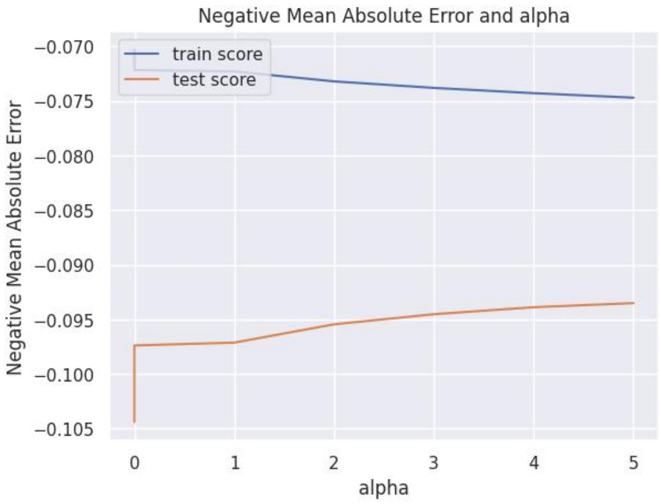
Question 1

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-

Ridge -

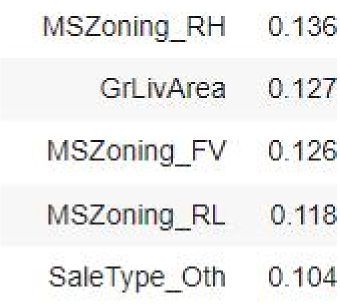


We can see Test score starts stabilizes after alpha=2, hence I selected this value

If we will increase alpha to 4 we can see based on chart Negative mean absolute error for Test will increase slightly

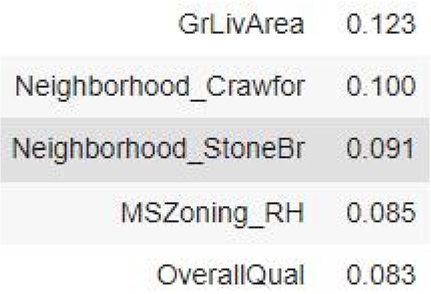
	Alpha	Number of variables	Test Score	Train Score	RMSE	Stage
Ridge	2	205	0.87	0.923	0.153	Optimal
Ridge	4	206	0.875	0.92	0.14	Double

Top 5 variable at Alpha=2 ->



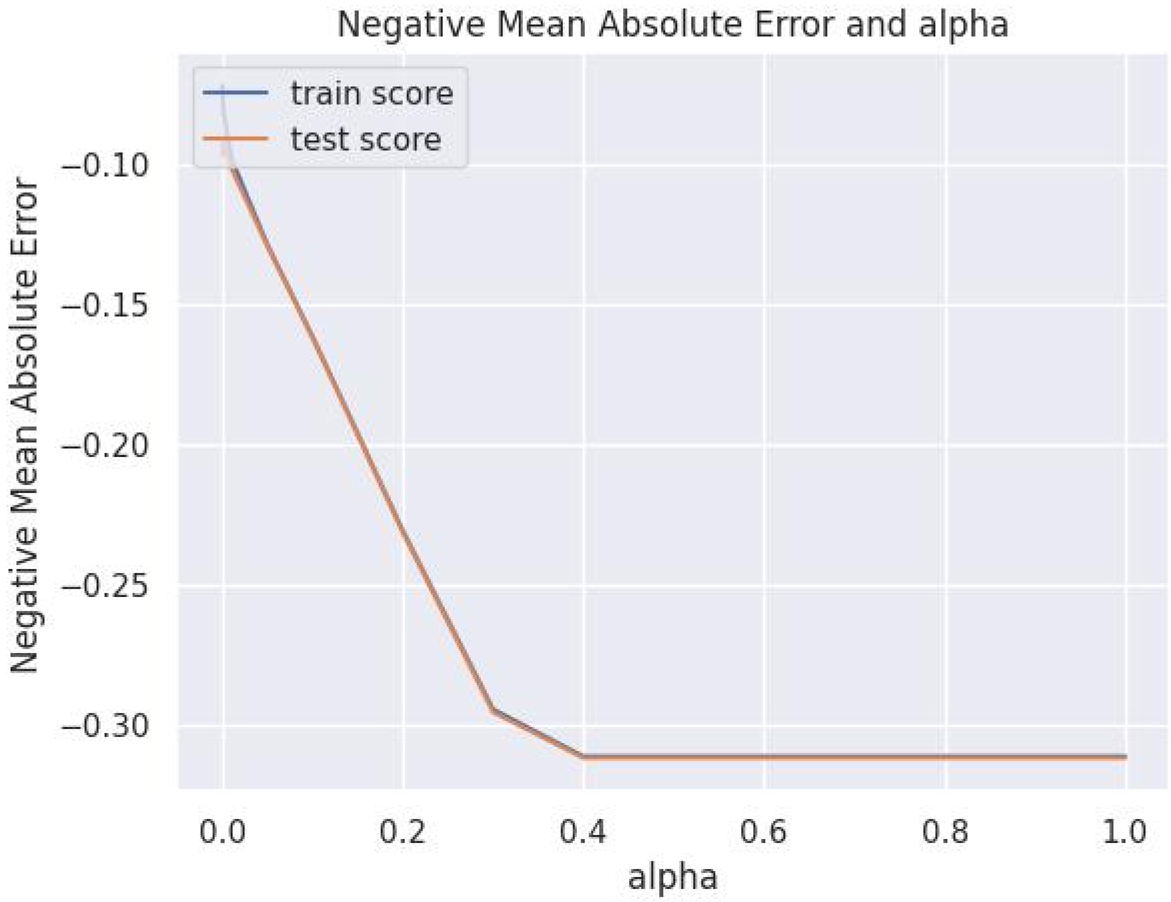
MSZoning_RH	0.136
GrLivArea	0.127
MSZoning_FV	0.126
MSZoning_RL	0.118
SaleType_Oth	0.104

Top 5 variable at Alpha=4 ->



GrLivArea	0.123
Neighborhood_Crawfor	0.100
Neighborhood_StoneBr	0.091
MSZoning_RH	0.085
OverallQual	0.083

Lasso -

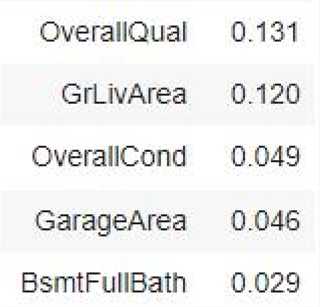


From the above graph we can see at alpha=0.4 Negative Mean Absolute Error is quite low,in order to balance the trade off between bias vs variance and top get the coefficients of smallest feature we will choose lower value of alpha

We can also see if we double the value of alpha, there is not much effect on Mean absolute error, but we are trying to keep the value of alpha low in-order to reduce the penalization

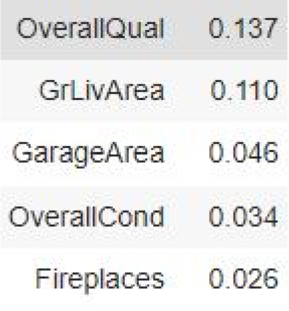
	Alpha	Number of variables	Test Score	Train Score	RMSE	Stage
Lasso	0.01	205	0.85	0.861	0.153	Optimal
Lasso	0.02	206	0.843	0.843	0.157	Double

Top 5 variable at Alpha=0.01 -> OverallQual,GrLivArea,OverallCond,GarageArea,BsmtFullBath



OverallQual	0.131
GrLivArea	0.120
OverallCond	0.049
GarageArea	0.046
BsmtFullBath	0.029

Top 5 variable at Alpha=0.02 -> OverallQual,GrLivArea,GarageArea,OverallCond,Fireplace



OverallQual	0.137
GrLivArea	0.110
GarageArea	0.046
OverallCond	0.034
Fireplaces	0.026

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 and Lasso Regression are both regularization techniques used to prevent over fitting in linear regression models. They add a penalty term to the cost function that constrains the model's coefficients.

Ridge Regression (L2 regularization) adds the squared magnitudes of the coefficients to the cost function, and its penalty term is proportional to the square of the L2 norm of the coefficient vector. It tends to reduce the coefficients towards zero, but it rarely sets them exactly to zero. This makes Ridge Regression useful when you have a lot of correlated features and want to include all of them in the model.

The penalty in Ridge is lambda times sum of square of the coefficients, hence the so coefficients that have greater value gets penalized.

As we increase the value of lambda the variance in model is dropped and bias remains.

Lasso Regression (L1 regularization), on the other hand, adds the absolute values of the coefficients to the cost function, and its penalty term is proportional to the L1 norm of the coefficient vector. Lasso tends to produce sparse models by driving some of the coefficients to exactly zero. This makes Lasso useful when you suspect that many of your features are irrelevant or redundant, and you want a simpler model with only the most important features.

In Lasso as Lambda value increases,coefficient values starts to reduce to 0, hence we need to select the value of lambda such a that we get optimal 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?

Answer-

New 5 most important predictors are

BsmtFullBath

Fireplaces

FullBath

TotalBsmtSF

LotArea

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-

To ensure that a model is robust and generalizable, you need to follow good practices during the model development and evaluation process

A model with high complexity can achieve high accuracy on the training data but might suffer from over fitting and low generalization (high variance). On the other hand, a model with too much regularization or too few features might have low accuracy on both training and testing data due to under fitting (high bias), hence a balance needs to made.

To strike a balance between accuracy and generalization, it's crucial to perform rigorous model evaluation, conduct cross-validation, and consider the implications of your modeling choices on the model's ability to perform well on unseen data. Sometimes, sacrificing a bit of accuracy for a more robust and generalizable model is preferable, especially in applications where reliability and stability are essential.