Advanced Regression Assignment

Problem Statement - Part II

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 developed :

- Ridge Regression Model --- **alpha = 1.0**
- Lasso Regression Model --- **alpha = 0.0001**

After doubling the alpha parameters:

- Ridge Regression Model --- **alpha = 2.0**
- Lasso Regression Model --- **alpha = 0.0002**

For Ridge Regression Model:

```
For Ridge Regression Model (Original Model, alpha=1.0):
 ============================================================

For Train Set:
R2 score: 0.9072
MSE score: 0.0022
MAE score: 0.0357
RMSE score: 0.047
 ============================================================

For Test Set:
R2 score: 0.8895
MSE score: 0.0029
MAE score: 0.0408
RMSE score: 0.054
 ============================================================
```

```
For Ridge Regression Model (New Model, alpha=2.0):
 ============================================================

For Train Set:
R2 score: 0.7999
MSE score: 0.0048
MAE score: 0.0522
RMSE score: 0.069
 ============================================================

For Test Set:
R2 score: 0.7716
MSE score: 0.006
MAE score: 0.0556
RMSE score: 0.0776
 ============================================================
```

```
For Ridge Regression (Doubled alpha model, alpha=2):
 ============================================================================================================================================================
The most important predictor variables after the change is implemented are as follows:

 ['BathroomNos', 'GarageCars', 'Neighborhood_NridgHt', 'LotArea', 'PorchArea', 'Neighborhood_StoneBr', 'Neighborhood_Crawfor',
'SaleType_New', 'OverallQual_8', 'Neighborhood_NoRidge']
 ============================================================================================================================================================
```

For Lasso Regression Model:

```
For Lasso Regression Model (Original Model: alpha=0.0001):
 ============================================================

For Train Set:
R2 score: 0.8
MSE score: 0.0048
MAE score: 0.0526
RMSE score: 0.069
 ============================================================

For Test Set:
R2 score: 0.7747
MSE score: 0.0059
MAE score: 0.0555
RMSE score: 0.0771
 ============================================================
```

```
For Lasso Regression Model (New Model: alpha=0.0002):
 ============================================================

For Train Set:
R2 score: 0.7966
MSE score: 0.0048
MAE score: 0.053
RMSE score: 0.0696
 ============================================================

For Test Set:
R2 score: 0.7751
MSE score: 0.0059
MAE score: 0.0551
RMSE score: 0.077
 ============================================================
```

```
For Lasso Regression (Doubled alpha model: alpha:0.0002):
 ==========================================================================================================================================
The most important predictor variables after the change is implemented are as follows:

 ['BathroomNos', 'GarageCars', 'LotArea', 'Neighborhood_NridgHt', 'PorchArea', 'Neighborhood_StoneBr', 'Neighborhood_Crawfor',
'SaleType_New', 'OverallQual_8', 'Neighborhood_NoRidge']
 ==========================================================================================================================================
```

#Please find the code block in the jupyter notebook.

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:

On comparing the ridge and lasso regression model on basis of the evaluation metrics we have the following:

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.9055	0.9032
R2 Score (Test)	0.8926	0.8971
MSE (Train)	0.0022	0.0023
MSE (Test)	0.0028	0.0027
MAE (Train)	0.0359	0.0363
MAE (Test)	0.0403	0.0394
RMSE (Train)	0.0474	0.0480
RMSE (Test)	0.0532	0.0521

It is evident that both the models perform nearly equally but on keen observation it can be seen that Lasso Model performs slightly better in the test set in all the parameters including R2 Score, MSE, MAE & RMSE, outperforming ridge model in the competition. Also Lasso model reduces the coefficients of less important features to 0 whereas Ridge model tends to take the coefficient to 0. Hence, Lasso Model is more preferable.

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:

The Top 5 Lasso Features were:

- TotalArea
- GrLivArea
- BuildingAge
- OverallQual_9
- OverallQual_10

After changing the incoming dataset, the evaluation metrics stands as follows:

```
For Lasso Regression Model (alpha=0.0001):
 ================================================================================

For Train Set:
R2 score: 0.8
MSE score: 0.0048
MAE score: 0.0526
RMSE score: 0.069
 ================================================================================

For Test Set:
R2 score: 0.7747
MSE score: 0.0059
MAE score: 0.0555
RMSE score: 0.0771
 ================================================================================
```

```
For Lasso Regression alpha:0.0001:
 ================================================================================
The Five most important predictor variables after     changing the data:

 ['BathroomNos', 'GarageCars', 'LotArea', 'Neighborhood_NridgHt', 'PorchArea']
 ================================================================================
```

Question 4

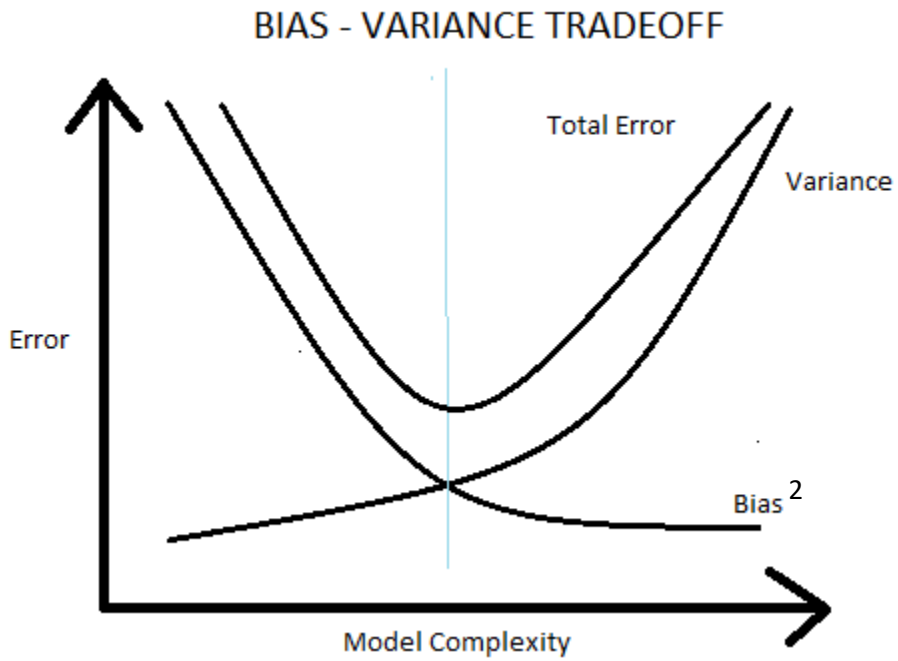
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

In general terms, a model's robustness is the degree by which its performance changes when exposed to a new unseen test data versus what has been observed when it was run on train data.

Accuracy and robustness lie at odds at most of the times as a model which is too accurate on the training data tends to learn all the corresponding features from the data and hence tends to over-fit, thus when it sees a new test data then the model's accuracy decreases.

To deal with this situation we use Bias-Variance Trade-off.



A model which is too simple will invite more bias and a model which is more complex and has more number of features will invite more variance. Thus we need to find the optimum balance between the two in order to make generalizable models.

The total error is given as:

Total Error = Bias^2 + Variance + Irreducible error

The aim of model building is to reduce this Total error. Here the regularisation measures find importance. Both the Lasso and Ridge regularisation methods help to achieve the same.