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?

Answer1:

Optimal value of alpha for ridge: 10
Optimal value of alpha for ridge: 100
After double of alpha for ridge and lasso i.e. 20 and 200
For Ridge: Coeff values are increasing as alpha will increase.
There is a drop in R2 score also of train and test ;
R2 score train : 0.7973 to 0.787
R2 score test : 0.790 to 0.785

For Lasso : As the value of alpha increases more features were removed from the model.
There is a drop in R2 score also of train and test ;
R2 score train : 0.797 to 0.793
R2 score test : 0.795 to 0.786

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:

We will choose Lasso as its giving feature selection option also. It has removed 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 five features are : OverallQual,1stFlrSF,2ndFlrSF, MSSubClass_90,MSSubClass_120. After dropping them model accuracy reduced from 79.7% and 79% to 48.5% and 39.6%. Now top most features are: MSSubClass_160, MSZoning_RL, LotConfig_CulDSac, Neighborhood_Crawfor, Neighborhood_NoRidge.

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:

To make model robust and generalisable 3 features are required:
1. Model accuracy should be > 70-75%: I our case its coming 80%(Train) and 81%(Test) which is correct.
2. P-value of all the features is < 0.05
3. VIF of all the features are < 5
Thus we are sure that model is robust and generalisable.