Advanced Linear Regression Assignment

Mandheer Singh
PGDML

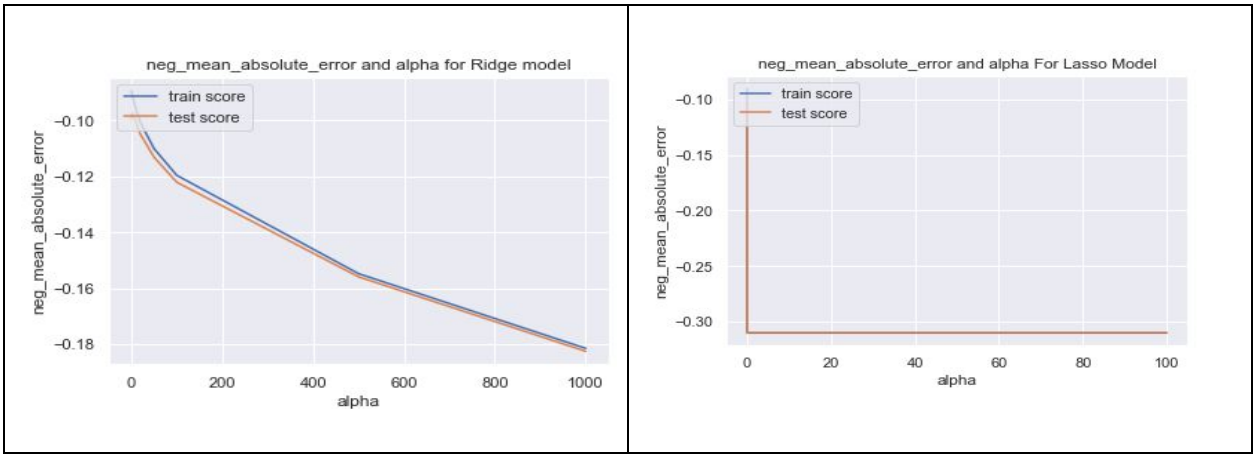
Question 1

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?

As per the solution submitted by me the optimal value for Ridge regression is **2.0** and optimal value of alpha for Lasso regression is **0.0001**. The results are shown in the following table:-

Model	Alpha Value	Train Score	Test Score
Ridge	Alpha= 2.0 (optimal)	89.49	84.49
	Alpha=4.0	89.00	84.94
Lasso	Alpha=0.0001 (optimal)	89.82	83.38
	Alpha=0.0002	89.64	83.82

Although the above results are looking comparable but look at the grid search graphs



In Ridge regression highest performance is at starting thus making value from 2.0 to 4.0

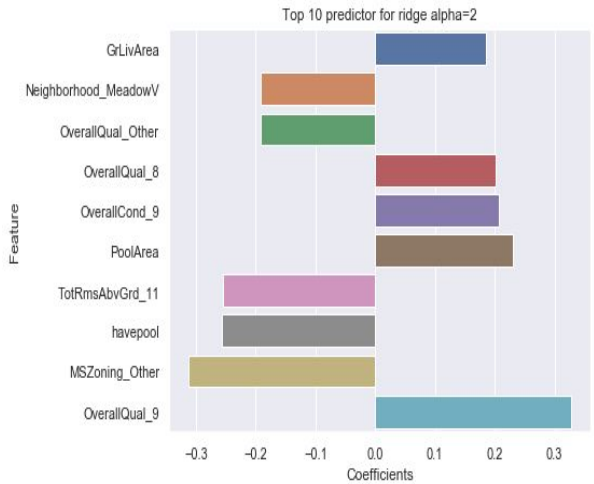
is not showing much difference in results but as we are increasing the value of alpha more the model performance is getting worse but the train and test scores are hugging each other means model is not suffering with the overfitting. Simply we focus more by making the model simple(increasing the alpha) will make error terms suffer.

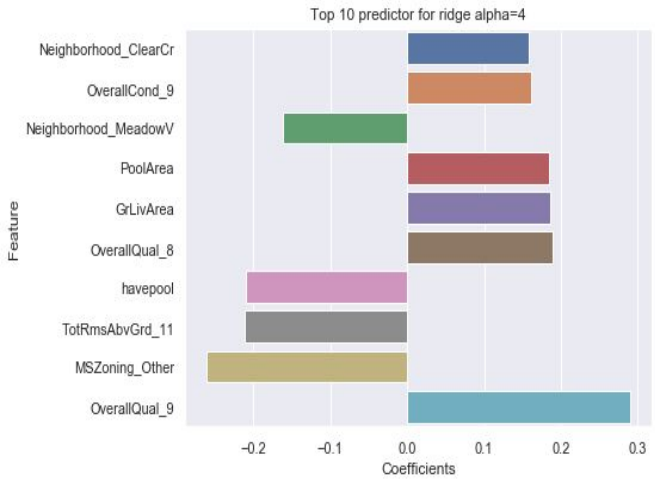
In lasso we are getting an L-shaped overlapped curved it means there is no overfitting although optimal alpha is 0.0001 means it is very less and as we increase alpha to higher values model performance goes down drastically.To make more sense out of this look at the below table which shows the corresponding results of ridge and lasso regression by increasing alpha little more-

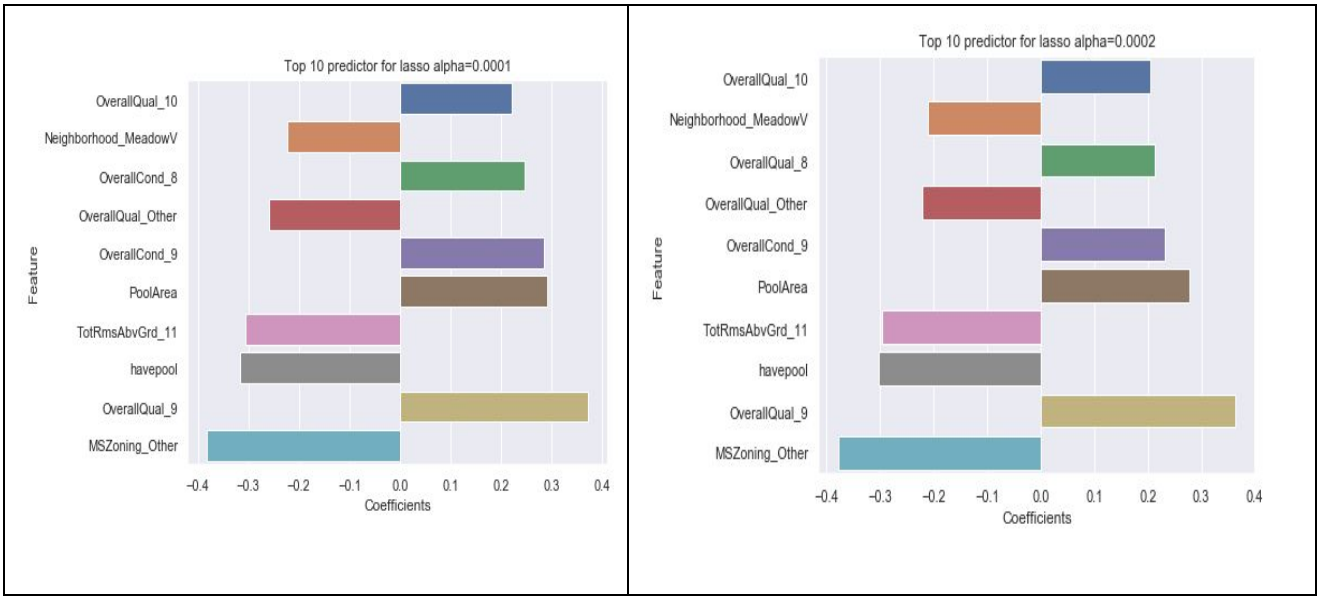
Model	Alpha Value	Train Score	Test Score
Ridge	Alpha= 20	86.39	84.03
	Alpha=100	81.49	80.02
Lasso	Alpha=0.01	75.53	75.26
	Alpha=0.3	0.00	-0.00000158

In ridge performance getting down slowly while in lasso results are absurd even for 0.3.

Most important predictors for optimal alpha and its double value is:-







We can see that predictors have not been changed, only the ordering is slightly changed.

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?

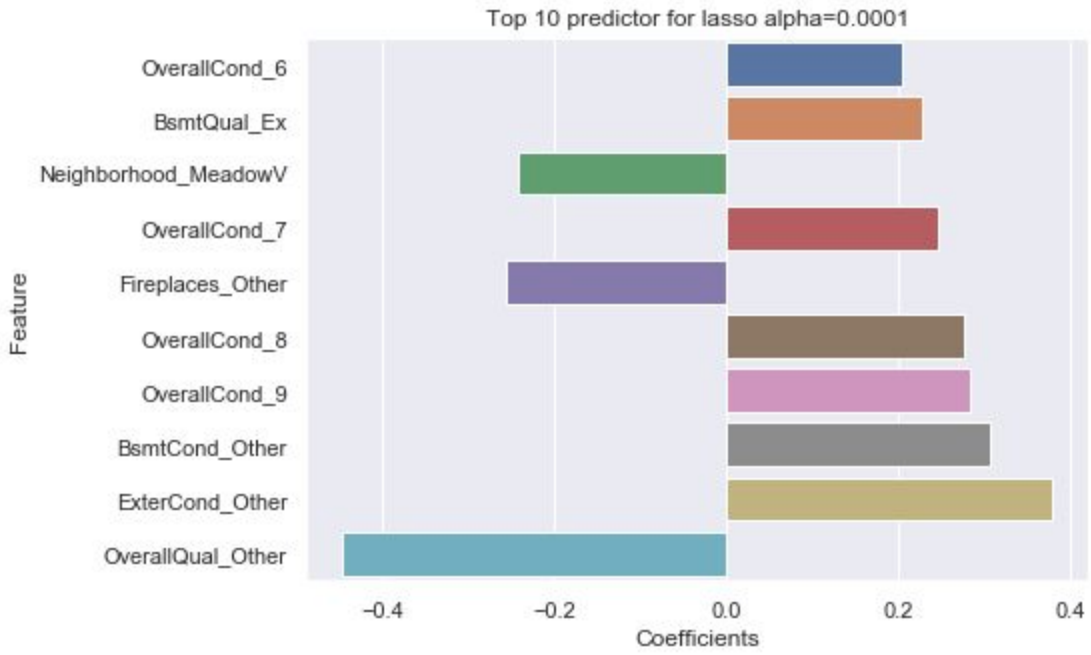
Answer-

From the previous question we have seen that performance of both ridge and lasso is quite comparable for the optimal alpha. The advantage of lasso is that it reduces the number of features but in our case after applying RFE we are 50 features to the both ridge and lasso model. Lasso is returning all 50 features, so as the Ridge. Thus i would rather use the Ridge regression in this case as it has a computational advantage over lasso.

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-



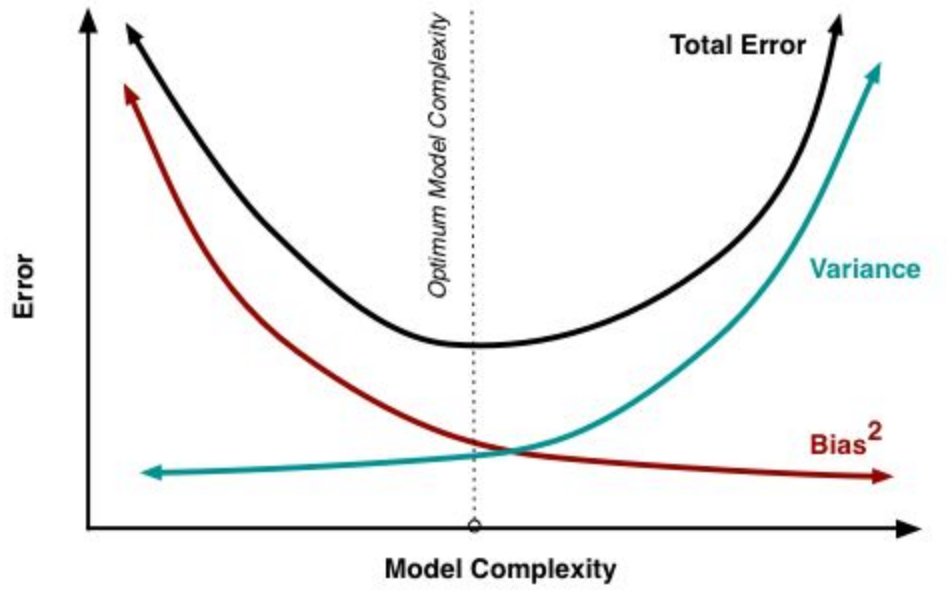
After removing Top 5 predictors for the lasso model we got next top 10 predictors as per the above figure. Thus next top 5 predictor will be:-

- OverallQual_Other
- ExterCond_Other
- BsmtCond_Other
- OverallCond_9
- OverallCond_8

Question 4
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer-

To make the model robust and generalizable we must take care of the bias-variance trade-off. If we focus more on reducing errors, the model may overfit i.e. the model will perform good on the training set but it will have bad performance on new data.Opposite of that if we focus more on making the model simple it may cost in terms of less accuracy of the model. Hence our goal should be to find a model which is in between i.e. simple and accurate.



In the above figure as model complexity increases variance will increase and bias will decrease. Thus we need optimal model complexity.This can be achieved using optimal alpha as we did in this exercise.