Garv Daga

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?

Optimal value of alpha for ridge and lasso regression:

Ridge Regression Alpha: 7

Lasso Regression Alpha: 0.001

When we double the value of R2 the score gets a bit weaker: -

Ridge (alpha = 7) :- Most Important - `Neighborhood_IDOTRR`

```
R-Squared (Train) = 0.92
R-Squared (Test) = 0.81
RSS (Train) = 13.88
RSS (Test) = 12.85
MSE (Train) = 0.01
MSE (Test) = 0.03
RMSE (Train) = 0.12
RMSE (Test) = 0.17
```

Ridge (alpha =14) :- Most Important - `Neighborhood_IDOTRR`

```
R-Squared (Train) = 0.91
R-Squared (Test) = 0.81
RSS (Train) = 14.36
RSS (Test) = 12.97
MSE (Train) = 0.01
MSE (Test) = 0.03
RMSE (Train) = 0.12
RMSE (Test) = 0.17
```

Lasso (alpha=0.001) :- Most Important - `Neighborhood_IDOTRR`

```
R-Squared (Train) = 0.91
R-Squared (Test) = 0.81
RSS (Train) = 15.08
RSS (Test) = 13.44
MSE (Train) = 0.01
MSE (Test) = 0.03
RMSE (Train) = 0.12
RMSE (Test) = 0.18
```

Lasso (alpha=0.002) :- **Most Important** - `OverallQual`

```
R-Squared (Train) = 0.90
R-Squared (Test) = 0.80
RSS (Train) = 16.59
RSS (Test) = 13.98
MSE (Train) = 0.02
MSE (Test) = 0.03
RMSE (Train) = 0.13
RMSE (Test) = 0.18
```

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?

Ans : Ridge is marginally better than Lasso but comes with some tradeoffs as it makes the model more complex as the coefficients or betas never touch zero. While , with a small tradeoff of accuracy Lasso makes the model simpler by reducing the coefficients to exactly zero.
I will choose Lasso Regression in this case.

Q3: After building the model, you realized 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:

If the top 5 features of Lasso are removed , we still get the alpha as 0.001.

New Metrics :-

```
# Checking the mtrics for train and test data
lasso_metrics = show_metrics(y_train, lasso_y_train_pred, y_test, lasso_y_pred)
```

✓ 0.0s

```
R-Squared (Train) = 0.89
R-Squared (Test) = 0.77
RSS (Train) = 18.36
RSS (Test) = 15.83
MSE (Train) = 0.02
MSE (Test) = 0.04
RMSE (Train) = 0.13
RMSE (Test) = 0.19
```

Top Variables :-

```
# lasso model parameters
lasso_model_parameters = list(lasso.coef_)
lasso_model_parameters.insert(0, lasso.intercept_)
lasso_model_parameters = [numpy.round(x, 3) for x in lasso_model_parameters]
cols = X.columns
cols = cols.insert(0, "constant")
lasso_model_parameters = list((zip(cols, lasso_model_parameters)))
lasso_model_parameters.sort(key = lambda x: abs(x[1]), reverse=True)
lasso_model_parameters
```

✓ 0.0s

```
[('constant', array([12.074])),
 ('Neighborhood_Crawfor', -0.127),
 ('1stFlrSF', 0.102),
 ('TotalBsmtSF', 0.083),
 ('OverallCond', 0.082),
 ('OverallQual', 0.067),
```

New 5 top variables :-

'Neighborhood_Crawfor', '1stFlrSF', 'TotalBsmtSF', 'OverallCond' and 'OverallQual'.

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:

To ensure that the model is robust and generalizable :-

1. We should focus on the data and try to remove noise. Outliers, wrong data types and unhandled null values might cause issues in our model predictions. This inturn helps the model to find the underlying pattern as the noise is reduced.
2. We have to closely monitor the Bais of the model. The model should not overfit the training data (too complex models tend to do so) and work poorly on test data.
3. Also we should check the variance of the error terms produced by the model, We should try to see if there are patterns (Heteroskedasticity) and try to remove them by better feature selection.
4. Maintaining the bais and variance might lead to trade off in the model score but we should focus on making a stable model which performs with the same accuracy consistently on unseen data.