Subjective Questions and Solutions

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?

Solution:

(Please refer ipynb for detailed code for the answers)

1. Optimal Value of Alpha:
   • The computed optimal value of alpha for Ridge Regression (Original Model): 2
   • The computed optimal value of alpha for Lasso Regression (Original Model): 0.002
2. Changes in the model, if you choose double the value of alpha for both ridge and lasso regression:
   (Please refer the jupyter (.ipynb) file for the code. Results are mentioned below)
   (i) Ridge Regression:
   Original Model (alpha=2), Doubled Alpha Model(alpha=4)

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

For Train Set:
R2 score: 0.9140467590425583
MSE score: 0.08595324095744175
MAE score: 0.20937283256494246
RMSE score: 0.2931778316268844

For Test Set:
R2 score: 0.8826471455614764
MSE score: 0.12629084639949664
MAE score: 0.25496240157816685
RMSE score: 0.35537423429322595
 ****************************************
```

```
For Ridge Regression Model (Doubled alpha model, alpha=2*2=4):
 ****************************************

For Train Set:
R2 score: 0.912770435968376
MSE score: 0.08722956403162396
MAE score: 0.21047879547179743
RMSE score: 0.295346515184493

For Test Set:
R2 score: 0.8832293629028906
MSE score: 0.12566428540796842
MAE score: 0.2538515727298641
RMSE score: 0.35449158721748025
 ****************************************
```

Observations:

• The test accuracy of the ridge regression model (alpha=2) is slightly higher in comparison to the test accuracy of the doubled alpha model (doubled alpha=4).
• MSE test scores comparing similar data of the original dataset and doubled alpha model gives us an idea that it is slightly smaller for the single alpha model than the doubled alpha model.
• Ridge Regression model (single alpha model) seems to perform better on the train and test data in comparison to the doubled alpha Ridge Regression model.
• Increase in the value of alpha in the model lead to a decrease in R2 score but an increase in the

MSE (causing more shrinkage of coefficient values). Thus, making the original (single) alpha model a better choice.

(ii) Lasso Regression:
Original Model (alpha=0.001), Doubled Alpha Model(alpha=0.002)

```
For Lasso Regression Model (Original Model: alpha=0.002):
 ****************************************

For Train Set:
R2 score: 0.9153651941577272
MSE score: 0.08463480584227281
MAE score: 0.20832015672340426
RMSE score: 0.29092061776758416

For Test Set:
R2 score: 0.8834784025370518
MSE score: 0.12539627806945317
MAE score: 0.2553770268160431
RMSE score: 0.3541133689504721
 ****************************************
```

```
For Lasso Regression Model: (Doubled alpha model: alpha:0.002*2 = 0.004)
 ****************************************

For Train Set:
R2 score: 0.9034800767367714
MSE score: 0.09651992326322859
MAE score: 0.22108183863993736
RMSE score: 0.3106765573119874

For Test Set:
R2 score: 0.882621176225094
MSE score: 0.12631879364873796
MAE score: 0.2552837418575564
RMSE score: 0.35541355298966576
 ****************************************
```

Observations:

- The test accuracy of the lasso regression model (alpha=0.002) is slightly higher in comparison to the test accuracy of the doubled alpha model (doubled alpha=0.004).
- MSE test scores comparing similar data of the original dataset and doubled alpha model gives us an idea that it is slightly smaller for the single alpha model than the doubled alpha model.
- Lasso Regression model (single alpha model) seems to perform better on the train and test data in comparison to the doubled alpha Lasso Regression model.
- Increase in the value of alpha in the model lead to a decrease in R2 score but an increase in the MSE (causing more shrinkage of coefficient values). In Lasso, the insignificant coefficients that have their values near to 0 correspond to 0 values; performing feature selection in the model. Thus, making the original (single) alpha model a better choice.

3. The most important predictor variables after the change is implemented. Top 10 features are as follows:

   (i) Ridge Regression Model (doubled alpha=4)

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

 ['MSZoning_FV', 'GrLivArea', 'OverallQual', 'Neighborhood_OldTown', 'GarageType_BuiltIn', 'Neighborhood_MeadowV', 'TotalBs
mtSF', 'MSZoning_RL', 'Neighborhood_Crawfor', 'Neighborhood_NridgHt']
 ********************************************************************************************************************************************
***
```

(ii) Lasso Regression Model (doubled alpha=0.004)

```
For Lasso Regression (Doubled alpha model: alpha:0.002*2 = 0.004):
 ********************************************************************************************************************************************************
***
The most important top10 predictor variables after the change is implemented are as follows:

 ['OverallQual', 'GrLivArea', 'TotalBsmtSF', 'MSZoning_FV', 'Foundation_PConc', 'Neighborhood_OldTown', 'd_SaleCondition',
'MSZoning_RL', 'GarageCars', 'Neighborhood_Crawfor']
 ********************************************************************************************************************************************************
***
```

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?

Solution:

Optimal Value of Alpha:

- The computed optimal value of alpha for Ridge Regression (Original Model): 2.0
- The computed optimal value of alpha for Lasso Regression (Original Model): 0.002

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

For Train Set:
R2 score: 0.9140467590425583
MSE score: 0.08595324095744175
MAE score: 0.20937283256494246
RMSE score: 0.2931778316268844

For Test Set:
R2 score: 0.8826471455614764
MSE score: 0.12629084639949664
MAE score: 0.25496240157816685
RMSE score: 0.35537423429322595
 ****************************************
```

```
For Lasso Regression Model (Original Model: alpha=0.002):
 ****************************************

For Train Set:
R2 score: 0.9153651941577272
MSE score: 0.08463480584227281
MAE score: 0.20832015672340426
RMSE score: 0.29092061776758416

For Test Set:
R2 score: 0.8834784025370518
MSE score: 0.12539627806945317
MAE score: 0.2553770268160431
RMSE score: 0.3541133689504721
 ****************************************
```

- The R2 test score on the Lasso Regression Model is slightly better than that of Ridge Regression Model. Moreover, the training
accuracy is slightly reduced; hence, making the model an optimal choice as it seems to perform better on the unseen data.
- The MSE for Test set (Lasso Regression) is slightly lower than that of the Ridge Regression Model; implies Lasso Regression performs
better on the unseen test data. Also, since Lasso helps in feature selection (the coefficient values of some of the insignificant
predictor variables became 0), implies Lasso Regression has a better edge over Ridge Regression. Therefore, the variables predicted
by Lasso can be applied in order to choose significant variables for predicting the price of a house in this analysis.

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 realized 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?

Solution:

(Please refer ipynb for detailed code for the answers)

Top five features in original Lasso Model (before removing) were as follows:

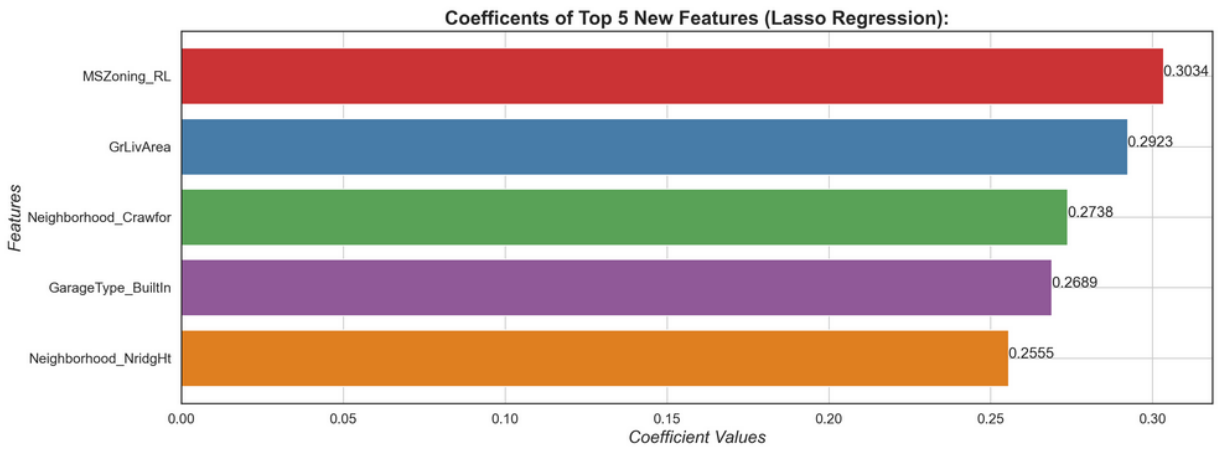
```
#Creating the list of top 5 features from Lasso Regression Model which we will be using later to answer 'Question 3'
top5_org_lasso_features = list(top10_lasso_df['Features'].iloc[0:5])
top5_org_lasso_features
```

```
['MSZoning_FV',
 'MSZoning_RL',
 'MSZoning_RM',
 'MSZoning_RH',
 'Neighborhood_MeadowV']
```

Top five predictor variables in the new model:

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

 ['MSZoning_RL', 'GrLivArea', 'Neighborhood_Crawfor', 'GarageType_BuiltIn', 'Neighborhood_NridgHt']
 ************************************************************************************************************************************************
***
```



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?

Solution:

The testing error of a model must be consistent with the training error, or the model must perform well with sufficient stability even after adding noise to the dataset. As a result, a model's robustness (or generalizability) is a measure of how well it can be applied to data sets other than the ones 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 Bias-
Variance Trade-off:

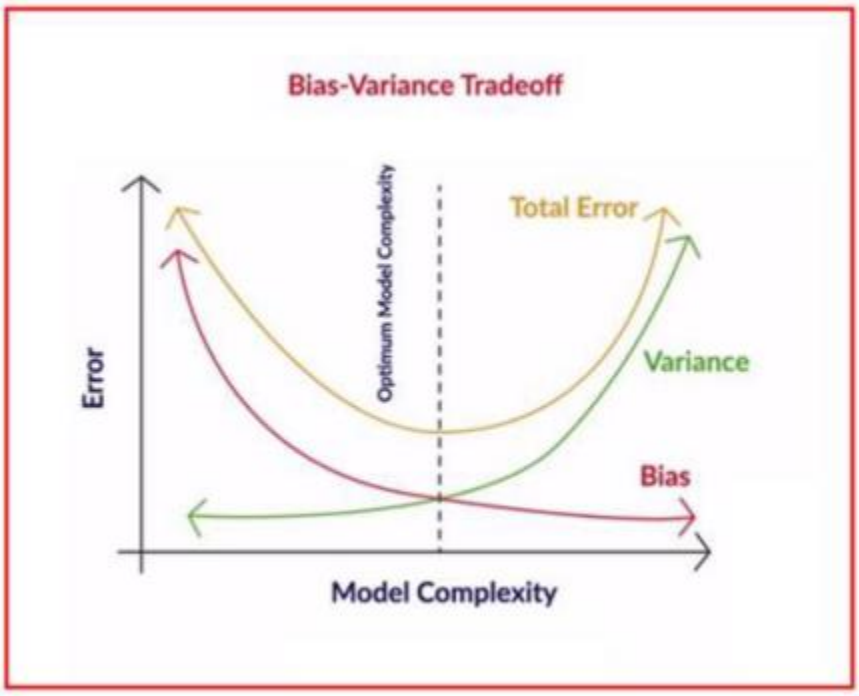
➢ A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive
to any changes in the training data.

➢ 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 allows you to determine how accurate a model is on test data. A sophisticated model can make an accurate job forecast if adequate training data is available. Models that are too naive, such as those that produce the same results for all test inputs and make no discrimination, have

a big bias since their predicted error across all test inputs is quite high. The degree of change in the model itself in relation to changes in the training data is referred to as variance.

As demonstrated in the graph below, the accuracy of the model may be maintained by maintaining a balance between Bias and Variance, as this reduces the total error.



As a result, accuracy and robustness may be at odds, as an overly accurate model can be prone to overfitting, causing it to be overly accurate on train data but fail when confronted with actual data, or vice versa.