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:

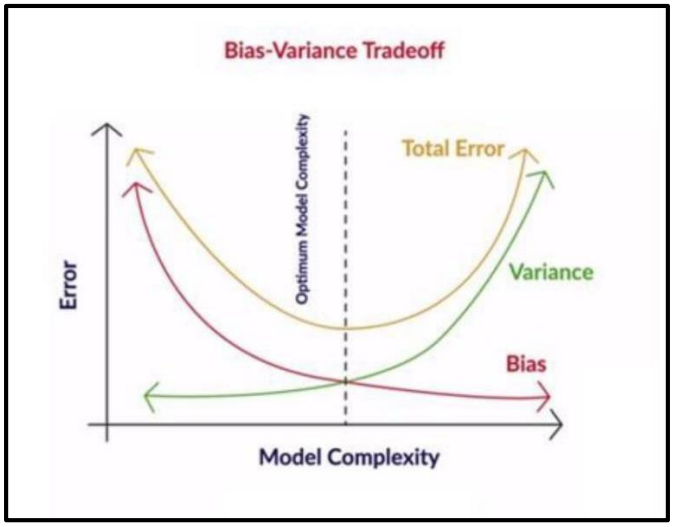
A model must either perform well with sufficient stability even after adding noise to the dataset, or its testing error must be consistent with its training error. As a result, a model's robustness (or generalizability) is a gauge of how well it performs when applied to unobserved data.

Regularization approaches allow us to manage the trade-off between model complexity and bias, which is directly related to the model's robustness. By punishing the coefficients for overcomplicating the model, regularisation aids in ensuring that just the ideal level of complexity is allowed. By making the model simpler, it aids in managing the robustness of the model. In order to maintain a careful balance between keeping the model basic and preventing it from becoming too naive to be of any value, one must ensure that the model is generalizable and robust.

A complex model will need to adjust for any tiny change in the dataset and is therefore very unstable and particularly sensitive to any changes in the training data. Making a model simple also results in the BiasVariance Trade-off.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.

You may assess bias by looking at how accurate the model is expected to be on test data. If there is enough training data, a complicated model can provide an accurate job forecast. Too naive models, such as those that produce the same answers for all test inputs and make no distinction at all, have a very big bias since their expected error for all test inputs is quite high. Variance is the extent to which the model itself changes in relation to changes in the training set of data.

As a result, the model's accuracy can be preserved by maintaining a balance between bias and variance, which reduces overall error as illustrated in the graph below.



Therefore, accuracy and robustness may be at conflict with one another since an overly accurate model may be susceptible to over fitting, which causes it to be overly accurate on test data but fail when it encounters real data, or the opposite may be true.