

Q. 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?

Ans. In the python code shared for Housing assignment, the optimum value of Lasso regression is 0.01 and Ridge regression is 1.002 .

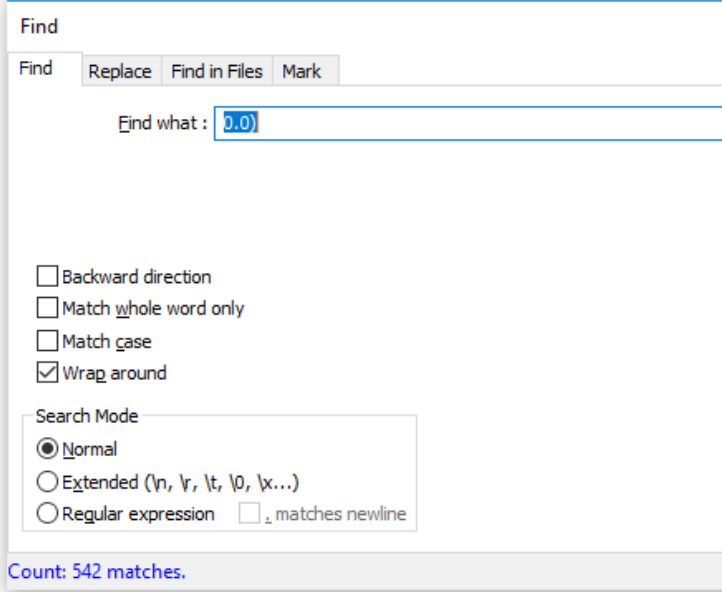
Lasso Regression: We can see that 530 coefficients are exactly 0. If we increase the value of alpha, our model will try to penalize more and try to make most of our coefficient value zero. As we can see below now 542 coefficients are exactly zero. Also R- squared decreases from 0.89 to 0.85

At alpha=0.01

```
1  [ ('constant', -0.492),
2    ('LotFrontage', 0.001),
3    ('LotArea', 0.06),
4    ('MasVnrArea', 0.024),
5    ('BsmtFinSF1', 0.093),
6    ('BsmtFinSF2', 0.0),
7    ('BsmtUnfSF', -0.0),
8    ('TotalBsmtSF', 0.147),
9    ('1stFlrSF', 0.0),
10   ('2ndFlrSF', 0.0),
11   ('LowQualFinSF', 0.0),
12   ('GrLivArea', 0.329),
13   ('TotRmsAbvGrd', -0.0),
14   ('Fireplaces', 0.07),
15   ('GarageArea', 0.111),
16   ('WoodDeckSF', 0.03),
17   ('OpenPorchSF', 0.024),
18   ('EnclosedPorch', 0.0),
19   ('3SsnPorch', 0.0),
20   ('ScreenPorch', 0.0),
21   ('PoolArea', 0.0),
22   ('MiscVal', 0.0),
23   ('SalePrice', -0.0),
24   ('MSSubClass_30', -0.0),
25   ('MSSubClass_40', 0.0),
26   ('MSSubClass_45', -0.0),
27   ('MSSubClass_50', 0.0),
28   ('MSSubClass_60', 0.0),
29   ('MSSubClass_70', -0.0),
30   ('MSSubClass_75', 0.0),
31   ('MSSubClass_80', -0.0),
32   ('MSSubClass_85', -0.0),
33   ('MSSubClass_90', 0.0),
34   ('MSSubClass_120', -0.0),
35   ('MSSubClass_160', -0.0),
36   ('MSSubClass_180', -0.0),
37   ('MSSubClass_190', 0.0),
38   ('MSZoning_FV', 0.0),
```

At $\alpha=0.02$ (When we double the value of α)

```
1  [ ('constant', -0.01),
2    ('LotFrontage', 0.0),
3    ('LotArea', 0.049),
4    ('MasVnrArea', 0.037),
5    ('BsmtFinSF1', 0.089),
6    ('BsmtFinSF2', 0.0),
7    ('BsmtUnfSF', -0.0),
8    ('TotalBsmtSF', 0.17),
9    ('1stFlrSF', 0.0),
10   ('2ndFlrSF', 0.0),
11   ('LowQualFinSF', 0.0),
12   ('GrLivArea', 0.312),
13   ('TotRmsAbvGrd', 0.0),
14   ('Fireplaces', 0.096),
15   ('GarageArea', 0.148),
16   ('WoodDeckSF', 0.036),
17   ('OpenPorchSF', 0.028),
18   ('EnclosedPorch', 0.0),
19   ('3SsnPorch', 0.0),
20   ('ScreenPorch', 0.0),
21   ('PoolArea', 0.0),
22   ('MiscVal', 0.0),
23   ('SalePrice', -0.0),
24   ('MSSubClass_30', -0.0),
25   ('MSSubClass_40', 0.0),
26   ('MSSubClass_45', -0.0),
27   ('MSSubClass_50', 0.0),
28   ('MSSubClass_60', 0.0),
29   ('MSSubClass_70', -0.0),
30   ('MSSubClass_75', 0.0),
31   ('MSSubClass_80', -0.0),
32   ('MSSubClass_85', -0.0),
33   ('MSSubClass_90', 0.0),
34   ('MSSubClass_120', -0.0),
35   ('MSSubClass_160', 0.0),
36   ('MSSubClass_180', -0.0),
37   ('MSSubClass_190', 0.0),
38   ('MSZoning_FV', -0.0),
```



Find

Find Replace Find in Files Mark

Find what: 0.0

☐ Backward direction
☐ Match whole word only
☐ Match case
☒ Wrap around

Search Mode

☒ Normal
☐ Extended (\n, \r, \t, \0, \x...)
☐ Regular expression ☐ _ matches newline

Count: 542 matches.

Ridge Regression: If we double the value of α , it will apply more penalty on the curve and try to make the model more generalizable. It will not try to fit every data from the data set.

On Increasing penalty the coefficients gets impacted.

For example the value of MSZoning_FV has decreased from 0.274 to 0.197.

Q.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. Our selection of model will depend upon what our aim is.

Lasso Regression

If our aim is to do feature selection, then we should use Lasso regression.

In this case we can see many Lasso coefficients are close to zero.

Ridge Regression

If our aim is to make the coefficients small ie reduce the coefficient magnitude, then we should use Ridge Regression.

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?

Below are the top 5 Lasso predictor variables.

('OverallQual_8', 0.402),

('GrLivArea', 0.329),

('OverallQual_7', 0.301),

('BedroomAbvGr_8', 0.198),

('Condition1_Feedr', 0.112),

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. A robust model is the one whose performance is not impacted by the variations in data.

A generalizable model is the which is able to adapt to the unseen data.

Any model should not overfit in order for it to be robust and generalizable. A model is said to overfit if it performs badly on the test data but performs well on the training data. These type of models have low bias and high variance.

A model should not be complex for it to be robust and generalizable.

A model which is too complex will have high accuracy. So to make the model more robust and generalizable, we will have to decrease variance which will lead to some bias.

There should be a balance between model accuracy and complexity. Regularization techniques like Ridge and Lasso Regression helps in achieving that.