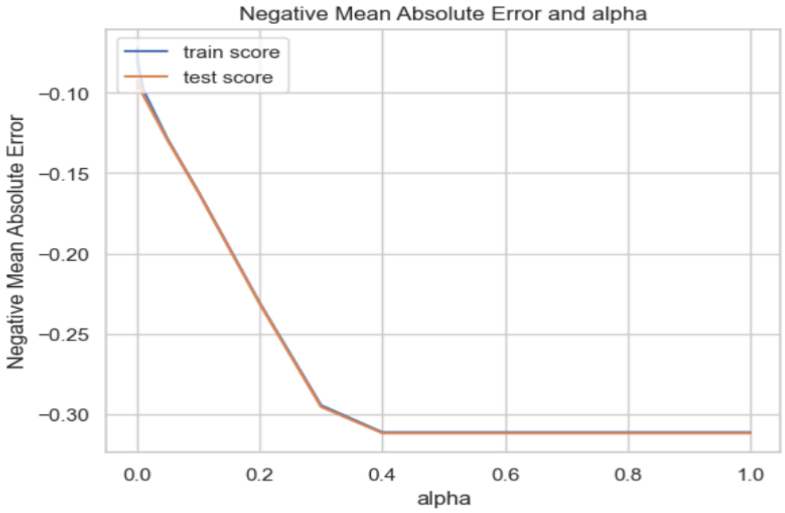
Question 1

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?

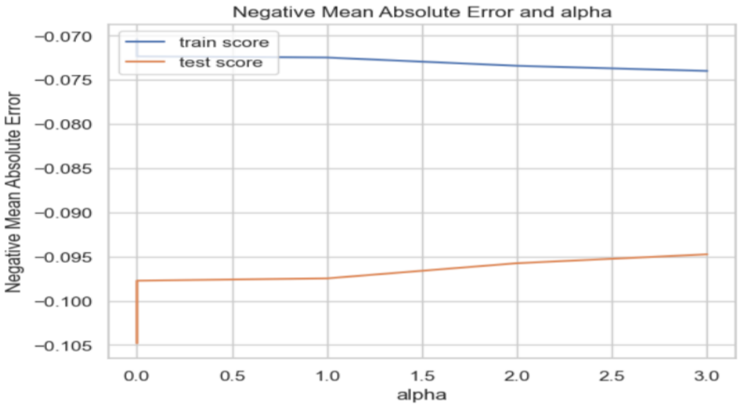
Answer :

For Lasso Regression : When we built our model and plotted the graph between alpha and negative mean absolute error the graph came as shown below



We see that initially as value of alpha increases the error comes down and than around value of alpha = 0.4 it almost stabilizes. However the intention of model building is not to minimize the negative mean absolute error but to strike a balance between variance and bias. Keeping this in mind we choose a very small value of alpha = 0.01 and obtained the coefficients. If we increase the value of alpha in Lasso regression we found that our r squared value was decreasing continuously and more and more variables were being rendered useless. This is not desired hence we choose a small value for alpha in Lasso.

For Ridge regression: When we built the model and plotted the graph between alpha and negative mean absolute error the below graph was obtained



From the above graph we notice that as alpha increases error in test set decrease till alpha = 2. Later it just subsides. For the train data it just is opposite. Hence an optimum value of alpha for Ridge regression can be chosen as 2.

Effect of doubling value of alpha

For Lasso : IF we double the value of alpha for Lasso, we observe that more variables are rendered useless as their coefficients become 0. Also the r squared value decreases. Hence the model is penalized if we double the value of alpha

For Ridge : IF we double the value of alpha we see that the error in train data will increase more and error for test set will also be more. Hence the model will be penalized due to more generalization.

List of variables : The list of most important variables is as follows

'OverallQual', 'GrLivArea', 'OverallCond', 'GarageArea', 'BsmtFullBath', 'Fireplaces', 'FullBath', 'TotalBsmtSF', 'LotArea', 'MSZoning_RL', 'WoodDeckSF', 'ScreenPorch', '1stFlrSF', 'BsmtFinSF1', 'HalfBath', 'KitchenAbvGr', 'MSSubClass', 'PoolArea', 'PropAge'

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 :

Ridge regression : Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. It uses a tuning parameter called lambda as penalty which is obtained by cross validation. As the value of lambda increases the variance is dropped but bias remains constant. It includes all the variables in the final model.

Lasso regression : Lasso regression also uses a tuning parameter called lambda. But as the value of lamda increases it shrinks the coefficients towards zero. By doing this it also does the variable selection. The final model will have selected variables on which the model majorly depends.

Out of both the approaches we prefer Lasso regression by sheer fact that it does variable selection as well without compromising on the correct predictions.

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 five most important variables in the model are

4	OverallQual	0.132
13	GrLivArea	0.117
5	OverallCond	0.049
21	GarageArea	0.045
14	BsmtFullBath	0.029

Now if we have to exclude these variables due to absence in data set and run the model again then the set of 5 most important variables after excluding above will be

20	Fireplaces	0.026
16	FullBath	0.022
9	TotalBsmtSF	0.017
3	LotArea	0.015
31	MSZoning_RL	0.010

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 :

In order for a model to be robust and generalisable we should always have a trade-off between bias and variance. We need to select the value of alpha such that without making model complex we can achieve the desired results. In other words for model to be robust and generalisable it must be very simple.

If we try to simplify the model too much than there is high chance of model’s accuracy going down. Thus we must avoid over simplification of model as well. Though the model should never be very complex.

A robust model will perform equally well on train and test data set.

Bias occurs when a model is unable to capture the complexity or true nature of the underlying data due to oversimplification or assumptions made during model development. It's important to note that bias is different from variance, which refers to the error that occurs when a model is too sensitive to variations in the training data and performs poorly on new data. A good model should strike a balance between bias and variance to achieve optimal performance.