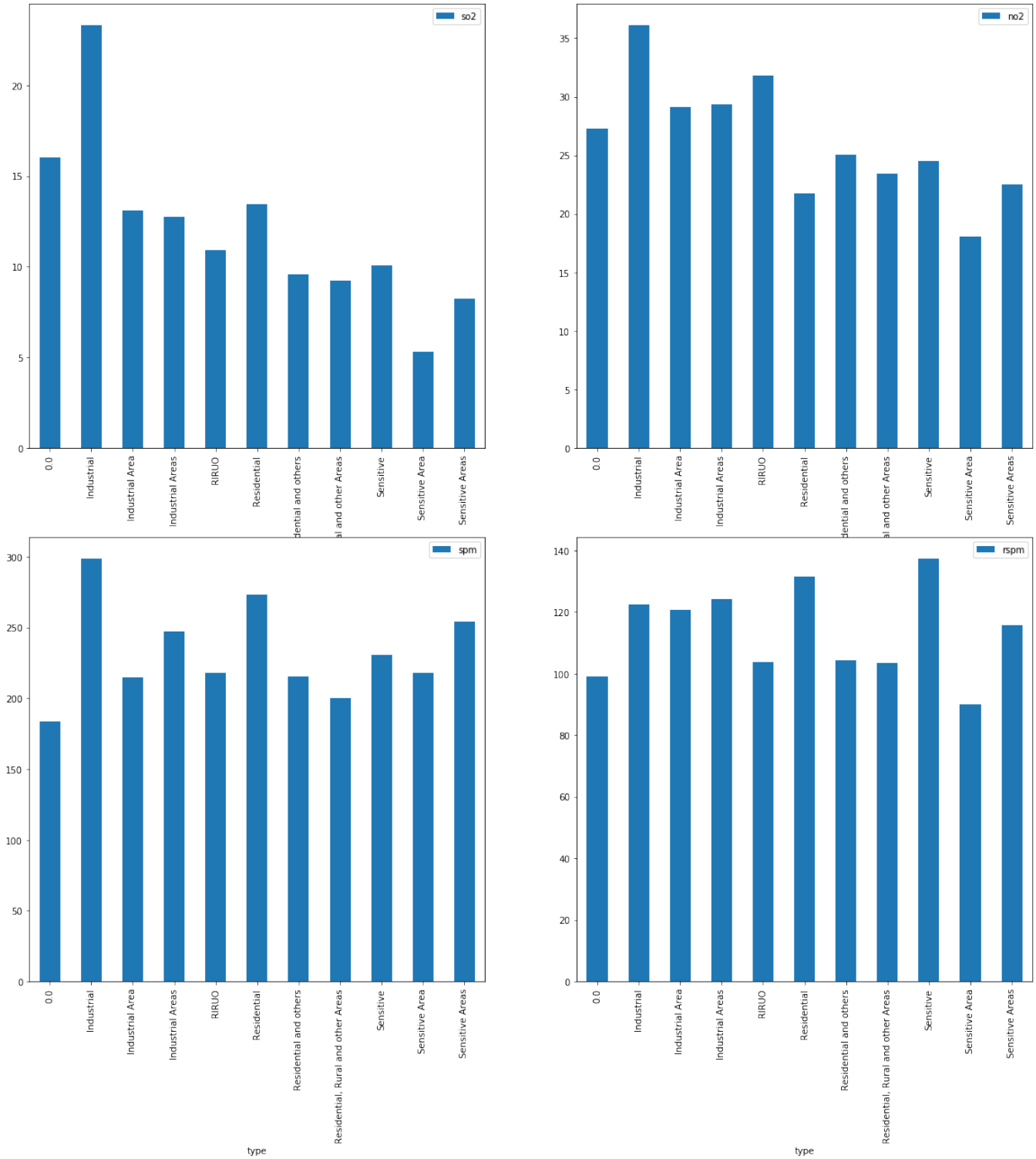


Figure 1: analysis of pollutant concentration with respect to type of the pollutant

MLE with and without intercept analysis

```
(348593, 4)
 MLE WITHOUT INTERCEPT RMSE:22.53
```

Figure 2: Snippets from code

```
(348593, 4) (5, 1)
(87149, 1) (87149, 1)
r2_Square:0.95
MLE WITH INTERCEPT RMSE:20.23
```

Figure 3: Snippets from code

We observe from the above figures that RMSE from MLE without intercept is more than RMSE calculated from MLE with intercept. It is due to the fact that MLE without intercept just considers the set of straight lines passing through origin and do not generalize to all straight line set (y = mx+c).

MLE with and without intercept analysis

```
 MLE MSE:36.11
 MAP MSE:31.50
[33]: Text(0,0.5,'MLE')
```

Figure 4: Snippets from code

We observe that RMSE MAP is less than RMSE MLE because of the fact that MAP reduces the loss by preventing the model to overfit. It acts similar to regularization and helps in penalizing the parameters that overfits the data and reduce the testing loss.

MLE with regularization

```
REGULARIZED MLE MSE:33.22
```

Figure 5: Snippets from code

We observe that MLE RMSE without regularization is 36.11 whereas, after regularization it is 33.22. Hence, It helps to prevent overfitting of the model by reducing the loss / testing error.

Training and Testing RMSE error plotted against the degree of polynomial

Below is the figure that depicts nature of training and testing RMSE (Root Mean Square error) with respect to degree of polynomial in polynomial regression. We observe that training error decreases with an increase in the degree of polynomial whereas, testing error first decreases and then rises as the degree of polynomial increases. It can be inferred that that due to rise in the degree of polynomial, the model tends to overfit and hence, it works well on the training set but not on the unseen data i.e. the test set. Overfitting can be reduced by the use of MAP Estimation or Regularization.

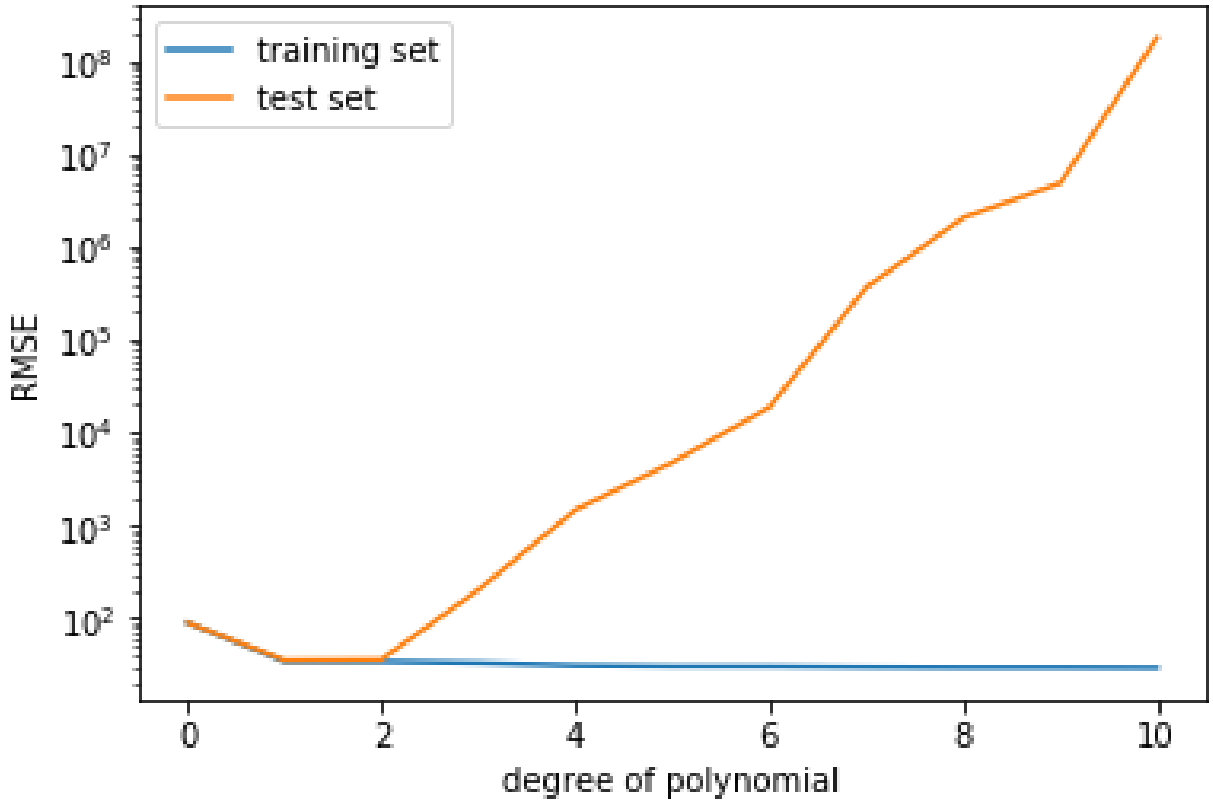


Figure 6: RMSE v/s degree of polynomial

RMSE testing error against the regularization parameter lambda

Below is the figure that plots testing error RMSE with respect to lambda (regularization parameter). It is used to reduce overfitting by penalizing the parameter values by forcing them close to the origin. Hence, it is evident that as lambda increases, overfitting is prevented and due to this, the testing error decreases with the increasing lambda.

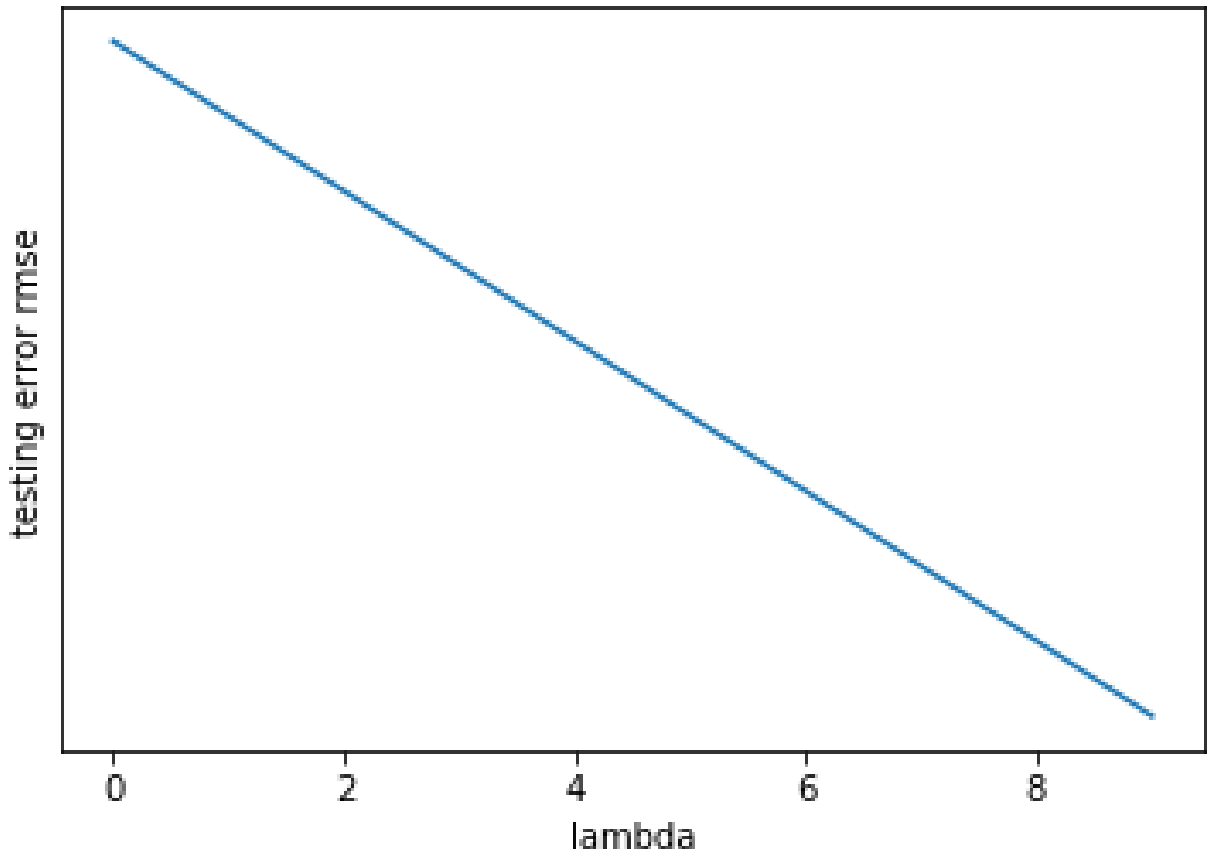


Figure 7: RMSE V/S Lambda(regularization parameter)

LINEAR REGRESSION

Below is the plot of Actual AQI v/s Predicted AQI using linear regression and as we see that the line is not entirely a straight line that depicts y=x, hence, there is significant amount of error present in both, actual and predicted values.

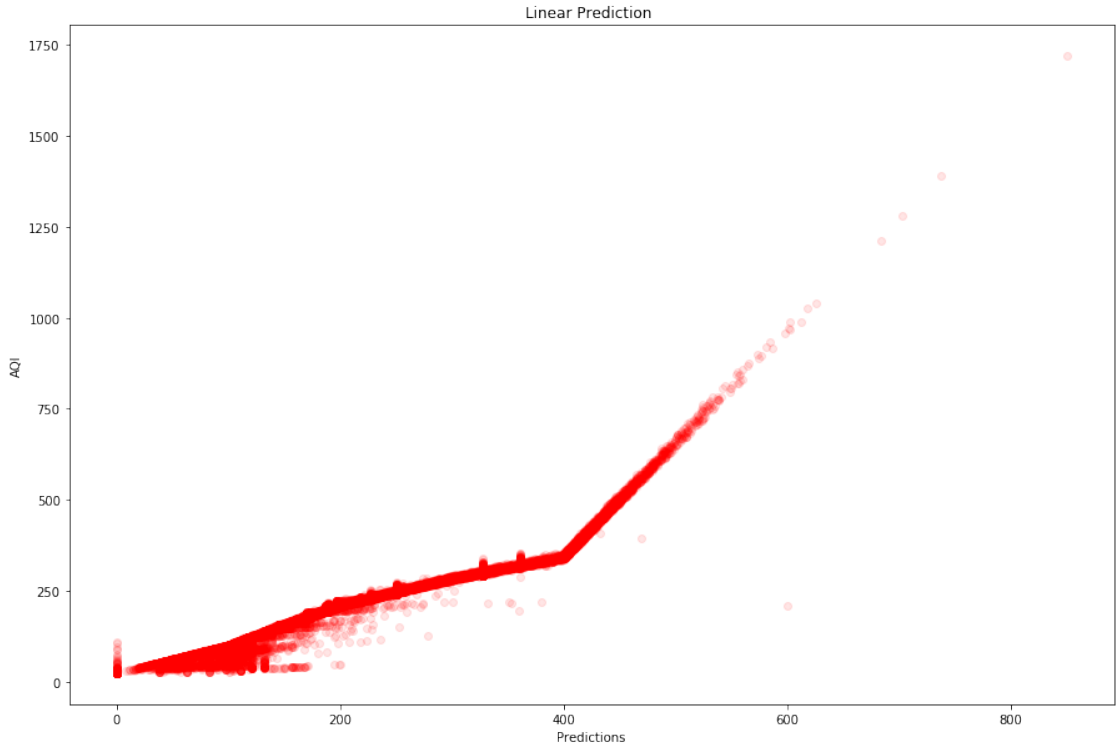


Figure 8: Actual AQI v/s predicted AQI

Below is the plot of Actual AQI, Predicted AQI plotted against each year using linear regression. They are not overlapping and as said in (a), there is significant amount of error in both values.

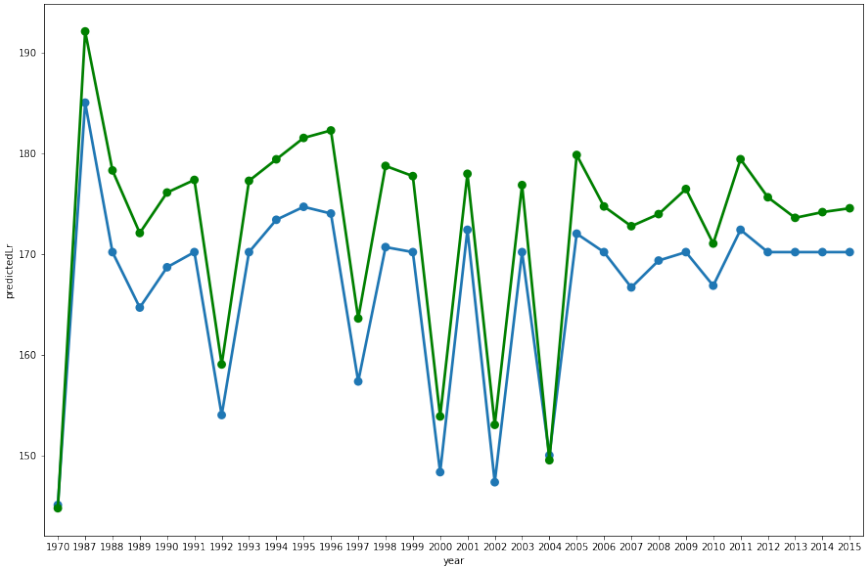


Figure 9: Actual v/s predicted value against each year

RANDOM FOREST REGRESSION

Below is the plot of Actual AQI v/s Predicted AQI using random forest regression and as we see that the line is nearly a straight line that depicts y=x, hence, there will not be much amount of error present in both, actual and predicted values.

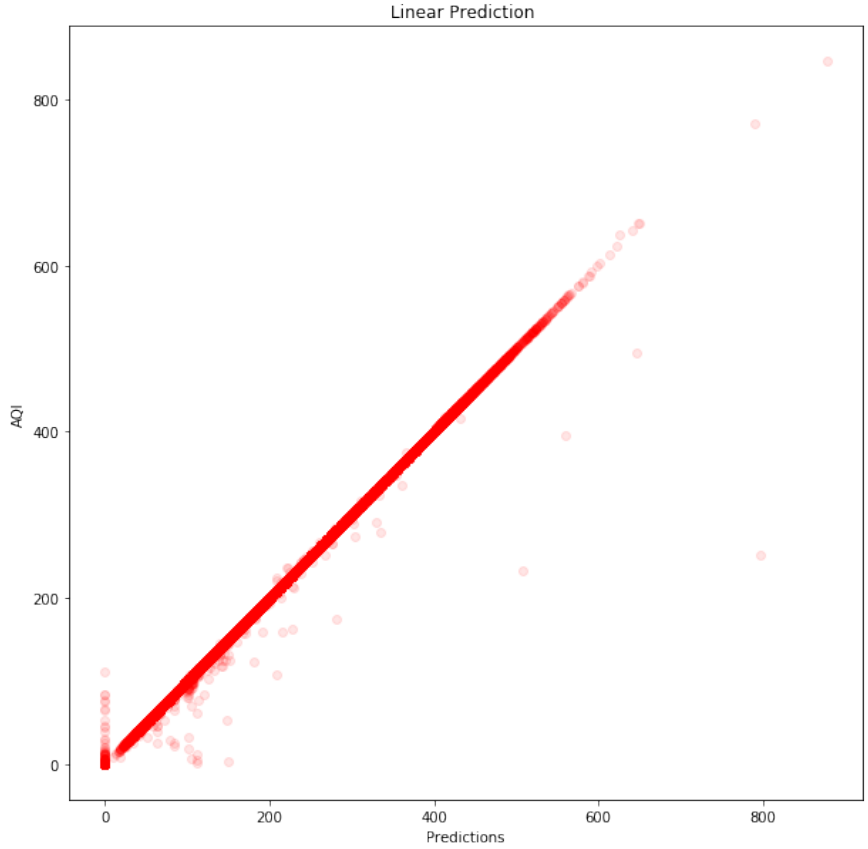


Figure 10: Actual AQI v/s predicted AQI

Below is the plot of Actual AQI, Predicted AQI plotted against each year using random forest regression. As visible, both the curves are overlapping and as said in (a), there is not much amount of error present in both the values. We infer that this model is better than the linear regression model.

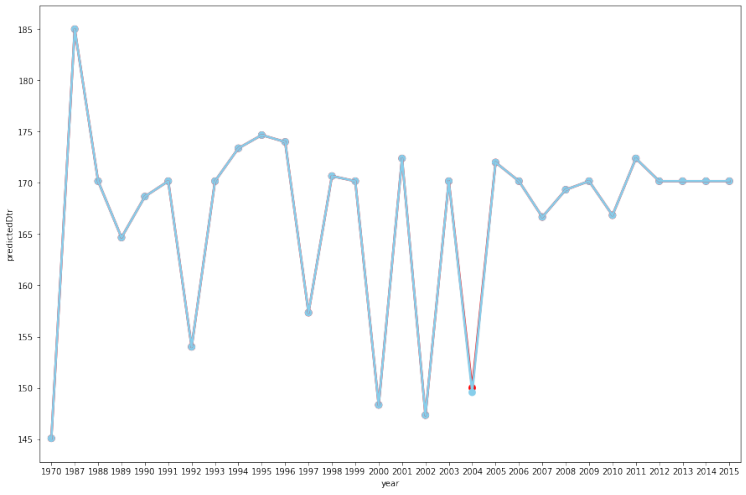


Figure 11: Actual v/s predicted value against each year

DECISION TREE REGRESSION

Below is the plot of Actual AQI v/s Predicted AQI using decision tree regression and as we see that the line is nearly a straight line that depicts y=x, hence, there will not be much amount of error present in both, actual and predicted values.

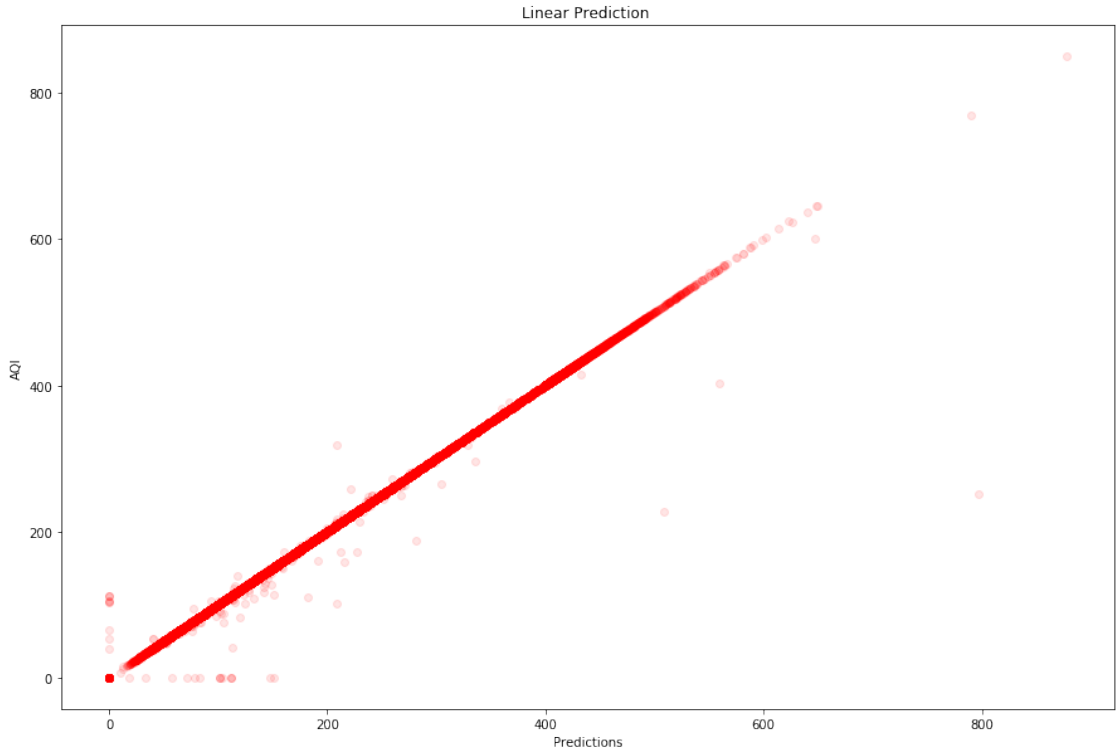


Figure 12: Actual AQI v/s predicted AQI

Below is the plot of Actual AQI, Predicted AQI plotted against each year using random forest regression. As visible, both the curves are overlapping and as said in (a), there is not much amount of error present. Hence, we infer that this model is also better than the linear regression.

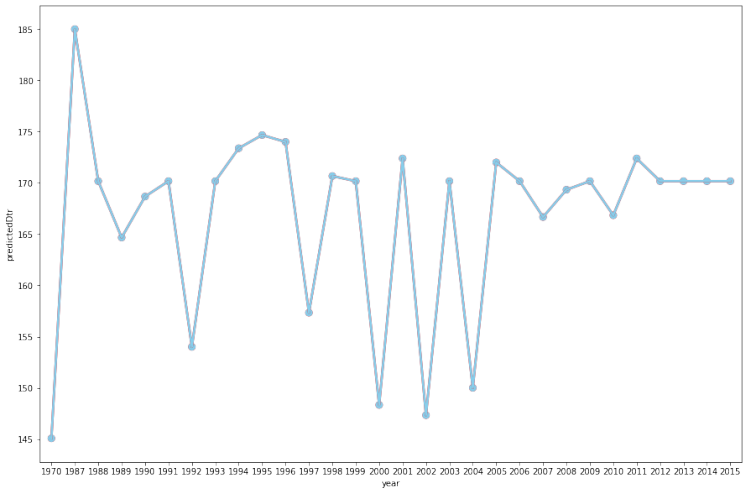


Figure 13: Actual v/s predicted value against each year

GRADIENT BOOSTING REGRESSION

Below is the plot of Actual AQI v/s Predicted AQI using gradient boosting regression and as we see that the line is nearly a straight line that depicts y=x, hence, there will not be much amount of error present in both, actual and predicted values.

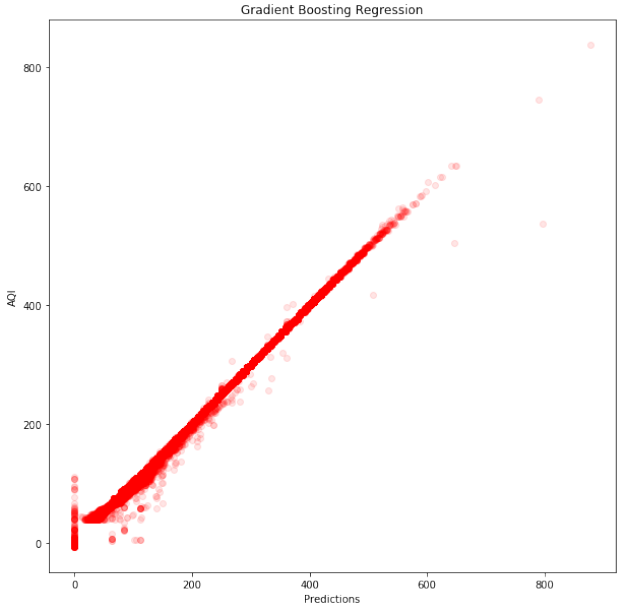


Figure 14: Actual AQI v/s predicted AQI

Below is the plot of Actual AQI, Predicted AQI plotted against each year using random forest regression. As visible, both the curves are not entirely overlapping but have similar values and hence, as said in (a), there is not much amount of error present in both the values. Hence, we can infer that this model is also better than the linear regression model but not better than random forest and decision tree regression.

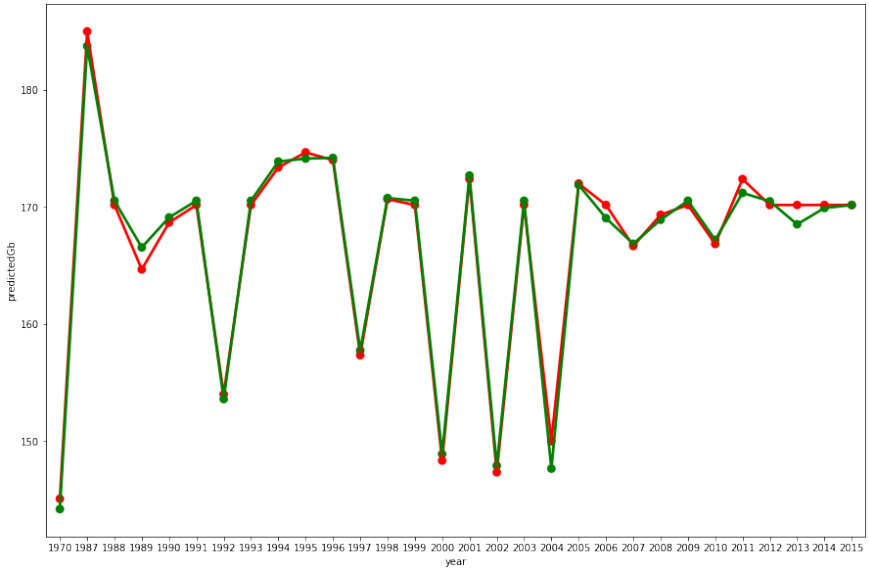


Figure 15: Actual v/s predicted value against each year

RMSE Analysis of the regression techniques Below is the graph of RMSE analysis of all the regression techniques present. Y axis contains all the regression techniques and x axis contains the RMSE error. RF (Random Forest regression) is performing best whereas LR is performing worst.

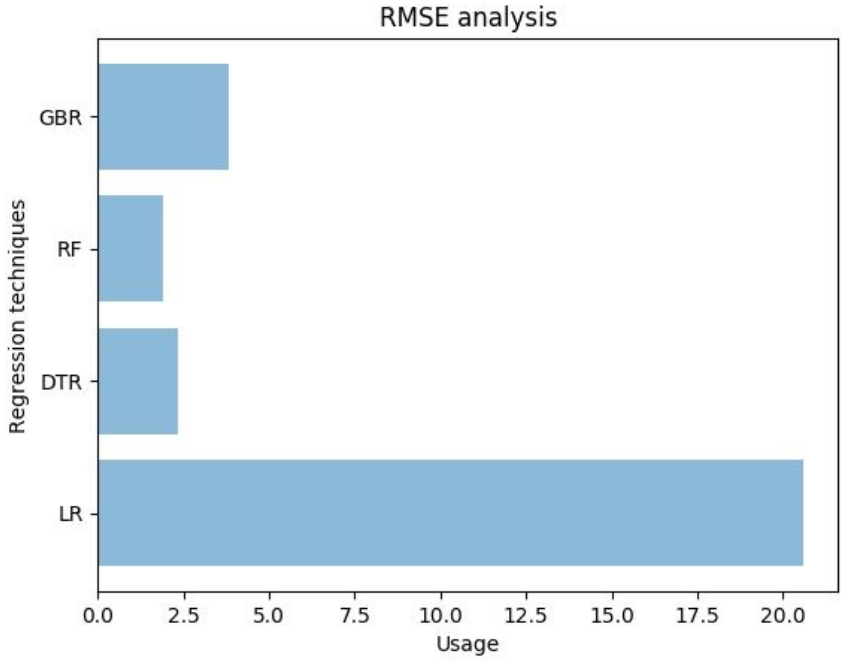


Figure 16: RMSE analysis of various regression techniques

CONCLUSION NOTES:

1. We inferred that MLE with intercept performs better than MLE without intercept (without θ_0)

2. We inferred that RMSE from MAP is lesser than RMSE from MLE because MAP helps in reducing overfitting and hence, reduces the loss and testing error.

3. We inferred that in polynomial regression, as degree of polynomial increases, the training error decreases whereas testing error first decreases a little and then increases. It is because as polynomial degree increases, model becomes more complex, thus giving rise to overfitting which fits the training set but does not work well on the unseen data (test set).

4. We inferred that as we increase the regularization parameter, lambda, the testing set error decreases which indicates that this is preventing overfitting and hence, fitting the model well on the unseen data (test data) as well.

5. We inferred that among the 4 regression techniques applied on our dataset, Random Forest Regression is best suited as it produces the least RMSE and fits the model well, whereas, Linear Regression performs worst among all regression techniques.