Lab Course Machine Learning **Kalaiselvan Panneerselvam**

Exercise 6 **278920**

1 Data Preprocessing

Regression Datasets

μ = 1 and σ = 0.05

Mean = (a+b)/2

SD = [sqrt(1/12)]*(b-a)

b = ([sqrt(12)/2]*SD) + mean

a = 0.9133

b = 1.0866

```
x = np.random.uniform(0.9133, 1.0866, size = (100, 1))
```

```
y=((1.3 *(np.square(x))) + (4.8 * x) + 8)
```

	x	y
0	0.979162	14.554806
1	0.997187	14.905441
2	0.893190	13.948001
3	1.082014	14.892356
4	0.910328	14.038140
5	0.957913	14.280122
6	1.025144	14.834796

Wine Dataset

```
#Winedata preprocessing
winered_data = pd.read_csv('winequality-red.csv', delimiter = ';')
```

- Convert any non-numeric values to numeric values

```
#Normalise wine dataset
for feature in winered_data.columns:
    if feature != "quality":
        winered_data[feature] = normalize(winered_data[feature])
winered_data.head(5)
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	0.247788	0.397260	0.00	0.068493	0.106845	0.140845	0.098940	0.567548	0.606299	0.137725	0.153846	5
1	0.283186	0.520548	0.00	0.116438	0.143573	0.338028	0.215548	0.494126	0.362205	0.209581	0.215385	5
2	0.283186	0.438356	0.04	0.095890	0.133556	0.197183	0.169611	0.508811	0.409449	0.191617	0.215385	5
3	0.584071	0.109589	0.56	0.068493	0.105175	0.225352	0.190813	0.582232	0.330709	0.149701	0.215385	6
4	0.247788	0.397260	0.00	0.068493	0.106845	0.140845	0.098940	0.567548	0.606299	0.137725	0.153846	5

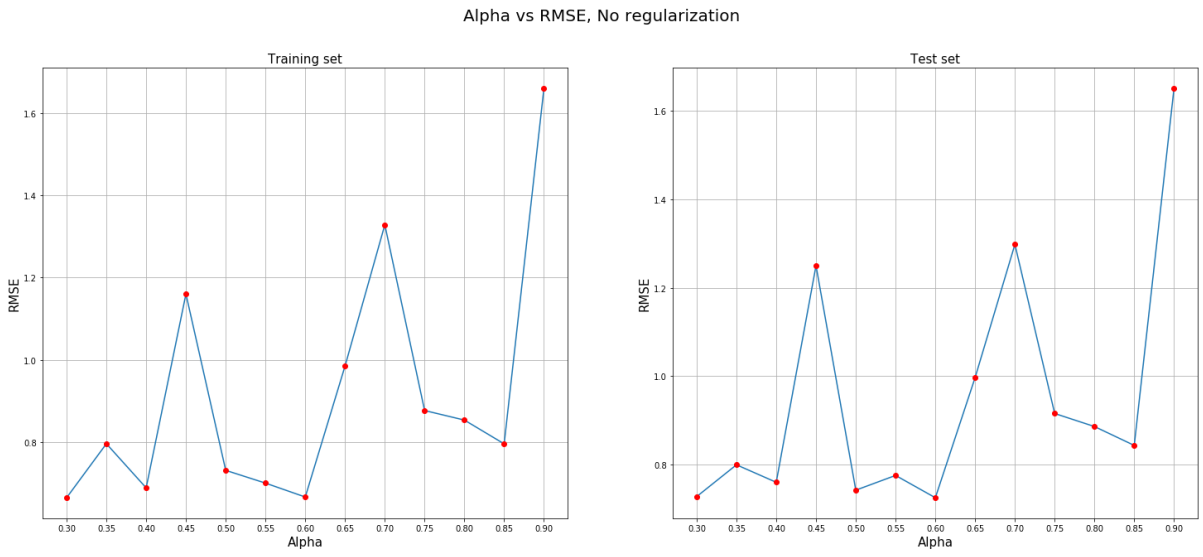
Exercise 1: Generalized Linear Models with Scikit Learn

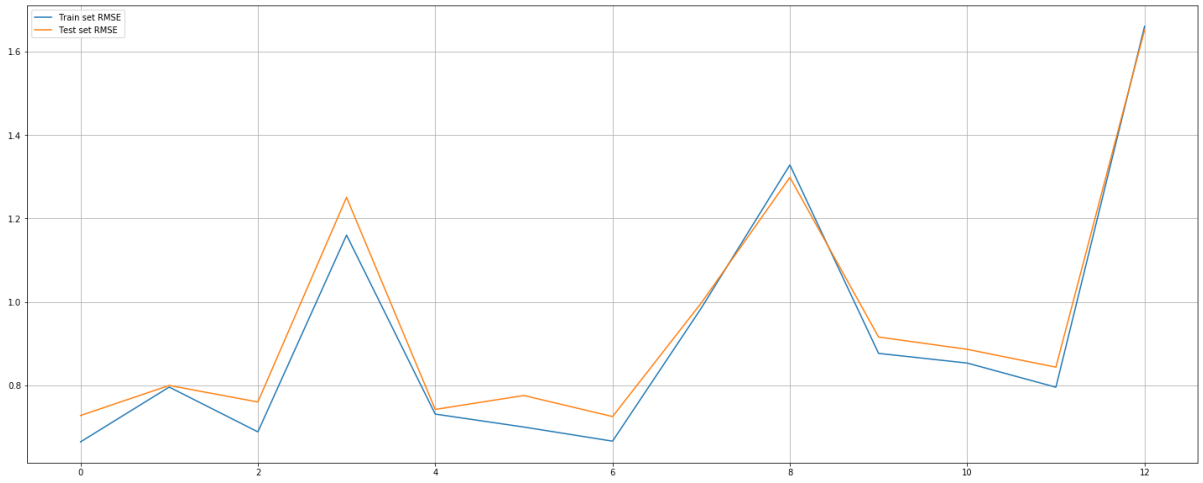
1.Split your data into Train and Test Splits.

```
#Split wine dataset
trainSet1, testSet1 = split_DataSet(winered_data, 0.8)
x_Train1 = trainSet1.as_matrix(columns = ['volatile acidity', 'chlorides', 'density', 'alcohol'])
x_Test1 = testSet1.as_matrix(columns = ['volatile acidity', 'chlorides', 'density', 'alcohol'])
y_Train1 = trainSet1['quality']
y_Test1 = testSet1['quality']
```

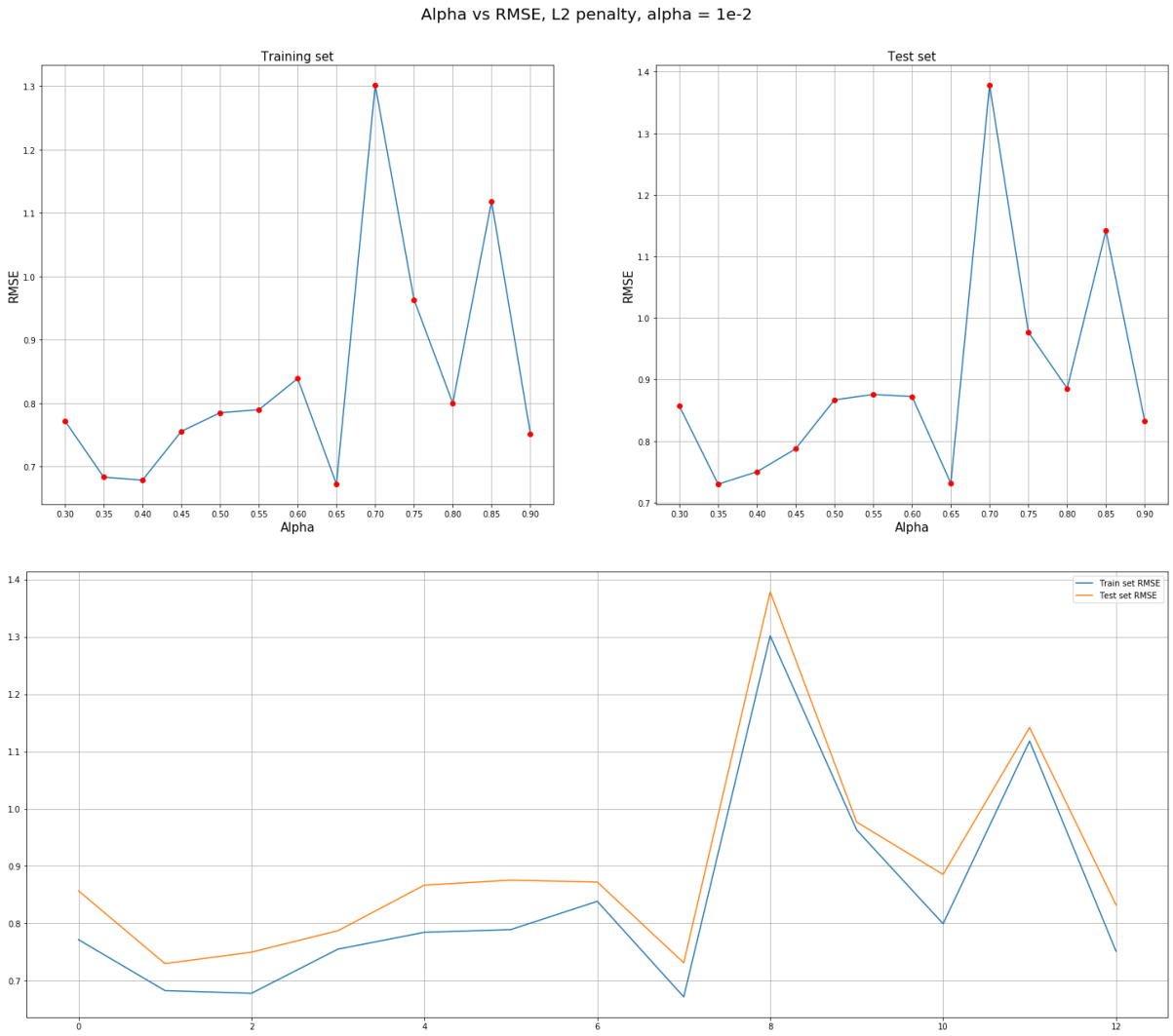
2. Plots for each model, pick three sets of hyperparameters and learn each model without CV

1. Ordinary Least Squares

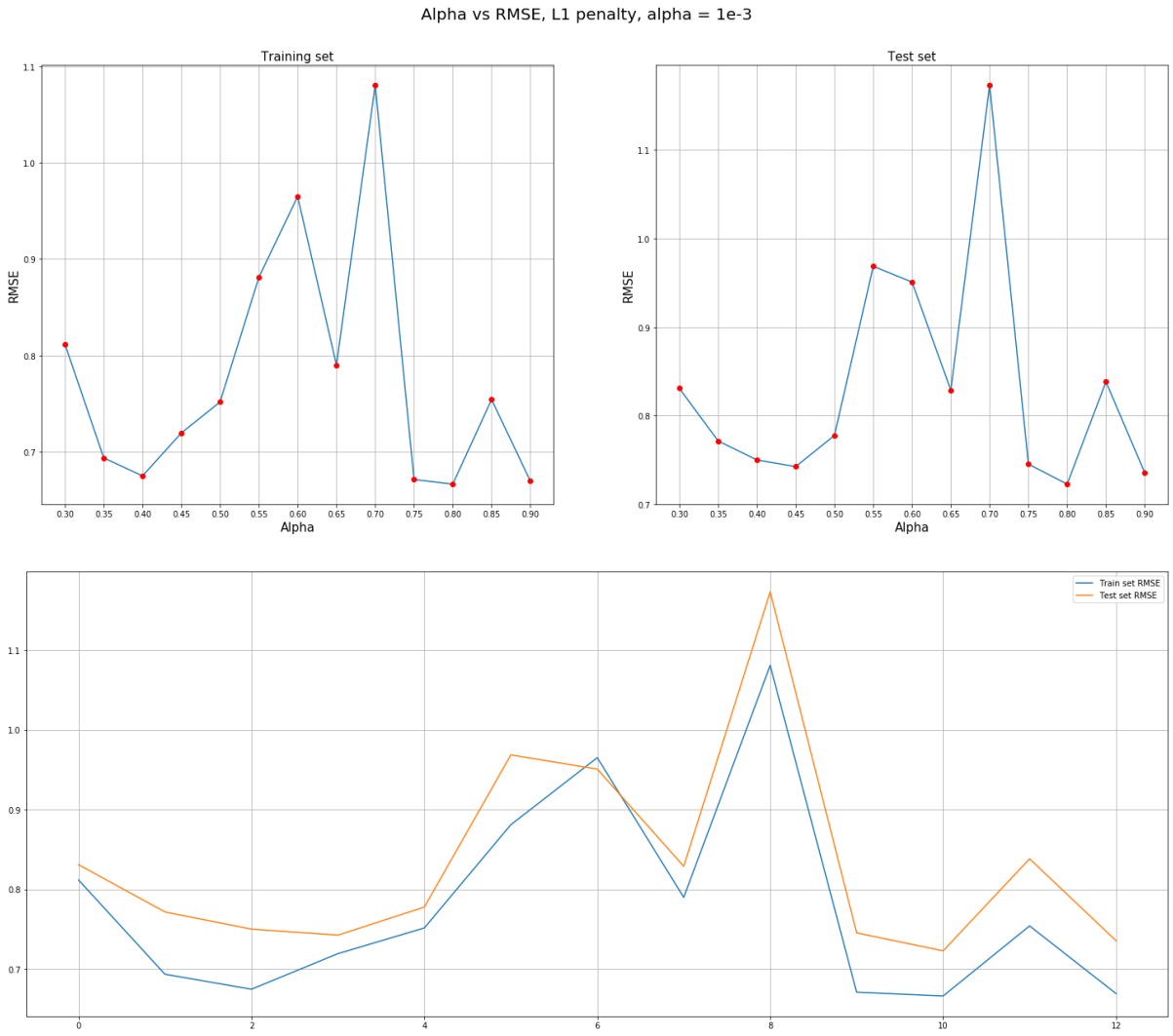




2. Ridge Regression



3. LASSO



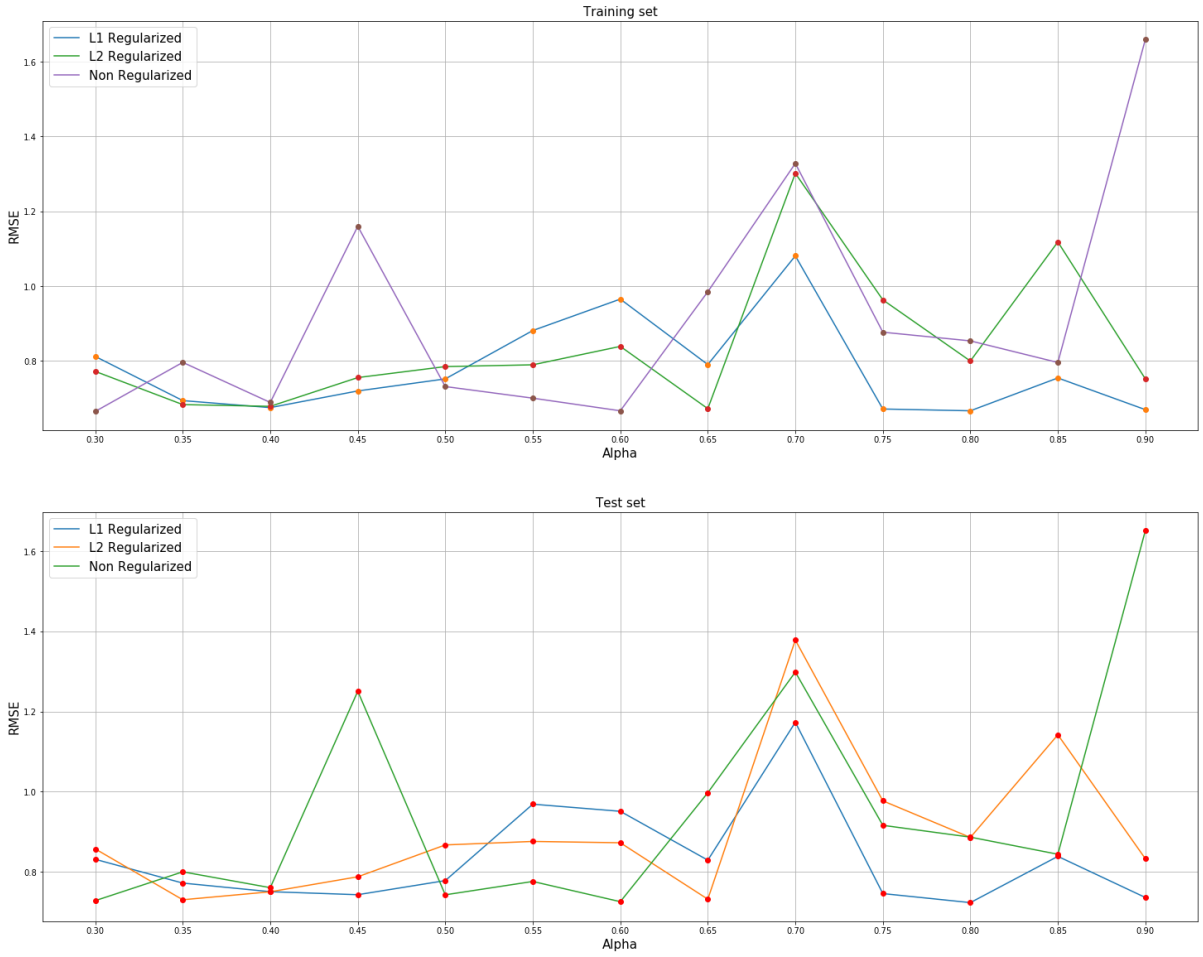
Regularized vs non-regularized models

Overfitting refers to a model that models the training data too well.

Underfitting refers to a model that models too simple to capture the trend in the data.

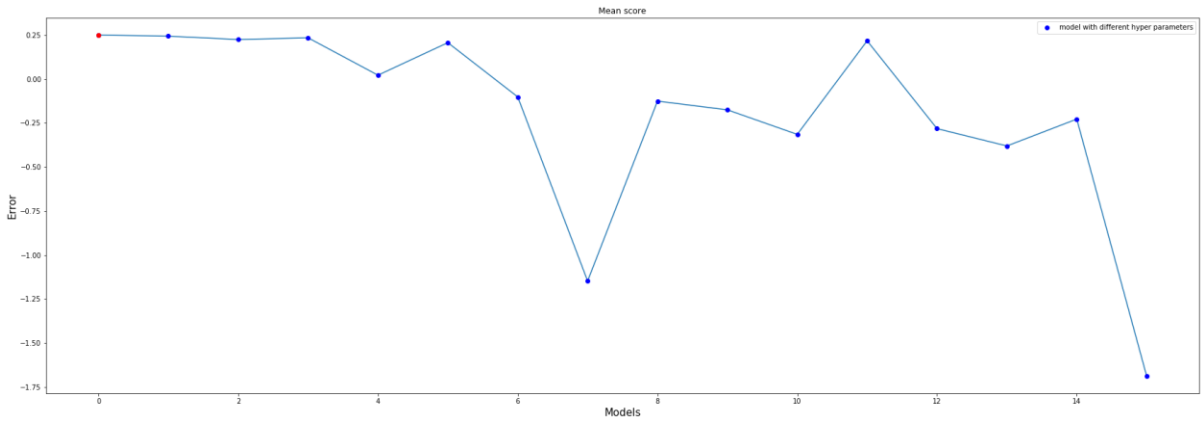
In the below plot, it can be seen that the non-regularized model is performing better on train set when compared with Lasso and Ridge regularized model. This is because of the model tends to **overfit**. It performs good on training set. But the second graph shows the results on test set. We see that the non-regularized model performs poorly on test set when compared to Lasso and Ridge regularized.

Alpha vs RMSE

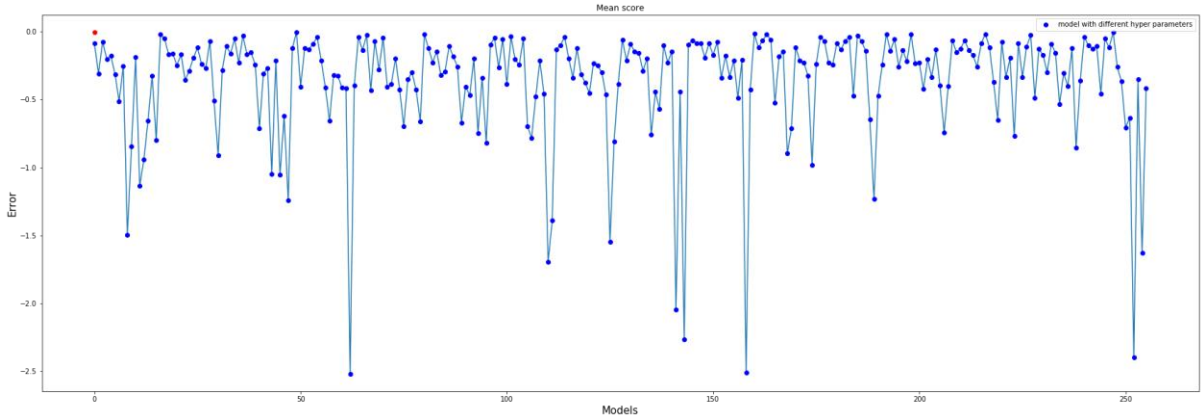


Plots for GridSearchCV

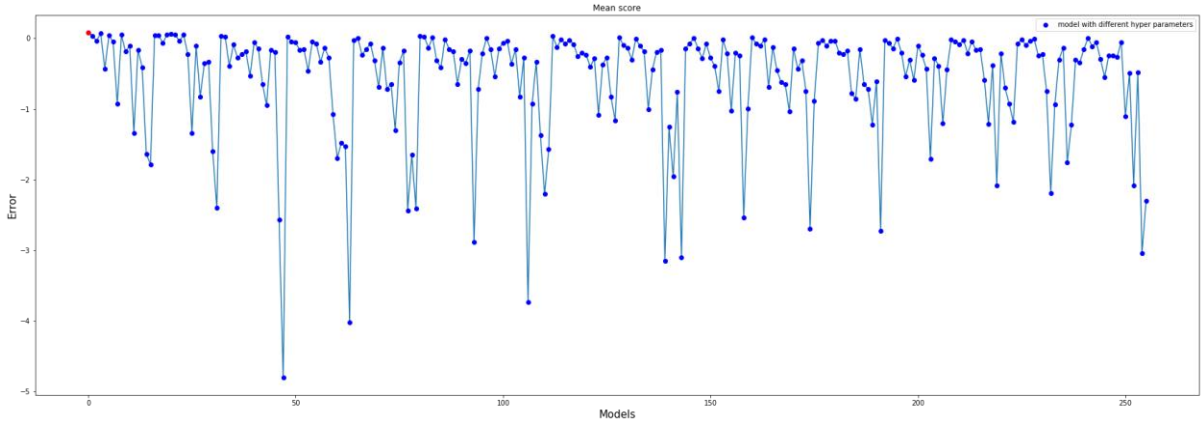
1. Ordinary Least Squares



2. Ridge Regression

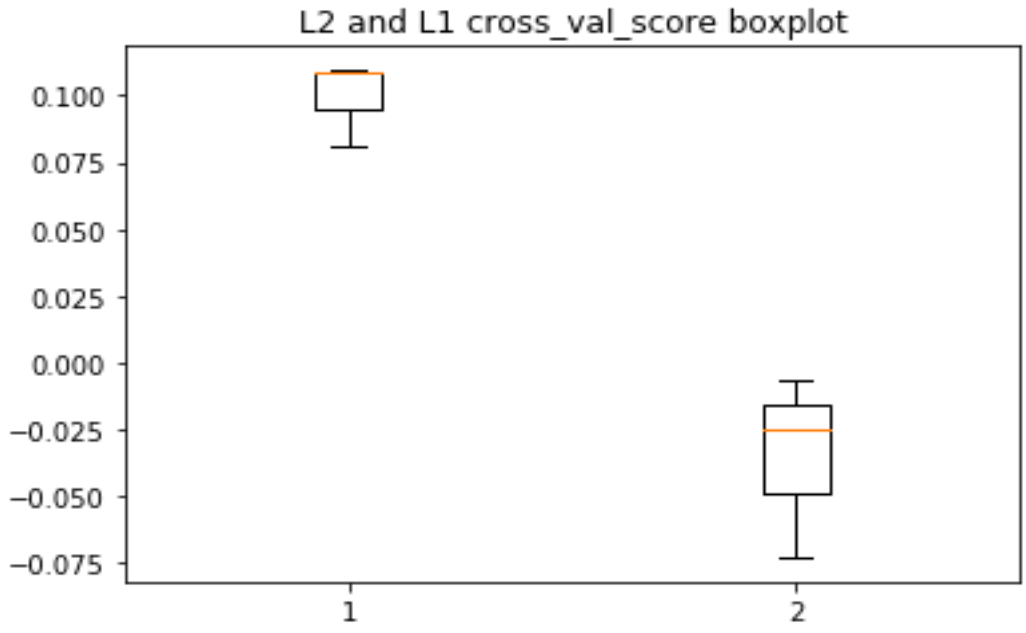


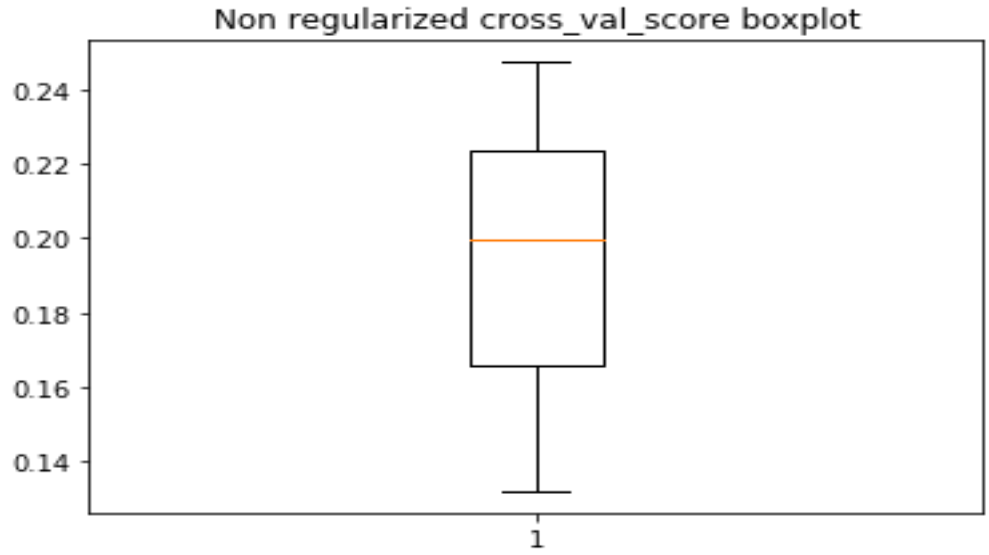
3. LASSO



Regularized vs non-regularized Boxplot for cross_val_score

In the below plot, from the cross validation score it can be seen that the Regularized model perfoms well than the non- regularied model.

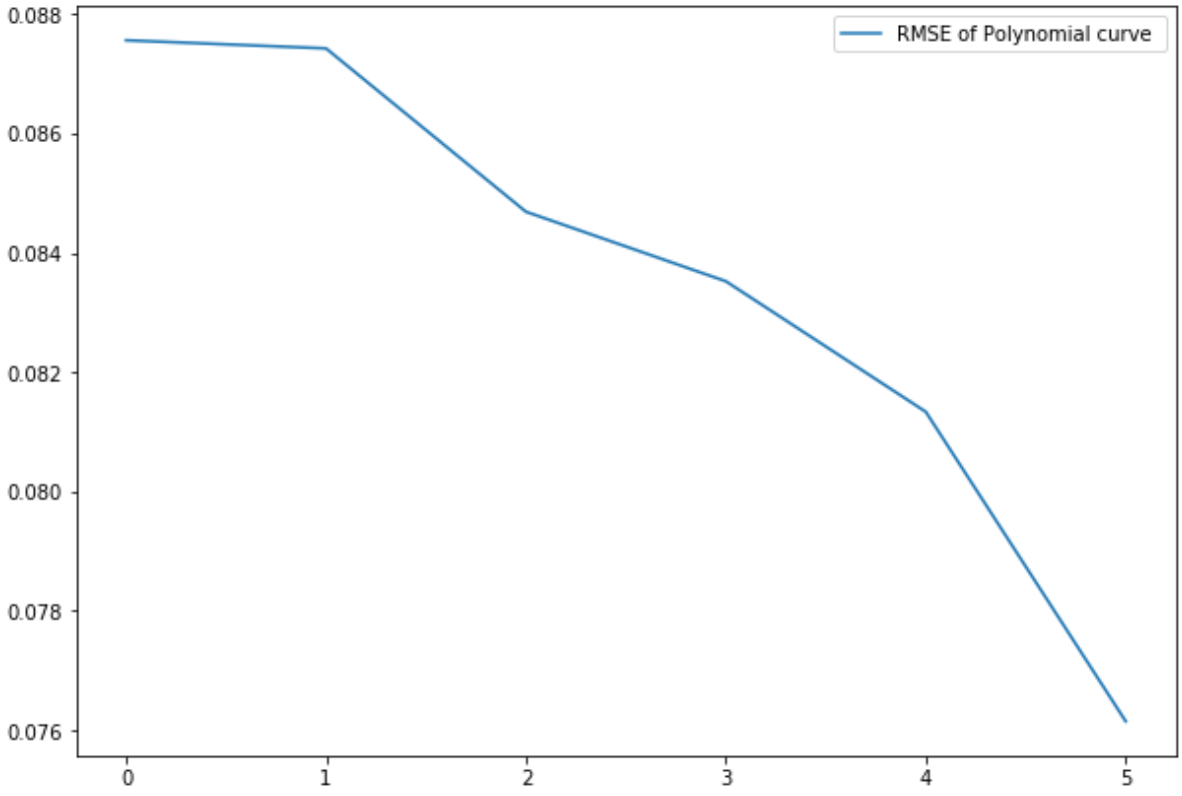


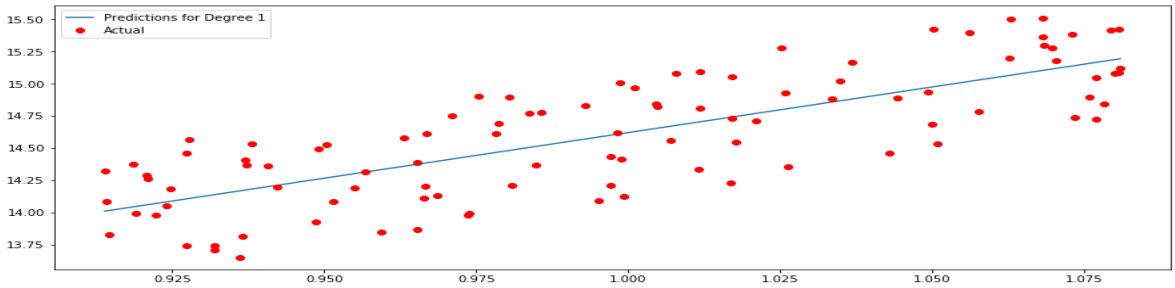


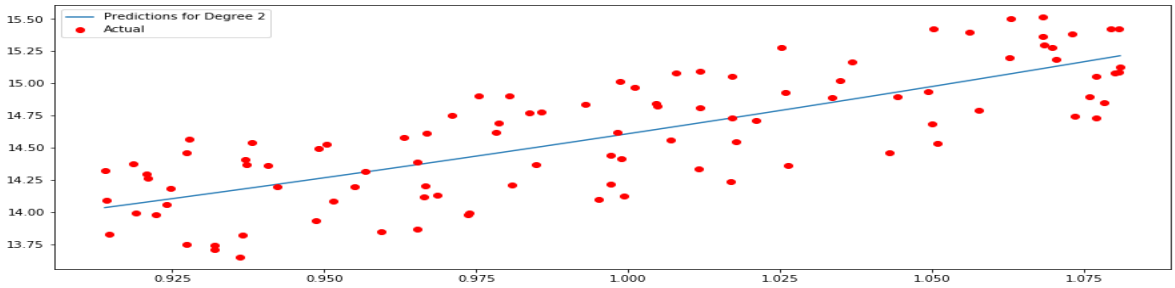
Polynomial Regression

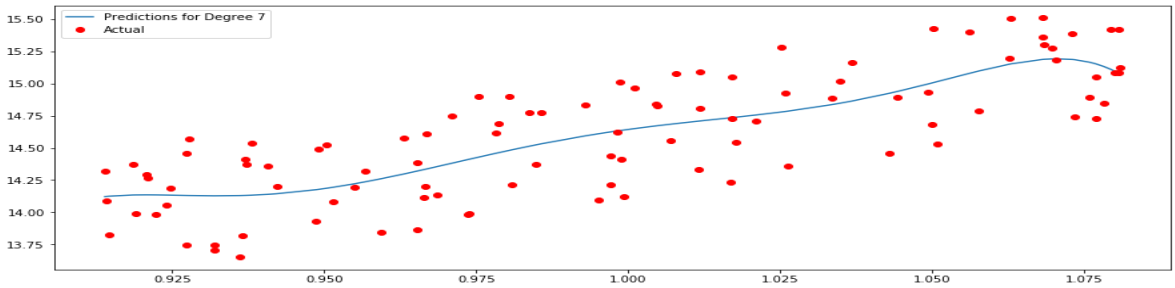
Task A : Prediction with high degree of polynomials

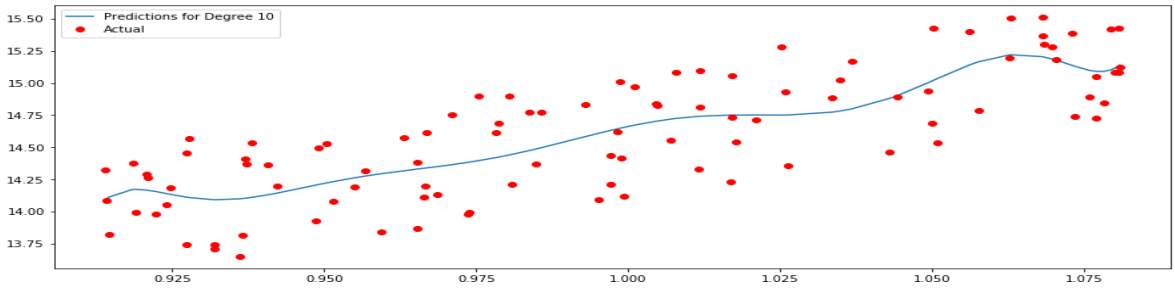
From the below gaph it can be seen that when the degree is increased from 1 to 100, the error is decreasing as it overfits the training set.

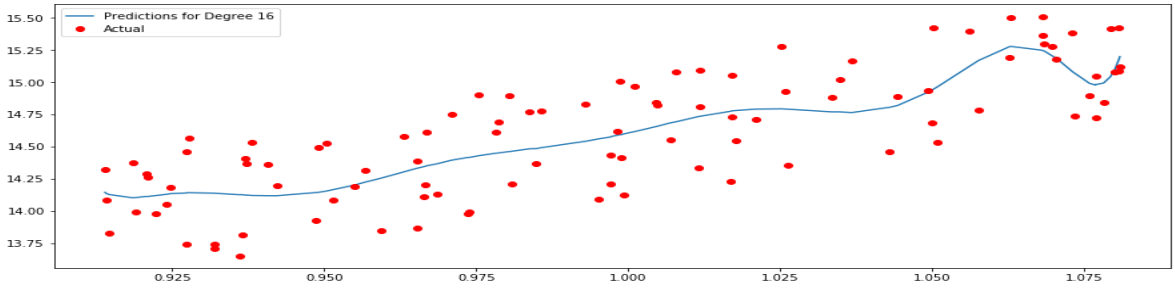


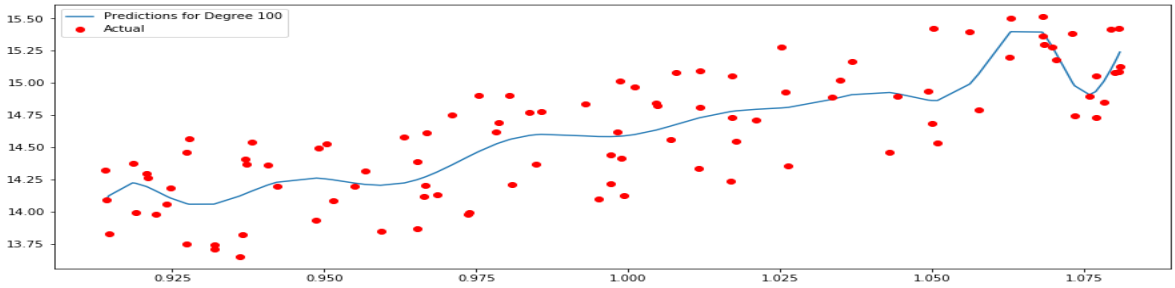

Predictions for Degree 1
Actual
15.50
15.25
15.00
14.75
14.50
14.25
14.00
13.75
0.925
0.950
0.975
1.000
1.025
1.050
1.075


Predictions for Degree 2
Actual
15.50
15.25
15.00
14.75
14.50
14.25
14.00
13.75
0.925
0.950
0.975
1.000
1.025
1.050
1.075

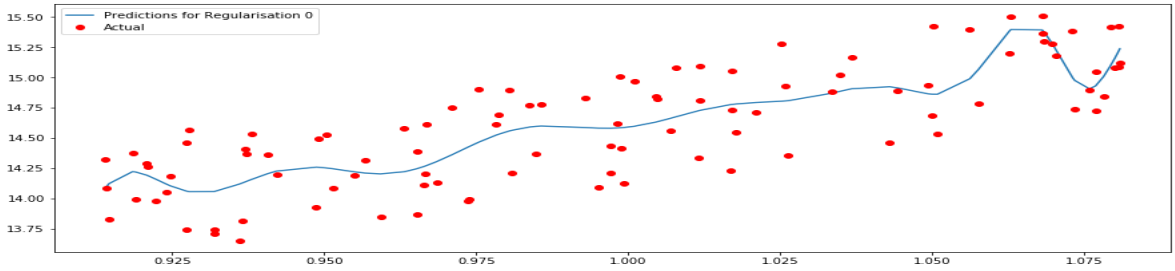

Predictions for Degree 7
Actual
15.50
15.25
15.00
14.75
14.50
14.25
14.00
13.75
0.925
0.950
0.975
1.000
1.025
1.050
1.075

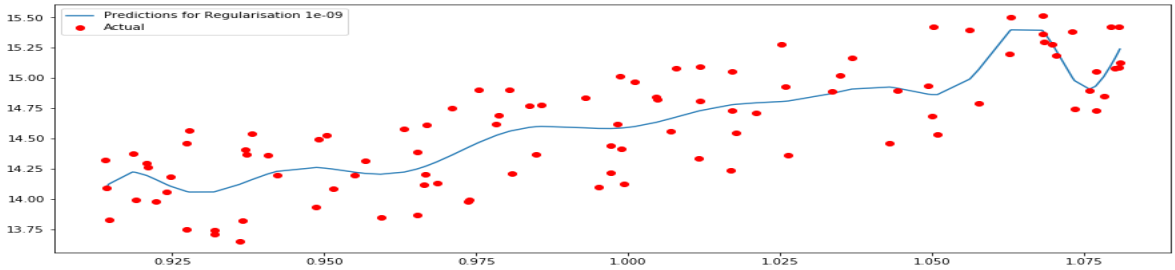

Predictions for Degree 10
Actual
15.50
15.25
15.00
14.75
14.50
14.25
14.00
13.75
0.925
0.950
0.975
1.000
1.025
1.050
1.075

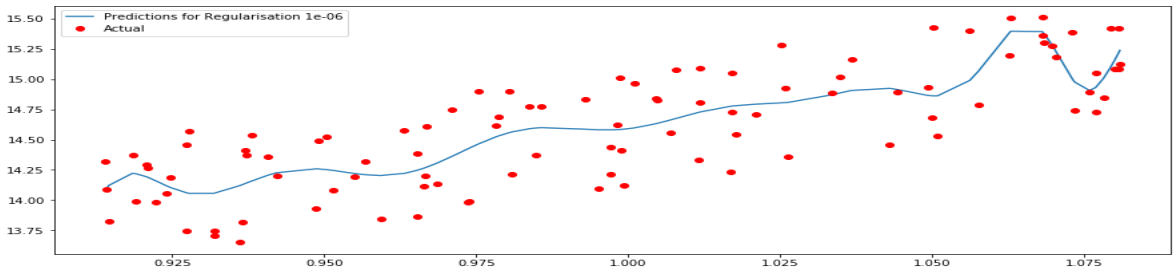

Predictions for Degree 16
Actual
15.50
15.25
15.00
14.75
14.50
14.25
14.00
13.75
0.925
0.950
0.975
1.000
1.025
1.050
1.075

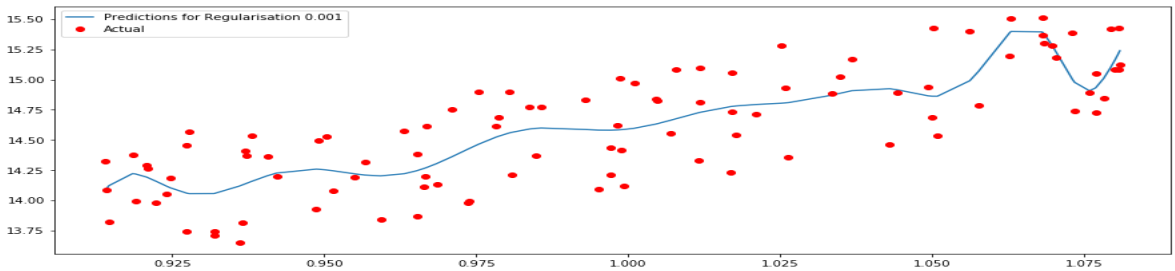

Predictions for Degree 100
Actual
15.50
15.25
15.00
14.75
14.50
14.25
14.00
13.75
0.925
0.950
0.975
1.000
1.025
1.050
1.075

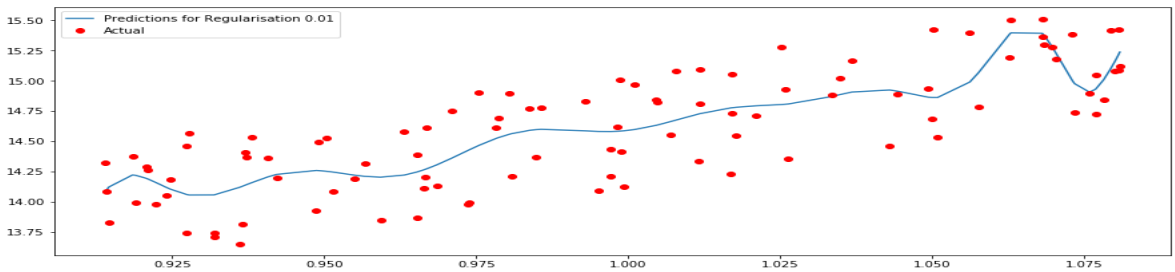
Task B: Effect of Regularization

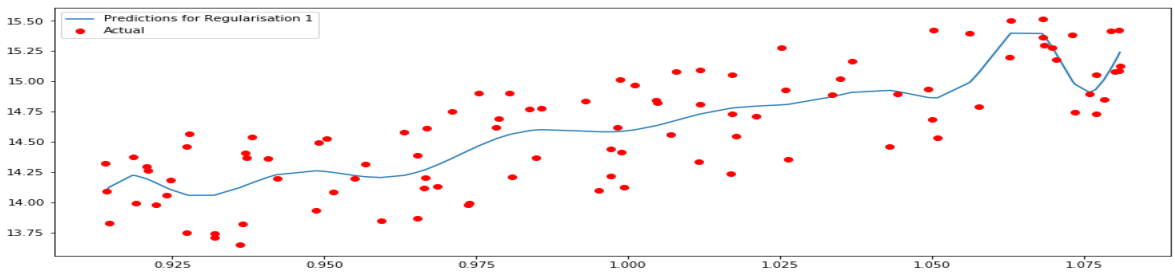


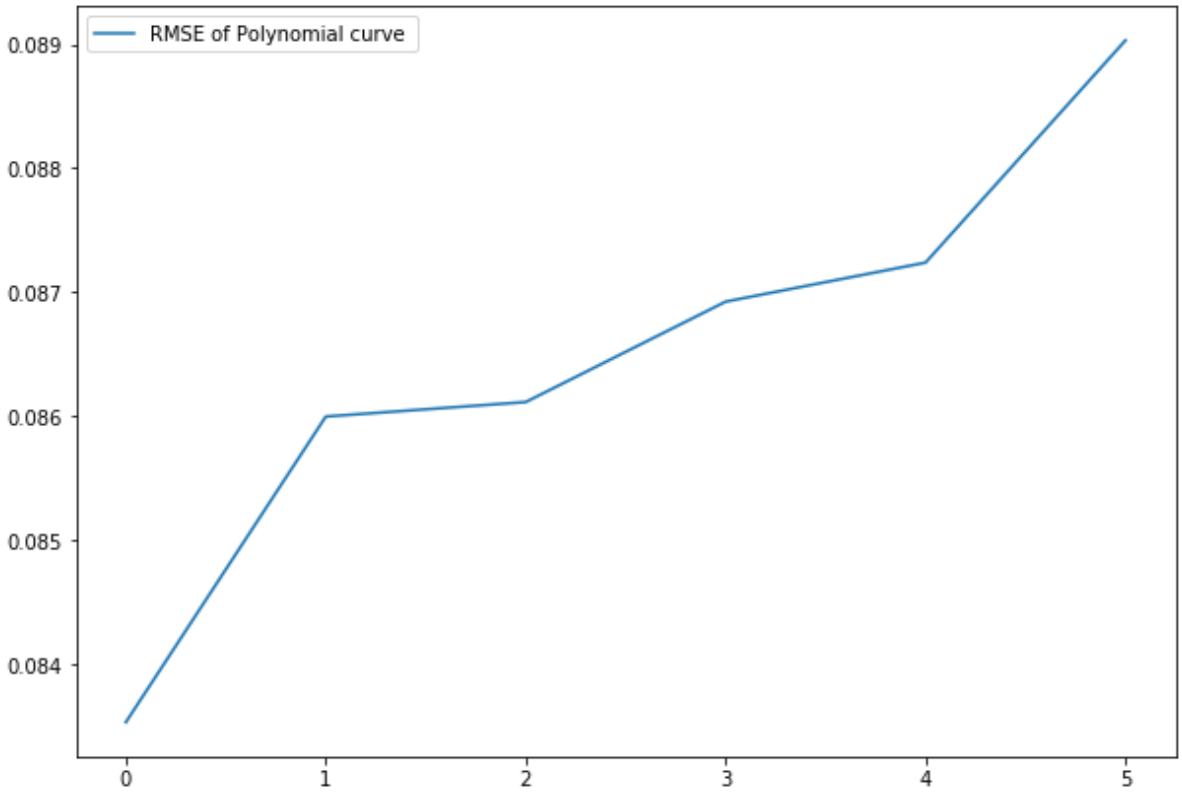












From the above graph it can be seen that when the degree is increased from 1 to 100, the error increases meaning the data is not overfit the training set. Thus, Regularization reduces the overfitting the data.