

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 –

Ridge Alpha value – 3, value of alpha for lasso: 0.0001

After make the double alpha for ridge and lasso i.e. **6 and 0.0002**

After making the changes we seeing below changes -

Ridge-

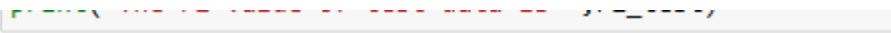
- When we double the alpha value there is a slight increase in the mean squared error whereas the r2 value of train and test remains almost same.

```
The output when alpha is 2:
The mean squared error value is  0.002186087311512901
The r2 value of train data is  0.91843530557298
The r2 value of test data is  0.9067943719752616

The output when alpha is 4:
The mean squared error value is  0.0022184899668293515
The r2 value of train data is  0.91843530557298
The r2 value of test data is  0.9067943719752616
```

Lasso-

- When we double the alpha there is a slight increase in the mean squared error, the r2 value of train slightly decreases whereas there is a huge fall in the r2 value of test thus making the model and prediction worse.



```
The output when alpha is 0.0001:
The mean squared error value is  0.002165519440146776
The r2 value of train data is  0.9241248183546377
The r2 value of test data is  0.9068657096423898

The output when alpha is 0.0002:
The mean squared error value is  0.0021671781536494778
The r2 value of train data is  0.91843530557298
The r2 value of test data is  0.9067943719752616
```

During Double Alpha value below are the top 10 features-

Ridge -

Top correlated features when alpha is 6 are :

	Coefficient
OverallQual	0.110406
Total_sqr_footage	0.108212
GrLivArea	0.093727
Neighborhood_StoneBr	0.072983
LotArea	0.056217
TotalBsmtSF	0.054989
OverallCond	0.053870
Neighborhood_NridgHt	0.049009
YearBuilt	0.041820
Total_porch_sf	0.040380

Lasso -

Top correlated features when alpha is 0.0002 are:

	Coefficient
Total_sqr_footage	0.200119
OverallQual	0.181269
YearBuilt	0.129526
Neighborhood_StoneBr	0.093954
OverallCond	0.086215
GrLivArea	0.080870
BsmtUnfSF	0.072364
LotArea	0.061624
Neighborhood_NridgHt	0.059778
Neighborhood_Crawfor	0.050292

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-

We preferring LASOO respect to Ridge because lasso have one advantage that it's provided to us a feature selection option that is not affecting the model accuracy. It is making the model simple, robust and generalized in nature.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.924239	0.924302	0.924125
1	R2 Score (Test)	0.892745	0.905981	0.906866
2	RSS (Train)	1.477075	1.475845	1.479294
3	RSS (Test)	0.910256	0.797922	0.790415
4	MSE (Train)	0.041686	0.041669	0.041717
5	MSE (Test)	0.049938	0.046756	0.046535

The RSS for the test data has reduced from **0.910** to **0.7979** and **0.7904** for Ridge and Lasso respectively (lower the value better)

The MSE for the test data has reduced from 0.0499 to **0.0467** and **0.0465** for Ridge and Lasso respectively (lower the value better)

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?

If we dropping the most 5 predictor variables of lasso model and again build the model then we got below 5 predictors -

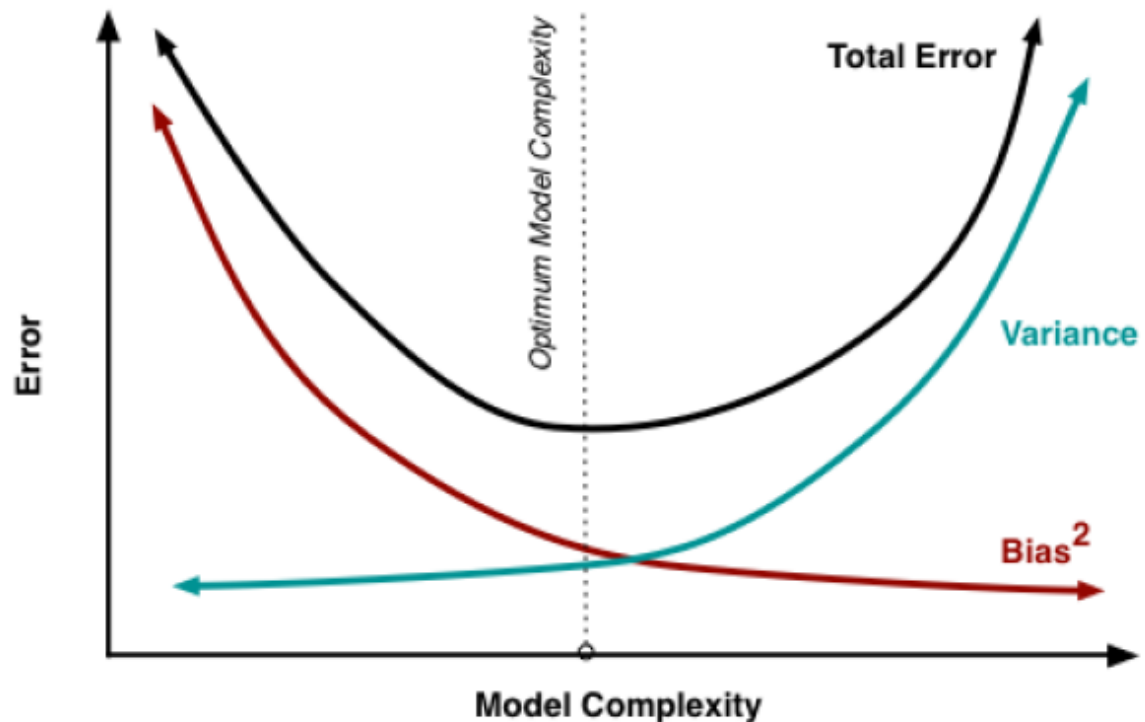
GrLivArea
TotalBsmtSF
GarageArea
LotArea
LandContour_HLS

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-

Model has simple as much as possible its accuracy though its accuracy will be decreased but it will be more robust. We also understand that using bias-variance trade off graph given below-



Sometime model facing issue of over and under fitting which one resolve using regularisation.

And other features are below –

Model accuracy should be $> 70-75\%$

P-value of all the features is < 0

VIF of all the features are < 5