Name: Gaurav Kumar, Student Id-21061492, Code link

A Comparative Analysis of Linear vs. Logistic Regression – on Air Quality Dataset

Introduction:

This report is about comparison and analysis between two supervised machine learning techniques, Linear Regression and logistic regression on, Air Quality Data Set (UCI) for Italy. The primary objective of this analysis is to predict the Surface Ozone level which is a secondary pollutant as it is formed in presence of various other air pollutants such a **CO, NO2, Benzene and other hydro carbons** particles levels provided in the dataset and also analyse the impact of these pollutant on the **AIR QUALITY INDEX (AQI)**. Data set link: Air Quality Data Set (UCI)

Methodology and choice of Target columns:

In our given dataset, for Linear Regression we will use the column (PT08.S5(O3)) ozone level depicts our target column or independent variable while other columns such as **CO(GT), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC), NOx(GT), PT08.S3(NOx), NO2(GT), PT08.S4(NO2)** are to be used as linearly dependent variable or Input. On other hand, for the Logistic Regression, we have used from the same dataset a calculated field **AQI_Range_Binary** where 0 indicates good air quality and 1 indicates Bad air quality. Calculation of **AQI and AQI_Range_Binary** and threshold value for the pollutants for Italy is done using doc in the Google scholar website. Link for the document

- **Linear Regression:** Here the output can be predicted using a best fit line given by formula **y = mx + c**, together with loss function such as **mean squared error (MSE)**. Our accuracy depends on minimizing the MSE as much.so that the predicted value is closest to the best fit line. That is why this is applied on continuous linearly correlated data. Since my output will also lie on/ closest to my best fit line.
- **Logistic Regression:** Here we use **S-curve or sigmoid function (sigmoid = 1 / (1+e^-x)),** which gives probability of an outcome **occurring as '0' or '1'**. The loss function here is **Cross-entropy** which measures the difference between two probability distributions for a given random variable or set of events. So for any value which tends towards (- ∞), outcome is '0' while for very large value i.e. tending towards (+ ∞) the outcome is '1', in a way it can be applied only to binary categorical data.

Exploratory Data Analysis:

1) Generated a pair plot **Fig_1 in References** to visualize correlations among data columns. Selected input columns for linear regression based on the identified linear relationships—choosing those with clear positive or negative trends in the pair plot graphs.
2) Created line graphs for the data columns, where we found linear continuous dependency between input and target column. **Fig_2 in References.**
3) Cumulative Distribution Function (CDF) of AQI Values for Different Classes **Fig_3 in References** shows that column AQI is a good candidate to use for logistic regression, as the value of AQI id clearly divided among 2 classes.

Data pre-processing:

a) **"-200"** was found in many columns, this value was handle by using Forward fill method since we are dealing with time series data.
b) **Missing values** were handled by filling it with Mean value.
c) **StandardScaler** was used to scale the input columns to bring uniformity in data before feeding the data in our model.
d) Divided the data into **train and test set, using train test spilt method.**

Model Performance evaluation:

Methodology	Input Data column	Target Column	Accuracy
Linear regression	'CO(GT)', 'NMHC(GT)', 'C6H6(GT)','NOx(GT)', 'NO2(GT)'	'PT08.S5(O3)'	80%

	'CO(GT)', 'NMHC(GT)', 'C6H6(GT)','NOx(GT)', 'NO2(GT)','PT08.S1(CO)','PT08.S2(NMHC)' ,'PT08.S3(NOx)','PT08.S4(NO2)','**AQI','T'**	'PT08.S5(O3)'	88.02%
Logistic regression	'CO(GT)', 'NMHC(GT)', 'C6H6(GT)','NOx(GT)'	AQI_Range_Binary	76%
	'CO(GT)', 'NMHC(GT)', 'C6H6(GT)','NOx(GT)', **'PT08.S5(O3)','T', 'RH', 'AH'**	AQI_Range_Binary	83%

```
Logistic Analysis Report
              precision    recall  f1-score   support

           0       0.83      0.86      0.85      1284
           1       0.82      0.78      0.80      1056

    accuracy                           0.83      2340
   macro avg       0.83      0.82      0.82      2340
weighted avg       0.83      0.83      0.83      2340

[0 0 0 ... 0 0 0]
Accuracy Score:0.83
Probabilty of dependent variable
0.453073950810539
```

```
MAE: 0.26666451538447505
MSE: 0.11888033733552154
RMSE: 0.3447902802219366
```

Observations and Comparative Discussion on the model Output: For linear regression we are achieving initial accuracy of **80%** and then when we included **'AQI','T'** columns as input to our model the accuracy significantly improved to **88%.** Whereas for logistic regression initial accuracy was **76%** based on our initial assumption of dependent variable but when we added **'PT08.S5(O3)','T', 'RH', 'AH'** data columns as input the model accuracy improved to **83%**.From the above observations we can clearly see that right choice of features for training our model improves the accuracy, for example in **Linear Regression** while calculating **AQI**, we considered concentration for **CO** and **NO2** only, but our model clearly showed a correlation between **Ozone level and AQI and Temperature**, even if initially we ignored the importance of these 2 input data. In **Logistic Regression** we see the same improvement in accuracy after adding **'PT08.S5(O3)','T', 'RH', 'AH'** columns thus giving a clear idea to the users regarding initial considerations of input feature selection.

We can summarize our findings as below about the Linear and logistic regression Methodology.

a) In Linear regression adding of new features increases accuracy, this means we should not **overlook potential non-linear relationship between the target and input variables** as it can improve the performance of model significantly.
b) Linear regression model is **more sensitive** to inclusion of new features, because of its linear nature while for logistic regression seems to be **less sensitive** highlighting that there can be **more** complex **nonlinear relationship** between input data and binary output.
c) Significant improvement in the accuracy score of linear regression, shows that the initial model was having **under fitting** and addition of features solved the issue while logistic regression model was initially close to **optimal fit**.
d) The results also emphasize on choosing enough number of input features especially in case of linear regression, else we can have a case of **overfitting**.
e) Feature selection is an important aspect in model accuracy for both these algorithm, so all the input data should be considered based on **mathematically derived association as well as experimentally derived relationship** which may not be apparent.

Conclusion: Logistic regression and linear regression are both very different solutions for different problems. With logistic regression, we can predict predict the categorical dependent variable while with linear regression we can predict linearly dependent variable, but we can make use of these 2 types of Machine learning algorithms in conjunction and careful considerations about feature selection, model the nature of the relationships captured can help any organizations to work on Air quality improvements for Italy.

References:

1) https://www.kdnuggets.com/2022/07/logistic-regression-work.html

2) https://kandadata.com/comparing-logistic-regression-and-ordinary-least-squares-linear-regression-key-differences-explained/ **3)** https://https://scholar.google.com/scholar **4)** https://www.sciencedirect.com/science/article/pii/S0048969799002429#bBIB3

Fig_1 :

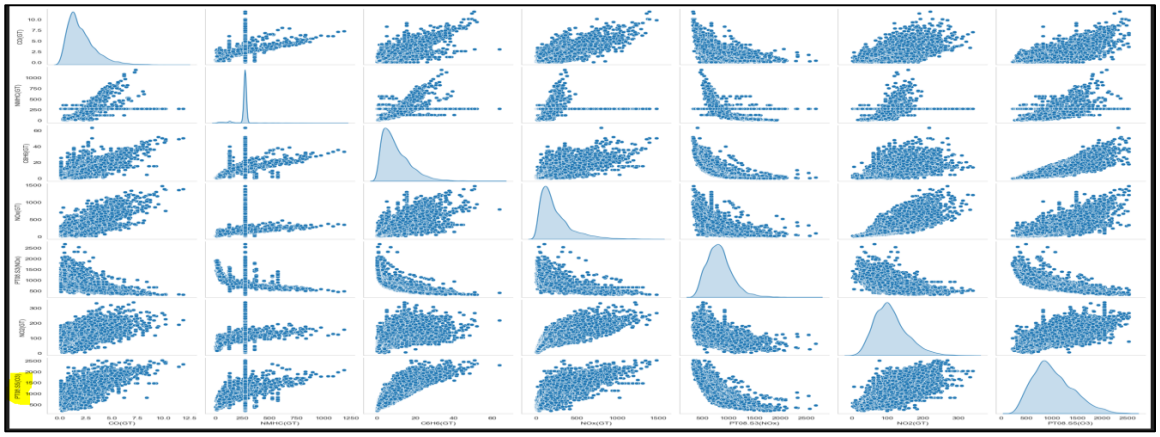

Fig_2:

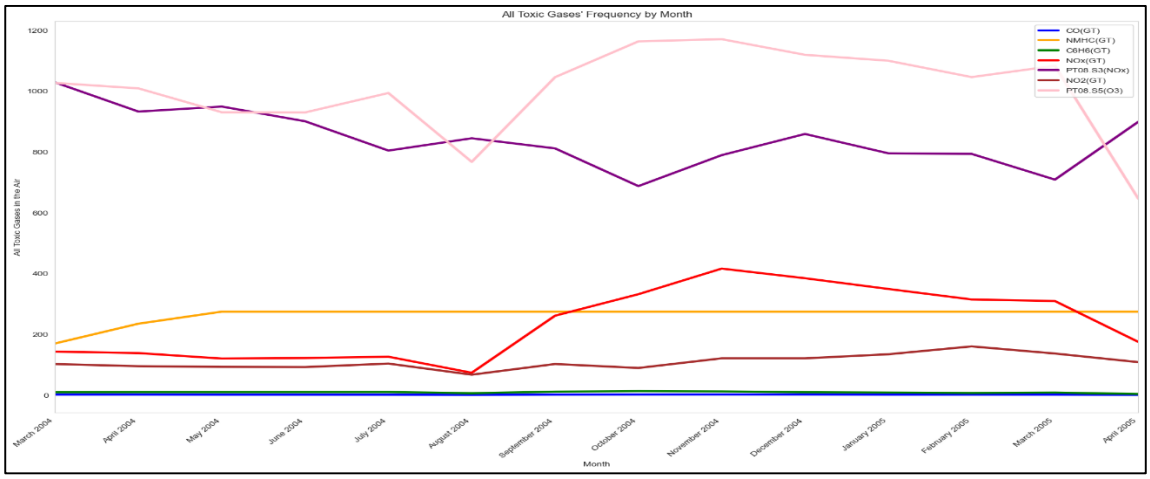

Fig_3