Subjective Questions

Q1. 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?

ANS: Optimal Value of Alpha:

Ridge: 5 => Optimal Value

If I make the value of alpha in ridge as 10 my difference in RMSE between train and test set is increasing and R2 score difference between train and test is decreasing.

At Alpha = 5

```
Root Mean Square Error train = 0.10429458889501306
Root Mean Square Error test = 0.14542198366647888


R-Square for training data 0.9257760983758252
R-Square for testing data 0.8881019622141497
```

At Alpha = 10

```
Root Mean Square Error train = 0.11114812490731465
Root Mean Square Error test = 0.15015383464442858


R-Square for training data 0.9157005953778842
R-Square for testing data 0.8807014411109009
```

Lasso: 0.001 => Optimal Value

If I make alpha as 0.002 we can see RMSE in both training and testing data is increasing and the R2 score in both training and testing is dropping. But the difference between R2 score of training and testing is less when alpha is 0.002

At Alpha = 0.001

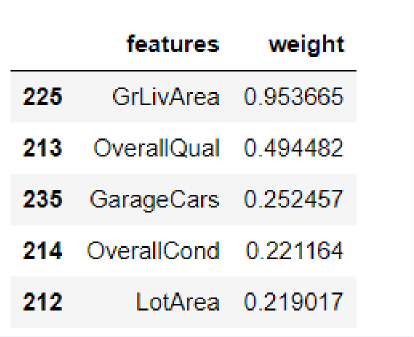
```
Root Mean Square Error train = 0.11393308774671904
Root Mean Square Error test = 0.14150783424647898


R-Square for training data 0.9114232051059425
R-Square for training data 0.8940445475272306
```

At Alpha = 0.002

```
Root Mean Square Error train = 0.12425109673137233
Root Mean Square Error test = 0.1505856370703493
Root Mean Square Error train = 0.12425109673137233
Root Mean Square Error test = 0.1505856370703493
```

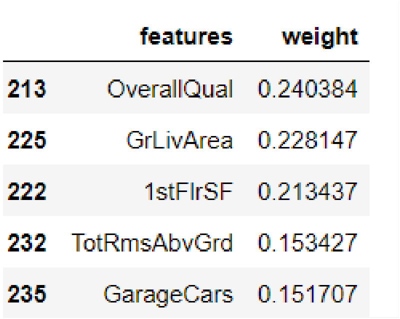
Most Important features after doubling the alpha

Lasso:



	features	weight
225	GrLivArea	0.953665
213	OverallQual	0.494482
235	GarageCars	0.252457
214	OverallCond	0.221164
212	LotArea	0.219017

Ridge:



	features	weight
213	OverallQual	0.240384
225	GrLivArea	0.228147
222	1stFlrSF	0.213437
232	TotRmsAbvGrd	0.153427
235	GarageCars	0.151707

Q2. 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?

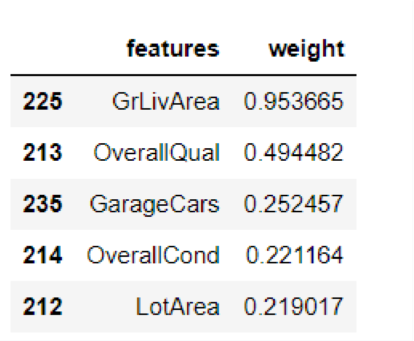
ANS:

I have determined Ridge and Lasso alpha value as 5 and 0.001 respectively. I will be choosing lasso as my model because in lasso I can see the difference between RMSE value and R2 score of training and testing data is less as compare to Ridge.

Q3: 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?

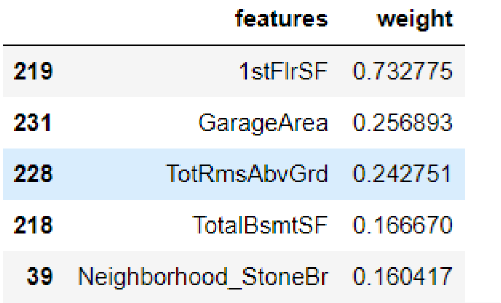
ANS:

Top 5 features initially:



	features	weight
225	GrLivArea	0.953665
213	OverallQual	0.494482
235	GarageCars	0.252457
214	OverallCond	0.221164
212	LotArea	0.219017

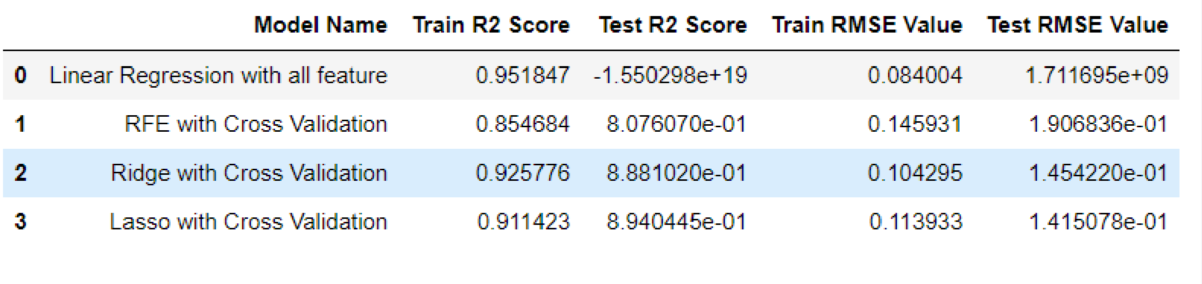
After removing above features top 5 features:



	features	weight
219	1stFlrSF	0.732775
231	GarageArea	0.256893
228	TotRmsAbvGrd	0.242751
218	TotalBsmtSF	0.166670
39	Neighborhood_StoneBr	0.160417

Q4. 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?

ANS:

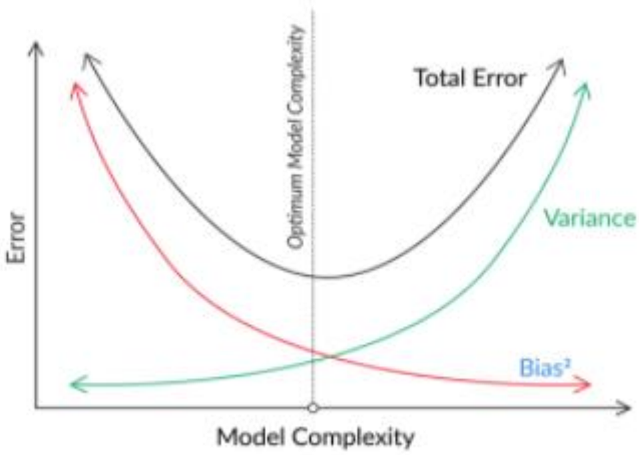


	Model Name	Train R2 Score	Test R2 Score	Train RMSE Value	Test RMSE Value
0	Linear Regression with all feature	0.951847	-1.550298e+19	0.084004	1.711695e+09
1	RFE with Cross Validation	0.854684	8.076070e-01	0.145931	1.906836e-01
2	Ridge with Cross Validation	0.925776	8.881020e-01	0.104295	1.454220e-01
3	Lasso with Cross Validation	0.911423	8.940445e-01	0.113933	1.415078e-01

A generalized and robust model is that model where it is performing well on testing or unseen data. From the above table we can see that first model i.e. Linear regression is not performing well and giving negative R2 score. After that we tried RFE model where it is performing well as compared to Linear regression model. But still the difference in training and testing R2 is more compare to Ridge and lasso. Ridge and Lasso both are performing good but if we talk about generalisation, we can see that Lasso is performing well on testing well on testing data as compare to ridge.

If a model is generalized its accuracy on training and testing set it comparable. It should not be like where the training accuracy is good but testing is very bad. That means model is overfitting. If model is not generalized it means it tried to learn pattern from training dataset so much that it got overfitted and now not being able to predict correctly when we are giving unseen data points.

The base of this is bias and variance. Our model should not learn too much training data, if it does so the variance is high and bias is low. We have to build a model where variance and bias both are low.


Optimum Model Complexity
Total Error
Variance
Error
Bias²
Model Complexity