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-

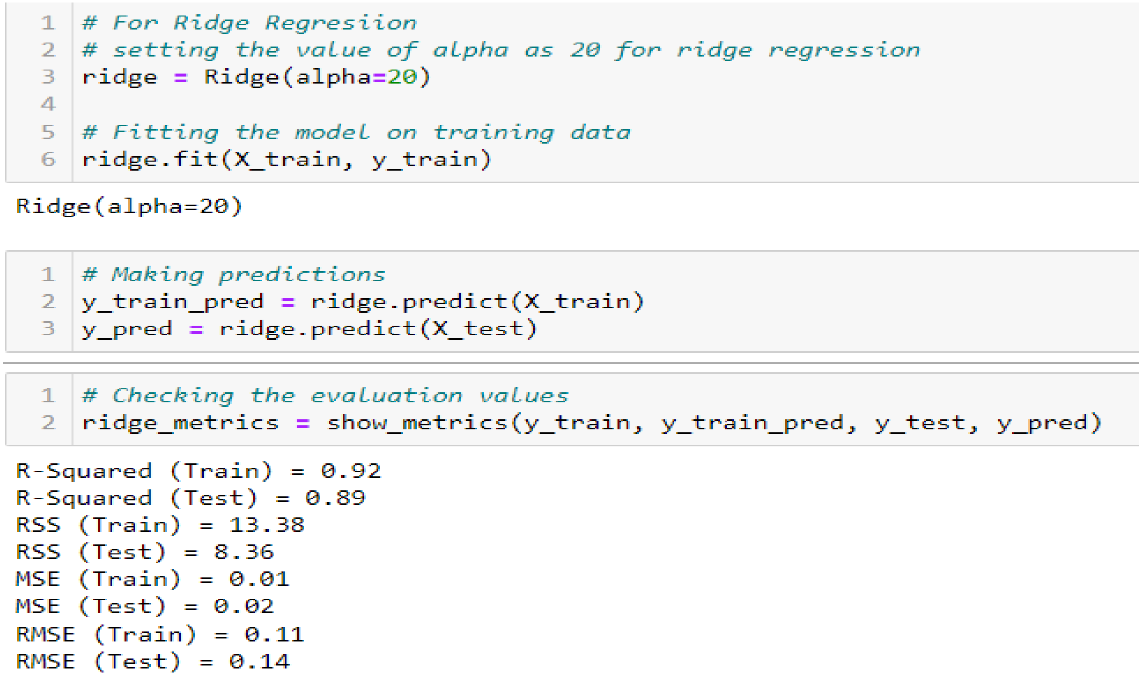
After performing Ridge and Lasso regression the alpha values we obtained are-

Optimal value of alpha for Lasso Regression = 0.001

Optimal value of alpha for Ridge Regression = 10

if we double them: alpha for lasso = 0.002 alpha for ridge = 20

For Ridge Regression we performed the changes on the model itself and found the below results.



```python
# For Ridge Regresion
# setting the value of alpha as 20 for ridge regression
ridge = Ridge(alpha=20)

# Fitting the model on training data
ridge.fit(X_train, y_train)
```

```
Ridge(alpha=20)
```

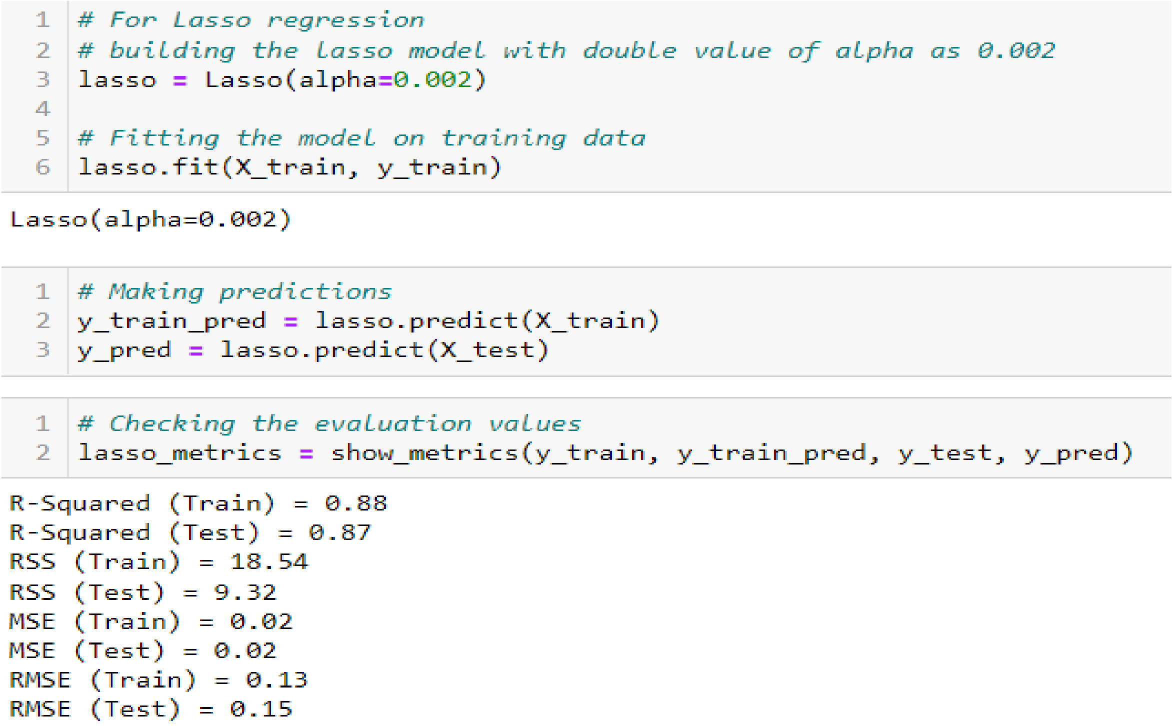
```python
# Making predictions
y_train_pred = ridge.predict(X_train)
y_pred = ridge.predict(X_test)
```

```python
# Checking the evaluation values
ridge_metrics = show_metrics(y_train, y_train_pred, y_test, y_pred)
```

```
R-Squared (Train) = 0.92
R-Squared (Test) = 0.89
RSS (Train) = 13.38
RSS (Test) = 8.36
MSE (Train) = 0.01
MSE (Test) = 0.02
RMSE (Train) = 0.11
RMSE (Test) = 0.14
```

- R-squared value and Mean Squared Error value for test data are the same as before when alpha was = 10 in our Ridge regression model i.e., 0.89 and 0.02.

For Lasso Regression as well, we applied the changes on the code of lasso model and found the below results:



```
# For Lasso regression
# building the lasso model with double value of alpha as 0.002
lasso = Lasso(alpha=0.002)

# Fitting the model on training data
lasso.fit(X_train, y_train)
```

```
Lasso(alpha=0.002)
```

```
# Making predictions
y_train_pred = lasso.predict(X_train)
y_pred = lasso.predict(X_test)
```

```
# Checking the evaluation values
lasso_metrics = show_metrics(y_train, y_train_pred, y_test, y_pred)
```

```
R-Squared (Train) = 0.88
R-Squared (Test) = 0.87
RSS (Train) = 18.54
RSS (Test) = 9.32
MSE (Train) = 0.02
MSE (Test) = 0.02
RMSE (Train) = 0.13
RMSE (Test) = 0.15
```

- The R-squared value has a very negligible change of 0.02 in the test data of Lasso regression model, however, the Mean Squared Error value is still the same.

The most Significant variables after doubling the values of alpha are -

- OverallQual_9
- GrLivArea
- OverallQual_8
- Neighborhood_Crawfor
- CentralAir_Y
- Functional_Typ
- GarageCars
- Exterior1st_BrkFace
- Neighborhood_NridgHt
- Condition1_Norm

Please, refer to the jupyter notebook for detailed procedure and explanation.

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 –

After performing both Lasso and Ridge regression on the dataset, we have very similar results. R- squared value for both the regression models was the same i.e., 0.89. I think both the models will work pretty well in predicting the significant variables and the 'SalePrice'.

However, given a chance, I would like to choose Lasso Regression model over Ridge Regression model, as Lasso Regression offers the scope of eliminating not- so important features and select the most relevant ones in case of a long list of predictors.

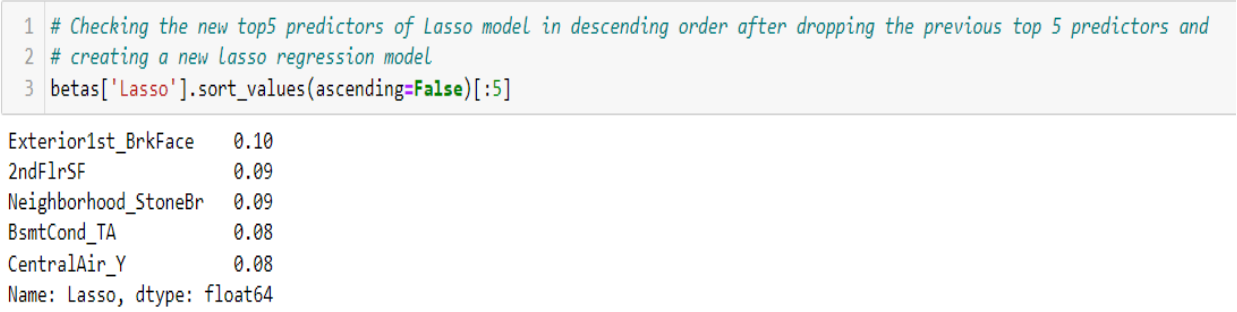
Question 3

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 –

After dropping the previous top 5 predictors and building a new Lasso regression model with the left-out predictors, we have our new top 5 significant variables, which are-

- Exterior1st_BrkFace, 2ndFlrSF, Neighborhood_StoneBr, BsmtCond_TA, and CentralAir_Y



```
# Checking the new top5 predictors of Lasso model in descending order after dropping the previous top 5 predictors and
# creating a new lasso regression model
betas['Lasso'].sort_values(ascending=False)[:5]
```

```
Exterior1st_BrkFace    0.10
2ndFlrSF               0.09
Neighborhood_StoneBr   0.09
BsmtCond_TA            0.08
CentralAir_Y           0.08
Name: Lasso, dtype: float64
```

Please, refer to the jupyter notebook for detailed procedure and explanation.

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 –

We can call a model robust and generalisable,

1 - If the performance of the model on unseen data remains the same as the performance of the model on the training data or vary very little on the unseen data as compared to the train data.

2 - Also, we need to make sure that the model doesn't overfit, that is, even a small change in the data should not affect the performance of the model and its prediction qualities widely. The simplest models can be considered as robust and generalisable as long as they are able to catch the patterns of data in both training and unseen data.

3 - If the model overfits, then to be exact, we can say it memorises all the patterns and variations in the train data which results in high accuracy. However, they fail drastically when an unseen data is fed to the model, the accuracy drops with a huge gap.

4 - In order to achieve a balance between the model complexity and high prediction accuracy we need to introduce bias or a penalty to our model so that we can regularize it. In regularization we basically penalize the large co-efficient which results in decreasing the accuracy a little. Lasso and Ridge regressions are two regularization techniques which can be used to achieve the same.

In Ridge regression, the loss function is modified to minimize the complexity of the model. We add a penalty parameter that is equivalent to the square of the magnitude of the coefficients.

Least Absolute Shrinkage and Selection Operator, or Lasso is also a modification of linear regression to regularize the model. In lasso, the loss function is modified to minimize the complexity of the model, here we limit the sum of the absolute values of the model coefficients.

The major differences between linear regression and regularized regression model are that in the regularized regression we focus on tuning a hyperparameter, lambda or alpha (penalty parameter). We first find the optimal alpha/lambda value in both ridge and lasso regression, then train/test the model with that penalty term.