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?

A: The optimal value of alpha for Ridge Regression is 0.2 and Lasso Regression is 0.0001. When we doubled the alpha, we observed that the r2_score for training set decreased slightly whereas we see a slight increase in testing set r2_score. Post doubling the alpha in both Ridge and Lasso regression we see the same predictor variables whereas their coefficients are slightly different. Please find the below screenshot of Coefficients post doubling alpha for ridge and lasso

[92]:

	Ridge	Lasso
LotArea	43962.405206	45088.992962
OverallQual	105546.758885	105435.826206
BsmtFinSF1	47751.119818	47249.531981
TotalBsmtSF	84154.681625	85567.433873
1stFlrSF	56854.315140	114346.148415
2ndFlrSF	43076.699432	76007.014013
GrLivArea	80555.423830	24086.169029
BedroomAbvGr	-45325.926361	-47809.295217
TotRmsAbvGrd	37129.335138	36408.530863
GarageCars	43072.331480	42899.741892
MSZoning_FV	51662.365719	63811.584524
MSZoning_RH	38608.185972	51684.531376
MSZoning_RL	46281.307999	58280.890391
MSZoning_RM	34890.210079	46970.121426
Street_Pave	45210.153949	45294.000360
Exterior2nd_CBlock	-58405.475666	-71573.902867
ExterQual_Gd	-39537.595299	-40840.618543
ExterQual_TA	-50341.456771	-51318.830450
Foundation_Wood	-36941.630763	-40829.768699
BsmtQual_Fa	-45151.239019	-45468.123734
BsmtQual_Gd	-37199.772591	-36956.824239
BsmtQual_TA	-42545.052121	-42223.732253
Heating_OthW	-50895.644373	-61560.663610
Functional_Sev	-83485.382650	-101231.305732
SaleType_Con	35950.254576	39098.021748

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?

A: Lasso regression will be the best choice to apply because the r2_score is slightly higher when compared to r2_score of ridge regression which helps in developing/achieving the robust model with more relevant features.

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?

A: The five most important predictor variables in the lasso model are

- 'LotArea'
- 'OverallQual'
- 'BsmtFinSF1'
- 'TotalBsmtSF'
- '1stFlrSF'

After dropping the above five predictor variables and performing lasso regression we found that the r2_score decreased in both train and test sets. Now the most important top 5 predictor variables in Lasso Model are

- 'GrLivArea'
- 'MSZoning_FV'
- 'MSZoning_RL'
- 'MSZoning_RH'
- 'GarageCars'

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?

A: To call a model to be generalised model there should not be much difference between the accuracy of seen (training) and unseen(testing) datasets. If the accuracy score of training set is higher than the accuracy of testing set then the model will become overfitted model because the model memorized the data and the results will be not accurate. The splitting of training and testing dataset will also play a key role in make the model more generalised, because we split the training and testing data as 90 and 10 then we are feeding or training the model on the complete data and testing on only some portion of it which will land into overfitting model.

Robustness of the model is not completely dependent on high test score. It always goes with assumption as test scores are lesser or near by training score but not higher. The robustness of the model will always be calculated based on it performance on unseen data. Apart from this to make a model robust enough we should concentrate on outliers in the data and make sure that outliers which are really showing impact on the model should be retained and others should be dropped from the dataset. While performing outlier analysis to make the model robust we should not build the complex model the model should always be simple and generalised.