Assignment Part-II Qazi March 21 Batch

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?

Answer 1 – As the alpha increases it gives more penalty to cost function and reduces the magnitude of coefficients. If alpha is too low then it has little effect on the magnitude of coefficients and the model remains overfit (low bias, high variance). If alpha is too high then in Ridge it make more and more coefficients closer to zero while in Lasso, coefficients will be zero, so model will be underfit (high bias, low variance). In order to deal with this problem of overfitting and underfitting, a optimal value of alpha should be selected. If a model is overfit then the optimal value of alpha is that value where there is significant reduction in variance while bias is little compromised.

Changes in the model if we double the alpha is shown here.

Results for Ridge Regression

Alpha = 50

```
R square train 0.8685162849411042
R square test 0.8615386367951177
RSS train 835974011528.6069
RSS test  382257422685.55786
MSE train 823619715.791731
MSE test 876737207.9943987
```

Alpha = 100

```
R square train 0.8559016037812839
R square test 0.8538643979335832
RSS train 916178207223.9111
RSS test  403444089495.5817
MSE train 902638627.806809
MSE test 925330480.4944535
```

As the alpha get doubles the residual sum of squares (RSS) and mean squared error (MSE) increases

Results for Lasso Regression

Alpha 100

```
R square train 0.8983072664388432
R square test 0.8666679372179392
RSS train 646562826281.1877
RSS test 368096698607.56055
MSE train 637007710.6218598
MSE test 844258483.0448636
```

Alpha 200

```
R square train 0.8811646622937619
R square test 0.869457395330661
RSS train 755555575297.9867
RSS test 360395697807.21045
MSE train 744389729.3576224
MSE test 826595637.1725011
```

The most important predictor variables after the change is implemented are –

```
Neighborhood_NoRidge                    Score: 35044.147022564335
Neighborhood_NridgHt                    Score: 30516.13376987832
Neighborhood_StoneBr                    Score: 23560.43856320335
GrLivArea                               Score: 23199.218081337058
OverallQual                             Score: 17724.884162979364
```

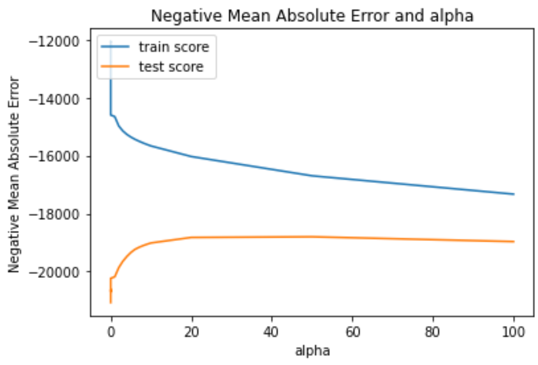
Question 2

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?

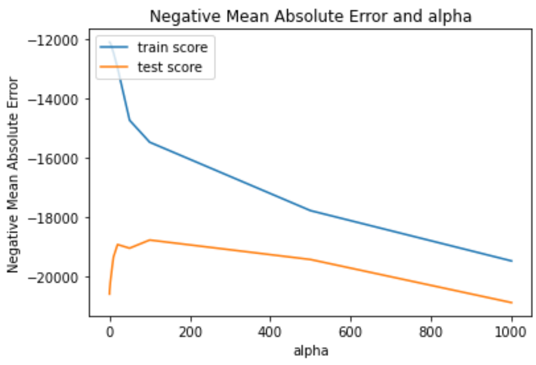
Answer 2 - Though both the regression worked decently well with R square more than 0.85 for test and train in each model. However the value of lambda for Ridge was 50 while lambda for Lasso was 100. And the difference in R square for train and test is less in Ridge than Lasso Model. So based on this I would like to go with Ridge regression as with increasing lambda the penalty increases.

Parameter	Ridge	Lasso
R square train	0.8685162849411042	0.8983072664388432
R square test	0.8615386367951177	0.8666679372179392
RSS train	835974011528.6069	646562826281.1877
RSS test	382257422685.55786	368096698607.56055
MSE train	823619715.791731	637007710.6218598
MSE test	876737207.9943987	844258483.0448636

Ridge



Lasso



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?

Answer 3 -

```
Most important 5 features that are affecting Sales pricing are:
Neighborhood_NoRidge                    Score: 36889.384059814445
RoofMatl_WdShngl                        Score: 35965.08719497574
Neighborhood_NridgHt                    Score: 31008.788577376166
Neighborhood_StoneBr                    Score: 25808.109274752827
GrLivArea                               Score: 25139.96911263669
```

If we remove these 5 from the data and again build the model then we get following parameters affecting the price most

```
Most important 5 features that are affecting Sales pricing are:
LotShape_IR3                            Score: -26400.39526179794
2ndFlrSF                                Score: 25599.716402737562
Exterior2nd_ImStucc                     Score: 23736.289853455568
BsmtQual_Gd                             Score: -21838.54286741414
KitchenQual_Gd                          Score: -21451.798131396677
```

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?

Answer 4– To make sure a model is robust and generalisable we must take into account the variance and bias model is showing. For a generalise model it should not be overfitted so that it can perform well on unseen data. If a model is overfitted then at first glance it seems performing well with the data but if there is little change in the data then its performance decrease drastically. One such example from practical life was also explained in upgrad lecture where a student practised all the questions of a particular entrance exam can perform better than the student who have learned the concept but not practise the same set of question. However if the question paper changes, the student with basic cleared can perform better. So similarly for our model it must be taken into account the optimal value for variance and bias.

If we make a model more generalise let suppose, we are okay with R square of 0.75 with our model then here the accuracy is not that high but the model can fit with more variant data. So with increasing generalisation, we compromise a bit on accuracy. It is decision of the person building the model, what accuracy he/she wants based on domain knowledge.