Question 1

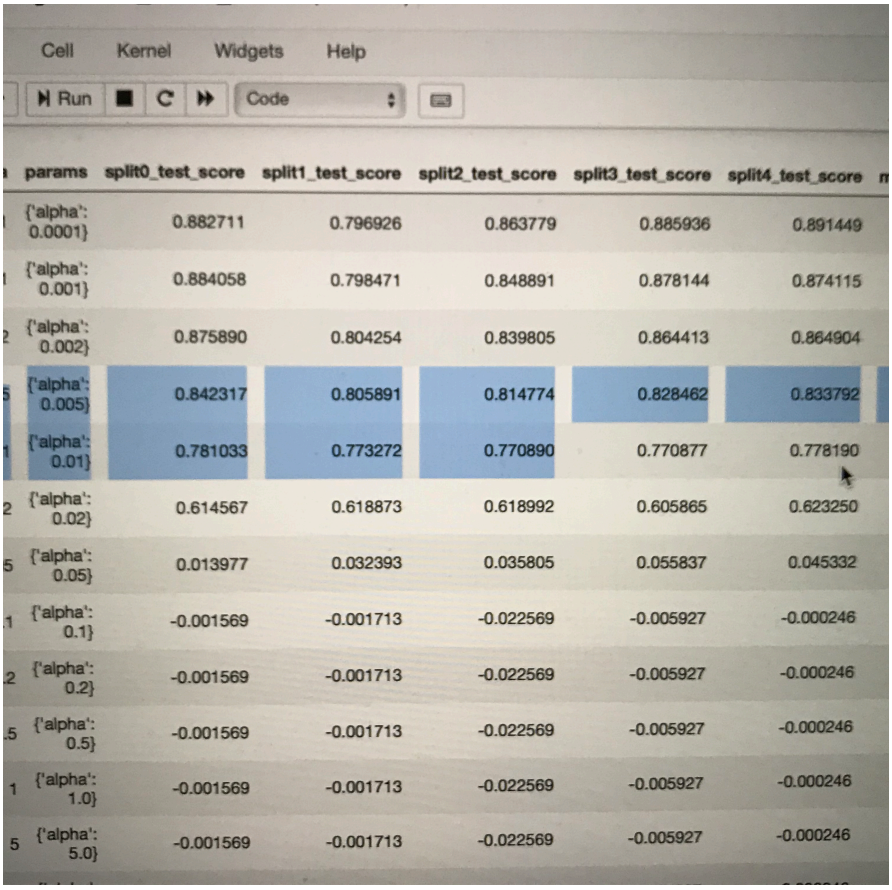
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Optimal values of Alpha for my model - ridge = 20 and lasso=0.005

If we double values then model will become more general due to increase in regularisation term. Model will have more bias as we see from below diagram(dropped R2). Most important predictors observed to be same (details shared below).. I have assumed business need accuracy above 80% for reference.

Lasso regression result:



	params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score
	{'alpha': 0.0001}	0.882711	0.796926	0.863779	0.885936	0.891449
	{'alpha': 0.001}	0.884058	0.798471	0.848891	0.878144	0.874115
2	{'alpha': 0.002}	0.875890	0.804254	0.839805	0.864413	0.864904
5	{'alpha': 0.005}	0.842317	0.805891	0.814774	0.828462	0.833792
1	{'alpha': 0.01}	0.781033	0.773272	0.770890	0.770877	0.778190
2	{'alpha': 0.02}	0.614567	0.618873	0.618992	0.605865	0.623250
5	{'alpha': 0.05}	0.013977	0.032393	0.035805	0.055837	0.045332
.1	{'alpha': 0.1}	-0.001569	-0.001713	-0.022569	-0.005927	-0.000246
.2	{'alpha': 0.2}	-0.001569	-0.001713	-0.022569	-0.005927	-0.000246
.5	{'alpha': 0.5}	-0.001569	-0.001713	-0.022569	-0.005927	-0.000246
1	{'alpha': 1.0}	-0.001569	-0.001713	-0.022569	-0.005927	-0.000246
5	{'alpha': 5.0}	-0.001569	-0.001713	-0.022569	-0.005927	-0.000246

Lasso model with alpha 0.005

```
('constant', 11.13),
 ('OverallQual', 0.813),
 ('totalSF', 0.74),
 ('GarageArea', 0.257),
 ('Fireplaces', 0.165),
 ('YearBuilt', -0.141),
 ('HeatingQC', 0.136),
 ('TotRmsAbvGrd', 0.135),
 ('MSZoning_RL', 0.065),
 ('MSZoning_RM', -0.031),
 ('CentralAir_Y', 0.027),
 ('WoodDeckSF', 0.02),
```

Lasso model with alpha 0.01

```
('constant', 11.233),
 ('OverallQual', 0.901),
 ('totalSF', 0.329),
```

```
 ('GarageArea', 0.203),
 ('Fireplaces', 0.144),
 ('TotRmsAbvGrd', 0.142),
 ('YearBuilt', -0.113),
 ('HeatingQC', 0.113),
 ('MSZoning_RL', 0.082),
 ('MSZoning_RM', -0.015),
```

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: We chose to apply Lasso regression as it made coeff of non significant predictor to 0. Lesser predictor variables brings inherit feature selection. It reduces complexity of model and helps business to look at relatively focussed predictors of target variable.

Ridge coeff o/p:

```
array([ 2.75290206e-01,  1.37300324e-01,  4.50393060e-01,
-2.27016906e-01,
        2.13640856e-01,  2.93098357e-01,  1.43203369e-01,
1.38918165e-01,
        3.42734084e-01,  1.82133723e-02,  3.61574833e-02,
6.36636778e-02,
       -2.07141932e-02,  7.69028525e-02, -2.14167395e-02,
-5.14651808e-02,
```

```
        1.11759223e-01, -5.68720303e-02,  1.05903216e-01,
-6.04598262e-02,
       -6.62605140e-03,  1.92144908e-02, -4.25656070e-04,
9.71749079e-02,
       -3.13231730e-02, -1.80388766e-02,  1.57528544e-02,
7.10217329e-03,
        4.51700515e-02,  3.25499274e-02])
```

Lasso coeff o/p:

```
array([ 0.14177216,  0.        ,  0.90095368,
-0.11318572,  0.14424651,
        0.20329005,  0.        ,  0.        ,
0.32892086,  0.1133552 ,
        0.        ,  0.        , -0.        ,
0.08213732, -0.01545134,
       -0.        ,  0.        , -0.        ,
0.        , -0.        ,
        0.        ,  0.        , -0.        ,
0.        , -0.        ,
       -0.        ,  0.        , -0.        ,
0.        ,  0.        ])
```

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:

Below are imp predictors after rebuilding.

```
('TotRmsAbvGrd', 0.558),
('FullBath', 0.376),
('MasVnrArea', 0.323),
('HeatingQC', 0.303),
('Neighborhood_StoneBr', 0.297),
('OpenPorchSF', 0.256),
```

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:

I have focussed on EDA in 1st part of assignment to remove outliers, making distribution normal, removing/correcting unbalanced features etc.. It will help avoid overfit and improve generalisation and accuracy on unseen data. In model building portion, I have used lasso/ridge regression and perform cross validation with various values of regularisation param to select alpha value which will give optimal bias/variance trade off.