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 of alpha for ridge and lasso regression are:

Ridge: 100

Lasso: 0.001

For all the models, the training score has decreased slightly and the testing score has increased slightly. The gap between train and test data is the very less.

The most important predictor variables after the change implemented are;
In Ridge:
GrLivArea, OverallQual, TotalBsmtSF, BsmtFinSF1, OverallCond

In Iasso:
GrLivArea, OverallQual, OverallCond, TotalBsmtSF, BsmtFinSF1

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. The r2 score is slightly higher for Lasso and the gap between training and testing is slightly lower. Hence, I would choose lasso. Lasso helps in reducing the features in the model, helping to create a simpler final model. This is important for creating a robust and generalizable model. It also has the lowest residual sum of squares of all of the created models.

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 predictor **lasso** variables are:

GrLivArea	0.136950
OverallQual	0.063599
OverallCond	0.042275
TotalBsmtSF	0.033984
BsmtFinSF1	0.032797

After removing them, the top predictor variables are:

2ndFlrSF	0.062053
FullBath	0.048194
GarageCars	0.045646
MSZoning_RL	0.041861
YearRemodAdd	0.039137

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. Generalization is important and test accuracy needs to be higher than the training score. However, the difference should not be exceedingly high. The model should generalize during training, but if the score is very high in training, and lower in testing, it means that the model has memorized the data, meaning that it is overfitting. Overall, there should't be large differences between the results. Low test scores could result from splitting the data set too early in the preprocessing step, so that some steps may be missed on the test data. Ensuring a model's robustness and generalizability involves several key steps. Firstly, it requires attention to data quality, ensuring the dataset is diverse and representative. Secondly, effective feature engineering is essential to extract relevant information. Thirdly, managing model complexity through techniques like regularization prevents overfitting and improves generalization. Furthermore, cross-validation assesses the model's performance on unseen data, while testing on diverse datasets confirms its adaptability across different scenarios. Ensemble methods like bagging and boosting can also enhance robustness by combining multiple models' predictions. Prioritizing robustness may sometimes lead to a trade-off with training accuracy, as simpler models or regularization techniques may reduce accuracy on the training data. However, this focus ultimately improves the model's performance on new, unseen data, ensuring its practical reliability and utility in real-world applications.