Available online at: http://www.ijmtst.com/vol6issue06.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 06, Issue No: 06, June 2020

Analysis & Demonstration of Impact of Air Pollution

Debjyoti Saha[1] | Shashikant Patil[2]

[1,2]EXTC Department SVKMs NMIMS Shirpur Campus

To Cite this Article

Debjyoti Saha and Shashikant Patil, "Analysis & Demonstration of Impact of Air Pollution", *International Journal for Modern Trends in Science and Technology*, Vol. 06, Issue 06, June 2020, pp.:87-91; https://doi.org/10.46501/IJMTST060619

Article Info

Received on 02-May-2020, Revised on 30-May-2020, Accepted on 05-June-2020, Published on 11-June-2020.

ABSTRACT

In this study we have analyzed the impact of air pollution in day to day life in all aspects. The main focus of this contribution is learning about modeling of data by supervised algorithms i.e. (Linear Regression (regression) and Logistic Regression (classification) and its consequences. This particular analysis ofAir Pollution Impact (India & US), and factors that affects AQI. The dataset we have used comprises concentration of pollutants and there is needof each of it for calculating the air quality index, so that is been calculated further in the process and has been utilized in analysis. Here we also seen the combination of the independent variables (Interaction effect) and its impact on dependent variable and the accuracy of the model variation as well as interdependence/ correlation (Multicollinearity) between various independent variable and its adverse effect on the dependent variable and on the given data model. The solution to the problems of multicollinearity is also been discussed in the following kernel i.e. Regularization and Stepwise Regression.

KEYWORDS: *Predictive Analysis, Regression, Algorithms, Data Analysis, Data Handling*

DOI: https://doi.org/10.46501/IJMTST060619

I. INTRODUCTION

As the world is upgrading in terms of technology and resources us the human beings are somehow neglecting the nature's miracle. As a result we are just building a model to destroy ourselves. This paper mainly focuses on that nature's forecast which we generally don't recognize at its best, the air. Air & water are the 2 main resources without which no animals, insects and even human beings can't survive.[1] So this paper fully demonstrates the culture and quality of air in India. As we know the air pollution with the use of factories and vehicles running of petrol & diesel is increasing day by day [2]. So we have analyze the dataset extracted from "**Kaggle**" & "**GitHub**" of those 2 countries and build a regression model to know how much and how this has increased with the appropriate cause[3][4].

II. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods [5]. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task[6][7][8].As this becomes the first step of analysis, we have imported some of the valuable libraries to support our program & the task of study[9].

```
In [3]: import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.preprocessing import Imputer
import matplotlib.pyplot as plt
%matplotlib inline
plt.rcParams['figure.figsize'] = (10, 7)
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_log_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.feature_selection import RFE
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
import statsmodels.formula.api as sm
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from statsmodels.regression.linear_model import OLS
from statsmodels.tools import add_constant
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn import metrics
from statsmodels.stats.outliers_influence import variance_inflation_factor
import warnings; warnings.simplefilter('ignore')
```

Fig 1: Libraries Imported

These libraries help us throughout the analysis and demonstration. Also these helps us to understand the model generated by regression [10]. Extracting the dataset we have generated a visualization bar graph for proper understanding of AQI in different states of India. The graph shows us the normal AQI of Indian States. It includes mostly all the reputed cities in the country India. The best way to study this particular graph is higher the value, more polluted the city is[11].

The scale of this normal AQI city graph is up to 200+, which is not a good indication for the environment. The disease related to respiratory and also relevant to some skin allergies are more likely to occur in those regions [12][13]. Thus, we will predict the index measure with the help of the ranges given by the respective governments of India and United States [14-22].

As we have analyze the graph of cities, now we will have a look to the state wise AQI in the country.

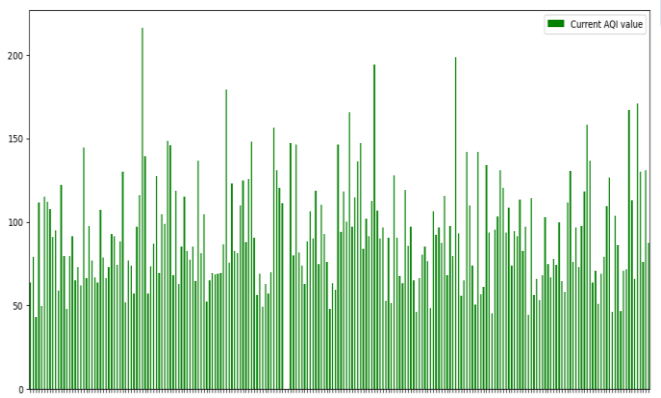

Fig 2: AQI Values Ranges

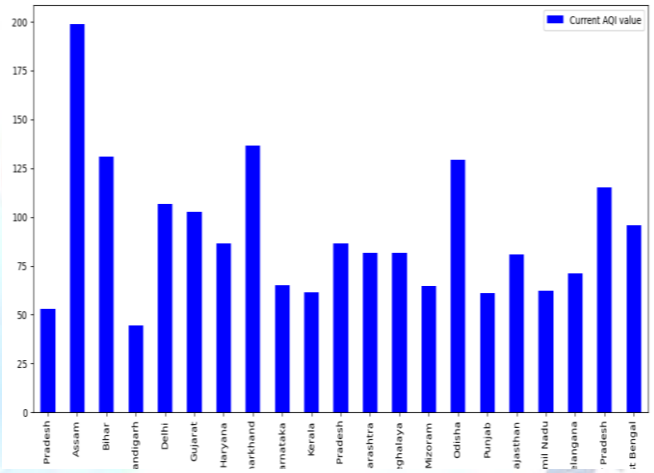

Fig 3: AQI Values Ranges (States)

III. METHODOLOGY

Data Cleaning & Aqi Values

As the dataset when received from the respective files, contained several null values in columns and rows. So proceeding with these null values will generate some disturbance and uneven in the model in our results.

The best to normalize the handling of null values is by using a mathematical function and chapter called as central tendency[23-26]. We extracted the column data and row data and calculated the mean value of that particular function present in the dataset and filled those null spaces with that mean values. This is the process of null value handling or data handling.

	City	State \
0	Amaravati	Andhra Pradesh
1	Rajamahendravaram	Andhra Pradesh
2	Tirupati	Andhra Pradesh
3	Visakhapatnam	Andhra Pradesh
4	Guwahati	Assam
5	Gaya	Bihar
6	Gaya	Bihar
7	Haripur	Bihar
8	Muzaffarpur	Bihar
9	Patna	Bihar

Fig 4: Null Values

These are cities with their respective states which are having null values. Handling the null values becomes very important because sometimes they behave abnormally and disrupt the visualization.

Current AQI value	
State	
Andhra Pradesh	53.150538
Assam	198.760000
Bihar	130.773684
Chandigarh	44.640000
Delhi	106.601542
Gujarat	102.522727
Haryana	86.364929
Jharkhand	136.600000
Karnataka	64.952941
Kerala	61.555556
Madhya Pradesh	86.573407
Maharashtra	81.536957
Meghalaya	81.833333
Mizoram	64.550000
Odisha	129.321429
Punjab	61.230337
Rajasthan	80.690678
Tamil Nadu	62.422222
Telangana	71.360825
Uttar Pradesh	115.253385
West Bengal	95.607692

Fig 5: State AQI Values

ALGORITHMS & MODELS

LINEAR REGRESSION

Linear regression is basically a linear approach to model the relationship shared between a scalar response (or dependent variable) i.e. AQI and one or more explanatory variables

Model-1

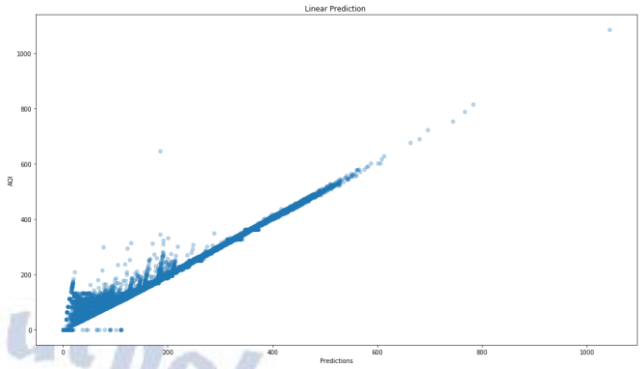

Fig 8 : LR Model-1

Model-2

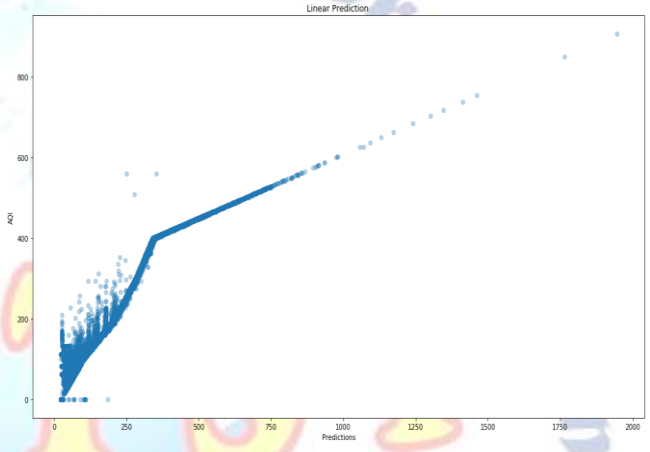

Fig 6: LR Model-2

Model-3

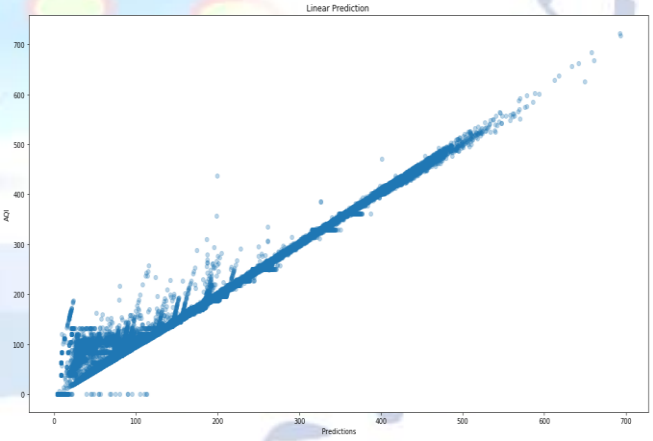

Fig 7: LR Model-3

LOGISTIC REGRESSION

Model-1

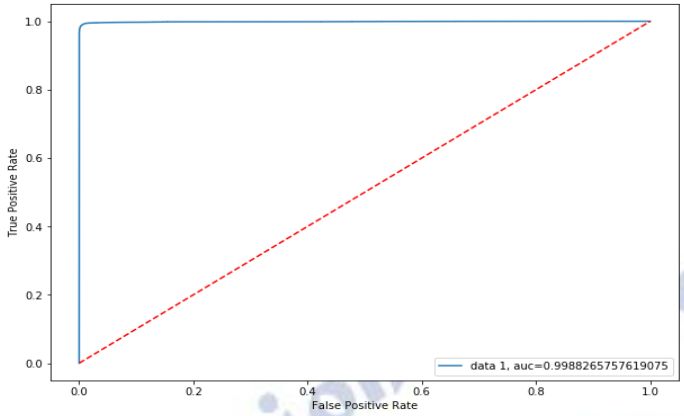

Fig 8: Logistic Model-1

Model-2

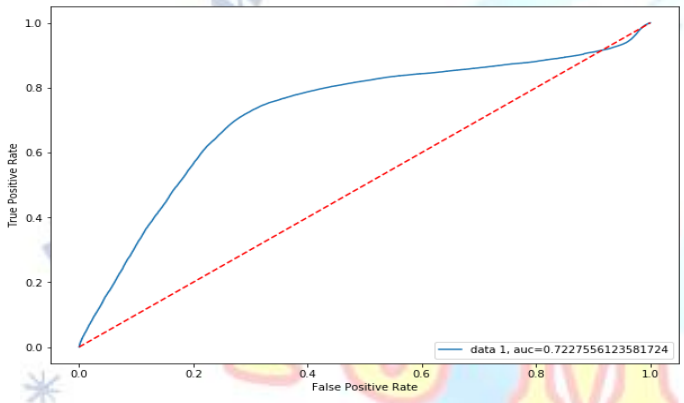

Fig 9: Logistic Model- 2

Model-3

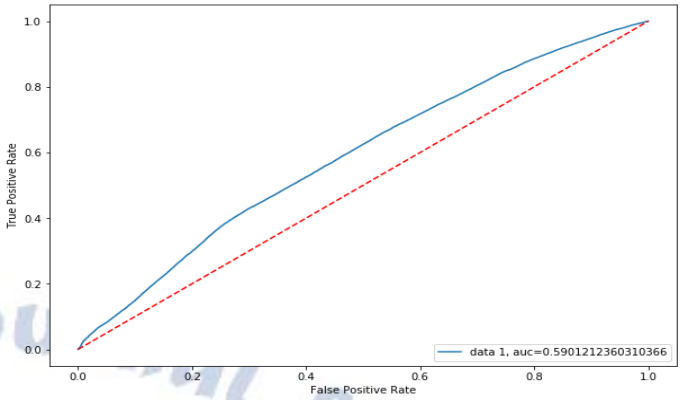

Fig 10: Logistic Model-3

IV. RESULTS & CONCLUSIONS

The result table is the appropriate summary table of the above models. This will tell us the model will bring the most accurate predictions and cure for this pollution. Here we have discussed and addressed various issues related to air pollution and air quality and the usage of various statistical model for deciding suitable model and selection of model as well as its effect along with variation in different parameters. In future we will use neural networks and machine learning approach for advancement in studies and betterment in outcome

OLS Regression Results

Dep. Variable:	AQI	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.998
Method:	Least Squares	F-statistic:	3.605e+07
Date:	Sat, 02 Mar 2019	Prob (F-statistic):	0.00
Time:	04:43:16	Log-Likelihood:	1.2978e+06
No. Observations:	348588	AIC:	2.596e+06
Df Residuals:	348584	BIC:	2.596e+06
Df Model:	4		
Covariance Type:	nonrobust		
Omnibus:	472975.883	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24702[3]613.534
Skew:	7.551	Prob(JB):	0.00
Kurtosis:	132.536	Cond. No.	24.7

Fig 11: Least Square Results

Model Summary

Results: Logit

Model:	Logit	Pseudo R-squared:	0.017
Dependent Variable:	type_label	AIC:	583668.1702
Date:	2019-03-02 04:45	BIC:	583778.0181
No. Observations:	435735	Log-Likelihood:	-2.9182e+05
Df Model:	9	LL-Null:	-2.9687e+05
Df Residuals:	435725	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	5.0000		

Fig 12: Logit Model Results

REFERENCES

[1] Ahad N, Qadir J, Ahsan N. Neural networks in wireless networks: techniques, applications and guidelines. J Network ComputAppl 2016;68:1–27.

[2] Anitescu C, Atroshchenko E, Alajlan N, Rabczuk T. Artificial neural network methods for the solution of second order boundary value problems. CMC: Comput Mater Continua 2019;59(1):345–59.

[3] Antanasijevi D, Pocajt V, Peri-Gruji A, Risti M. Urban population exposure to tropospheric ozone: a multi-country forecasting of SOMO35 using artificial neural networks. Environ Pollut 2019;244:288–94.

[4] Athanasopoulos G, Hyndman RJ, Kourentzes N, Petropoulos F. Forecasting with temporal hierarchies. Eur J Oper Res 2017;262(1):60–74.

[5] Atsalakis GS, Atsalaki IG, Pasiouras F, Zopounidis C. Bitcoin price forecasting with neuro-fuzzy techniques. Eur J Oper Res 2019.

[6] Belytschko T, Lu YY, Gu L. Crack propagation by element-free Galerkin methods. EngFractMech 1995;51(2):295–315.

[7] Bittencourt TN, Wawrzynek PA, Ingraffea AR, Sousa JL. Quasi-automatic simulation of crack propagation for 2d LEFM problems. EngFractMech 1996;55(2):321–34.

[8] Blanc SM, Setzer T. Analytical debiasing of corporate cash flow forecasts. Eur J Oper Res 2015;243(3):1004–15.

[9] Cha Y-J, Choi W, Büyüköztürk O. Deep learning-based crack damage detection using convolutional neural networks. Comput-Aided Civil Infrastructure Eng 2017;32(5):361–78.

[10] Chang C, Mear ME. A boundary element method for two dimensional linear elastic fracture analysis. Int J Fract 1996;74(3):219–51.

[11] Chaphalkar NB, Iyer KC, Patil SK. Prediction of outcome of construction dispute claims using multilayer perceptron neural network model. Int J Project Manage 2015;33(8):1827–35.

[12] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 2011;12:2121–59.

[13] Duflot M, Nguyen-Dang H. A meshless method with enriched weight functions for fatigue crack growth. Int J Numer Meth Eng 2004;59(14):1945–61.

[14] Erdogan F, Sih GC. On the crack extension in plates under plane loading and transverse shear. J Basic Eng 1963;85(4):519–25.

[15] Fakhri M, ShahniDezfoulian R. Pavement structural evaluation based on roughness and surface distress survey using neural network model. Constr Build Mater 2019;204:768–80.

[16] Fan Y, Li W, Gatebe CK, Jamet C, Zibordi G, Schroeder T, et al. Atmospheric correction over coastal waters using multilayer neural networks. Remote Sens Environ 2017;199:218–40.

[17] Fan Z, Wu Y, Lu J, Li W. Automatic Pavement Crack Detection Based on Structured Prediction with the Convolutional Neural Network; 2018. arXiv:1802.02208 [cs]. arXiv: <1802.02208>.

[18] Fathi E, MalekiShoja B. Chapter 9 - deep neural networks for natural language processing. In: Gudivada VN, Rao CR, editors. Handbook of statistics. Computational analysis and understanding of natural languages: principles, methods and applications, vol. 38. Elsevier; 2018. p. 229–316.

[19] Forman RG, Kearney VE, Engle RM. Numerical analysis of crack propagation in cyclic-loaded structures. J Basic Eng 1967;89(3):459–63.

[20] Ghorashi SS, Valizadeh N, Mohammadi S. Extended isogeometric analysis for simulation of stationary and propagating cracks. Int J Numer Meth Eng 2012;89(9):1069–101.

[21] http://joshlawman.com/metrics-classification-report-breakdown-precision-recall-f1/

[22] https://github.com/nikbearbrown/INFO_6105/tree/master/Week_2

[23] https://www.kaggle.com/anbarivan/indian-air-quality-analysis-prediction-using-ml

[24] https://datascience.stackexchange.com/questions/937/does-scikit-learn-have-forward-selection-stepwise-regression-algorithm

[25] https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

[26] https://www.kaggle.com/marcogdepinto/feature-engineering-eda-data-cleaning-tutorial

[27] https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/