Subjective Questions

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 values found for Ridge and Lasso are 1500 and 500 respectively. By doubling the value for below observations were found:

Lasso:

By increasing the alpha two times we can observe that the training R2 Score has dropped a bit, also the most important predictor variables remained the same but the coefficients increased for most of the predictor variables.

Ridge:

In Ridge By increasing the alpha two times we can observe that the training R2 Score hasn't changed much, also the most important predictor variables remained the same with addition of new predictor variable Neighborhood_NoRidge.

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 has a score of 91.34 but since it is using almost all the features then R2 score can increase because of that whereas Lasso model is just using 62 feature variables with score 91.24 and has reduced most of the features to 0 , hence we will go with the Lasso 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: The top 5 predictor variables that we got during the training of Lasso model are:

- GrLivArea
- OverallQual
- BsmtFinSF1
- BsmtQual
- KitchenQual

After dropping above variables, we found below features to be in the top 5 list:

- BsmtUnfSF
- TotalBsmtSF
- BsmtFinSF2
- MasVnrType_BrkFace
- Neighborhood_NPkVill

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: A model can be considered generalised and robust if it works well on unseen data such that the test set accuracy is not very less than the training accuracy but is almost at par with the training accuracy. In addition to this model should be able to handle outliers. If the model is not robust, it cannot be trusted for predictive analysis.

A model suffering from underfitting has high bias and low variance, which can be treated by training the model on more data while a model suffering from overfitting has high variance and low bias, which can be handled by using regularization, so basically we want a model that as low bias and low variance so that it works well on the data outside the training data.