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. The optimal value of alpha (Ridge Model) : 3
   The optimal value of alpha (Lasso Model) : 0.006

2. Pls refer to the Ipynb Notebook for the calculation regarding the same

RIDGE REGRESSION

```
For Ridge Regression Model (Original Model, alpha=3.0):
 ****************************************

For Train Set:
R2 score: 0.9096736794426218
MSE score: 0.09032632055737826
MAE score: 0.21328703059683138
RMSE score: 0.30054337550073906

For Test Set:
R2 score: 0.8852491134043146
MSE score: 0.11385202736883343
MAE score: 0.2361829366123452
RMSE score: 0.33741966061395035
 ****************************************
```

```
For Ridge Regression Model (Doubled alpha model, alpha3*3=9):
 ****************************************

For Train Set:
R2 score: 0.9066699216586404
MSE score: 0.09333007834135966
MAE score: 0.21658338787483453
RMSE score: 0.30549971905283263

For Test Set:
R2 score: 0.885470656694236
MSE score: 0.11363221945748903
MAE score: 0.23597865066579454
RMSE score: 0.33709378436495835
 ****************************************
```

Observations :

- The R2 Train Accuracy is slightly higher when lower Alpha Model, similarly for the MAE Score which is slightly better in training, which explains why this alpha has been chosen
- MAE is higher for higher Alpha model which is our test measurement
  Other Significant Changes are not seen in the model for the Test Data on increasing Alpha

LASSO REGRESSION

```
For Lasso Regression Model (Original Model: alpha=0.0006):
 ****************************************

For Train Set:
R2 score: 0.9085574638360507
MSE score: 0.09144253616394922
MAE score: 0.2150839979509765
RMSE score: 0.3023946695362688

For Test Set:
R2 score: 0.8859459960943489
MSE score: 0.11316060345523694
MAE score: 0.2361529382068441
RMSE score: 0.3363935246927874
 ****************************************
```

```
For Lasso Regression Model: (Doubled alpha model: alpha:0.001*2 = 0.0012)
 ****************************************

For Train Set:
R2 score: 0.9058443893093867
MSE score: 0.09415561069061332
MAE score: 0.2173018211538171
RMSE score: 0.30684786245078083

For Test Set:
R2 score: 0.8845325928962782
MSE score: 0.1145629352747454
MAE score: 0.2376162295487515
RMSE score: 0.33847146892278146
 ****************************************
```

Observations :

- The R2 Train Accuracy is slightly higher when lower Alpha Model, similar for the MAE Score which is slightly better in training, which explains why this alpha has been chosen
- The test data has better metrics in all areas, which means that the Alpha we have selected is the best fit, and overfitting has not happened in this scenario

3. **The Most Important Predictor Variables after the change is implemented:**

```
For Lasso Regression (Doubled alpha model: alpha:0.006*2 = 0.0012):
 ************************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['HouseStyle_2.5Fin', 'HouseStyle_SFoyer', 'TotalBsmtSF', 'MSZoning_RM', 'GarageType_BuiltIn', 'Neighborhood_NPkVill', 'Exterior1st_Stone', 'Exterior1st_BrkFace', 'MSSubCl
ass_70', 'o_OverallQual']
 ************************************************************************************************************************

For Ridge Regression (Doubled alpha model, alpha=3*3=9):
 ************************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['GrLivArea', 'MSSubClass_160', 'Neighborhood_Crawfor', 'PropertyAge', 'Neighborhood_Somerst', 'o_OverallQual', 'TotalBsmtSF', 'MSZoning_RL', 'o_OverallCond', 'MSZoning_F
V']
 ************************************************************************************************************************
```

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

```
For Ridge Regression Model (Original Model, alpha=3.0):
 ****************************************

For Train Set:
R2 score: 0.9096736794426218
MSE score: 0.09032632055737826
MAE score: 0.21328703059683138
RMSE score: 0.30054337550073906

For Test Set:
R2 score: 0.8852491134043146
MSE score: 0.11385202736883343
MAE score: 0.2361829366123452
RMSE score: 0.33741966061395035
 ****************************************
```

```
For Lasso Regression Model (Original Model: alpha=0.0006):
 ****************************************

For Train Set:
R2 score: 0.9085574638360507
MSE score: 0.09144253616394922
MAE score: 0.2150839979509765
RMSE score: 0.3023946695362688

For Test Set:
R2 score: 0.8859459960943489
MSE score: 0.11316060345523694
MAE score: 0.2361529382068441
RMSE score: 0.3363935246927874
 ****************************************
```

Observations:

- The R2 Fit of the Lasso Model is a bit less than the Ridge Model in Train, but the R2 Fit is better in Test
- The MSE, and MAE Errors are also higher but the Test Values are lower. This indicates that the Lasso Model has fit the model a bit better by compromising a bit of bias for lower variance.
- We can also see that the L1 Norm would have simplified the model by using lower values coeffecients, which makes it a more simpler model.
- Moreover, while choosing a type of regression in the real world, an analyst has to deal with the lurking and confounding dangers of outliers, non-normality of errors and overfitting especially in sparse datasets among others. Using L2 norm (Ridge) results in exposing the analyst to such risks. Hence, use of L1 norm (Lasso) could be quite beneficial as it is quite robust to fend off such risks to a large extent, thereby resulting in better and robust regression models.

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 Features to be removed are :
**********************************************************************************************************************************
['MSSubClass_160', 'GrLivArea', 'Neighborhood_Crawfor', 'Neighborhood_Somerst', 'MSZoning_RL']
```

The top 5 features after removing the top Features are

```
For Lasso Regression (Remove 5 features model: alpha:0.0001):
 **********************************************************************************************************************************
The most important top10 predictor variables after the change is implemented are as follows:

 ['HouseStyle_2.5Fin', 'HouseStyle_SFoyer', 'TotalBsmtSF', 'MSZoning_RM', 'GarageType_BuiltIn', 'Neighborhood_NPkVill', 'Exterior1st_Stone', 'Exterior1st_BrkFace', 'MSSubCl
ass_70', 'o_OverallQual']
 **********************************************************************************************************************************
```

Question 4

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

Answer :

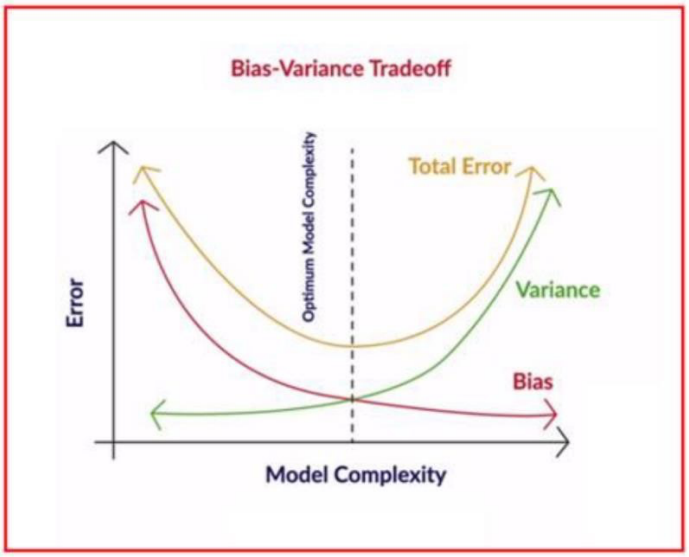
Robustness of a model implies, either the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise to the dataset. Thus, the robustness (or generalizability) of a model is a measure of its successful application to data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

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. It helps in controlling the robustness of the model by making the model optimal simpler. Therefore, in order to make the model more robust and generalizable, one need to make sure that there is a delicate balance between keeping the model simple and not making it too naive to be of any use. Also, making a model simple leads to BiasVariance 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 changes in the training data.
- ➢ A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

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.



Thus, accuracy and robustness may be at the odds to each other as too much accurate model can be prey to over fitting hence it can be too much accurate on train data but fails when it faces the actual data or vice versa.