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

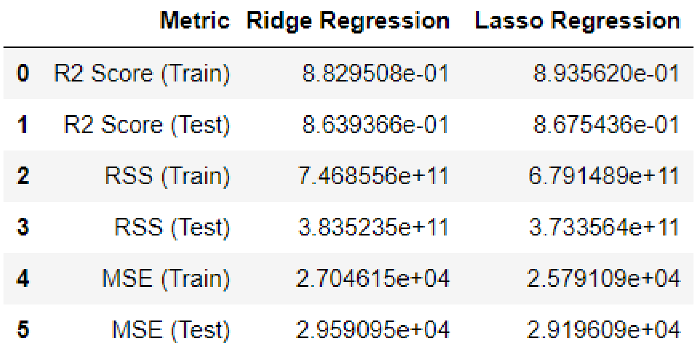
The optimal value of alpha for ridge regression is 20

The optimal value of alpha for lasso regression is 100.

The most important predictors for these optimal value of alpha are:-

OverallQuality - Excellent
Neighborhood - NoRidge
OverallQual - Very Excellent
Neighborhood - Northridge Heights
GrLivArea
BsmtExposure – Gd

The R2 score, RSS and MSE are as follows:-



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.829508e-01	8.935620e-01
1	R2 Score (Test)	8.639366e-01	8.675436e-01
2	RSS (Train)	7.468556e+11	6.791489e+11
3	RSS (Test)	3.835235e+11	3.733564e+11
4	MSE (Train)	2.704615e+04	2.579109e+04
5	MSE (Test)	2.959095e+04	2.919609e+04

After we doubled the value of alpha for both ridge and lasso, the important predictors(in order) are:-

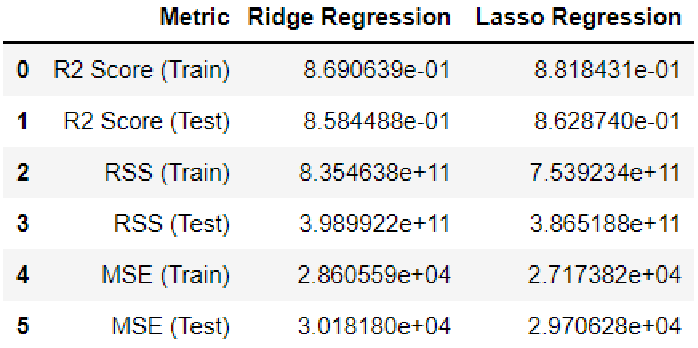
OverallQuality - Excellent
Neighborhood – NoRidge

GrLivArea

Neighborhood - Northridge Heights

2ndFlrSF

The R2 score, RSS and MSE after we double the alpha are as follows:-



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.690639e-01	8.818431e-01
1	R2 Score (Test)	8.584488e-01	8.628740e-01
2	RSS (Train)	8.354638e+11	7.539234e+11
3	RSS (Test)	3.989922e+11	3.865188e+11
4	MSE (Train)	2.860559e+04	2.717382e+04
5	MSE (Test)	3.018180e+04	2.970628e+04

There is a marginal drop in R2 test score and marginal increase in MSE(test) score after increasing the alpha.

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

We will choose to apply Lasso regression. The R2 score(Test) of Lasso regression is better than Ridge regression and also MSE score of Lasso regression is less than Ridge regression which means that the prediction using lasso regression would be better. Also, many of the coefficient in lasso regression are 0 which means it will be less complex when compared with Ridge regression.

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

The five most important predictor variables after deleting 5 most important predictors are:-

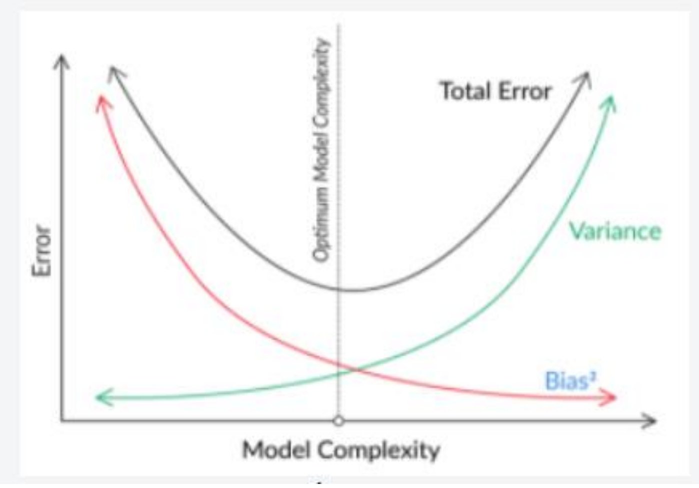
2ndFlrSF, 1stFlrSF, BsmtExposure_Gd, Neighborhood_Crawfor, OverallQual_VeryGood

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 make the Model robust and generalizable, the model should not overfit the training data i.e. it should not mug up the entire training data but should be able to identify underlying patterns. The model complexity should neither be too high, which would lead to overfitting, nor too low, which would lead to a model with high bias (a biased model) that does not even identify necessary patterns in the data. To overcome the problem of overfitting, we add a regularizing term which tries to push the model coefficients towards 0 and tries to keep the model complexity to minimum. Adding too much weightage to the regularize term might lead to the problem of underfitting. So a balance needs to be maintained.



The implication of using regularization is that it reduces bias marginally but has a very high impact on reducing the model variance.