IBM MACHINE LEARNING: REGRESSION PROJECT

Regression of NASA Aerofoil Self Noise Dataset

1. Objective

Regression is used to find relation between features and labels in machine learning. In a simple way to describe, it is a mathematical model which finds out the relation between x and y i.e., finds out the function describing the relation between x and y. Regression is one of the main machine learning algorithms used in the world.

In this project, we use the NASA Airfoil Self Noise dataset, where we are going to find the relation between Noise of the Airfoil obtained when the Airfoil is subjugated to different angle of attack, free stream velocity, chord length and other parameters.

2. Data:

The NASA data set comprises different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack. The span of the airfoil and the observer position were the same in all of the experiments. There are a total of 1503 observations. Following is the descriptions of the data features and labels.

- Input features:

    1. **f**: Frequency in Hertz [Hz].
    2. **alpha**: Angle of attack (AoA, α), in degrees [°].
    3. **c**: Chord length, in meters [m].
    4. **U_infinity**: Free-stream velocity, in meters per second [m/s].
    5. **delta**: Suction side displacement thickness (δ), in meters [m].

- Output:
    1. **SSPL**: Scaled sound pressure level, in decibels [dB].

Figure 1. Wikipedia image of Flow over a Airfoil during testing.

2.1 Data Exploration:

A good way to explore the data is to use histograms and see the distributions of features. Following are the histograms of features.

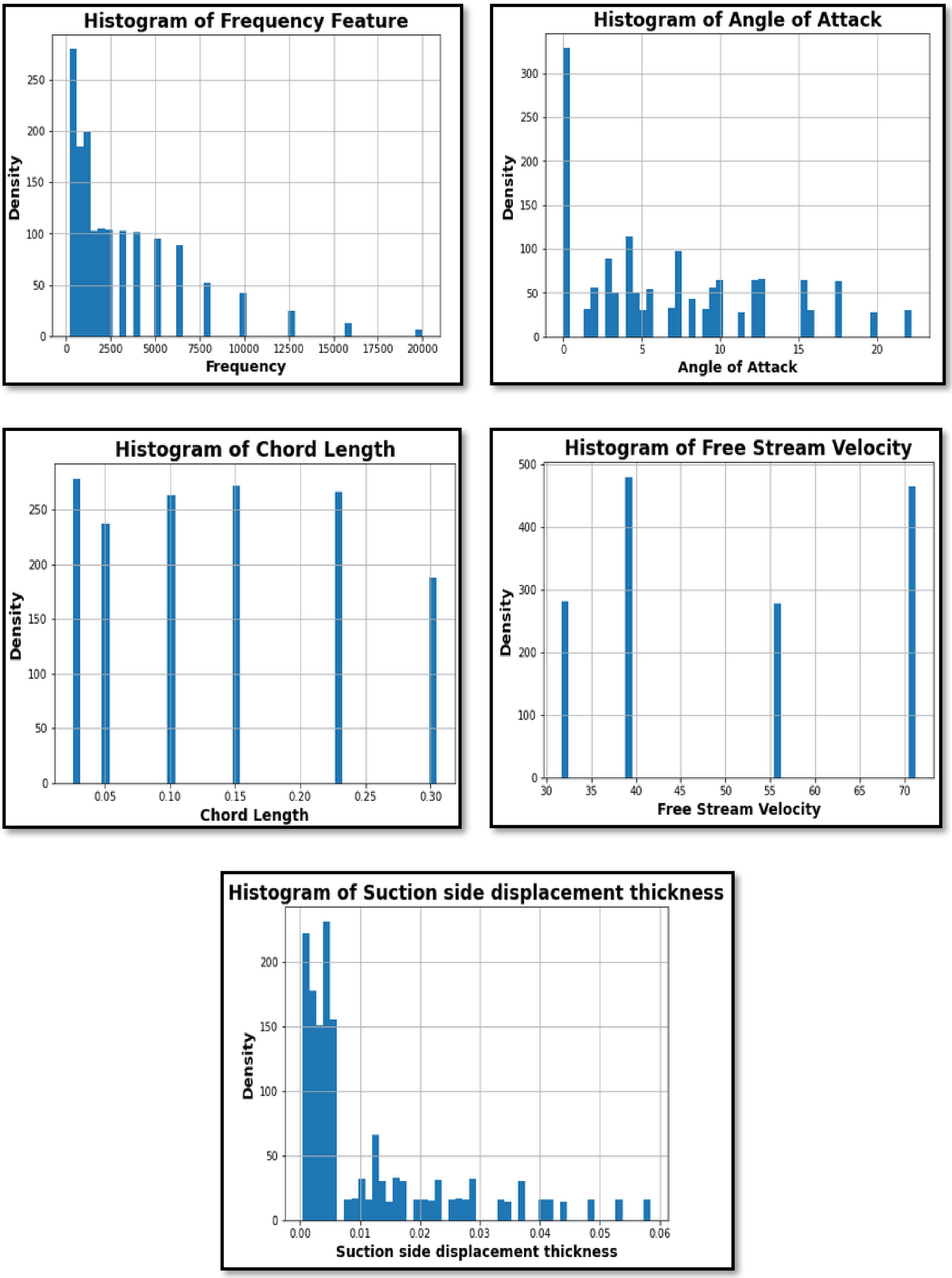


Figure 2. Histograms of Features.

3. Exploratory Data Analysis

One of the assumptions in regression is that the residuals obtained are assumed to be normally distributed. Statistically this means that the mean is centred around zero and standard deviation is one (unit variance). The distribution is in the shape of a bell curve.

3.1 Hypothesis:

Null Hypothesis: The target values are normally distributed.

Alternate Hypothesis: The target values are not normally distributed.

Statisticians say if $p > 0.05$, then the dataset is normally distributed. If observation is not normally distributed, so we reject the Null Hypothesis.

3.2 Determine Normality

1. If the target variable is Normally distributed, it obtains better results.

2. If our target is not normally distributed, we can apply a transformation to it and then fit our regression to predict the transformed values

3. To see if target values are normally distributed, check:

a. Visually

b. Statistically

1. Visual Test using Histogram of the Label Feature of SSPL Noise:

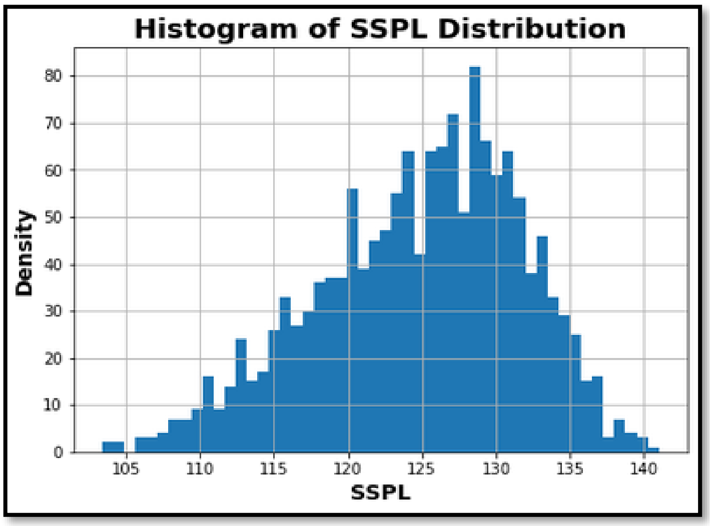
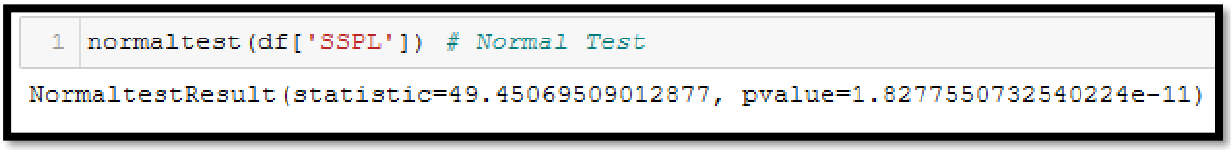


Figure 3. Histogram of SSPL Output Label.

Visually inspecting we can say that the histogram is not normally distributed and is skewed towards the left.

2. Statistic Test using p-value:

A normal test is conducted on the ‘SSPL’ label and the p-value obtained is less than 0.05.



```
normaltest(df['SSPL']) # Normal Test
NormaltestResult(statistic=49.45069509012877, pvalue=1.8277550732540224e-11)
```

Figure 4. P-value of 'SSPL'.

Thus, the desired output ‘SSPL’ is not normally distributed.

3.3 Transformations

Transformation of the output label can make the it normally distributed. Some transformation which can make a non-normally distributed to a normally distributed are:

a. Log Transformation

b. Square Root Transformation

c. Box-Cox Transformation

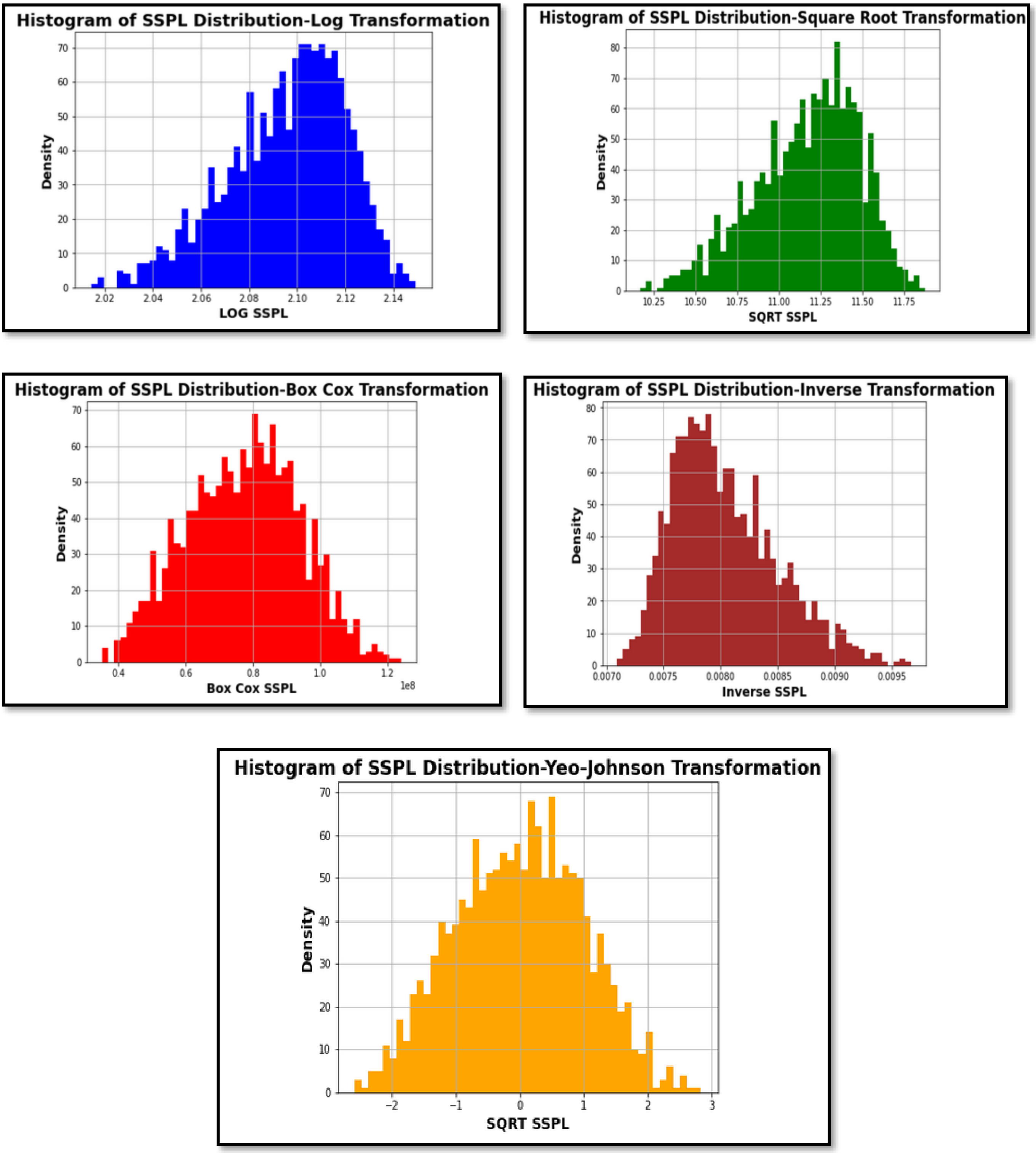
d. Inverse Transformation

e. Yeo-Johnson Transformation

We try all the above transformation and conduct a normal test after transformation. Following table describes the p-value obtained during different transformations

Serial No	Transformation	p-value
1	Log	1.721815737185587e-15
2	Square Root	3.5078566279140157e-13
3	Box-Cox	1.7951647448346226e-08
4	Inverse	5.301185393463406e-22
5	Yeo-Johnson	1.63207892e-07

It looks like all the transformation were not able to obtain a normal distribution, with Yeo-Johnson Transformation performing the best since p-value is close to 0.05 when compared with others. The transformations are also visually inspected using histograms.



Histogram of SSPL Distribution-Log Transformation

Histogram of SSPL Distribution-Square Root Transformation

Histogram of SSPL Distribution-Box Cox Transformation

Histogram of SSPL Distribution-Inverse Transformation

Histogram of SSPL Distribution-Yeo-Johnson Transformation

Figure 5. Histograms of "SSPL" after various transformations.

Since none of the transformation were able to make the output label normally distributed, we reject the null hypothesis that the targe values are normally distributed and go ahead with the alternate hypothesis without any transformations.

4. Methodology

In this project, we will use three different type of regression models. Given below are the details of the models:

1. Simple Multiple Linear Regression with 3-split-cross validation.

2. Lasso Regression with polynomial degree 2 with 3-split-cross validation.

3. Ridge Regression with polynomial degree 3 with 3-split-cross validation.

5. Results and Discussion

1. Simple Multiple Linear Regression:

The dataset is modelled with Simple Multiple Linear regression with 3 split Cross Validation. The data is normalised and then fitted using the regression model. Following are the R2-Score on the 3-test-split data

Test set	R2-Score
1	0.4917
2	0.4753
3	0.5428

Test Set	Coefficient of Frequency	Coefficient of AOA	Coefficient of Chord Length	Coefficient of Free Stream velocity	Coefficient of SSD
1	-4.1277473	-2.780564	-3.39462717	1.47274716	-1.82517
2	-4.08915661	-2.26732	-3.55020604	1.50560166	-2.032543
3	-3.90993173	-2.45077	-3.0517319	1.69212149	-1.93178

2. Lasso Regression with polynomial degree 2 with 3-split-cross validation:

In this regression model we transform the features to a degree 2 polynomial and normalize it using standard scaler. The using pipeline, we fit the model to a Lasso Regression model with 3-split-cross-validation. The hyper parameter alpha is varied from 1e-09 to 1.0 with 25 points in between. Following graph shows which alpha gets the highest R-2 Score.

The alpha with the highest score was:

Alpha(hyperparameter) = 0.03162277660168379

R2- Score = 0.6183693810486752

Following table summarises alphas and its respective scores.

	Alphas	Scores
0	1.00E-09	0.617022
1	2.37E-09	0.617022
2	5.62E-09	0.617022
3	1.33E-08	0.617022
4	3.16E-08	0.617022
5	7.50E-08	0.617022
6	1.78E-07	0.617022
7	4.22E-07	0.617022
8	1.00E-06	0.617022
9	2.37E-06	0.617022
10	5.62E-06	0.617022
11	1.33E-05	0.617023
12	3.16E-05	0.617025
13	7.50E-05	0.617029
14	0.000178	0.61704
15	0.000422	0.617063
16	0.001	0.617113
17	0.002371	0.617196
18	0.005623	0.617262
19	0.013335	0.61786
20	0.031623	0.618369
21	0.074989	0.618327
22	0.177828	0.608829
23	0.421697	0.559904
24	1	0.376013

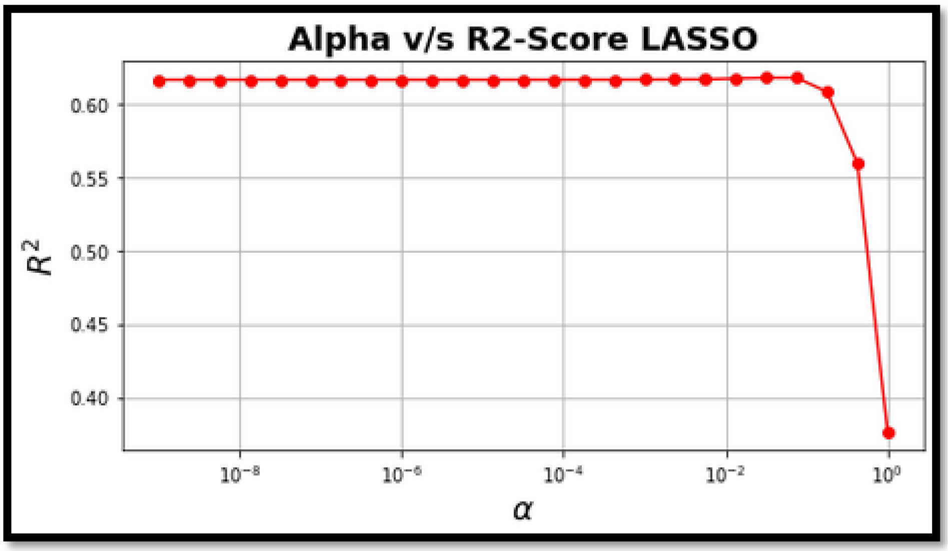


Figure 6. Alpha v/s R2-Score for Lasso Regression with polynomial degree 2.

3. Ridge Regression with polynomial degree 3 with 3-split-cross validation.

In this regression model we transform the features to a degree 3 polynomial and normalize it using standard scaler. The using pipeline, we fit the model to a Riidge Regression model with 3-split-cross-validation. The hyper parameter alpha is varied from 0.001 to 10 with 25 points in between. Following graph shows which alpha gets the highest R-2 Score.

The alpha with the highest score was:

Alpha(hyperparameter) = 2.154434690031882 with R2- Score = 0.7196946904189027

Serial No	Alphas	Scores
0	0.001	0.718974
1	0.001468	0.718977
2	0.002154	0.71898
3	0.003162	0.718985
4	0.004642	0.718993
5	0.006813	0.719003
6	0.01	0.719016
7	0.014678	0.719034
8	0.021544	0.719056
9	0.031623	0.719083
10	0.046416	0.719116
11	0.068129	0.719152
12	0.1	0.719193
13	0.14678	0.719236
14	0.215443	0.719283
15	0.316228	0.719333
16	0.464159	0.719388
17	0.681292	0.719451
18	1	0.719526
19	1.467799	0.719613
20	2.154435	0.719695
21	3.162278	0.719714
22	4.641589	0.71954
23	6.812921	0.718931
24	10	0.717505

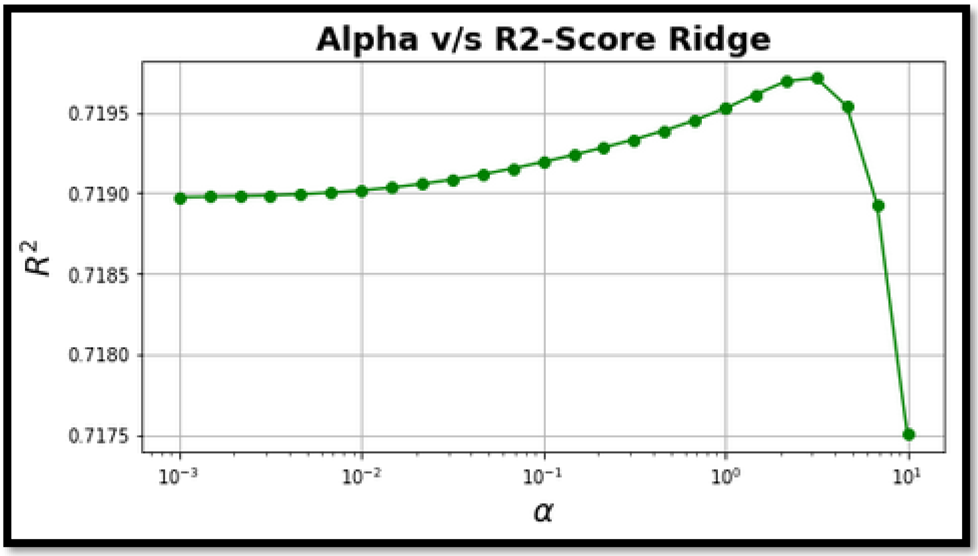


Figure 7 . Alpha v/s R2-Score for Ridge Regression with polynomial degree 3.

6. Conclusion

Three regression models we used to analyse the airfoil dataset. Clearly model 3 i.e., ridge regression with polynomial degree 3 features performs better than the other two models. The models can be further enhanced by using higher degree polynomial features with inclusion of regularization to get better R2-Scores.

7. Reference

1. Data Reference:

- https://archive.ics.uci.edu/ml/datasets/airfoil+self-noise
- https://www.kaggle.com/fedesoriano/airfoil-selfnoise-dataset