Advanced Regression Assignment Questions/Answers

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 for Alpha for ridge is 0.7 and optimal value for Lasso is 0.0001. if we double the alpha values , we get slight change in Train R2 score and Test score , increasing test score . Both Training and Test score reduces more if we go higher and higher on alpha value (4 times or 6 times). Most important predictor variables still remains same which is LotArea, OverallQual , YearBuilt,TotalBsmtSF etc..

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: Optimal value of alpha is 0.5 and Lasso is .0001 , in the models I have built , may be due to EDA , some drops did due to RFE , some drops due to data skew etc,.. I don't get much difference between linear regression Vs ridge Vs Lasso. Lasso Regression seems to best with higher test score and less variance

Best is to choose lasso here , since it gives better test results with less predictors to have simpler model and less explain ability on parameters.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.894356	0.893703	0.892221
1	R2 Score (Test)	0.835637	0.839047	0.840042
2	RSS (Train)	2.221321	2.235055	2.266206
3	RSS (Test)	1.392899	1.364001	1.355571
4	MSE (Train)	0.056616	0.056791	0.057185
5	MSE (Test)	0.068483	0.067769	0.067559

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: Lasso has removed few parameters 5 or 6 for my model , but all of them are Neighborhood or similar derived categorical variable due to encoding of category variable. Dropping a part of the encoded category variable and re modelling is not correct. We either drop whole category predictor variable or keep it to avoid creating errors in model due to partial drops. Tried dropping whole Neighborhood predictor or other categorical it yields poorer result on train and test score .

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: Models need to be kept simpler – Occam’s Razor and avoid taking too many parameters leading to overfit but poor test scores due to variance. Regularization and RFE methods are important to be followed to keep the model more robust and generalisable with unseen data. If we don’t do Feature elimination and regularizations techniques, models will perform poorly with unseen data and accuracy of the model will be low and unpredictable due to more bias or more variance.