**Assignment**

**Advanced Regression**

**Part – II**

**Student Name: Hafizullah Mahmudi**

**Student Id: APFE19M00734**

**IIIT Role Number: DDS1950112**

**Kabul, Afghanistan**

**April 1, 2020, 2020**

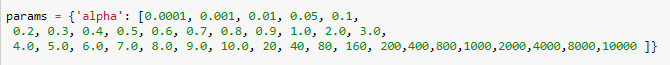
**Institute: upGrad/IIIT-B**

**Question 1**

**What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

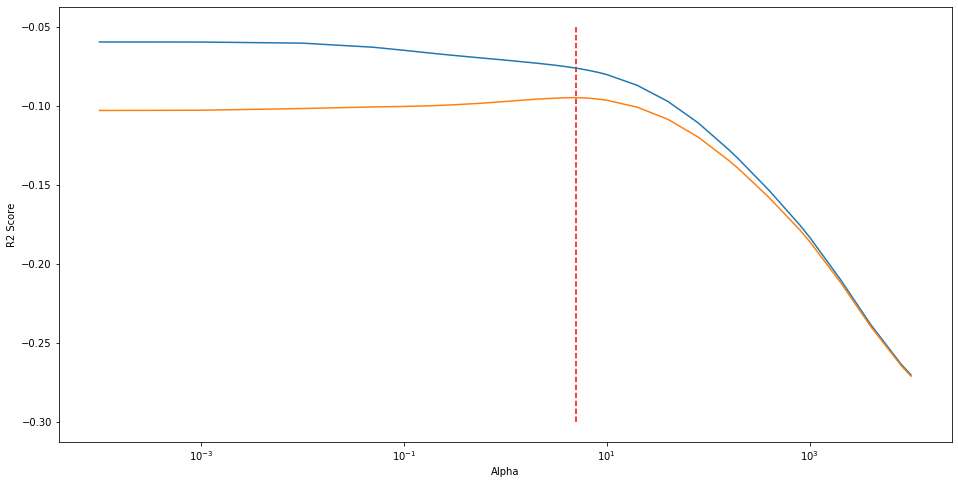
**Answer:**

After parameter tuning for the following hyper parameters for Ridge and Lasso Regressions:



The following value was identified as optimal value. For Ridge regression the lambda (alpha) was 5.0 and for Lasso regression the lambda(alpha) was 0.001

Following is the graph of r2 score of train and test for the above list of alphas for ridge regression at alpha 5.0



**And following is the scores for ridge regression with the alpha is as following**

- Ridge regression linear model intercept: 11.42714636940864

- R-squared score (r2\_scor) (training): 0.9200

- R-squared score (r2\_score) (test): 0.8850

- Number of non-zero features: 257.0000

- RMSE (training): 0.11141145172132126

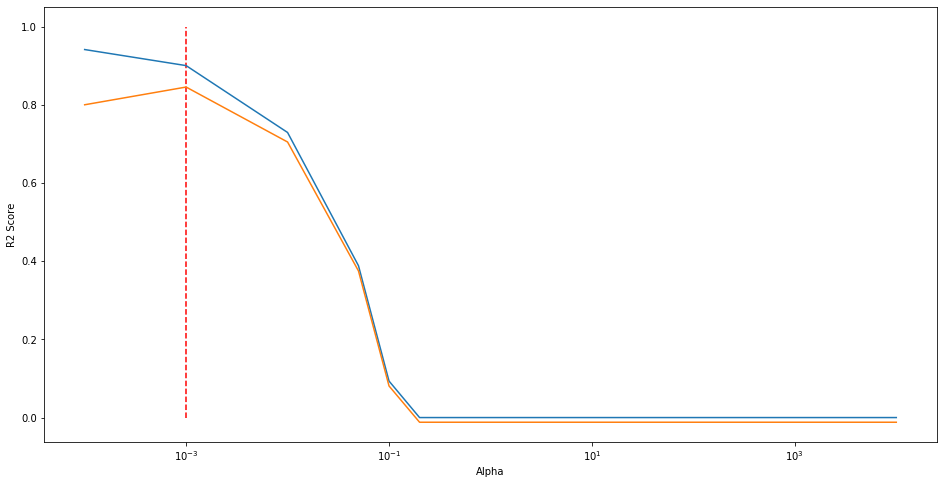
- RMSE (test): 0.13970093286444146

- Best score -0.09476075333644439

**And following are the most important predictor variables using Ridge regression**

* GarageCars, 0.1980416
* GrLivArea, 0.1929001
* 2ndFlrSF, 0.1877744
* OverallQual\_9, 0.1614598
* OverallQual\_3, -0.1489834

**Following is the graph of r2 score of train and test for the above list of alphas for Lasso regression at alpha 0.001**



**And following is the scores for ridge regression with the alpha is as following**

- Lasso regression linear model intercept: 11.512920130925382

- R-squared score (training): 0.898

- R-squared score (test): 0.884

- Number of non-zero features: 96.000

- RMSE (training): 0.12570161560105642

- RMSE (test): 0.14014205975358185

- Best Score: 0.8460068690021971

**And following are the most important predictor variables using Lasso regression**

* GrLivArea, 0.9308829
* GarageCars, 0.3189173
* OverallQual\_3, -0.2064934
* OverallQual\_9, 0.1965392
* OverallCond\_3, -0.1717684

**Now doubling the alpha for ridge from 5.0 to 10 and for lass from 0.001 to 0.002 will get the following important predictors:**

**Ridge Regression:**

* GarageCars, 0.1816128
* GrLivArea, 0.1544637
* 2ndFlrSF, 0.1454925
* Fireplaces, 0.1427075
* OverallQual\_9, 0.1407809

**Lasso Regression:**

* GrLivArea, 0.8710633
* GarageCars, 0.3586118
* OverallQual\_9, 0.2119916
* Fireplaces, 0.1902302
* OverallQual\_3, -0.1689119

As seen above, the top 5 most important predictors are the same but the weight of predictor is different between ridge and lasso regression.

With alpha to 0.002, the lasso regression has penalized more variable than alpha 0.001.

**Question 2**

**You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?**

**Answer:**

For Ridge regression regularization, the optimal value was 5.0 and for Lasso it was 0.001. I choose the 0.001 for Lasso for the following reasons:

Lasso Regression would help in feature elimination and the model will be more robust.

The r2 score 89.8% for train and 88.4% for test. The model prediction capability for test is 884% which is a good. The r2 score can help measure how well the future samples are likely to be predicted by the model.

The RMSE gap between training and test in Lasso is much smaller than Ridge. It means that there is slightly higher bias but lower variance.

On the other hand, r2 score gap between training test in lasso is smaller than Ridge. It means that the model is robust and generalizes well.

**The code for calculating these metrics are available at the end of the notebook for assignment part I under question 1.**

**Question 3**

**After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?**

**Answer:**

The most important predictor variables now are the following:

* 1stFlrSF, 0.5117664
* GarageArea, 0.3665628
* 2ndFlrSF, 0.3105457
* Fireplaces, 0.1914727
* FullBath, 0.1586906

**The code for calculating these predictors are available at the end of notebook for assignment part I under question 3.**

**Question 4**

**How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?**

**Answer:**

When a model performance is impacted by outliers or when a model performs well on train and bad on test then the model is not robust and a model is not generalizable or when there is a big gap in between train and test score.

A model can give accurate predication when it gives good result for data sets other than the one used during training. The outlier analysis is important needs to be done and removed those outliers that does not make sense and make the model unstable. Removing outlier will increase the accuracy of the model.

And when a model performs best well in training set with RSS of 0 and perform bad on test then the model is overfit. To resolve the issue, we need to add some bias to the model and decrease variance so that it performs well both on train and test.