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

Optimal value of alpha for ridge is 1.01

Optimal value of alpha for lasso is 0 (I.e. no regularization)

When the alpha for ridge is doubled, in this case to 2.02 the top five most important predictor variables changes. As alpha for lasso is 0, we don't get in this case, but in general the number of variables with non-zero coefficients will reduce.

In case of ridge after the alpha is doubled the most important predictor variables are as follows:

MSZoning_RL, MSZoning_RH, MSZoning_FV , MSZoning_RM, Neighborhood_Crawfo, Foundation_Stone ,Exterior2nd_Brk Cmn, Neighborhood_NoRidge, KitchenAbvGr, CentralAir_Y.

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

For my case the lasso coefficient came out to be zero. Therefore it is just OLS. If we compare the coefficient values we see the ridge coefficients are much lower and thus the model is simpler. So, I will choose ridge.

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

The top five most important variables now are:

1. Condition1_PosA
2. Exterior2nd_Brk Cmn
3. Neighborhood_Crawfor
4. BldgType_Twnhs
5. CentralAir_Y

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 ensure that a model is robust and generalisable we can take the following steps:

1. *Make the model as simple as possible and no more (Ockham's razor*). This implies that we should always select a simpler model with less parameters and low value of parameters. But we must also ensure that bias-variance trade-off is accomplished, thus the model should be as much simple as needed but not too simple which will cause the bias to rise
2. *Use models with more constraints, I.e., more simplifying assumptions*. More constrained models span a lower dimensional space and hence are more generalizable.
1. Use cross-validation to ensure performance across unseen data.

The implications of all the above on model accuracy is that the difference between training accuracy and test accuracy decreases with robust models. This means that accuracy over unseen data increases.

This happens as simpler models try to capture the general trends in training data and ignore the random noise as well as the fluctuations specific to the training set. So, they are able to generalise easily over unseen data having similar general trends while ignoring the specificities.