Q1

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:

The actual implemented Ridge and Lasso alpha values are

Ridge = 4.7
Lass = 0.0001

After doubling the alpha value i.e

Ridge = 9.4

The R2 Score becomes 0.92 on train data and 0.89 on test data.

Important variables
GrLiveArea, MSZoning_FV, MSZoning_RL, MSZoning_RH, MSZoning_RM

Lasso = 0.0002

The R2 Score becomes 0.93 on train data and 0.89 on test data.

Important variables
GrLiveArea, MSZoning_FV, MSZoning_RL, MSZoning_RH, MSZoning_RM

In both the cases, the important variables remain same.

Q2:

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:

The lambda value of Ridge is 4.7 and Lassos's 0.0001

Ridge
Train R2 Score 0.93
Test R2 Score 0.90

Lasso

Train R2 Score 0.94
Test R2 Score 0.89

I will choose Lasso as the train score is higher than Lasso. Also, Lasso has done feature selection by making lots of feature's coeff 0. This makes the model easy to interpret.

Q3:

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:

If the five important predictor variables aren't in the incoming data, then next five important predictor can be picked up in this case and those are

1. MSZoning_FV
2. RoofMatl_Membran
3. RoofMatl_Metal
4. MSZoning_RH
5. MSZoning_RL

Q4:

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:

Generalization refers to a model's ability to adapt to new and never seen data and to do so we divide the data into train and test set.

If we see here that in both the modes (Ridge and Lasso), the test R2 Score is really close to the train R2 Score and this implies that the model is quite robust to the never unseen data.