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:

Optimal lambda values : Lasso - 0.001, Ridge - 0.5

If Lambda in Lasso is chosen to be 0.002, The train and test accuracies have reduced by approx 5-6% each and LowQualFinSF, BsmtHalfBath, BsmtFinSF1, GarageArea, OpenPorchSF, LotArea are the top 5 features

If Lambda in Ridge is chosen to be 1, The train and test accuracies haven't changed much and WoodDeckSF, LowQualFinSF, 1stFlrSF, MasVnrArea, GarageArea are the top 5 features

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:

Its better to choose Lasso over Ridge as from the model we can see that Lasso regularizes accuracies across Train & Test better

Moreover Lasso provides feature selection too, gives us only the necessary set of variables by making the coefficients of the non relevant variables as 0

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:

The top features after removing the top 5 variables from the previous iteration are:
BsmtFullBath, OverallQual_V_Good, LotArea, OverallQual_Fair, WoodDeckSF

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:

We can make sure a model is robust by different means by visualizing them through plots, looking at the train and test errors

A model needs to be robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Model's real performance can only be assessed when we predict against a data that it has not seen