Advance Regression Assignment Part 2

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:

With the Ridge Regression Alpha value is 10.0, please see in below snippet

In [88]:
```
# Will perform Ridge with the help of above generic function
params = {'alpha': [0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0,
                    9.0, 10.0, 20, 50, 100, 500, 1000 ]}

ridge_final_model, y_test_predicted = build_model(X_train_rfe, y_train, X_test_rfe, params, model='ridge')
```

```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
Optimum alpha for ridge is 10.000000
ridge  Regression with  10.0
==================================
R2 score (train) :  0.9166751151617185
R2 score (test) :  0.8704201226046138
RMSE (train) :  0.11304414693007461
RMSE (test) :  0.15390088041290242
```

With Lasso Regression Alpha value is 0.001

In [89]:
```
#Lasso Regression
params = {'alpha': [0.000001, 0.00001,0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 500, 1000, 10000]}

lasso_final_model, y_test_predicted = build_model(X_train_rfe, y_train, X_test_rfe, params, model='lasso')
```

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Optimum alpha for lasso is 0.001000
lasso  Regression with  0.001
==================================
R2 score (train) :  0.9157339730212566
R2 score (test) :  0.8745163217343286
RMSE (train) :  0.11368076274295139
RMSE (test) :  0.15144883684391258
```

The Most important predicted variable is below with Lasso value doubled (0.002).

In [101]:
```
model_coefficients.sort_values(by='Lasso (alpha = 0.0002)', ascending=False).head(1)
```

Out[101]:

	Ridge (alpha=20.0)	Lasso (alpha=0.002)	Ridge (alpha = 20.0)	Lasso (alpha = 0.0002)
1stFlrSF	0.125737	0.126662	0.122502	0.128749

The Most Important predicted variables is below with Ridge Value 20.0

```
In [103]: model_coefficients.sort_values(by='Ridge (alpha = 20.0)', ascending=False).head(1)
```

Out[103]:

	Ridge (alpha=20.0)	Lasso (alpha=0.002)	Ridge (alpha = 20.0)	Lasso (alpha = 0.0002)
1stFlrSF	0.125737	0.126662	0.122502	0.128749

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: We will choose Lasso Regression as R2_score is greater on Test data as compare to Ridge.

3) After building the model, you realized 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 below snippet is for Top 5 important predictor variable

```
In [101]: # Top 5 featues in Lasso final model

          model_coefficients.sort_values(by='Lasso (alpha=0.001)', ascending=False).head(5)
```

Out[101]:

	Ridge (alpha=10.0)	Lasso (alpha=0.001)	Ridge (alpha = 20.0)	Lasso (alpha = 0.0002)	Lasso (alpha = 0.002)
1stFlrSF	0.125737	0.126662	0.122502	0.124616	0.124616
2ndFlrSF	0.106067	0.105968	0.103016	0.102877	0.102877
OverallQual	0.078235	0.080717	0.078339	0.083190	0.083190
OverallCond	0.048387	0.049037	0.047556	0.047834	0.047834
SaleCondition_Partial	0.034283	0.033529	0.033619	0.031926	0.031926

Now the above variable needs to dropped and created another model

And below is the result.

In [106]:
```
model_coeff = pd.DataFrame(index=X_test_new.columns)
model_coeff.rows = X_test_new.columns
model_coeff['Lasso'] = lasso_model.coef_
model_coeff.sort_values(by='Lasso', ascending=False).head(5)
```

Out[106]:

	Lasso
GarageArea	0.077538
KitchenQual	0.063253
LotArea	0.060798
Fireplaces	0.060263
BsmtQual	0.044828

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?

Answer: A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not give to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations)or can be used IQR as well. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis