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?

Answer: The optimal value for alpha in Ridge and lasso regression in my case are –

➢ Ridge = 500
➢ Lasso = 0.01

If I doubled the value of alpha (i.e. 1000 for ridge and 0.02 for lasso) then my lasso regression will be a better fit for my model as the R2 for test data set in lasso regression would be 0.87 and for ridge R2 will be slightly less, which is 0.86.

The most important predictor variable would be 'YearBuilt' , even after the alpha is doubled.

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?

Answer: From the analyses we will select Lasso regression as the optimal method over Ridge and linear regression as the R2 is 0.90 for training data and 0.87 for test data, which is slightly higher than Ridge regression.

Also, Lasso regression helps in feature selection, we will use only features that are non-zero as determined in Lasso regression and manipulate essential features.

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?

Answer: The next most important 5 variables are 'SaleCondition_AdjLand', 'SaleCondition_Alloca', 'SaleCondition_Family', 'SaleCondition_Normal', 'SaleCondition_Partial'.

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?

Answer:

We can say that model is robust and generalisable as the residuals are equally distributed across the 0 axis and the histogram of error terms are perfectly normalized.
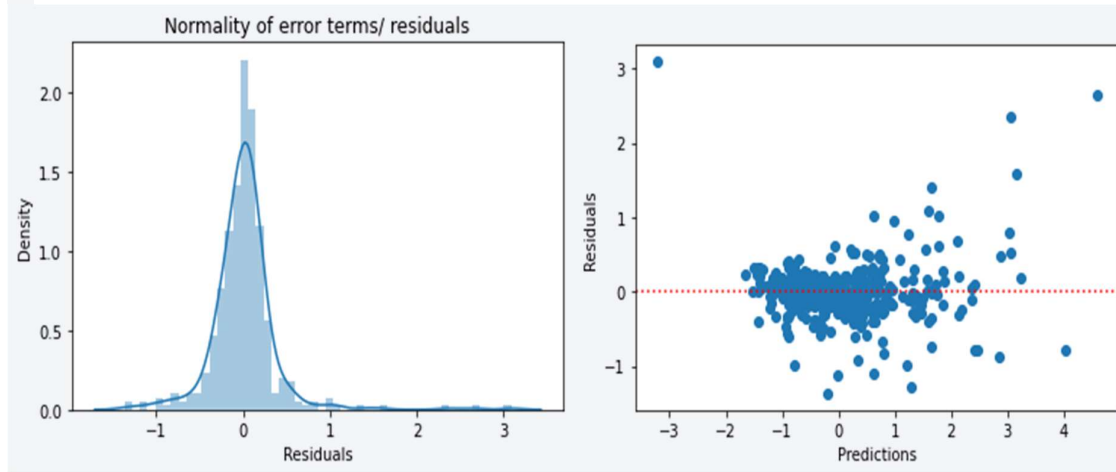
The accuracy of the model could be known from the Root mean square error values, which in case after lasso regression comes around 0.38 for test data.

Git Repository link:
https://github.com/harishivan/SurpriseHousing.git