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: The optimal value of alpha for lasso is 20 and for ridge is 50.

After making the alpha for lasso as 40 :

a. The r2 score for train data decreased and for test data there is an increase.
b. The RMSE also decreases.

After making the alpha for ridge as 100 :

a. The r2 score for both train and test data decreased slightly.
b. The RMSE increases by a slight amount.

The most important variables after the change is implemented are below. These have the positive effect on the model i.e. they increase the overall r2_square of the model.

a. PoolQC_No Pool
b. RoofMatl_WdShngl
c. PoolArea
d. Neighborhood_StoneBr
e. Neighborhood_NoRidge
f. 2ndFlrSF
g. Street_Pave
h. Neighborhood_NridgHt
i. SaleType_New
j. Condition1_Norm

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 2: The optimal value of alpha for lasso is 20 and for ridge is 50.

The r2 square value of test data for ridge is more than lasso and RMSE is also lesser but the number of coefficients used in lasso are quite less as compared to ridge and it would have a significant improvement in computation complexity of the model, hence as a trade-off it is better to use lasso regression technique.

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 3:

Current 5 most important variables are:

1. PoolQC_No Pool
2. PoolArea
3. RoofMatl_WdShngl
4. Neighborhood_StoneBr
5. RoofMatl_Membran

If these aren't available in the incoming data then the next 5 most important variables will be:

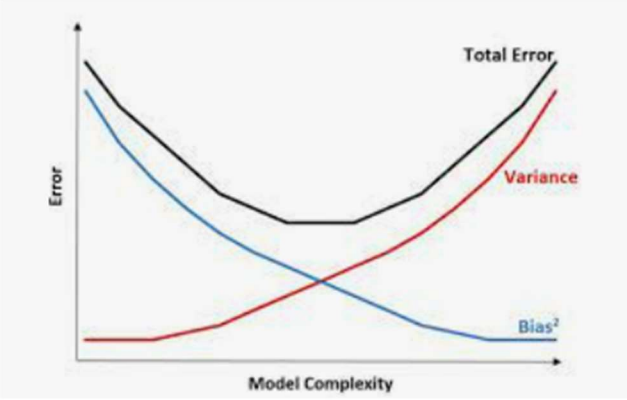
1. Street_Pave
2. Neighborhood_NoRidge
3. 2ndFlrSF
4. Functional_Typ
5. SaleType_New

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 4:

The model should be as simple as possible to make it robust and generalisable. It should adhere to the Bias-Variance trade-off rule.



As per the above figure the point where model is robust is where the bias and variance curve intersect. This point will give us the location where model will behave optimally in terms of accuracy, both for the test and train data. This means that the model should be generalized so that the test accuracy should not be lesser than the training score. The outliers should also be removed and only those are retained which doesn't alter the model accuracy.