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 values of Alpha:
   - The computed optimal value of Alpha for Ridge Regression (Original Model) is 10
   - The computed optimal value of Alpha for Lasso Regression (Original Model) is 0.0004

2. Changes in the model if you choose double the values of alpha for ridge and lasso:
   - Ridge Regression:
     Original Model (alpha is 10) & Doubled Alpha Model (alpha is 20)

```
For Ridge Regression Model (Original Model, alpha=10)
 ****************************************

For Train Set:
R2 score: 0.9178724453245637
MSE score: 0.01059018649480147
MAE score: 0.0719744676316332
RMSE score: 0.10290863177985349

For Test Set:
R2 score: 0.9061419322503745
MSE score: 0.013532056542967956
MAE score: 0.0777355990757612
RMSE score: 0.11632736798779536
```

```
For Ridge Regression Model (Doubled alpha model, alpha=10*2=20)
 ****************************************

For Train Set:
R2 score: 0.9174534845036636
MSE score: 0.010644210667868362
MAE score: 0.0721308123809888
RMSE score: 0.1031707839839766

For Test Set:
R2 score: 0.9064886778155941
MSE score: 0.013482064243881956
MAE score: 0.07780697383931404
RMSE score: 0.11611229152799439
```

Observations:

- The test accuracy of the ridge regression model (alpha=20) is bit higher than the ridge regression model (alpha=10).
- The MSE for test set of the doubled model is slightly higher than the original model MSE.
- Ridge Regression model (Doubled Alpha Model) is performing better on the train and test data in comparison to the single alpha model.
- Increase in the value of alpha in the model, increases the R2 score but decrease in the MSE (causing higher penalty to the coefficients). The doubled alpha model is performing better than the original alpha model.

- Lasso Regression:
  Original Model (alpha is 0.0004) & Doubled Alpha Model (alpha is 0.0008)

```
For Lasso Regression Model (Original Model: alpha=0.0004)
 ****************************************

For Train Set:
R2 score: 0.9177743523155614
MSE score: 0.010602835395201288
MAE score: 0.07195535678240425
RMSE score: 0.10297007038553138

For Test Set:
R2 score: 0.9075697845256938
MSE score: 0.013326194881973947
MAE score: 0.07720019976804227
RMSE score: 0.11543913929848033
```

```
For Lasso Regression Model: (Doubled alpha model: alpha:0.0004*2 = 0.0008)
 ************************************

For Train Set:
R2 score: 0.9170249806997015
MSE score: 0.0106994653898143
MAE score: 0.07218900206278267
RMSE score: 0.10343822015973739

For Test Set:
R2 score: 0.9090910821335023
MSE score: 0.013106860670849322
MAE score: 0.0769991824014096
RMSE score: 0.11448519847932012
```

Observations:

- The test accuracy of the lasso regression doubled alpha model (alpha=0.0008) is slightly higher compared to the test accuracy of the original alpha model (alpha=0.0004).
- MSE test scores comparing the similar data of the doubled alpha model is slightly lower than the original alpha model.
- Lasso Regression model (doubled alpha model) seems to be performing better than when comparing to the original alpha model.
- Increase in the alpha in the model lead to increase in the R2 score but decrease in the MSE value (causing more penalty to the coefficients). In Lasso, the insignificant coefficients that have their values near to the zero correspond to zero values (performing feature selection in the model). Thus, making the doubled alpha model a better choice.

3. The most important predictor variables after the change is implemented. Top 10 features are as follows:
    - Ridge Regression (doubled alpha=20)

```
For Ridge Regression (Doubled alpha model: alpha:10*2 = 20):
 ************************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['LotFrontage', 'OverallQual', 'OverallCond', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF',
'LowQualFinSF']
```

    - Lasso Regression (doubled alpha=0.0008)

```
For Lasso Regression (Doubled alpha model: alpha:0.0004*2 = 0.0008):
 ************************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['LotFrontage', 'OverallQual', 'OverallCond', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF',
'LowQualFinSF']
```

The predictor for both the Ridge and Lasso Regression are same but the coefficients are different.

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:

Optimal values of Alpha:

- The computed optimal value of Alpha for Ridge Regression (Original Model) is 10
- The computed optimal value of Alpha for Lasso Regression (Original Model) is 0.0004

```
For Ridge Regression Model (Original Model, alpha=10)
 ************************************

For Train Set:
R2 score: 0.9178724453245637
MSE score: 0.01059018649480147
MAE score: 0.0719744676316332
RMSE score: 0.10290863177985349

For Test Set:
R2 score: 0.9061419322503745
MSE score: 0.013532056542967956
MAE score: 0.0777355990757612
RMSE score: 0.11632736798779536
```

```
For Lasso Regression Model (Original Model: alpha=0.0004)
 ************************************

For Train Set:
R2 score: 0.9177743523155614
MSE score: 0.010602835395201288
MAE score: 0.07195535678240425
RMSE score: 0.10297007038553138

For Test Set:
R2 score: 0.9075697845256938
MSE score: 0.013326194881973947
MAE score: 0.07720019976804227
RMSE score: 0.11543913929848033
```

Observations:

- The R2 test score of the Lasso Regression model is slightly better than that of Ridge Regression model. On the other hand, the training accuracy has slightly reduced. Thus, making the lasso model an optimal choice as it will perform better on the unseen data
- The MSE for Test set (Lasso Regression) is slightly lower than the Ridge Regression model. This shows that the Lasso Regression will perform better on the unseen data. Since Lasso helps in feature selection (the coefficient values of some of the insignificant predictor variables become zero). Hence, shows that the Lasso Regression has a better edge than Ridge Regression. Therefore, the variables predicted by Lasso can be applied in order to choose significant variables for predicting the price of a house.

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 (before dropping) are below:

```
['LotFrontage', 'OverallQual', 'OverallCond', 'BsmtFinSF1', 'BsmtFinSF2']
```

Top 5 features variables in the new Model are below:

```
['GrLivArea',
 'TotalBsmtSF',
 'GarageType_Attchd',
 'GarageArea',
 'SaleCondition_Partial']
```

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:

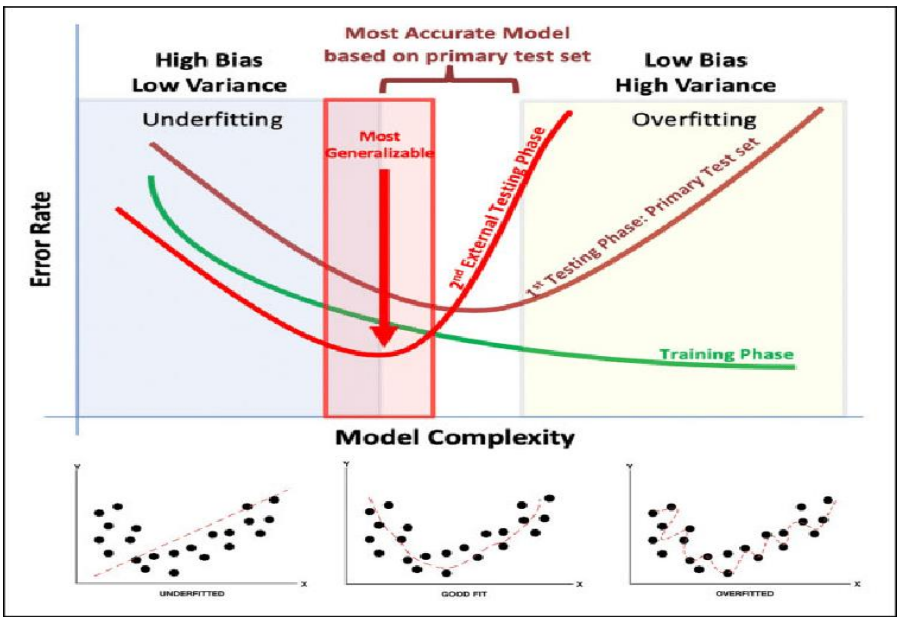
Robustness of a model implies, either the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise in the dataset. On the other hand, robustness (or generalisable) of a model is a measure of its successful application to the dataset other than the one used for training and testing.

By implementing regularisation techniques, we can control the trade-off between model complexity and bias which is directly connected to the robustness of the model. Regularization, helps in penalizing the coefficients for making the model too complex. And allowing only the optimum amount of complexity to the model. It helps in controlling the robustness of the model by making the model simpler. Therefore, in order to make model more robust and generalisable, one need to make sure that the balance is created between the bias and complexity. Also, making a model simple lead to Bias-Variance Trade-off.

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any change in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change widely even if more points are added or removed.

Bias helps to quantify; how accurate the model is likely to respond on the test data. A complex model can do an accurate prediction provided there has to be enough training data. Model which are too naive for e.g., one that gives same results for all the test inputs and make 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.

Accuracy of the model can be maintained by keeping the balance between the Bias and Variance as it minimizes the total error.



Thus, accuracy and robustness may be at the odds to each other as too much accurate model can be a prey to overfitting, which means that it can be too much accurate on the train data but fails when it faces the unseen data or vice versa.