Subjective Question

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:

Optimal value of Ridge regression:8.0

Optimal value of Lasso regression:0.001

If we double the alpha value of lasso regression, more coefficient becomes zero. For Ridge regression, it moves towards zero.

Important variable after the change:

1. 1stFlrSF

2. 2ndFlrSF

3. OverallQual,

4. OverallCond

5.SaleCondition_Partial

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:

Reason to select Lasso Regression:

- Lasso Regression produced slightly r2 score on test data than Ridge Regression. Choosing Lasso as the final model.

- Lasso Regression gave us a simpler model with more coefficients as zero values.

Subjective Question

Question3:

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:

After deleting five most important variable, now we have new predictable variables:

1. TotRmsAbvGrd
2. GarageArea
3. ExterQual
4. Fireplaces
5. BsmtQual

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:

If the accuracy of the test and unseen dataset is almost same as of the train dataset, then we can say that the model is robust and generalised. The bias-variance trade-off can also be used to understand it.

Variance in the model indicated the degree of changes in the model itself with respect to changes in the training data. High variance means it perform very well on train dataset but it perform very poorly on unseen dataset.

Bias indicates the accuracy of the model on test dataset.

The simpler the model, the greater the bias, but the lower the variance and generalizability and the vice-versa. To avoid overfitting and underfitting of dataset ,it is important to maintain balance of variance and bias.

1. Generally we try to maintain the train and test dataset accuracy >80% and there should not be more than 5% variation in train and test dataset.
2. P-value of all feature should be less than 0.05 .
3. VIF for all the feature should less than 5%.