Problem Statement – II

Name: Chandrasekaran B
Email: b.chandrasekaran2@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: 4.0
Lasso Regression: 0.0008

i) Ridge Regression:

Changes in Model after doubled alpha rate:

```
Ridge Regression Model (alpha=4):


For Train Set:
R2 score: 0.9155917554585343
MSE score: 0.08440824454146563
RMSE score: 0.2905309700212107

For Test Set:
R2 score: 0.8838394363239533
MSE score: 0.13277783265126003
RMSE score: 0.3643869271135561
```

```
Ridge Regression Model with doubled alpha rate (alpha=8):


For Train Set:
R2 score: 0.9144691354622632
MSE score: 0.08553086453773687
RMSE score: 0.29245660282807234

For Test Set:
R2 score: 0.8844449598207201
MSE score: 0.13208568641009397
RMSE score: 0.36343594540179147
```

Most Important Variables after change is implemented:

	Features	Coefficient
0	Neighborhood_Crawfor	0.3192
1	BldgType_Twnhs	-0.3094
2	OverallQual	0.2585
3	MSZoning_FV	0.2460
4	GrLivArea	0.2367
5	BldgType_Duplex	-0.2217
6	MSZoning_RL	0.1781
7	Neighborhood_NridgHt	0.1749
8	selling_Age	-0.1741
9	BsmtExposure_Gd	0.1634

Observations after doubling alpha rate

- ▪ The test accuracy of the new model increases slightly
- ▪ There is a reduction in MSE
- ▪ Overall the old model seems to perform better, since it has a good training and test score.

(ii) Lasso Regression:

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


For Train Set:
R2 score: 0.9405093314944781
MSE score: 0.059490668505521864
RMSE score: 0.24390708990417204

For Test Set:
R2 score: 0.8955132345880263
MSE score: 0.11943404726266275
RMSE score: 0.3455923136625911
```

```
Lasso Regression Model with doubled alpha rate (alpha=0.0016):


For Train Set:
R2 score: 0.9365333297401836
MSE score: 0.06346667025981641
RMSE score: 0.25192592216724424

For Test Set:
R2 score: 0.8935283322513238
MSE score: 0.12170289842825102
RMSE score: 0.34885942502425105
```

Most Important Variables after change is implemented:

	Features	Coefficient
0	GrLivArea	0.3129
1	Neighborhood_Crawfor	0.3069
2	Neighborhood_StoneBr	0.2406
3	OverallQual	0.1872
4	Exterior1st_BrkFace	0.1851
5	Neighborhood_BrkSide	0.1477
6	OverallCond	0.1368
7	selling_Age	-0.1352
8	SaleCondition_Normal	0.1260
9	MSZoning_RM	-0.1244

Observations after doubling alpha rate:

- ▪ Both train and test accuracy decrease after doubling the alpha rate
- ▪ The initial model is the better model here also.

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:

Ridge Regression: 4.0
Lasso Regression: 0.0008

- 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)

```
Ridge Regression Model (alpha=4):


For Train Set:
R2 score: 0.9155917554585343
MSE score: 0.08440824454146563
RMSE score: 0.2905309700212107

For Test Set:
R2 score: 0.8838394363239533
MSE score: 0.13277783265126003
RMSE score: 0.3643869271135561
```

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


For Train Set:
R2 score: 0.9405093314944781
MSE score: 0.059490668505521864
RMSE score: 0.24390708990417204

For Test Set:
R2 score: 0.8955132345880263
MSE score: 0.11943404726266275
RMSE score: 0.3455923136625911
```

- 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,

Top 5 Features Dropped:

	Features	Coefficient
0	Neighborhood_Crawfor	0.3217
1	GrLivArea	0.2942
2	Neighborhood_StoneBr	0.2902
3	Exterior1st_BrkFace	0.2108
4	MSZoning_FV	0.1850

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

	Features	Coefficient
0	2ndFlrSF	0.2930
1	Neighborhood_MeadowV	-0.2693
2	1stFlrSF	0.2328
3	Neighborhood_Edwards	-0.2226
4	OverallQual	0.1802

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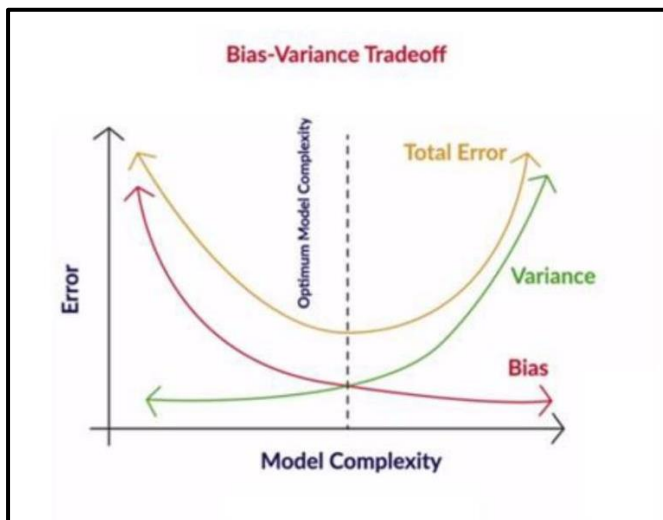
Thus, accuracy and robustness may be at the odds to each other as too much accurate model can be prey to over fitting hence it can be too much accurate on train data but fails when it faces the actual data or vice versa.