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:

- Optimal value of Alpha
  - Ridge: 4
  - Lasso: .001
- Change in model if we double the values of alpha in Ridge and Lasso

Ridge	Lasso
For Ridge Regression Model (Doubled alpha model, alpha=4*2=8): ********** For Train Set: R2 score: 0.8979907381055817 MSE score: 0.10200926189441831 MAE score: 0.21738027947892058 RMSE score: 0.31938888818244493 For Test Set: R2 score: 0.8918476817745514 MSE score: 0.1004938891015741 MAE score: 0.2259149762361879 RMSE score: 0.31700771142288336 **********	For Lasso Regression Model: (Doubled alpha model: alpha:0.001*2 = 0.002) ********** For Train Set: R2 score: 0.8962362283965272 MSE score: 0.10376377160347287 MAE score: 0.2192566387666148 RMSE score: 0.3221238451333165 For Test Set: R2 score: 0.8922281355114673 MSE score: 0.10014037586881841 MAE score: 0.22777127811534406 RMSE score: 0.316449641916085 **********
For Ridge Regression Model (Original Model, alpha=4): ********** For Train Set: R2 score: 0.8979907381055817 MSE score: 0.10200926189441831 MAE score: 0.21738027947892058 RMSE score: 0.31938888818244493 For Test Set: R2 score: 0.8918476817745514 MSE score: 0.1004938891015741 MAE score: 0.2259149762361879 RMSE score: 0.31700771142288336 **********	For Lasso Regression Model (Original Model: alpha=0.001): ********** For Train Set: R2 score: 0.8991175807585077 MSE score: 0.10088241924149234 MAE score: 0.21678903513244718 RMSE score: 0.3176199289111002 For Test Set: R2 score: 0.8903426824542664 MSE score: 0.10189231714520969 MAE score: 0.228504671845209 RMSE score: 0.31920575988727035 **********

Observations:

- In case of Ridge changing alpha from 4 to 8 has not changed any results in test dataset
- Incase of Lasso changing alpha from .001 to .002 has slightly decreased the test scores. Which is very minimal

Below are the most important predictor variables post doubling the alpha

```
For Ridge Regression (Doubled alpha model, alpha=4*2=8):
 ****************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['Neighborhood_StoneBr', 'GrLivArea', 'Neighborhood_NridgHt', 'Neighborhood_Crawfor', 'MSSubClass_160', 'Exterior1
st_BrkFace', 'MSSubClass_120', 'Neighborhood_BrkSide', 'OverallQual', 'AgeofProperty']
 ****************************************************************************************************
```

```
For Lasso Regression (Doubled alpha model: alpha:0.001*2 = 0.002):
 ****************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['Neighborhood_StoneBr', 'Neighborhood_NridgHt', 'Neighborhood_Crawfor', 'GrLivArea', 'MSSubClass_160', 'MSSubClas
s_120', 'Exterior1st_BrkFace', 'Neighborhood_BrkSide', 'MSSubClass_90', 'AgeofProperty']
 ****************************************************************************************************
```

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:

If we compare all the scores of test data then

- R2 score of Lasso is slightly lower than Ridge
- MSE score is slight higher for Lasso compare to Ridge
- MAE score is slight higher for lasso than Ridge
- RMSE score for lasso is slight higher than Ridge

As the results are almost same upto 2nd decimal points so its not easy to select one based on score. But would recommend Lasso as it does feature elimination and becomes easy to explain.

Ridge	Lasso
For Ridge Regression Model (Original Model, alpha=4): ********** For Train Set: R2 score: 0.8979907381055817 MSE score: 0.10200926189441831 MAE score: 0.21738027947892058 RMSE score: 0.31938888818244493 For Test Set: R2 score: 0.8918476817745514 MSE score: 0.1004938891015741 MAE score: 0.2259149762361879 RMSE score: 0.31700771142288336 **********	For Lasso Regression Model (Original Model: alpha=0.001): ********** For Train Set: R2 score: 0.8991175807585077 MSE score: 0.10088241924149234 MAE score: 0.21678903513244718 RMSE score: 0.3176199289111002 For Test Set: R2 score: 0.8903426824542664 MSE score: 0.10189231714520969 MAE score: 0.228504671845209 RMSE score: 0.31920575988727035 **********

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:

```
Top 5 features in original lasso model (dropped):
 ['Neighborhood_StoneBr', 'Neighborhood_NridgHt', 'Neighborhood_Crawfor', 'GrLivArea', 'MSSubClass_160']
```

Below are the list of top 5 features by removing the first 5 features

```
For New Lasso Regression Model (After eliminating the top5 features from the original model):
**********
The top5 new most important predictor variables are as follows:

 ['Exterior1st_BrkComm', 'Exterior1st_AsphShn', 'Neighborhood_MeadowV', 'Foundation_Slab', 'Neighborhood_Gilbert']
**********
```

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?

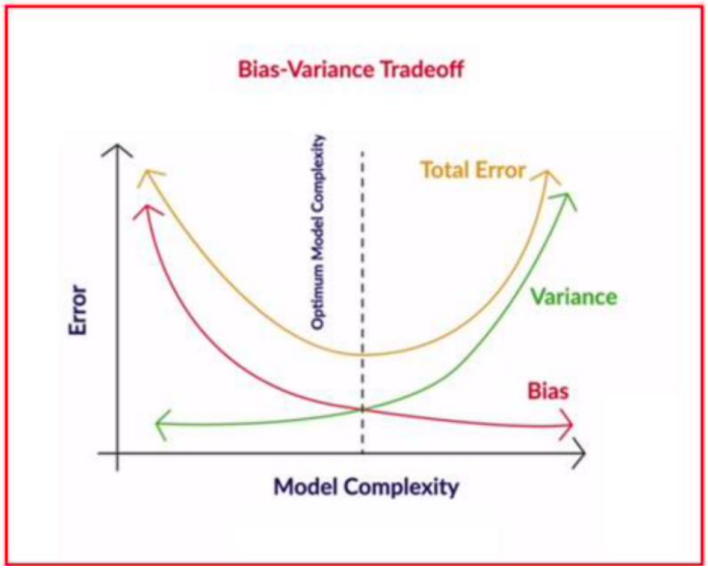
Answer:

Robustness of a model implies, either the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise to the dataset. By the implementing regularization techniques, we can control the trade-off between model complexity and bias which is directly connected the robustness of the model. Regularization, helps in penalizing the coefficients for making the model too complex; thereby allowing only the optimal amount of complexity to the model.

Ensuring that a model is robust and generalizable involves several practices and considerations:

- Cross-Validation: Use techniques like k-fold cross-validation to assess how well the model performs on unseen data. By splitting the dataset into multiple subsets and training the model on different combinations of training and validation sets, you can evaluate its performance across various data samples.
- Regularization: Apply techniques like Ridge or Lasso regression to prevent overfitting and improve the model's ability to generalize to new data. Regularization penalizes overly complex models, helping to reduce variance and improve robustness.

Bias helps you quantify, how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there has to be enough training data. Models that are too naïve, for e.g., one that gives same results for all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high. Variance is the degree of changes in the model itself with respect to changes in the training data. Thus, accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.



Hence, there exists a trade-off between accuracy and robustness, wherein an excessively accurate model is susceptible to overfitting. Consequently, while such a model may demonstrate high accuracy on the training data, its performance may falter when applied to real-world data, or conversely