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 1 :

Optimal value (as per Grid Search CV) for :
Ridge Regression – **0.6**
Top predictor variable – ***GrLivArea*** : ‘Above Ground Living Area’
Lasso Regression – **50**
Top predictor variable – ***GrLivArea*** : ‘Above Ground Living Area’

When the model is fitted again after doubling the Alpha value (1.2, 100) there is no difference in either Ridge or Lasso Regression,
Still the top predictor variable is ‘**GrLivArea’.**

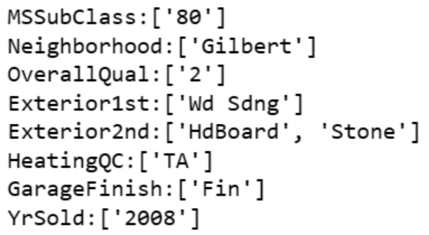
Question 2:

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 2:

First, lets compare both models:

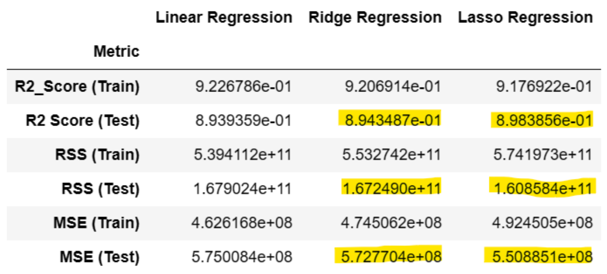
- Lasso Regression removed following features:



```
MSSubClass:['80']
Neighborhood:['Gilbert']
OverallQual:['2']
Exterior1st:['Wd Sdng']
Exterior2nd:['HdBoard', 'Stone']
HeatingQC:['TA']
GarageFinish:['Fin']
YrSold:['2008']
```

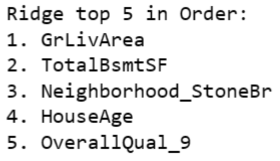
  - And Ridge removed None

- Alpha value:
  - Ridge : 0.6
  - Lasso : 50
- Scores:

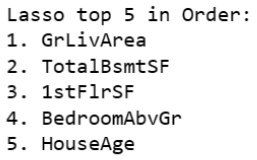


Metric	Linear Regression	Ridge Regression	Lasso Regression
R2_Score (Train)	9.226786e-01	9.206914e-01	9.176922e-01
R2 Score (Test)	8.939359e-01	8.943487e-01	8.983856e-01
RSS (Train)	5.394112e+11	5.532742e+11	5.741973e+11
RSS (Test)	1.679024e+11	1.672490e+11	1.608584e+11
MSE (Train)	4.626168e+08	4.745062e+08	4.924505e+08
MSE (Test)	5.750084e+08	5.727704e+08	5.508851e+08

- Top 5 features:



```
Ridge top 5 in Order:
1. GrLivArea
2. TotalBsmtSF
3. Neighborhood_StoneBr
4. HouseAge
5. OverallQual_9
```



```
Lasso top 5 in Order:
1. GrLivArea
2. TotalBsmtSF
3. 1stFlrSF
4. BedroomAbvGr
5. HouseAge
```

Conclusion:

- Lasso's Regression's alpha value is higher than Ridge but still it is not very high.
- But, with this higher alpha value, it is still selecting better top 5 features from Ridge Regression.
- Lasso’s test scores are slightly better than Ridge scores.
- And after performing lasso some features are also eliminated which can not be done in Ridge regression.

Thus, my choice of model would be ‘***Lasso Regression***’ for this problem.

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 3:

Initially below are the top 5 predictor variable:

1. GrLivArea
2. TotalBsmtSF
3. 1stFlrSF
4. BedroomAbvGr
5. HouseAge

When above five predictor variables are not available, then, on creating another Lasso model below are the next Top 5 predictor variable:

1. TotRmsAbvGrd
2. BsmtFinSF1
3. Neighborhood_StoneBr
4. BsmtUnfSF
5. OverallQual_9

Question 4:

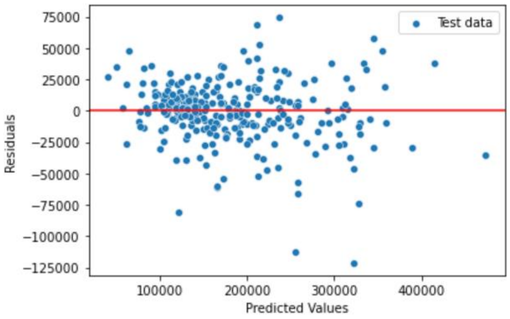
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer 4:

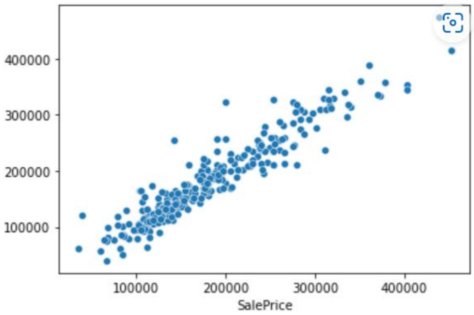
To ensure that the model is Robust and Generalisable, i have performed Cross validation in Grid Search CV to tune hyperparameter, in this case Alpha value.

So the after performing modelling with optimal alpha value (50), below are following observations:

1. final scores:
    - Train Scores:
        - R2 Score: 92.3
        - RSS: 53.9
        - MSE: 46.3
    - Test Scores:
        - R2 Score: 89.5
        - RSS: 16.7
        - MSE: 57.2
2. Residual Analysis : All the points are randomly scattered and no pattern is observed



3. Actual vs Predicted values follow a linear pattern.



Above scores can be considered as Good scores for test data.
So, from above observations i can say the model is 'Robust' and 'Generalisable'.

The implications of a Robust and Generalisable model are that, the model is most likely to perform well on Unseen data, and this is because of following reasons:

- The model is trained on variety of dataset and can identify patterns in a dataset, as it was trained using Cross-Validation
- Model is not relying on specific pattern.
- According to the scores the model is also not Overfitting on Training data.