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:

Currently the optimal alpha after doing Cross Validation with % folds results at,

Ridge – 0.4

Lasso - 0.0001

After making changes to double the alpha value for Ridge and lasso, certain changes were observed in the model.

R squared value increase in ridge and lasso model when alpha is doubled.

MSE is reduced in ridge and lasso model when alpha value is doubled.

The five most important predictor variables after the change are

Ridge - 'GrLivArea', 'OverallQual', 'LotArea', 'OverallCond', 'TotRmsAbv Grd'

Lasso - 'GrLivArea', 'OverallQual', 'LotArea', 'OverallCond', 'YearBuilt'

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:

After determining the optimal alpha values for ridge and lasso regression using Cross Validation of 5 folds. Certain checks regarding the metrics of the model were done. Results from those checks were,

R-squared MSE

Ridge- 0.87 0.0189

Lasso- 0.88 0.0186

These checks were conducted using the test set of the data. So obviously from the results ,we could see that Lasso has higher R-squared and lower MSE.

Hence Lasso Regression is chosen as the final model.

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?

ANSWER 3:

After removing the first five important features

'GrLivArea', 'OverallQual', 'LotArea', 'OverallCond', 'YearBuilt'

, the five important predictor variables are

'TotRmsAbvGrd', 'TotalBsmtSF', 'FullBath', 'GarageQual', 'GarageArea'

The next best model has reduced R squared (0.84) and increased MSE (0.023).

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?

ANSWER 4:

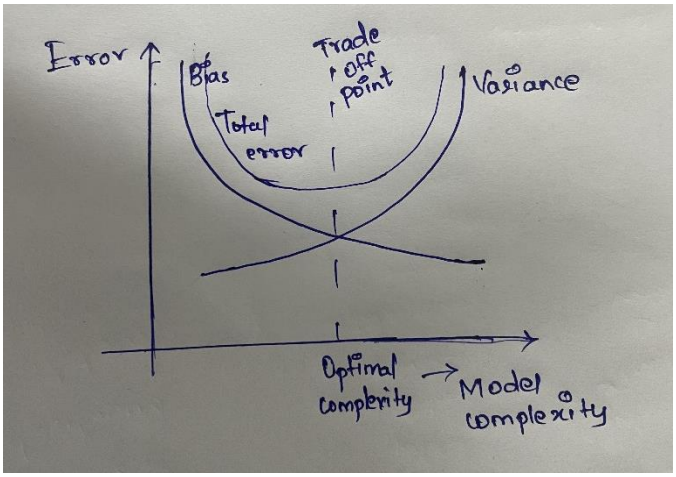
Making sure the model is robust and generalizable is dependent on the factor that the performance the model offers in training data should be on par with real world

data, here we assume that to be our validation data, if data size is less use cross validation to test validation data.

Generally, we check and build the model to have low error on training data, then test on validation data to check whether we are going in right direction. When we try to reduce the training loss by a large amount, we are at the risk of overfitting and having a large loss in validation data. If that happens even though our model has good training accuracy it will not be considered robust and generalizable.

Now to avoid this we will do regularization steps to bring the variance down. (Variance is the changes in output of model when training data changes).We could use Ridge and Lasso Regularization. We could simplify the model. We could increase the data.

These steps will reduce variance but increase bias (error in training data). We must find a tradeoff point between bias and variance to get a robust model.



The implications these have on accuracy of the model are, when the model becomes robust, it does not memorize the training data, so training accuracy reduces at the cost of increasing validation accuracy.

This happens because the coefficients of the model become less precise and around zero, when we do any kind of regularization.