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?

Answer: Below are the optimal value for ridge and lasso regression:

- Optimal value of alpha for Ridge Regression is 4,
- Optimal value of alpha for Lasso Regression is 0.0001

On increasing values of alpha, we are driving model more towards generalization, as the coefficients values decrease, residual sum of square increase therefore, compromise in bias for reduction in variance. On replacing the value for alpha by double value causes a drop in R2 scores for both ridge and lasso regression.

In ridge regression, on increasing regularization r2 score decreased slightly from 89.2% to 88.3% for test data, also Root Mean Squared Error value increased.

In Lasso regression, on increasing alpha by factor of 2, r2 score dropped from 0.896 of 0.882 for test data.

Most important predictor variables after making are:

- Neighborhood_StoneBr
- MasVnrArea
- GrLivArea
- TotalBsmtSF
- BsmtExposure_Gd

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: I will be opting for Ridge regression, as it can shrink the coefficients of all features, maintaining their presence in the model while preventing overemphasis on any single feature.

Although Lasso feature selection can help in identifying important features and utilising them for analysis, but any change in dataset in can make model unstable if the variables have correlation, therefore it will be better to opt for 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: After dropping top 5 features of Lasso regression new top features are as follows:

- 1stFlrSF
- 2ndFlrSF
- Neighborhood_NridgHt
- BsmtFinSF1
- OverallQual

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 model has to be robust so that any variation in the data does not affect its performance. A generalizable model is able to adapt properly to new data, providing predictions based on the training data.

To make sure a model is robust and generalizable, we need to avoid overfitting. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.

In other words, the model should be simple, in order to be robust and generalizable.
If we look at it for Accuracy, a complicated model will have a high accuracy due to overfitting in model. So, to make our model more robust and generalizable in order to reduce overfitting, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease.