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 value of alpha for ridge regression is :- 100

Optimal value of alpha for lasso regression is:- 0.01

On doubling the value of alpha for ridge and lasso, the R-squared value showed slight drop. When ridge was doubled, there was no change in the most important predictor variable and it remained: **OverallQual**.

But when Lasso was doubled, "**OverallQual**" became the most important predictor variable. Earlier, **2ndFlrSF** was the most important predictor.

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 my opinion, Lasso is better suitable for this problem as we have large number of predictors. Lasso provides automatic feature selection. With the removal of redundant features, it has helped in making the model simple and more accurate. Lasso has slightly higher R-squared value as well.

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:- Earlier the five most important features were:

'2ndFlrSF', '1stFlrSF', 'OverallQual', 'BsmtFinSF1', 'TotalBsmtSF'

After dropping these five features, and rebuilding the lasso model, the new five most important features were:-

'GrLivArea', 'ExterQual', ' BsmtQual', 'KitchenQual', 'BsmtExposure'

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 is robust and generalisable if it strikes a correct balance between bias and variance. Model has to be simple, with less number of features. The features should have low VIF, low p-value, and the R-Squared value should be high.

Low VIF :- indicates that features don't have strong correlation between them. Usually VIF < 5 is expected to be achieved.

Low p-value:- Low p-value is required for the statistical significance of the model. A p-value of 0.05 or lower is generally expected.

High R-Squared value:- High R-Squared value for both the training and test data directly implies that the model accuracy is good and it explains most of the variations.