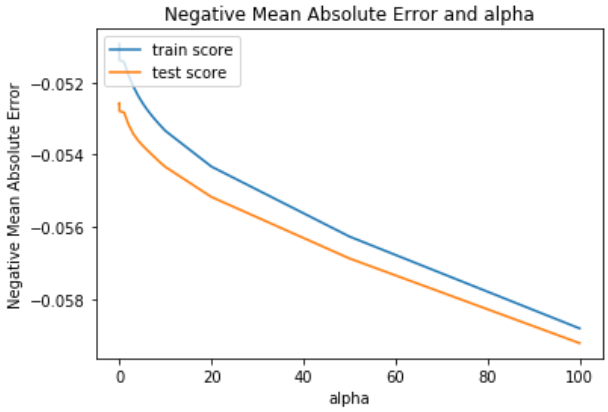
1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: The optimal value obtained for my model is:

Ridge: alpha= 10

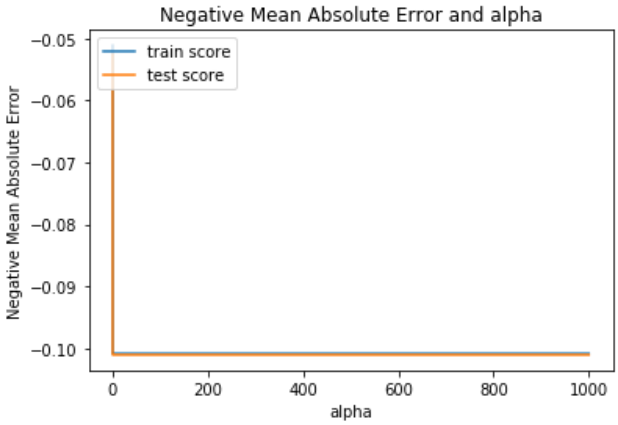
If I choose alpha=20 then, the values of the coefficients are getting decreased while the number of important variables remains the same.



Ridge

Lasso: alpha= 0.002

If I choose alpha=0.004 then, the values of the coefficients are getting decreased while the number of important variables too gets decreased.



Lasso

OverallQual will be the most important predictor variable, followed by GarageCars_3, TotRmsAbvGrd,BsmtExposure_Gd etc...(As mentioned in the Jupyter Notebook.)

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?

Ans: I will go with the Lasso regression (alpha=0.002), as it makes the coefficients of the useless predictors of "SalePrice", as zero. It reduces the complexity of the model, makes it more robust, less biased and with low variance. It makes the model more generic and avoids overfitting.

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?

Ans: Yes, agreed. Previously the five most important predictor variables are coming to be OverallQual, GarageCars_3, TotRmsAbvGrd,CentralAir_Y,BsmtExposure_Gd. But after dropping these columns and re-taking the ridge and lasso regression, the five most important fac

tor are coming to be: GrLivArea ,GarageArea ,BsmtFinType1_GLQ,Fireplaces_2,GarageType_Attchd. (As shown in the Jupyter notebook)

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?

Ans: The lesser the number of independent variables, more robust the model. In this case the model can be used for any other training data (in the model range, ofcourse) which will still give a good estimation of the equation. As the model data isn't overfitting, the accuracy will be comparatively low. But if the adjusted R squared is above 60% for most of the training data it gets fed, then the model is good to go.