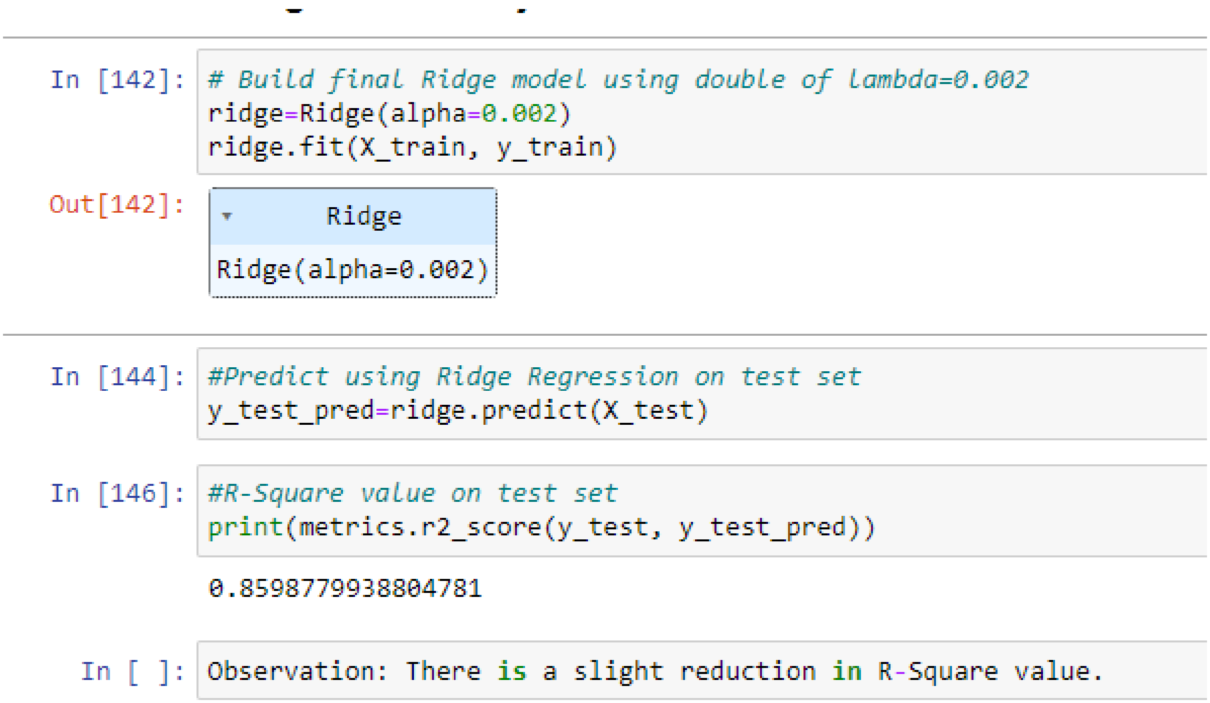
Advanced Regression – Subjective Questions

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?

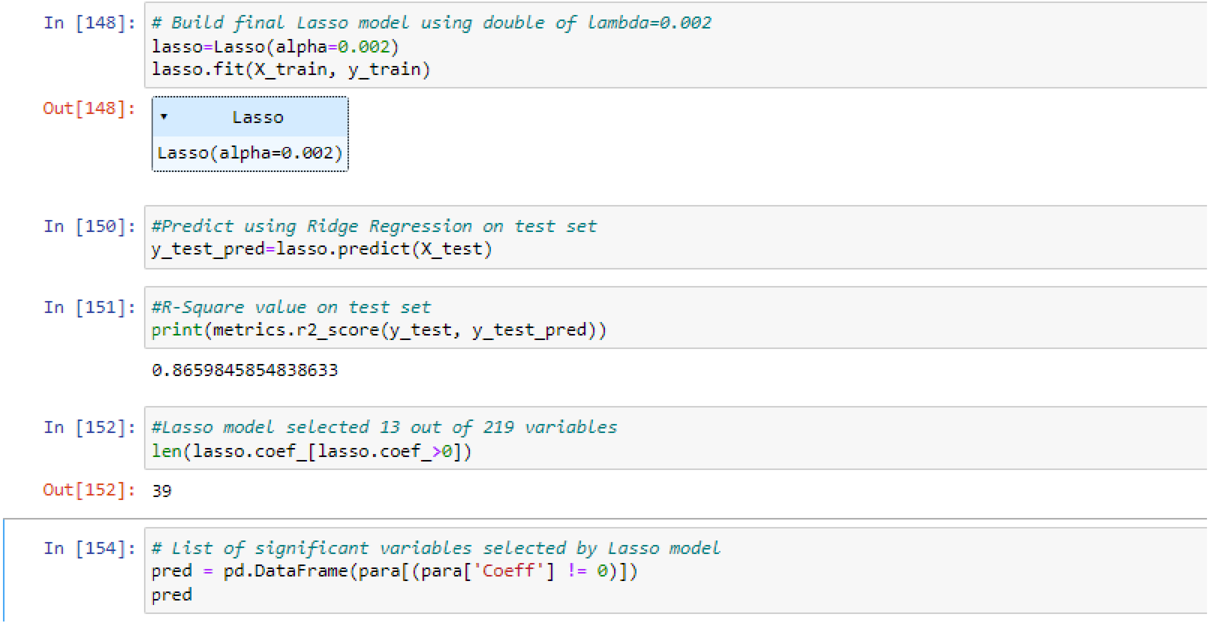
Answer

Ridge

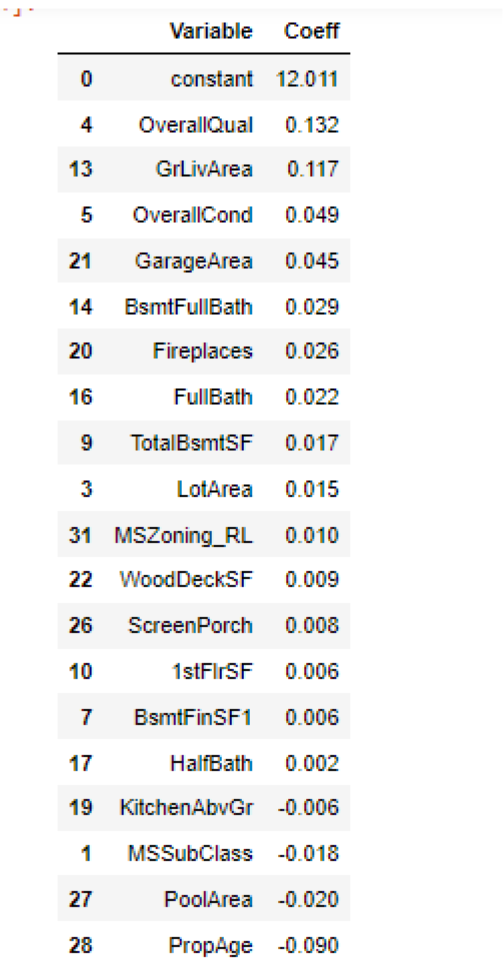


```
In [142]: # Build final Ridge model using double of lambda=0.002
          ridge=Ridge(alpha=0.002)
          ridge.fit(X_train, y_train)

Out[142]:  ▾       Ridge
           Ridge(alpha=0.002)

In [144]: #Predict using Ridge Regression on test set
          y_test_pred=ridge.predict(X_test)

In [146]: #R-Square value on test set
          print(metrics.r2_score(y_test, y_test_pred))

          0.8598779938804781

In [ ]: Observation: There is a slight reduction in R-Square value.
```

Lasso



```
In [148]: # Build final Lasso model using double of lambda=0.002
          lasso=Lasso(alpha=0.002)
          lasso.fit(X_train, y_train)

Out[148]: ▾      Lasso
          Lasso(alpha=0.002)

In [150]: #Predict using Ridge Regression on test set
          y_test_pred=lasso.predict(X_test)

In [151]: #R-Square value on test set
          print(metrics.r2_score(y_test, y_test_pred))

          0.8659845854838633

In [152]: #Lasso model selected 13 out of 219 variables
          len(lasso.coef_[lasso.coef_>0])

Out[152]: 39

In [154]: # List of significant variables selected by Lasso model
          pred = pd.DataFrame(para[(para['Coeff'] != 0)])
          pred
```



	Variable	Coeff
0	constant	12.011
4	OverallQual	0.132
13	GrLivArea	0.117
5	OverallCond	0.049
21	GarageArea	0.045
14	BsmtFullBath	0.029
20	Fireplaces	0.026
16	FullBath	0.022
9	TotalBsmtSF	0.017
3	LotArea	0.015
31	MSZoning_RL	0.010
22	WoodDeckSF	0.009
26	ScreenPorch	0.008
10	1stFlrSF	0.006
7	BsmtFinSF1	0.006
17	HalfBath	0.002
19	KitchenAbvGr	-0.006
1	MSSubClass	-0.018
27	PoolArea	-0.020
28	PropAge	-0.090

Above variable are the most important predictors now.

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

I will prefer Lasso because of following-

Simpler model with less variable

Model is giving decent performance.

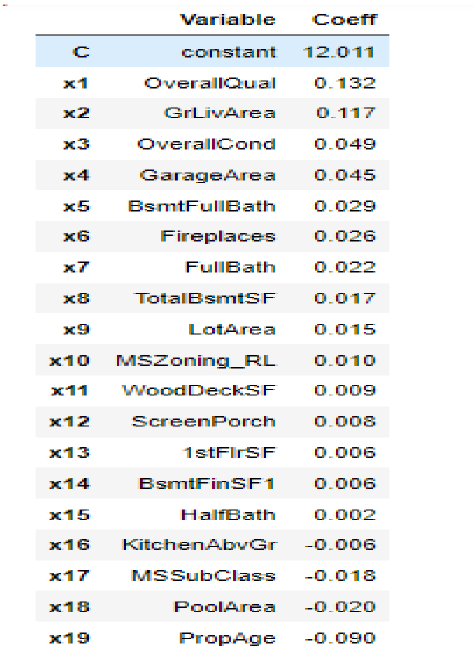
Efficiently solved high dimensionality problem by shrinking insignificant coefficients to zero.

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

From X6 till X10



	Variable	Coeff
C	constant	12.011
x1	OverallQual	0.132
x2	GrLivArea	0.117
x3	OverallCond	0.049
x4	GarageArea	0.045
x5	BsmtFullBath	0.029
x6	Fireplaces	0.026
x7	FullBath	0.022
x8	TotalBsmtSF	0.017
x9	LotArea	0.015
x10	MSZoning_RL	0.010
x11	WoodDeckSF	0.009
x12	ScreenPorch	0.008
x13	1stFlrSF	0.006
x14	BsmtFinSF1	0.006
x15	HalfBath	0.002
x16	KitchenAbvGr	-0.006
x17	MSSubClass	-0.018
x18	PoolArea	-0.020
x19	PropAge	-0.090

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

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.