UPGRAD SUBJECTIVE QUESTION ADVANCED REGRESSION

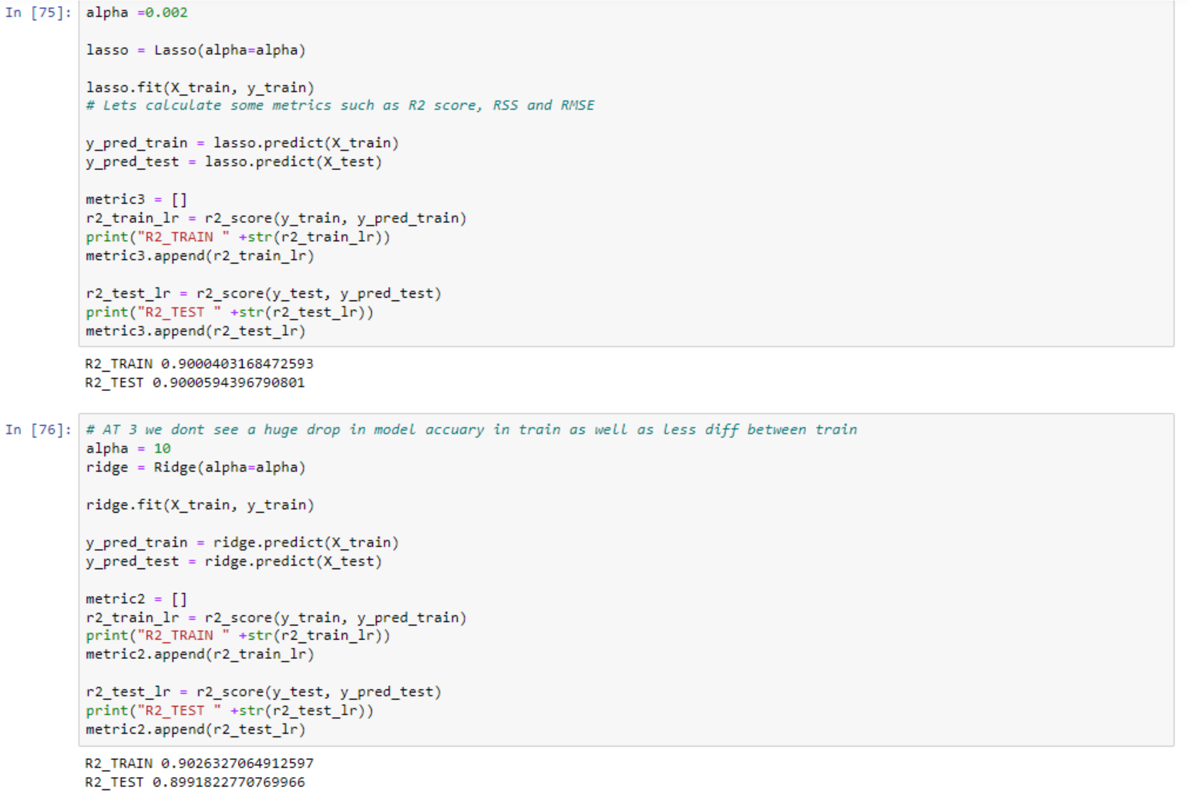
Question-1

1. What is the optimal value of alpha for ridge and lasso regression?

ANS The Optimal Value of Alpha for Ridge is 5 and lasso is 0.001

2. What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

ANS:- if we double the value of ridge it will be 10 and for lasso it will be 0.002 as alpha values are very small so the difference will wont be much.



```
In [75]: alpha =0.002

         lasso = Lasso(alpha=alpha)

         lasso.fit(X_train, y_train)
         # Lets calculate some metrics such as R2 score, RSS and RMSE

         y_pred_train = lasso.predict(X_train)
         y_pred_test = lasso.predict(X_test)

         metric3 = []
         r2_train_lr = r2_score(y_train, y_pred_train)
         print("R2_TRAIN " +str(r2_train_lr))
         metric3.append(r2_train_lr)

         r2_test_lr = r2_score(y_test, y_pred_test)
         print("R2_TEST " +str(r2_test_lr))
         metric3.append(r2_test_lr)

         R2_TRAIN 0.9000403168472593
         R2_TEST 0.9000594396790801

In [76]: # AT 3 we dont see a huge drop in model accuary in train as well as less diff between train
         alpha = 10
         ridge = Ridge(alpha=alpha)

         ridge.fit(X_train, y_train)

         y_pred_train = ridge.predict(X_train)
         y_pred_test = ridge.predict(X_test)

         metric2 = []
         r2_train_lr = r2_score(y_train, y_pred_train)
         print("R2_TRAIN " +str(r2_train_lr))
         metric2.append(r2_train_lr)

         r2_test_lr = r2_score(y_test, y_pred_test)
         print("R2_TEST " +str(r2_test_lr))
         metric2.append(r2_test_lr)

         R2_TRAIN 0.9026327064912597
         R2_TEST 0.8991822770769966
```

3. What will be the most important predictor variables after the change is implemented?

Ans. :- SaleCondition For Ridge with Coeff 0.26

GrLivArea for Lasso with Coeff 0.30

Question-2

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?

Ans :- We got it for ridge and lasso we will be chossing lasso as the accuracy is same for both the method as lasso will help me to remove the variable and it penalise more so that’s the reason to chose lasso.

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?

Ans:- We will be dropping Sale

SaleCondition_Partial

Neighborhood_ClearCr

Neighborhood_Crawfor

Neighborhood_StoneBr

GrLivArea

NOW THE MOST PROMINNENT VARIABLES ARE:

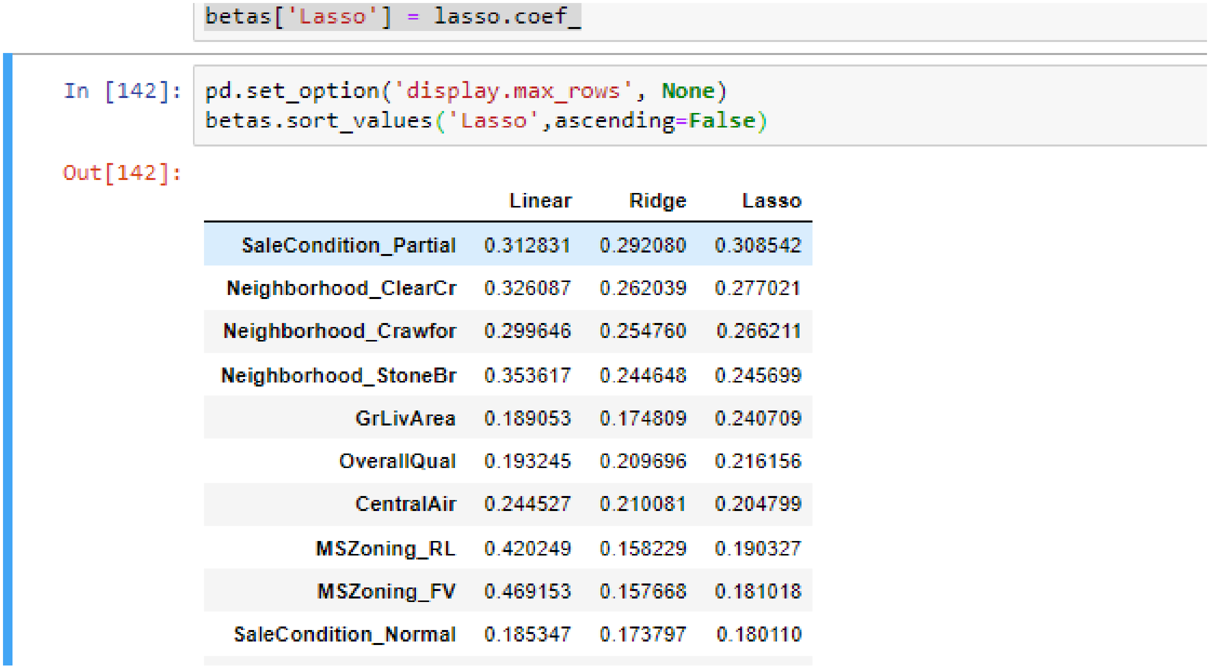
OverallQual

CentralAir

MSZoning_RL

MSZoning_FV

SaleCondition_Normal



```
betas['Lasso'] = lasso.coef_
```

```
In [142]: pd.set_option('display.max_rows', None)
          betas.sort_values('Lasso',ascending=False)
```

Out[142]:

	Linear	Ridge	Lasso
SaleCondition_Partial	0.312831	0.292080	0.308542
Neighborhood_ClearCr	0.326087	0.262039	0.277021
Neighborhood_Crawfor	0.299646	0.254760	0.266211
Neighborhood_StoneBr	0.353617	0.244648	0.245699
GrLivArea	0.189053	0.174809	0.240709
OverallQual	0.193245	0.209696	0.216156
CentralAir	0.244527	0.210081	0.204799
MSZoning_RL	0.420249	0.158229	0.190327
MSZoning_FV	0.469153	0.157668	0.181018
SaleCondition_Normal	0.185347	0.173797	0.180110

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?

Ans.:- As Per, Occam’s Razor— given two models that show similar ’performance’ in the finite training or test data, we should pick the one that makes fewer on the test data due to following reasons:-

▪ Simpler models are usually more ’generic’ and are more widely applicable and widely used

▪ Simpler models require less training samples for good training than the more complex model and hence are easier to train.

▪ Simpler models are more robust and good. o Complex models tend to change wildly with changes in the training data set it may lead to overfitting o Simple models have low variance, high bias and complex models have low bias, high variance o Simpler models make more errors in the training set.

• A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.

• A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed. Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Thus accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error