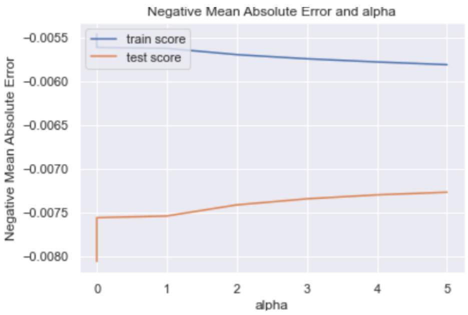
Question 1

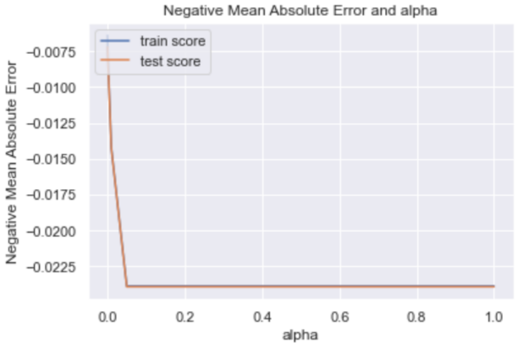
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

In Ridge Regression, when plotting the curve between Alpha and NMAE, its shown from the graph that at alpha = 1 the training error starts decreasing and test error increases. As the value of Alpha goes up the train keeps decreasing and the test keeps increasing. So alpha = 1 is the optimal value where the test error is minimum.



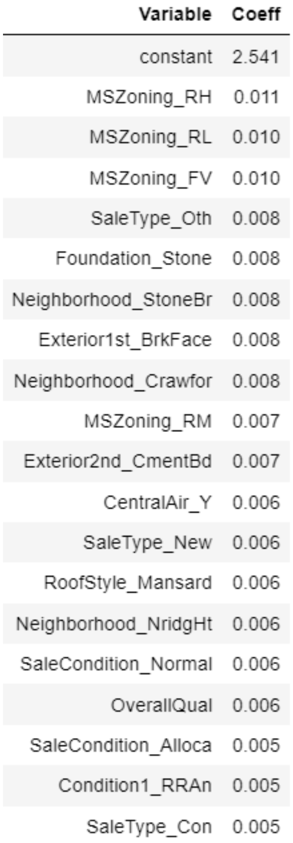
In Lasso Regression, although as seen in the graph, test score is starting to stabilize at 0.05 but at this value of alpha, the model penalizes more and makes mostly all the coefficient values to zero. So I used a very small value of alpha = 0.001 to keep the important coefficient alive.



As we see the trend in the graph, as the alpha increases penalty will be increased by the model so when we double the alpha in both Ridge and Lasso, the model will reduce the coefficient more.

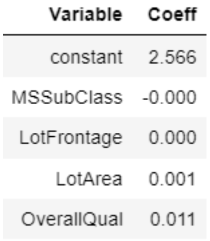
When applying double the values of alpha for Ridge and Lasso, following important predictor variables are found:

Ridge @ alpha = 2



Variable	Coeff
constant	2.541
MSZoning_RH	0.011
MSZoning_RL	0.010
MSZoning_FV	0.010
SaleType_Oth	0.008
Foundation_Stone	0.008
Neighborhood_StoneBr	0.008
Exterior1st_BrkFace	0.008
Neighborhood_Crawfor	0.008
MSZoning_RM	0.007
Exterior2nd_CmentBd	0.007
CentralAir_Y	0.006
SaleType_New	0.006
RoofStyle_Mansard	0.006
Neighborhood_NridgHt	0.006
SaleCondition_Normal	0.006
OverallQual	0.006
SaleCondition_Alloca	0.005
Condition1_RRAn	0.005
SaleType_Con	0.005

Lasso @ alpha = 0.002



Variable	Coeff
constant	2.566
MSSubClass	-0.000
LotFrontage	0.000
LotArea	0.001
OverallQual	0.011

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?

<u>Answer:</u>

Regularization of the coefficient improves the accuracy of the model and maintains the variance between the actual and predicted model.

In this case, Ridge regression uses lambda as the penalty value and regularizes the coefficient with the help of cross-validation. Those coefficients which have higher values of lambda get more penalized. Normally we chose those values of lambda which keeps the low variance in the model. Ridge regression makes the coefficient near zero and keeps all the coefficients in the final model, unlike in lasso regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is the absolute value of the magnitude of coefficients which is identified by cross-validation. In Lasso regression, as the lambda value increases, lasso keeps reducing the coefficient and make it towards zero, and this this makes the variables exactly equal to 0. This property makes Lasso a feature selection. When the lambda value is small it performs simple linear regression and as the lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

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?

<u>Answer:</u>

The five most important predictor variables after excluding the previous ones are:

```
BsmtFullBath   LotArea   TotalBsmtSF   WoodDeckSF   FullBath
```

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?

<u>Answer:</u>

The model should be as simple as possible, though its accuracy will decrease it will be more robust and generalizable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias is an error in the model, when the model is weak to learn from the data. High bias means the model is unable to learn details in the data. Model performs poorly on training and testing data.

Variance is an error in the model when the model tries to overlearn from the data. High variance means the model performs exceptionally well on training data as it has very well trained on this data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance between Bias and Variance to avoid overfitting and under-fitting of data.