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:

To Regularized any model we need a cost function

Cost function = RSS + Penalty

Penalty for lasso= Alpha * (sum of numeric value of all coefficient of regression)

Penalty for ridge= Alpha * (sum of square value of all coefficient of regression)

When alpha is zero: As there is no regularization, hence model is overfit.

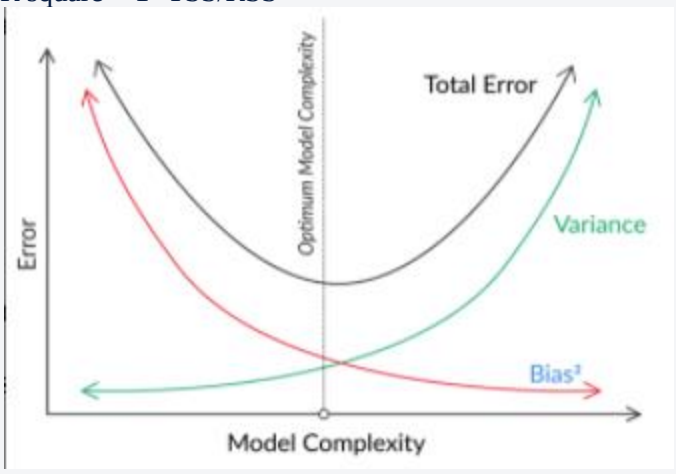
Optimal value of alpha: When we increase the value of alpha from zero, cost will start increasing, a point when there is relative difference between train R square and test R square is least, is the optimal value of alpha for ridge or lasso regression.

Or in other words when TSS is the least

TSS: Sum of square of Total Error

RSS: Sum of square of Residual

R square = 1- TSS/RSS



From the above figure, model is over fit, when it has high variance (most complex). At that point the regularization is not present there, can say value of lambda is zero. When we start the increasing the value

of lambda, model complexity and TSS will go down, then there is point when TSS is the least (when Variance and Bias are in the balanced state) lambda will be the optimal.

Double the value of alpha: By doubling alpha, Model's complexity will reduce. Model moves from current optimal fit state to under fit state.

After the regularization, top 5 important predictor variables were

1. GrLivArea
2. OverallQual
3. TotalBsmtSF
4. YearBuilt
5. YearRemodAdd

==========

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:,

Following are optimal lambda values of my program's output.

```
<===========Ridge Start=========================>
opt lampda====> 1235.2708512209517
opt_r2_score_train  0.8053248568894361
opt_r2_score_test  0.8053248568894361
<==========Ridge End=========================>
<==========Lasso Start=========================>
opt lampda====> 0.023000000000000007
opt_r2_score_train  0.8772105414811167
opt_r2_score_test  0.8772145660073668
<==========Lasso End=========================>
```

I would like to choose lasso model because had zero out many of the coefficients and r2_score is higher for train & test compare to ridge. Lasso model became simpler after zeroing many of the coefficients.

==========

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 2: following will be five most important predictor variables after dropping

1. 1stFlrSF
2. 2ndFlrSF
3. FullBath
4. Fireplaces
5. GarageCars

==========

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 any model robust and generalisable is that model must not too much trained on the training data set and when it perform on the test data set it's perform very bad. This state of the model is called the overfitted model. In this state model have least error on the training data and high complexity. To avoid this we required regularization. Required penalty on the complexity. Lasso & Ridge regression helps us on this.

Regularization reduces the accuracy of model on the training data set but do better on the training data set. Regularization help to move model from overfit to optimal fit. It reduces the accuracy on the training data set as overfit model had learned the noise of the data and due to this it performs badly on the test data set or unseen data set

==========