Question 1

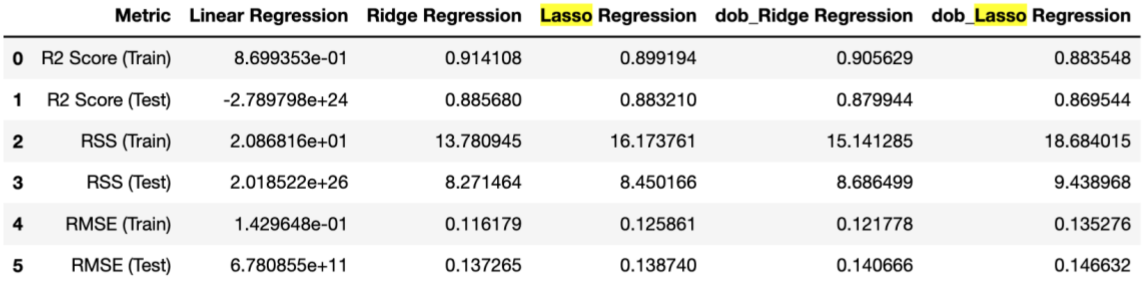
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Optimal values of alpha found through cross validation are given below.

- ridge - 4.6
- lasso - 0.001

After doubling the alpha values, there is no considerable change in the performance of the models. There is a slight increase in the rmse scores for both ridge and lasso. There is 2.4% increase in root mean square error in case of ridge and 5% increase in the case of Lasso. Also the r2 scores have been slightly reduced.



	Metric	Linear Regression	Ridge Regression	Lasso Regression	dob_Ridge Regression	dob_Lasso Regression
0	R2 Score (Train)	8.699353e-01	0.914108	0.899194	0.905629	0.883548
1	R2 Score (Test)	-2.789798e+24	0.885680	0.883210	0.879944	0.869544
2	RSS (Train)	2.086816e+01	13.780945	16.173761	15.141285	18.684015
3	RSS (Test)	2.018522e+26	8.271464	8.450166	8.686499	9.438968
4	RMSE (Train)	1.429648e-01	0.116179	0.125861	0.121778	0.135276
5	RMSE (Test)	6.780855e+11	0.137265	0.138740	0.140666	0.146632

Most important predictor variables remain as **GrLivArea** for lasso model before and after doubling optimum alpha values.

Important variables after doubling alpha for Lasso: GrLivArea, OverallQual, GarageCars, TotRmsAbvGrd, OverallCond

Most important predictor variables remain as **OverallQual** for ridge model before and after doubling optimum alpha values.

Important variables after doubling alpha for Ridge: OverallQual, GrLivArea , TotRmsAbvGrd , 1stFlrSF, OverallCond

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:

Even though Ridge regression has better metrics compared to Lasso in terms of RMSE values and r2 score I am choosing Lasso model over ridge because of two reasons.

1. The Lasso model has handled overfitting very well. As Lasso regression model gives similar performance for both train and test data, it will perform well for unseen data compared to Ridge regression model. There is only 1% difference in r2 scores of train and test sets for lasso.
2. There are many variables present in the dataset and lasso helps to reduce the number of features. So to improve efficiency and to reduce multicollinearity, it's better to select lasso model as it also helps us in feature selection.

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:

'1stFlrSF', '2ndFlrSF', 'GarageArea', 'BsmtQual', 'FullBath' are the new most important 5 predictor variables.

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:

A model becomes robust and generalisable when the variance of the model is less with the changes in data. By making the model as simple as possible we can ensure that the model is not overfitting. At the same time it should not be underfit also by not recognising important patterns in the data. A model that’s not overfit will perform better with the unseen data. We can identify this by partitioning the data to train and test set and checking accuracy of the model for train data and test data.

An overfit model gives a good training accuracy and performs worst with test data. When model complexity increases model tries to memorise the data and fail to capture patterns in the data. As model is not learning patterns in the data, it becomes difficult for the model to give correct predictions when encountered with unseen data. For a good model both train and test accuracy must be close enough.

Overfitting can be dealt using regularization techniques.