Assignment 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?

[Ans]

1. optimal value of alpha:- ridge (0.7) lasso (20).
2. changes in the model if you choose double the value of alpha for both ridge and lasso:- R^2 value of both Ridge & Lasso decreased .
3. Top 5 most important predictor variables after the change is implemented-
    a. Ridge - GrLivArea, OverallQual, MasVnrArea, Fireplaces, FullBath (refer to the last part of the code).
    b. Lasso - GrLivArea, OverallQual, MasVnrArea, Fireplaces, FullBath (refer to the last part of the code).

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]

Lasso regression would be a better option (a) It has slightly better r2 (b) it would make the coefficients to 0 for all the irrelevant columns (Feature Elimination). This makes model more robust.

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]

five most important predictor variables –

GrLivArea, OverallQual, MasVnrArea, Fireplaces, FullBath

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]
The model should be as simple as possible as this will increase its robustness and more generic , though accuracy will be reduced. The trade-off between bias and variance can also be used to understand it. A simple model will have more bias but less variance and is more generic. a robust and generalizable model will perform similarly on both training and test data, i.e., the accuracy does not change much for training and test data.