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:

Ridge:

Lambda = 30
R2 score (train) : 0.9095
R2 score (test) : 0.8896
RMSE (train) : 0.1202
RMSE (test) : 0.1324

Lasso:

Lambda = 0.00125
R2 score (train) : 0.9092
R2 score (test) : 0.8923
RMSE (train) : 0.1204
RMSE (test) : 0.1308

After make the double alpha for ridge and lasso i.e. 60 and 0.0025

For Ridge: Coefficient values of features have decreased as alpha increased. r2_score of train data is also drop from .909 to 0.908.

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 0.9092 to 0.9077 for train data.

Top Features: 1stFlrSF, 2ndFlrSF, OverallQual, YearBuilt, OverallCond

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 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 5 features are 1stFlrSF, 2ndFlrSF, OverallQual, YearBuilt, OverallCond

After dropping them r2 score on train set reduced from 0.9077 to 0.8753 and on test set from 0.89 to 0.85.

Now topmost features are: BsmtFinSF1, TotRmsAbvGrd, FullBath, BsmtUnfSF, GarageArea.

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 less 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.