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?

Answer:
Optimal value of alpha for ridge and lasso regression:
Ridge: 100
Ridge after RFE: 20
Lasso: 0.001

Changes in the model if you choose double the value of alpha for both ridge and lasso:
It was observed that the training score has decreased and the testing score has increased slightly for all the models after doubling the alpha. For Ridge with RFE model the change was more pronounced. Here, the gap between train and test data was the smallest.

The most important predictor variables after the change is implemented:
'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'YearBuilt' are the most important predictor variables.

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:
Based on the analysis, the Lasso regression model seems to be the most appropriate choice. This is because it achieves a slightly higher r2 score and exhibits a slightly lower gap between training and testing than the other models that were evaluated. In addition to these performance-related advantages, Lasso regression also assists in reducing the number of features in the final model, resulting in a more simplified and interpretable model. This attribute is key in developing a robust and generalizable model. Moreover, Lasso regression also produced the lowest residual sum of squares (RSS) among all the models that were created, further indicating its superiority.

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 :
Initially, the five most important predictor variables were: 'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'YearBuilt

After removing the above variables, the new set of most important predictor variables were :

BsmtFinSF1, BsmtUnfSF, MSZoning_RL, 2ndFlrSF, MSZoning_RM

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:
In machine learning, generalization is a crucial aspect to consider, and the test accuracy should be higher than the training score while also minimizing the difference between the two scores. The model should generalize well during the training phase, and if the difference between the scores is excessively high, it could indicate overfitting. However, it is important to note that low test scores may result from splitting the data set too early in the preprocessing step, which could result in essential steps being missed during the testing phase. The robustness of a model is not solely based on high test scores, but it is also dependent on the assumption that the training scores are higher than the test scores. Both scores need to be adequate for the particular business case and the model's expected usage. It is also vital to consider the values obtained for both training and testing to ensure that the model performs well on unseen data. This means that the data should contain some outliers to aid in making predictions. As we have seen in the assignment, the accuracy of the model will depend on how the data is processed and how features are selected. There may be no perfect model, but various steps can be taken to ensure that the model developed is appropriate for the specific context and the unique requirements of the business case.