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 1

Part 1 *– What is the optimal value of alpha for ridge and lasso regression?*

Optimal value of alpha for Ridge = 1.0 and Optimal value of alpha for Lasso = 0.0001

Part 2 *– What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?*

The accuracy of train and test scores have reduced slightly and in return the error terms have increased a bit in both the cases of Lasso and Ridge.

```
For Lasso Regression Model (Original Model: alp
 ******************************************

For Train Set:

R2 score: 0.9304
MSE score: 0.0016
MAE score: 0.0281
RMSE score: 0.1676

For Test Set:

R2 score: 0.8929
```

```
For Lasso Regression Model (Doubled Alpha Model: al
 ******************************************

For Train Set:

R2 score: 0.9244
MSE score: 0.0018
MAE score: 0.029
RMSE score: 0.1703

For Test Set:

R2 score: 0.8909
```

```
For Ridge Regression Model (Original Model: a
 ****************************************

For Train Set:

R2 score: 0.9334
MSE score: 0.0016
MAE score: 0.0276
RMSE score: 0.1661

For Test Set:

R2 score: 0.888
```

```
For Ridge Regression Model (Doubled Alpha Model:
 ****************************************

For Train Set:

R2 score: 0.9312
MSE score: 0.0016
MAE score: 0.0279
RMSE score: 0.167

For Test Set:

R2 score: 0.8869
```

Part 3 *- What will be the most important predictor variables after the change is implemented?*

Top 10 Predictors for Lasso –

GrLivArea, OverallQual, OverallCond, Age_Built_Sold, TotalBsmtSF, LotArea, Neighborhood_Crawfor, GarageCars, Neighborhood_Somerst, Neighborhood_NridgHt

Top 10 Predictors for Ridge –

GrLivArea, OverallQual, OverallCond, TotalBsmtSF, MSZoning_FV, MSZoning_RL, MSZoning_RH, Age_Built_Sold, 1stFlrSF, Neighborhood_StoneBr

(we are considering Lasso predictors over Ridge here as well)

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 2

The test accuracy is better for Lasso and in return the error is less. Also, the train test accuracy difference is less for Lasso.

Lasso helps in feature selection as the non important features coefficients are 0.

Due to above reasons I would select the Lasso model.

```
For Ridge Regression Model (Original Model: al For Lasso Regression Model (Original Model: alp
 ****************************************      ****************************************

For Train Set:                                 For Train Set:

R2 score: 0.9334                               R2 score: 0.9304
MSE score: 0.0016                              MSE score: 0.0016
MAE score: 0.0276                              MAE score: 0.0281
RMSE score: 0.1661                             RMSE score: 0.1676

For Test Set:                                  For Test Set:

R2 score: 0.888                                R2 score: 0.8929
```

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 3

The optimal alpha still remains the same i.e. 0.0001. The accuracy for both train and test scores have significantly reduced thus increasing the error's.

The top 5 predictor variables now are:

1stFlrSF, 2ndFlrSF, Neighborhood_Somerst, Exterior1st_BrkComm, Neighborhood_MeadowV

```
For Lasso Regression Model (Removed top 5 Features Model: a
 ************************************************

For Train Set:

R2 score: 0.9111
MSE score: 0.0021
MAE score: 0.0322
RMSE score: 0.1794

For Test Set:

R2 score: 0.8593
```

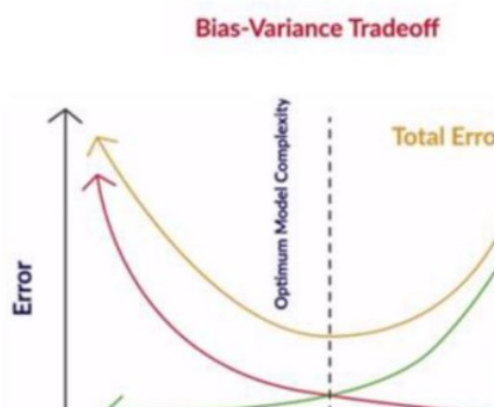
```
For Lasso Regression Model (Original Model: alp
 ********************************************

For Train Set:

R2 score: 0.9304
MSE score: 0.0016
MAE score: 0.0281
RMSE score: 0.1676

For Test Set:

R2 score: 0.8929
```

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 :

The model is robust when it performs well on unseen data. The difference in accuracy score for both test and training score should be similar. When the model is complex then the bias is low however the variance is high i.e. with change in training data the model changes. And if the model is too naïve the bias is too high. Hence there should be an optimum trade-off between the bias and variance error(as shown in below figure).



To achieve this we use various regularization techniques like ridge and lasso which penalizes the model and tries to get the beta coefficients as low as possible. Infact the lasso some of the insignificant beta's are forced to zero reducing the total number of parameters and hence the model complexity.

The accuracy and robustness should also have an optimum balance as they are at odds to each other. Too much accurate models can lead to overfitting and fails when it comes across unseen data.