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:

As per the model developed -

1. alpha for Ridge - 8.0

2. alpha for Lasso - 0.0025

If we choose to double alpha for ridge and lasso -

For Ridge -

Top 10 Features after doubling -
['GrLivArea', 'RL', 'OverallQual', 'FV', 'GarageCars', 'RM', 'OverallCond', 'TotRmsAbvGrd', 'BsmtFullBath', 'RH']

Top 10 Features before doubling -
['RL', 'GrLivArea', 'OverallQual', 'RM', 'FV', 'GarageCars', 'OverallCond', 'RH', 'BsmtFullBath', 'TotRmsAbvGrd']

Observations for Ridge -

1. Same set of top 10 features are there in Ridge.. However their relative importance has changed. Also, their coefficients have changed.
2. Train error increased and Test error remains flat

For Lasso -

Top 10 Features after doubling -
['OverallQual', 'GrLivArea', 'GarageCars', 'OverallCond', 'RL', 'BsmtFullBath', 'KitchenQual', 'TotalBsmtSF', 'FV', 'TotRmsAbvGrd']

In [ ]:

Top 10 Features before doubling -
['OverallQual', 'GrLivArea', 'GarageCars', 'OverallCond', 'RL', 'FV', 'BsmtFullBath', 'TotRmsAbvGrd', 'KitchenQual', 'LotArea']

Observations for lasso -

3. Change in the Top10 feature list has happened. For example lotArea was earlier in top 10 features but after doubling it disappeared.
4. # of variables with zero coefficient increased for lasso
5. Train prediction error increases and Test error stayed flat

After doubling alpha -

Most important feature for Ridge - 'GrLivArea'

Most important feature for Lasso - 'OverallQual'

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-
Ridge and Lasso both have produced almost identical R2 on test set ~ 0.87. While Lasso was computationally tough, post computation it has resulted in feature selection as well. Less features means - less complexity. As per Occam's Razor - with all other things the same, we should go for models with lesser complexity.

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:
'TotalBsmtSF', '2ndFlrSF', 'GarageArea', 'TotRmsAbvGrd', 'KitchenQual'

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 -
For the model to be robust and generalizable it should learn the pattern in the training data and not try to remember it. For this to happen, model should look at the variation that can be explained and ignore the variations that are either very specific to the Data or information capture process (such as a not so perfect thermometer used to measure temperature).

While model building, we should ensure the following -

1. The model parameter should not have multicollinearity
2. Cost function should not be designed only to minimize the prediction error.
3. Cost function should penalize complexity
4. Ensure that the assumptions of linear regression are valid
5. Hyperparameter tuning to lower complexity / overfitting

The accuracy of the model is always measured on the Test Dataset. The moment we penalize the cost function for complexity, our model will try to find a balance between lower prediction error and lower complexity. Both the objectives by themselves lead to poor results, but together they lead to optimal outcomes, i.e. high test accuracy.