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:
##Optimal Value of alpha for ridge and lasso regression observed in the model are as follows :

optimal_alpha_ridge = 5.0 #(Computed: For Ridge Regression)
optimal_alpha_lasso = 0.001 #(Computed: For Lasso Regression)

Changes in the model, if the values of alpha are doubled
ARidge Regression
For Ridge Regression Model (Original Model, alpha=5.0):
****

For Train Set:
R2 score: 0.891712413808081
MSE score: 0.1082875861919191
MAE score: 0.228675341222659
RMSE score: 0.3290707920674807

For Test Set:
R2 score: 0.8814796440865079
MSE score: 0.14105927847051095
MAE score: 0.23760191414721143
RMSE score: 0.3755785916030238
****
After Doubling the Alpha
For Ridge Regression Model (Doubled alpha model, alpha=5*2=10):
****

For Train Set:
R2 score: 0.8885146749529405
MSE score: 0.1114853250470595
MAE score: 0.23032302231958063
RMSE score: 0.3338941824097262

For Test Set:
R2 score: 0.8802846986589267
MSE score: 0.1424814657270973

MAE score: 0.23813172897915758
RMSE score: 0.3774671717210615
****

Observations :

1. Test accuracy of the original model with alpha value of 5 is slightly higher than new model.
2. MSE, MAE and RMSE are slightly lower for the original model with alpha 5
3. Train and test data r2 scores are better in original model
4. Increase in value of alpha in the model leads to decrease in R2 score and increase in Error terms, making the original alpha model a better choice

B. Lasso Regression

For Lasso Regression Model (Original Model: alpha=0.001):
****

For Train Set:
R2 score: 0.8911452854668185
MSE score: 0.10885471453318159
MAE score: 0.230058968350782
RMSE score: 0.32993137852162774

For Test Set:
R2 score: 0.8823974629412124
MSE score: 0.13996691872847866
MAE score: 0.23839444448391264
RMSE score: 0.3741215293570776
****

For Lasso Regression Model: (Doubled alpha model: alpha:0.001*2 = 0.002)
****

For Train Set:
R2 score: 0.8872361608101954
MSE score: 0.11276383918980457
MAE score: 0.23266369677452692
RMSE score: 0.3358032745370488

For Test Set:
R2 score: 0.8810587131585922
MSE score: 0.14156025750082246
MAE score: 0.238903497800735
RMSE score: 0.37624494348870985
****

Observations

1. Increase in value of alpha causes decrease in the value of R2 Score and increase in the

value of MSE, MAE and RMSE, making original model with alpha value of 0.001 a better choice

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?

A. Ridge Regression

For Ridge Regression Model (Original Model, alpha=5.0):

****

For Train Set:

R2 score: 0.891712413808081

MSE score: 0.1082875861919191

MAE score: 0.228675341222659

RMSE score: 0.3290707920674807

For Test Set:

R2 score: 0.8814796440865079

MSE score: 0.14105927847051095

MAE score: 0.23760191414721143

RMSE score: 0.3755785916030238

****

B. Lasso Regression

For Lasso Regression Model (Original Model: alpha=0.001):

****

For Train Set:

R2 score: 0.8911452854668185

MSE score: 0.10885471453318159

MAE score: 0.230058968350782

RMSE score: 0.32993137852162774

For Test Set:

R2 score: 0.8823974629412124

MSE score: 0.13996691872847866

MAE score: 0.23839444448391264

RMSE score: 0.3741215293570776

****

Considering these results, we can select Lasso.

Explanation for the lasso Model selection :

1. R2 score with Lasso Regression is slightly better than Ridge Regression Mode.
2. Training accuracy is slightly lesser in Lasso
3. Lasso seems work a bit better on unseen data
4. Also Lasso help in feature selection by reducing the coefficient values of insignificant

features towards 0.

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?

Ans

1. Top 5 Original features are
    a. 'Neighborhood_StoneBr',
    b. 'Neighborhood_NridgHt',
    c. 'GrLivArea',
    d. 'Neighborhood_NoRidge',
    e. 'SaleCondition_Partial'
2. Top 5 features after removing above features are
    a. ['GrLivArea',
    b. 'Foundation_Slab',
    c. 'MSSubClass_90',
    d. 'MSSubClass_160',
    e. 'MSSubClass_30']

Question-4:

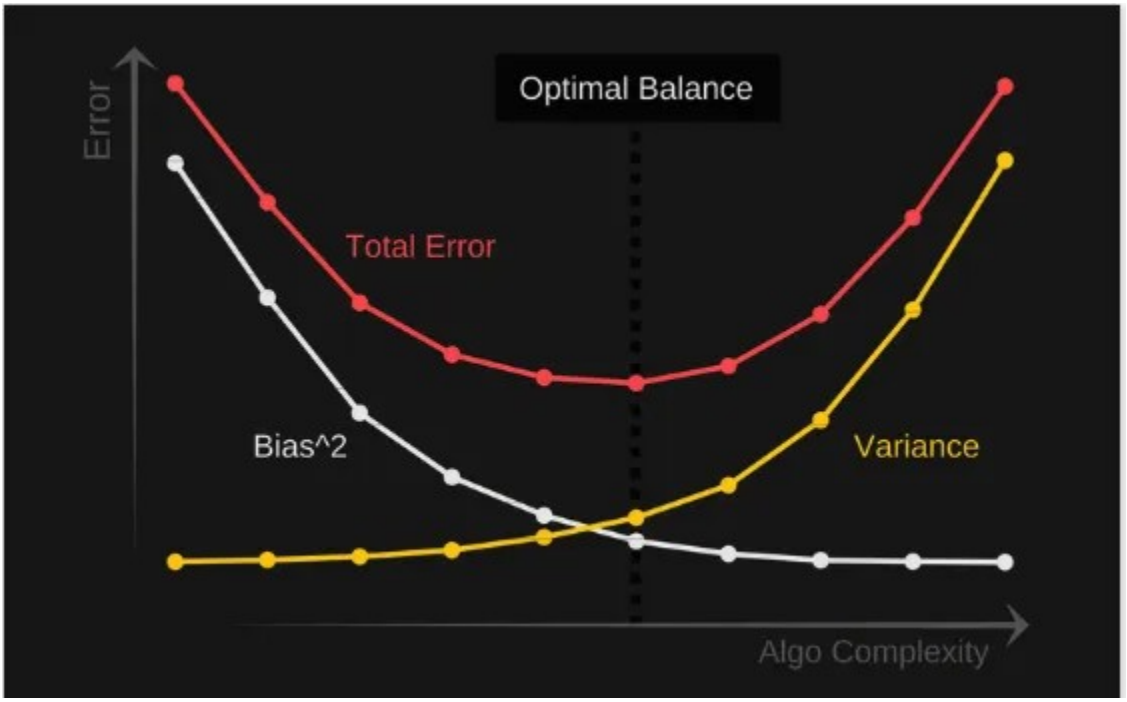
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

A model needs to be made robust and generalizable so that they are able to work fine even with the unseen and untrained data.If Model is too complex and remembers all the training data, it will work really great on the train dataset but may fail predicting the real work data.

Also model should be kept simple. So that the model performs equally well on the training and test data. Hence there are couple of terms :Bias and variance.

Bias refers to the difference between the predicted values of the model and the true values. In other words, bias measures how much the predictions of the model differ from the actual values of the data. A model with high bias tends to underfit the data and is unable to capture the underlying patterns in the data.

Variance refers to the variability of the model's predictions for different training sets. In other words, variance measures how much the predictions of the model vary for different data points. A model with high variance tends to overfit the data and is too complex, capturing noise in the data and not generalizing well to new data.



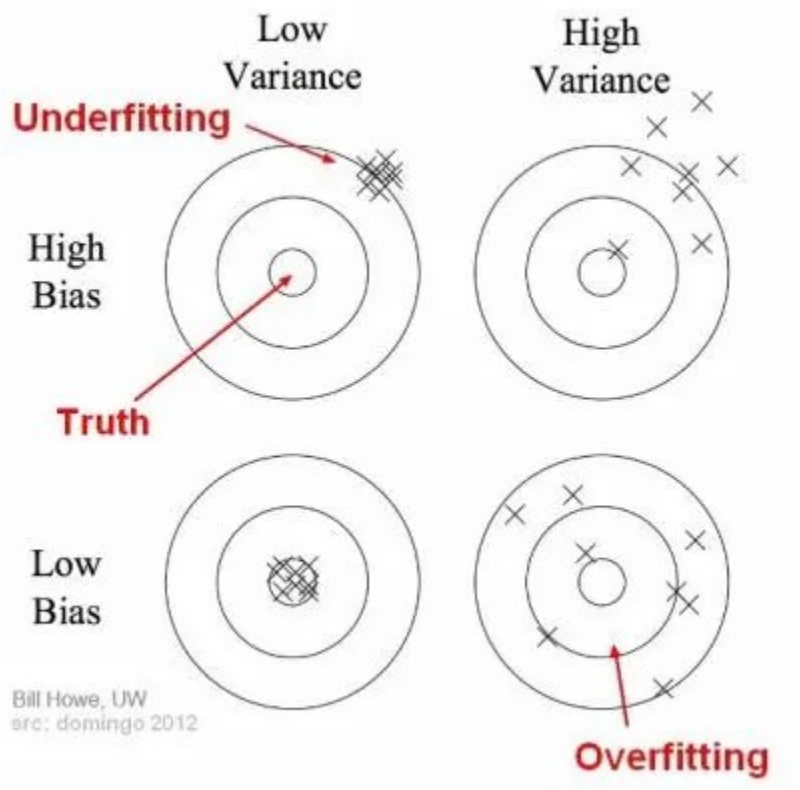
The above image shows the relationship between model complexity and error.

High Bias, Low Variance:The Left side (lower value of x) with Low Model Complexity shows high bias and low variance. The model is too simple and fails to capture the true relationship between the input and output data. As a result, it has a high training error and a high test error. This type of model is said to underfit the data.

Low Bias, High Variance: The Right side (higher value of x) with High Model Complexity shows a model with low bias and high variance. The model is too complex and captures noise in the data. As a result, it has low training error but a high test error . This type of model is said to overfit the data.

Balanced Model: Theideal value lies for low variance and low bias.. As a result, it has low training error and low test error.
Overall, it is important to strike a balance between bias and variance when training a machine learning model to achieve good performance and generalization.

The diagram below throws more light on the cases described above.


Low
Variance
High
Variance
Underfitting
High
Bias
Truth
Low
Bias
Bill Howe, UW
src: domingo 2012
Overfitting