Advanced Regression Subjective Questions

Ashutosh Kulkarni

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:

1. Optimal Value of Alpha:

   - The computed optimal value of alpha for Ridge Regression (Original Model): 3.0
   - The computed optimal value of alpha for Lasso Regression (Original Model): 0.0001

2. Changes in the model, if you choose double the value of alpha for both ridge and lasso regression:

Ridge Regression:

- Original Model (alpha=3)
- Doubled Alpha Model(alpha=6)

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

For Train Set:
R2 score: 0.9180922801233271
MSE score: 0.011524046417113963
MAE score: 0.07909235706433
RMSE score: 0.10735011139777156

For Test Set:
R2 score: 0.9000270308416499
MSE score: 0.01390852662943933
MAE score: 0.080325529172808
RMSE score: 0.11793441664518178
 ----------------------------------------
```

```
For Ridge Regression Model
(Doubled alpha model, alpha = 3*2 = 6):
 ----------------------------------------

For Train Set:
R2 score: 0.9178295902600888
MSE score: 0.011561005694967397
MAE score: 0.07922248113596868
RMSE score: 0.10752211723625701

For Test Set:
R2 score: 0.8999642581769162
MSE score: 0.013917259742863885
MAE score: 0.08047387138123295
RMSE score: 0.11797143613122578
 ----------------------------------------
```

Observations:

- The test accuracy of the ridge regression model (alpha=3) is slightly higher in comparison to the test accuracy of the doubled alpha model (doubled alpha=6).
- 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.

Lasso Regression:

- Original Model (alpha=0.0001)
- Doubled Alpha Model(alpha=0.0002)

```
For Lasso Regression Model
(Original Model: alpha=0.0001):
 ------------------------------------

For Train Set:
R2 score: 0.9181942256349852
MSE score: 0.011509703143853103
MAE score: 0.07899458578410501
RMSE score: 0.10728328455007846

For Test Set:
R2 score: 0.9004460067890003
MSE score: 0.013850237492186763
MAE score: 0.08011295181374324
RMSE score: 0.11768703196268807
 ------------------------------------
```

```
For Lasso Regression Model:
(Doubled alpha model: alpha:0.0001*2 = 0.0002)
 ------------------------------------------

For Train Set:
R2 score: 0.9180766869812588
MSE score: 0.011526240301808285
MAE score: 0.0790901771714888
RMSE score: 0.10736032927393752

For Test Set:
R2 score: 0.9008830204072753
MSE score: 0.013789438902344154
MAE score: 0.079947254083032
RMSE score: 0.11742844162443847
 ------------------------------------------
```

Observations:

- The test accuracy of the lasso regression model (doubled alpha=0.0002) is slightly higher in comparison to the test accuracy of the single alpha model (alpha=0.0001).
- 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 doubled alpha model than the single alpha model.
- Lasso Regression model (Doubled alpha model) seems to perform better on the train and test data in comparison to the single alpha Lasso Regression model.
- Increase in the value of alpha in the model lead to a increase in R2 score and an decrease in the MSE. 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 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 Model (doubled alpha=6)

```
For Ridge Regression (Doubled alpha model, alpha = 3*2 = 6):
 ------------------------------------------------------------------------------------------------------------
The most important top10 predictor variables after the change is implemented are as follows:

 ['GrLivArea', 'MSZoning_RL', 'OverallQual', 'MSZoning_RM', 'MSZoning_FV', 'TotalBsmtSF', 'OverallCond', 'Foundation_
PConc', 'GarageArea', 'Exterior2nd_CmentBd']
 ------------------------------------------------------------------------------------------------------------
```

Lasso Regression Model (doubled alpha=0.0002)

```
For Lasso Regression (Doubled alpha model: alpha:0.0001*2 = 0.0002):
 ----------------------------------------------------------------------------------------------------------
The most important top10 predictor variables after the change is implemented are as follows:

 ['MSZoning_RL', 'GrLivArea', 'MSZoning_RM', 'MSZoning_FV', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'Foundation_
PConc', 'GarageArea', 'MSZoning_RH']
 ----------------------------------------------------------------------------------------------------------
```

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): 3.0
- The computed optimal value of alpha for Lasso Regression (Original Model): 0.0001

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

For Train Set:
R2 score: 0.9180922801233271
MSE score: 0.011524046417113963
MAE score: 0.07909235706433
RMSE score: 0.10735011139777156

For Test Set:
R2 score: 0.9000270308416499
MSE score: 0.01390852662943933
MAE score: 0.080325529172808
RMSE score: 0.11793441664518178
 ----------------------------------------
```

```
For Lasso Regression Model
(Original Model: alpha=0.0001):
 ----------------------------------------

For Train Set:
R2 score: 0.9181942256349852
MSE score: 0.011509703143853103
MAE score: 0.07899458578410501
RMSE score: 0.10728328455007846

For Test Set:
R2 score: 0.9004460067890003
MSE score: 0.013850237492186763
MAE score: 0.08011295181374324
RMSE score: 0.11768703196268807
 ----------------------------------------
```

- The R2 test score on the Lasso Regression Model is slightly better than that of Ridge Regression Model. Hence, making the Lasso Regression 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.

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?

Solution:

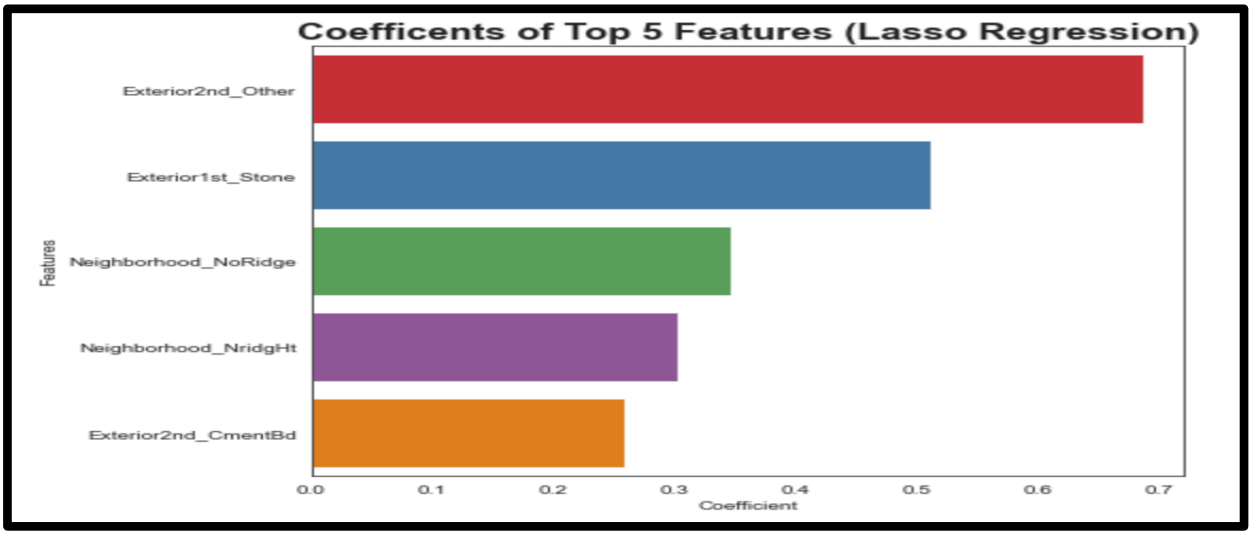
Top five features in original Lasso Model (before removing) were as follows:

```
Top 5 features in original lasso model:
 ['MSZoning_RL', 'GrLivArea', 'MSZoning_RM', 'MSZoning_FV', 'OverallQual']
```

Top five predictor variables after removing the aforementioned top 5 predictors from the original lasso model:

```
------------------------------------------------------------------------------------------------------------
The most important top 5 predictor variables after the change is implemented are as follows:

['Exterior2nd_Other', 'Exterior1st_Stone', 'Neighborhood_NoRidge', 'Neighborhood_NridgHt', 'Exterior2nd_CmentBd']
------------------------------------------------------------------------------------------------------------
```



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?

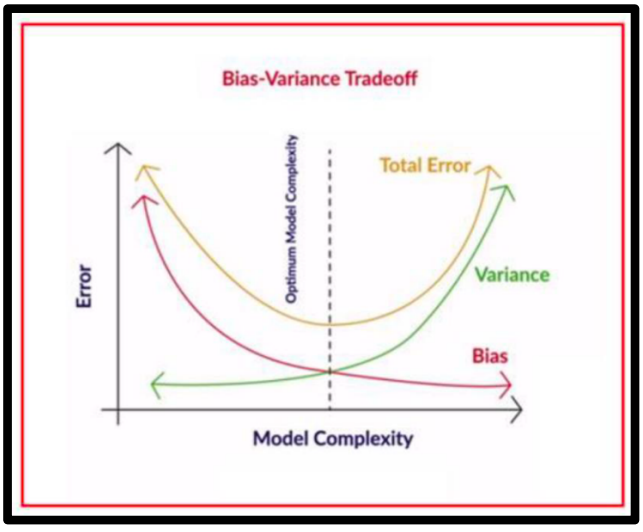
Solution:

The Robustness of a model implies, either the testing error of the model is consistent with the training error or the model performs well with enough stability even after adding some noise to the dataset. Thus, the robustness 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 related to 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 helps to 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.