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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer – As both ridge and lasso regression penalize the loss function, the alpha value controls this penalization. Higher the value of alpha greater the penalization.

For Lasso, which primarily reduces the magnitude of the coefficients. A higher value of alpha would mean the coefficient is extremely low (or the coefficient penalized aggressively) which can cause underfitting issues.

For Ridge, higher values of alpha would reduce the feature to zero quicker and hence reduce the dimension of the overall datasets. This can cause relevant features to be reduced to zero which in turn impacts the model learning.

So if the alpha value is doubled the model would penalise the cost function faster.

Cross-validation techniques can be used to identify the optimal values of alpha. From my experience, alpha values ranging from 0.01 to 0.001 should do the job. In the assignment problem statement, the optimal value of alpha is 0.01.

The most important predictor variable in the current model is '**OverallQual**' as it has the highest co-efficient values.

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 to and why?

Answer – In the current assignment, the optimal value of lambda for ridge and lasso has been identified. While the train accuracies are similar for both ridge and lasso, but Ridge accuracy is slightly better on the test set when compared with the lasso.

The RMSE score of the ridge is better than the lasso regression then it gives me more confidence to go with the ridge instead of the 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?

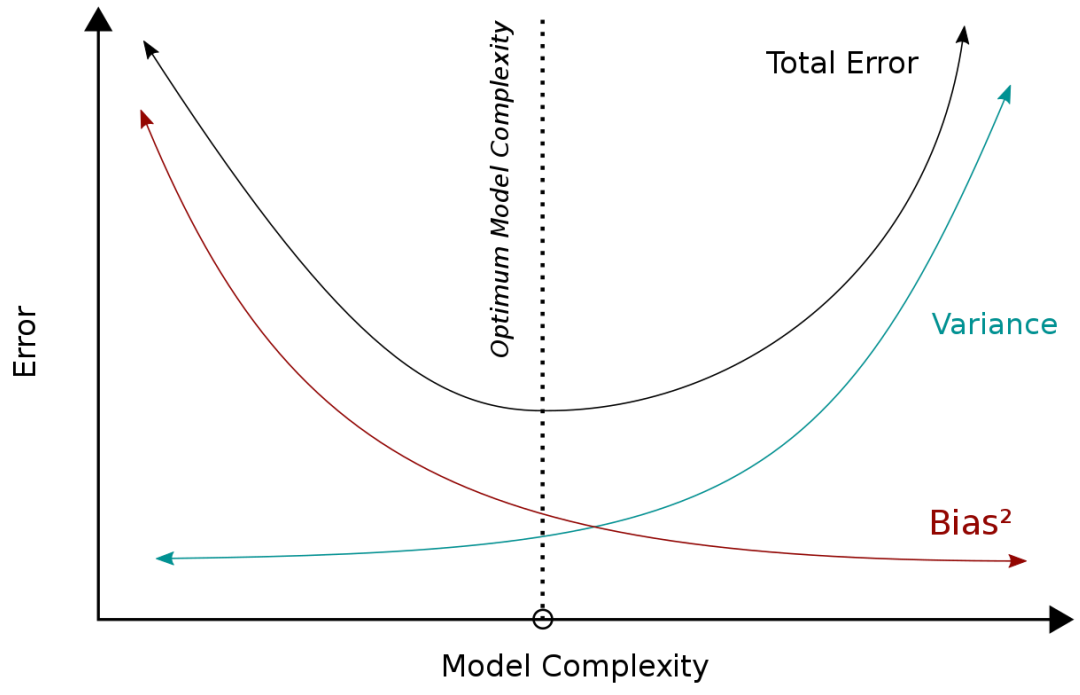
Answer - The next 5 important features are as follows -

1. 'GarageArea',
2. 'YearBuilt',
3. 'LotArea',
4. '2ndFlrSF',
5. 'YearRemodAdd',

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 - To ensure a model is robust and generalisable, you can think of the following image. Basically, you need to find the middle ground between a complex and a simple model.



A complex model tends to have high variance as it tries to remember everything. This would lead to low bias but a change in the input data or something a model has not seen previously can make it vulnerable.

A simple model on the other hand low variance as it tries to learn generally but it can't be accurate. Hence low accuracy would mean high bias.

There can be many reasons for having a complex and simple model but to find a middle point there are a few techniques. One of them is regularization. Regularization resolves the overfitting and underfitting issues by penalizing the loss function. With regularization in place/implemented, your testing accuracies improve as your model leans better as your overall loss is reduced. It minimized the co efficient estimation to zero to reduce the capacity of the model. In other words, it removes extra weight in the model.