Advanced Regression Assignment

Q1. 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?

A1. The Optimal Values of alpha for lasso regression comes out to be 0.001 while for ridge regression it comes out to be 20. If we take twice the value of alpha i.e. 0.002 for Lasso and 40 for Ridge, the r-square decreases slightly which is what is expected because as we increase alpha, the model emphasises more on regularisation than on minimising the error. Because the change was very less, most of the significant predictors have remained the same. Some new additions were – FullBath, FireplaceQu_no_fireplace and ExterQual_TA.

Q2. 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?

A2. The model that I will choose is Ridge, as this model gives me the better r-square value of 0.81 vs 0.78 I get from Lasso. The one thing good about Lasso is that most of the features have coefficients as 0 and hence feature selection has automatically taken place but with ridge too, these features have very less coefficients.

Q3. 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?

A3. The top 5 most significant variables I had using the Ridge Model was - 'TotRmsAbvGrd','TotalBsmtSF','1stFlrSF','GrLivArea','OverallQual'

After removing them and applying Ridge again there was a drop in r square value from 0.81 to 0.76. The most significant variables I got were- 'OverallCond','BsmtFinSF2','2ndFlrSF','BsmtUnfSF','BsmtFinSF1'.

Q4. 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?

A4. It is very important that a model is generalisable. It makes sure that the results we are getting will be more of less stable. If the model is not generalisable and is very complex i.e. it overfits the training data, any small change in the data can make it predict very different results. Hence the accuracy can highly change. Usually in such models the accuracy on training data is very high but on test data it dives down.

Some things which make the model less stable are -

1. If a model is too complex, there are very high chances of overfitting which basically results in high variability in results leading to a less robust model. Usually in such a scenario the r square on training set is good but on test set it performs badly.
2. Due to high multicollinearity, even though the predictive power of the model remains the same, the inferences drawn might change rapidly. We would not be able to find out which are the most significant features.

Things to improve stability and generalisability -

1. Making the model less complex helps - We can remove features which are not significant (Feature Selection) or make the coefficients of the model simpler(Regression).
2. In regression, the regression coefficient Alpha(hyperparameter) decides the amount of generalisability. If we increase alpha, the model penalises the model for complexity and we get a less complex and more generalisable model. While if we decrease alpha, the emphasis is more on accuracy and we may get a more complex and less stable model(chances of over fitting are high).