Advanced Regression Assignment

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 of alpha for Ridge is 10 and for Lasso is 0.001. This is based on test R-squared values that these two algorithms produce. At these values of alpha they produce the highest R-squared on test data, around 0.72 for both ridge and lasso. I have not considered training set MSE for this as there is very little difference between training set MSEs for different alphas.

If a single value of alpha is to be chosen for both the models, then 0.001 is optimal as beyond that Lasso's test R2 results start to decay significantly. Ridge produces a R2 of .69 for alpha of 0.001, which is acceptable, and lasso produces R2 of 0.72.

Increasing the alpha to double the value will increase the regularization so that coefficients will be penalized even more and will shrink down further. This will produce a sparser model that is more generalizable, perhaps at a greater risk of 'underfitting'. Bias will increase and model variance will get reduced. For Lasso, more features may have a coefficient of 0.

The most important predictor variables are likely to be the same, though their order may change in terms of magnitude of coefficients.

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?

We will apply whichever produces a higher r-squared on test data. In case of this assignment, ridge regression performs slightly better than Lasso. It is

possibly die to the fact the lasso makes the coefficients too sparse and can even reduce them to zero.

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?

The 5 most variable predicted by lasso regression with multicollinear variables removed are **'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'GarageArea', 'Fireplaces'**

If multicollinear variables are **not** removed from the data while data cleansing and EDA, then if 5 most important variables are not included, then the next 5 most important variables are their close cousins which are highly correlated wit the removed variables. These are **1stFlrSF, 2ndFlrSF, GarageCars, BsmtQual_Excellent, YearBuilt**

A	B	C	D	E	F	G
5 Variables Removed				All Variables Intact		
Column	Ridge	Lasso		Column	Ridge	Lasso
1stFlrSF	0.253709	0.551345		GrLivArea	0.100207	0.346735
2ndFlrSF	0.124039	0.241411		OverallQual	0.128287	0.284135
GarageCars	0.164008	0.174799		TotalBsmtSF	0.106983	0.144076
BsmtQual_Ex	0.107671	0.112535		GarageArea	0.095801	0.125627
YearBuilt	0.03606	0.054241		Fireplaces	0.107275	0.106035
WoodDeckSF	0.069185	0.042339		GarageCars	0.082556	0.072243
ExterQual_Gd	0.024804	0.040712		BsmtQual_Ex	0.070706	0.057937

If 5 most imp variables and their multicollinear counterparts are both removed from the data, then these are the 5 most imp variables:

	Ridge	Lasso
MasVnrArea	0.249731	0.271208
FullBath	0.217273	0.223424
BsmtQual_Ex	0.171530	0.184553
WoodDeckSF	0.109737	0.109306
OpenPorchSF	0.105278	0.104372

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?

The model complexity should be at the sweet spot of trade off between bias and variance. Simpler models have higher bias, i.e. the errors are either positively or negatively lopsided and model does not capture the full nuance of the data. On the other hand, overly complex model have high variance, i.e. they have over-learnt the training data and if the training date changes, the model changes drastically. Such models usually perform poorly on unseen data. The model should not be so complex that it performs poorly on unseen data, nor so simple that it just doesn't have the predictive power.

Techniques like ridge and lasso regression are used to obtain the right spot where model complexity is just right and it neither underfits nore overfits.