Problem Statement - Part II

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: Optimal value of alpha for ridge: **10**

Optimal value of alpha for ridge: **100**

After making the double alpha for ridge and lasso i.e. **20 and 200**

For Ridge: Coeff values are increasing as alpha will increase. r2_score of train data is also drop from 0.83 to 0.81.

For Lasso: As alpha value increased more features removed from model. But r2score is also dropped by 1% in both test and train data.

Top 5 features :

Neighborhood_NoRidge, Neighborhood_NridgHt,
2ndFlrSF, OverallQual, Neighborhood_Somerst

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: I will choose **Lasso** as its giving feature selection option. It has removed unwanted features from model without affecting the model accuracy. Which makes are model generalized and simple and accurate.

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: Top 5 features are ***Neighborhood_NoRidge, Neighborhood_NridgHt, 2ndFlrSF, OverallQual, Neighborhood_Somerst***. After dropping them model accuracy reduced from 81 and 79% to 58% and 59%.
Next top 5 features after dropping 5 main are:
LotConfig_CulDSac, Exterior2nd_BrkFace, Neighborhood_StoneBr, 1stFlrSF, HouseStyle_1Story

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: One must consider below 3 features to make model robust and generalisable:

1. **Model accuracy should be > 70-75%:** In our case its coming 81%(Train) and 79%(Test) which is correct.
2. **P-value of all the features is < 0.05**
3. **VIF of all the features are < 5**

Hence, we are sure that model is robust and generalisable.