PROBLEM STATEMENT – PART II

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: Alpha = 0.01 has been used in the case of lasso regression as, going above that value the model starts penalizing more and more, bringing down the r2 score as it tries to make more and more coefficient values zero. This is precisely the reason why 0.4 hasn’t been chosen as the alpha value, although it was suggested by the negative mean absolute error graph.

Alpha = 2 has been used in the case of ridge regression, as at this value the test error is minimum. Alpha values for both lasso and ridge regression are chosen based on the negative mean absolute error plots.

Doubling the alpha value for lasso regression will make alpha = 0.02. This decreases the r2 score while increasing the number of variable coefficients which have been turned to zero. For ridge regression however, doubling the alpha value produces more or less similar r2 scores. Curiously, the important variable do change due to a shift in coefficient values.

The important variables in case of lasso regression after the change:

- OverallQual
- GrLivArea
- OverallCond
- Fireplaces
- GarageArea
- TotalBsmtSF
- TotRmsAbvGrd
- GarageCars
- LotArea
- BsmtFullBath
- 1stFlrSF
- WoodDeckSF
- PoolArea
- AgeP
- MSSubClass

The important variables in case of ridge regression after the change:

- RoofMatl_CompShg
- Neighborhood_Crawfor
- MSZoning_RH
- Neighborhood_StoneBr
- MSZoning_FV
- CentralAir_Y
- OverallQual
- MSZoning_RL

- Exterior1st_BrkFace
- RoofMatl_WdShngl

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: In order for increasing the accuracy of the predictive abilities of a model, coefficients need to be regularized, which essentially means a reduction in variance. Lambda value, for lasso regression acts as the penalty upon the absolute value of the coefficients, cross-validated. Whereas, for ridge regression, it is the penalty upon square of coefficient values, again cross-validated.

I would prefer ridge regression over lasso regression, based on the lambda values. It is optimum if all variables are considered by my model, so I'll go with ridge regression.

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 most important predictor variables will be:

- BsmtFullBath
- Fireplaces
- LotArea
- TotRmsAbvGrd
- GarageArea

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: A simple model is the best model. Complexity brings in other variables that will only make the job tougher. We can also look at the 'Bias-Variance' tradeoff. Simpler models have low variance and are more biased while complex model are more variable but less generalizable.

In order to get a robust and generalised model, we have to find a happy medium, which will have accuracy and ability to adapt to variance in the data.