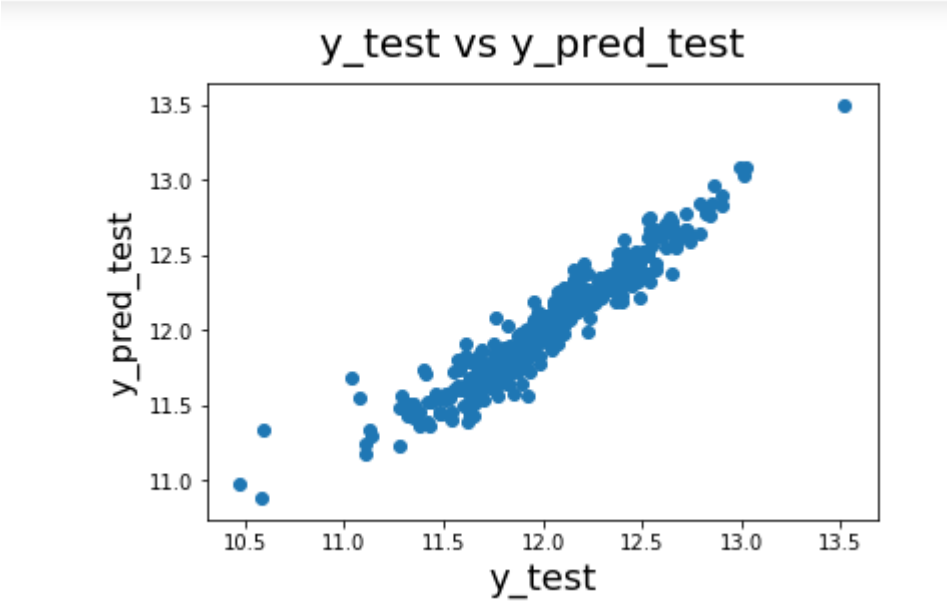
Assignment Part-II

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?

Answer: As per the model building assignment (Part I) the optimal value **for ridge regression** is **15** with the **ridge score of -0.013449534771624779**



The optimal value for Lasso Regression **is [0.08632293 0.1360321 0.09900142 0.12121709 0.12498837 0.10886999**

0.1428732 0.09987998 0.10029023 0.10661001 0.09077802]

Lasso RMSE: **0.1106 (0.0175)**

If we double the value of alpha , then for Ridge Regression, since we have normalized features the impact would be the coefficients will be halved while the others will remain unchanged and in case of Lasso Regression the more coefficients will be moving towards zero as the value of alpha increases or is doubled.

The most important predictor variables are: Year Built,GrLivArea, OverallQual etc.

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: For Ridge Regression the alpha value suitable given by the best_params_ function is 15 for the best alpha, so the same value 15 is being used in the Ridge Regression as out of all the 11 values provided to the ridge in the parameters, the best suited was for the 15 with a ridge score (-0.013449534771624779)

As for Lasso Regression, we think that this is the best suited model as this performs feature selection as well by moving the coefficients to 0, and the model will become more robust, and the mean score is also **0.1106**

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:

Currently with this models prediction, the top 5 features for predicting house prices were :

- GrLivArea
- Year Built
- OverallQual
- GarageCars
- Bath

Suppose, now if we are creating another model excluding these 5 important variables then in that case, my derived variables will become the most important features and they could be:

- Total Area
- Total Bathrooms
- Total Porch Area
- Overall Condition
- Year Remod etc.

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:

Our model should be robust and generalizable, and for this we need to make sure that the data does not contain and outliers and should have been treated if present just like our case, where we have removed the SalePrice outlier where for a house area less than maximum of house area the price was maximum and was definitely an outlier hence impacting our predictions which was then treated/removed.

The model should be generalisable in the sense that the accuracy of our models both Ridge and Lasso is not too abnormal for the train and the test sets.
The robust model can be trusted for predictive analysis whereas a non –robust model cannot be trusted.
Therefore, the accuracy, the outlier treatment/removal al needs to be taken care of while thinking about the model's robustness or accuracy.