Problem Statement - Part II

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?

Answer 1:

Ridge:
Lambda = 2
R2 score (train): 0.917904
R2 score (test): 0.879608
MSE (train): 0.113599
MSE (test): 0.1407

Lasso:
Lambda = 0.001
R2 score (train): 0.9169
R2 score (test): 0.8779
MSE (train): 0.1142
MSE (test): 0.1417

After make the double alpha for ridge and lasso i.e., 4 and 0.002

For Ridge: Coefficient values of features have decreased as alpha increased. r2_score of train data is also drop from .917 to 0.915.
For Lasso: As alpha value increased more features got their coefficient 0 and were removed from model. Coefficient values of features have decreased as alpha increased. R2_score dropped from 0.9169 to 0.873 for train data.
Top Features: OverallQual_Poor, BsmtQual_No Basement, OverallQual_Excellent, Neighborhood_MeadowV, OverallCond_fair

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?

Answer 2:

For final model Lasso has been chosen because of higher r2 test score and lower RMSE value. Furthermore, Lasso reduced the coefficients of 24 features to 0 which were not significant and reduced overfitting. Thus, removing unwanted features from model without affecting the model accuracy. Which makes are model generalized and simple and accurate.

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?

Answer 3:

Top Features: OverallQual_Poor, BsmtQual_No Basement, OverallQual_Excellent, Neighborhood_MeadowV, OverallCond_fair

After dropping them r2 score on train set reduced from 0.91 to 0.889 and on test set from 0.87 to 0.85.

Now topmost features are: Neighborhood_NridgHt, MSSubClass_1-STORY 1945 & OLDER, GrLivArea, Neighborhood_Edwards, Neighborhood_Crawfor.

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?

Answer 4:

Occam's Rule: The simplest explanation is the best one.

When simpler and fewer complex models are used accuracy will decrease slightly but the model will be more robust and generalisable. It will be immune to small changes in data.

It can be also understood using the Bias-Variance trade-off. Higher bias will make the model more robust but will also lead to decrease in accuracy and higher bias will lead to under fitting of the data. So, there should be balance between bias and variance to build a robust model which does not take a toll on accuracy/r2 score.