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

Ans: The optimal value of alpha for ridge is 2, when we plot the curve between negative mean absolute error and alpha, we see that alpha increase from 0 then error term decreases and train error is showing increasing trend, so we picked alpha 2 as for ridge regression.

For lasso optimal value of alpha 0.01, when we increase alpha the model try to penalize and most of the coefficient value zero.

When we double the alpha value for ridge, the model Predictors are same but the coefficient of these predictor has changed, the model will apply more penalty on the curve. Similarly when we increase alpha for lasso more coefficient of the variable will reduced to zero. R2 also decreases.

The most important variable for ridge: MSZoning_FV, MSZoning_RL, Neighborhood_Crawfor, MSZoning_RH, MSZoning_RM, SaleCondition_Partial, Neighborhood_StoneBr, GrLivArea, OverallQual

The most important variable for lasso: GrLivArea, OverallQual, OverallCond, TotalBsmtSF, BsmtFinSF1

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?

Ans: Ridge regression includes all variables in final model unlike Lasso Regression.Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. So better to go with Lasso Regression model.

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?

Ans: Those 5 most important predictor variables that will be excluded are :- GrLivArea, OverallQual, OverallCond, TotalBsmtSF, 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?

Ans: The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data.