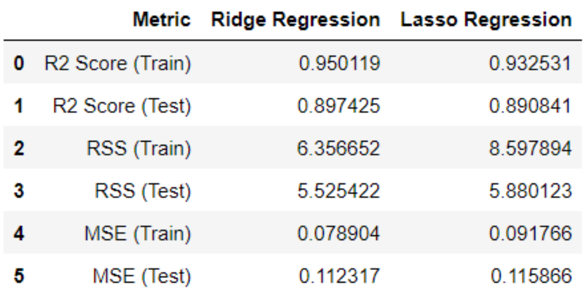
Question 1

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?

Optimal value of alpha for ridge – 8

Optimal value of alpha for lasso – 0.001

Below the comparison table for the optimal alpha

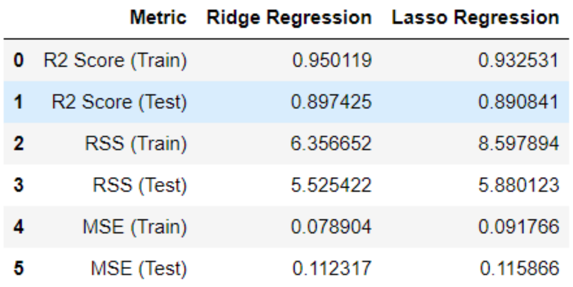


	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.950119	0.932531
1	R2 Score (Test)	0.897425	0.890841
2	RSS (Train)	6.356652	8.597894
3	RSS (Test)	5.525422	5.880123
4	MSE (Train)	0.078904	0.091766
5	MSE (Test)	0.112317	0.115866

Below the comparison table for double the alpha value

Optimal value of alpha for ridge – 8

Optimal value of alpha for lasso – 0.001



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.950119	0.932531
1	R2 Score (Test)	0.897425	0.890841
2	RSS (Train)	6.356652	8.597894
3	RSS (Test)	5.525422	5.880123
4	MSE (Train)	0.078904	0.091766
5	MSE (Test)	0.112317	0.115866

There are not much significant changes after double the alpha values. Below are the most important predictor variables.

GrLivArea
TotalBsmtSF
OverallQual_9
OverallQual_8
YearBuilt
Neighborhood_Crawfor
GarageArea
YearRemodAdd

LotArea
BsmtFinSF1
Neighborhood_Somerst
OverallQual_7

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?

Train vs test – R2 score and mean root square errors of Lasso model are lesser than the Ridge model. Hence, I am choosing Lasso model and it’s lambda (hyper parameter) from this analysis. As well Lasso allow us to choose predictor variables with ease, many unimportant feature’s coefficient became zero in Lasso outputs.

Question 3

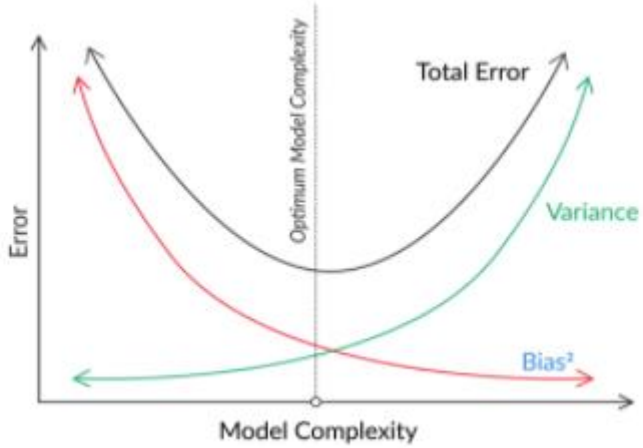
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Condition2_PosA
1stFlrSF
2ndFlrSF
BsmtFinSF1
SaleType_ConLD

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?

Overfitting is a major challenge in regression models. Model performs well with train data and doesn’t perform well on test data. These models show high variance and low bias which leads to less performance on unseen data. Finding the model with optimum value of variance and bias for given data becomes robust and this model can be generalizable.



Looking at the above graph, we must find the lowest total error which falls at intersection of the variance and bias. That is low bias and low variance, such that model identifies the all the patterns and perform well with unseen data. This can be possible by adding penalty term in cost function and find the optimum hyperparameter(penalty) which is called regularization.

Accuracy of the model can be increased by adding more data, feature scaling, relevant feature selection, regularization and hyperparameter tuning, etc.,