Advanced Regression Subjective Questions

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:

According to the model built,

- Optimal Value of alpha for Ridge Regression is - 7.0
- Optimal Value of alpha for Lasso Regression is - 0.0004

Doubling the Alpha Values

Ridge Model

```
[73]: #Ridge Regression
ridge2x = Ridge(alpha=14)

ridge2x.fit(x_train, y_train)

y_pred_r_train = ridge2x.predict(x_train)
y_pred_r_test = ridge2x.predict(x_test)

metric_ri_2x = displayR2_RSS_MSE(y_pred_r_train, y_pred_r_test)
```

```
R2_Train : 0.8730889759035121
R2_Test : 0.8304667285979249

RSS_Train : 1.5058084124715343
RSS_Test : 0.805113640358311

MSE_Train : 0.001560423225359103
MSE_Test : 0.001944718938063553
```

```
574]: #Lasso Regression
lasso2x = Lasso(alpha=0.0008)

lasso2x.fit(x_train, y_train)

y_pred_l_train = lasso2x.predict(x_train)
y_pred_l_test = lasso2x.predict(x_test)

metric_la_2x = displayR2_RSS_MSE(y_pred_l_train,y_pred_l_test)
```

```
R2_Train : 0.8404789610386157
R2_Test : 0.8075377002866366

RSS_Train : 1.892728580076908
RSS_Test : 0.9140036139953919

MSE_Train : 0.0019613767669190755
MSE_Test : 0.002207738198056502
```

```
In [577]: rg_2x_metric = pd.Series(metric_ri_2x, name = 'Ridge 2x')
ls_2x_metric = pd.Series(metric_la_2x, name = 'Lasso 2x')

#final_metric = pd.concat([final_metric, pd.Series(metric_ri_2x, name = 'Ridge 2x'), pd.Series(metric_la_2x, name = '
final_metric_2x = pd.concat([lr_metric, rg_metric, ls_metric,rg_2x_metric,ls_2x_metric], axis = 1)

final_metric_2x
```

Out[577]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge 2x	Lasso 2x
0	R2 Score (Train)	9.026104e-01	0.886441	0.869213	0.873089	0.840479
1	R2 Score (Test)	-4.548738e+17	0.834696	0.832862	0.830467	0.807538
2	RSS (Train)	1.155535e+00	1.347389	1.551793	1.505808	1.892729
3	RSS (Test)	2.160196e+18	0.785029	0.793738	0.805114	0.914004
4	MSE (Train)	1.197445e-03	0.001396	0.001608	0.001560	0.001961
5	MSE (Test)	5.217865e+15	0.001896	0.001917	0.001945	0.002208

```
In [ ]:
```

Once we double the alpha values both Ridge and lasso regression performance slightly decreased.

Important Variables after the Change

Ridge_2x_Imp Variables	Lasso_2x_Imp_Variables
KitchenQual_TA	KitchenQual_TA
KitchenQual_Gd	KitchenQual_Gd
MSSubClass	KitchenQual_Gd

GarageCars	GrLivArea
OverallQual	Neighborhood_NoRidge
GrLivArea	Neighborhood_NoRidge

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:

I would prefer to go with Lasso Regression, because of the following reasons.

1. The accuracy of the test set is better when compared to Ridge
2. No. of features is reduced, which reduces the complexity of the 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:

Code

```
#Removing above columns from X_train
x_train_copy = x_train #copying the dataset
x_train_copy.drop(['GrLivArea','OverallQual','GarageCars','Neighborhood_NoRidge
','TotRmsAbvGrd'], axis=1,inplace=True)
#Removing above columns from x_test
x_test_copy = x_test #copying the dataset
x_test_copy.drop(['GrLivArea','OverallQual','GarageCars','Neighborhood_NoRidge',
'TotRmsAbvGrd'], axis=1,inplace=True)
#Lasso Model
lassoQ3 = Lasso(alpha=0.0004)
lassoQ3.fit(x_train_copy, y_train)
```

```
y_pred_3_train = lassoQ3.predict(x_train_copy)
y_pred_3_test = lassoQ3.predict(x_test_copy)
metric_la_Q3 = displayR2_RSS_MSE(y_pred_3_train,y_pred_3_test)
ls_3_metric = pd.Series(lassoQ3.coef_,index=x_train_copy.columns)
ls_3_metric.sort_values()[:5]
```

According to above data after removing earlier top 5 variables and the current 5 most important variables are:
1. "1stFlrSF"
2. "2ndFlrSF"
3. "GarageArea"
4. "OverallCond"
5. "FullBath"

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:

A robust and generalized model will always have a balance between the model complexity and Simplicity and accuracy of the model on the unseen data(test data) Implications are as follows:

1. Trade-off/Balance between bias and variance
2. Reducing the overfitting of data
3. Performance of model on unseen data