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?

The optimal value of alpha are given below

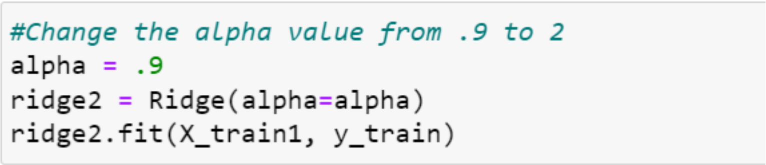
Ridge –

Alpha – 0.9

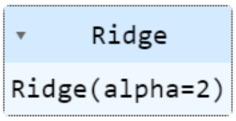
Lasso –

Alpha – 10

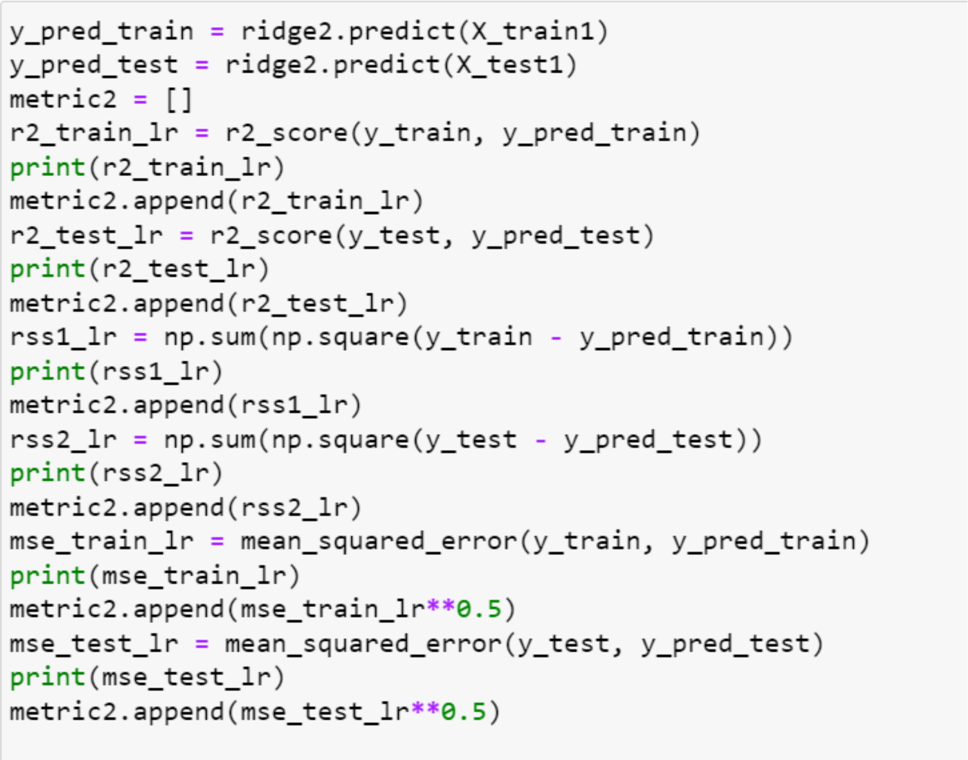
Ridge regression



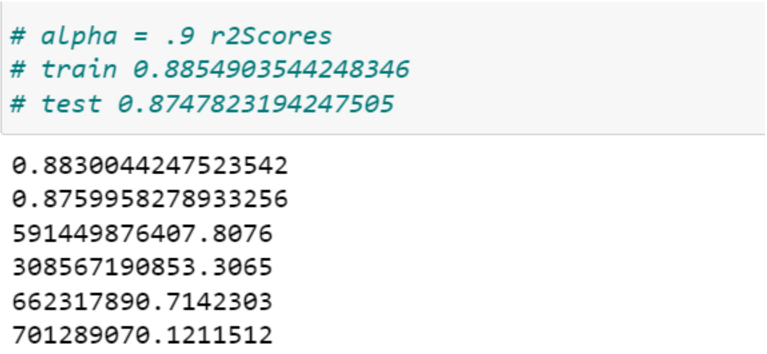
```
#Change the alpha value from .9 to 2
alpha = .9
ridge2 = Ridge(alpha=alpha)
ridge2.fit(X_train1, y_train)
```



```
Ridge
Ridge(alpha=2)
```



```
y_pred_train = ridge2.predict(X_train1)
y_pred_test = ridge2.predict(X_test1)
metric2 = []
r2_train_lr = r2_score(y_train, y_pred_train)
print(r2_train_lr)
metric2.append(r2_train_lr)
r2_test_lr = r2_score(y_test, y_pred_test)
print(r2_test_lr)
metric2.append(r2_test_lr)
rss1_lr = np.sum(np.square(y_train - y_pred_train))
print(rss1_lr)
metric2.append(rss1_lr)
rss2_lr = np.sum(np.square(y_test - y_pred_test))
print(rss2_lr)
metric2.append(rss2_lr)
mse_train_lr = mean_squared_error(y_train, y_pred_train)
print(mse_train_lr)
metric2.append(mse_train_lr**0.5)
mse_test_lr = mean_squared_error(y_test, y_pred_test)
print(mse_test_lr)
metric2.append(mse_test_lr**0.5)
```



```
# alpha = .9 r2Scores
# train 0.8854903544248346
# test 0.8747823194247505
```

```
0.8830044247523542
0.8759958278933256
591449876407.8076
308567190853.3065
662317890.7142303
701289070.1211512
```

R2score on training data has decreased but it has increased on testing data.

Lasso – After increasing alpha to 20 below are the R2Scores comaparison

# Lasso R2score for alpha = 10

    # train 0.8868752003873537

    # test 0.8704652562911501

# Lasso R2score for alpha = 20

    # train 0.886295598112146
    # test 0.8724174772773864

R2score of training data has decrease and it has increase on testing data

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?

The r2_score of lasso is slightly higher for the test dataset so we will choose lasso regression to solve this problem.

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?

Five most important predictor variables are -

11stFlrSF-First Floor square feet

GrLivArea - Above grade (ground) living area square feet

Street_Pave - Pave road access to property

RoofMatl_Metal - Roof material_Metal

RoofStyle_Shed - Type of roof(Shed)

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?

The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Robust estimators minimize the sum of the absolute values of the errors instead of the sum of squares

Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.