Problem Statement – II

Name: Sukumar A
Email: sukuramram@gmail.com

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:

Optimal Value of Alpha:

Ridge Regression: 0.01
Lasso Regression: 0.0001

i) Ridge Regression:

Changes in Model after doubled

```
Ridge Regression Model (alpha=0.01):

For Train Set:
R2 score: 0.9189132221391941
MSE score: 0.08108677786080591
RMSE score: 0.2847574017664965

For Test Set:
R2 score: 0.896325195911666
MSE score: 0.1185059313742948
RMSE score: 0.34424690466915575
```

```
Ridge Regression Model with doubled alpha rate (alpha=0.02):

For Train Set:
R2 score: 0.918911638036245
MSE score: 0.08108836196375499
RMSE score: 0.28476018324856267

For Test Set:
R2 score: 0.8962235579513032
MSE score: 0.11862210908267533
RMSE score: 0.3444156051671807
```

Most Important Variables after change is implemented:

	Features
0	MSZoning_FV
1	MSZoning_RL
2	MSZoning_RH
3	MSZoning_RM
4	Foundation_Stone
5	SaleCondition_AdjLand
6	Neighborhood_BrDale
7	Neighborhood_MeadowV
8	Neighborhood_Crawfor
9	BldgType_Duplex

Observations after doubling alpha rate

- No major changes

Lasso Regression:

```
Lasso Regression Model (alpha=0.0001):

For Train Set:
R2 score: 0.9480497231003272
MSE score: 0.05195027689967285
RMSE score: 0.22792603383482293

For Test Set:
R2 score: 0.8999506385188403
MSE score: 0.11436185358620292
RMSE score: 0.33817429468574767
```

```
Lasso Regression Model (alpha=0.0008):

For Train Set:
R2 score: 0.9400516572013397
MSE score: 0.059948342798660355
RMSE score: 0.2448435067520892

For Test Set:
R2 score: 0.8938502771857096
MSE score: 0.12133489788427966
RMSE score: 0.34833159185505935
```

Most Important Variables after change is implemented:

Features	Coefficient
Neighborhood_Crawfor	0.3335
Neighborhood_StoneBr	0.3062
GrLivArea	0.2923
Neighborhood_MeadowV	-0.2463
Exterior1st_BrkFace	0.2032
OverallQual	0.1696
CentralAir_Y	0.1539
SaleCondition_Normal	0.1456
Condition1_Norm	0.1420
Neighborhood_Somerst	0.1396

Observations after doubling alpha rate:

- Both train and test accuracy decrease slightly after doubling the alpha rate

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:

```
Ridge Regression Model (alpha=0.01):

For Train Set:
R2 score: 0.9189132221391941
MSE score: 0.08108677786080591
RMSE score: 0.2847574017664965

For Test Set:
R2 score: 0.896325195911666
MSE score: 0.1185059313742948
RMSE score: 0.34424690466915575
```

```
Lasso Regression Model (alpha=0.0001):

For Train Set:
R2 score: 0.9480497231003272
MSE score: 0.05195027689967285
RMSE score: 0.22792603383482293

For Test Set:
R2 score: 0.8999506385188403
MSE score: 0.11436185358620292
RMSE score: 0.33817429468574767
```

Optimal Value of Alpha:

Ridge Regression: 0.01
Lasso Regression: 0.0001

- 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. (Image is attached)
- 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 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:

After Building the Initial lasso model the top 5 features are shown in the image below,

Features	Coefficient
MSZoning_FV	0.7503
SaleCondition_AdjLand	0.6920
MSZoning_RH	0.6511
MSZoning_RL	0.5850
MSZoning_RM	0.4806

Top 5 Features Dropped:

The above features were dropped and new model was built, now the new top 5 potential features and their coefficients are,

Features	Coefficient
Neighborhood_Crawfor	0.3335
Neighborhood_StoneBr	0.3062
GrLivArea	0.2923
Neighborhood_MeadowV	-0.2463
Exterior1st_BrkFace	0.2032

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:

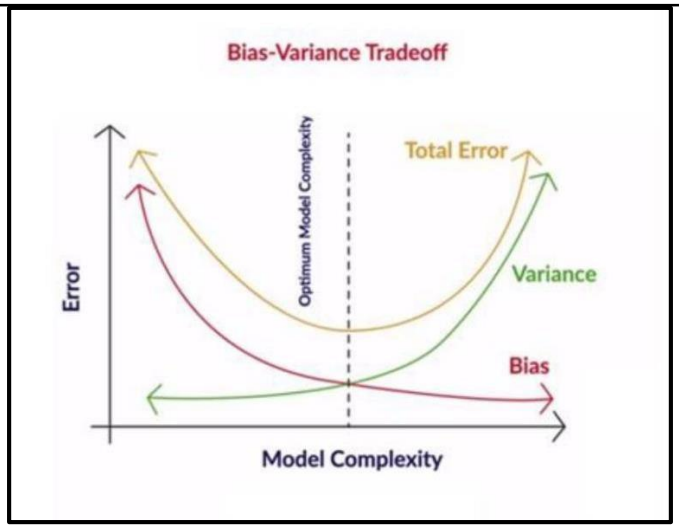
Robustness of a model implies, either the testing error of the model is consistent with the training error, the model performs well with enough stability even after adding some noise to the dataset. Thus, the robustness (or generalizability) of a model is a measure of its successful application to unseen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

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 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.