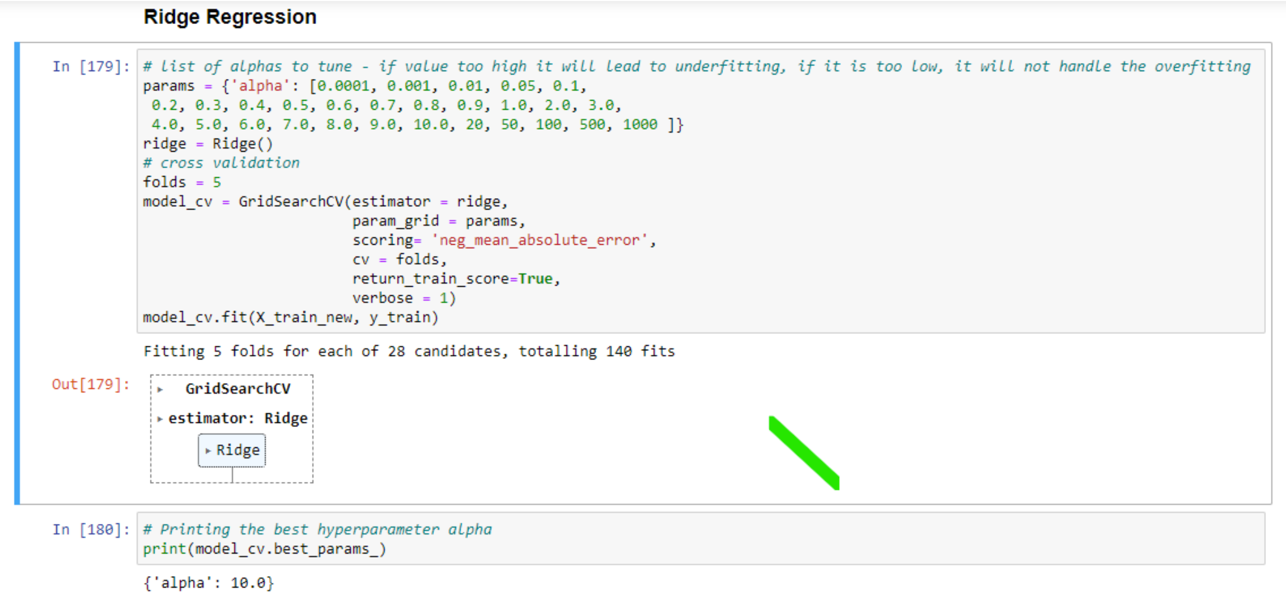
Question 1 : What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Optimal Value for alpha for ridge regression: 10



Ridge Regression

In [179]:
```
# list of alphas to tune - if value too high it will lead to underfitting, if it is too low, it will not handle the overfitting
params = {'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1,
 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000 ]}
ridge = Ridge()
# cross validation
folds = 5
model_cv = GridSearchCV(estimator = ridge,
                        param_grid = params,
                        scoring= 'neg_mean_absolute_error',
                        cv = folds,
                        return_train_score=True,
                        verbose = 1)
model_cv.fit(X_train_new, y_train)
```

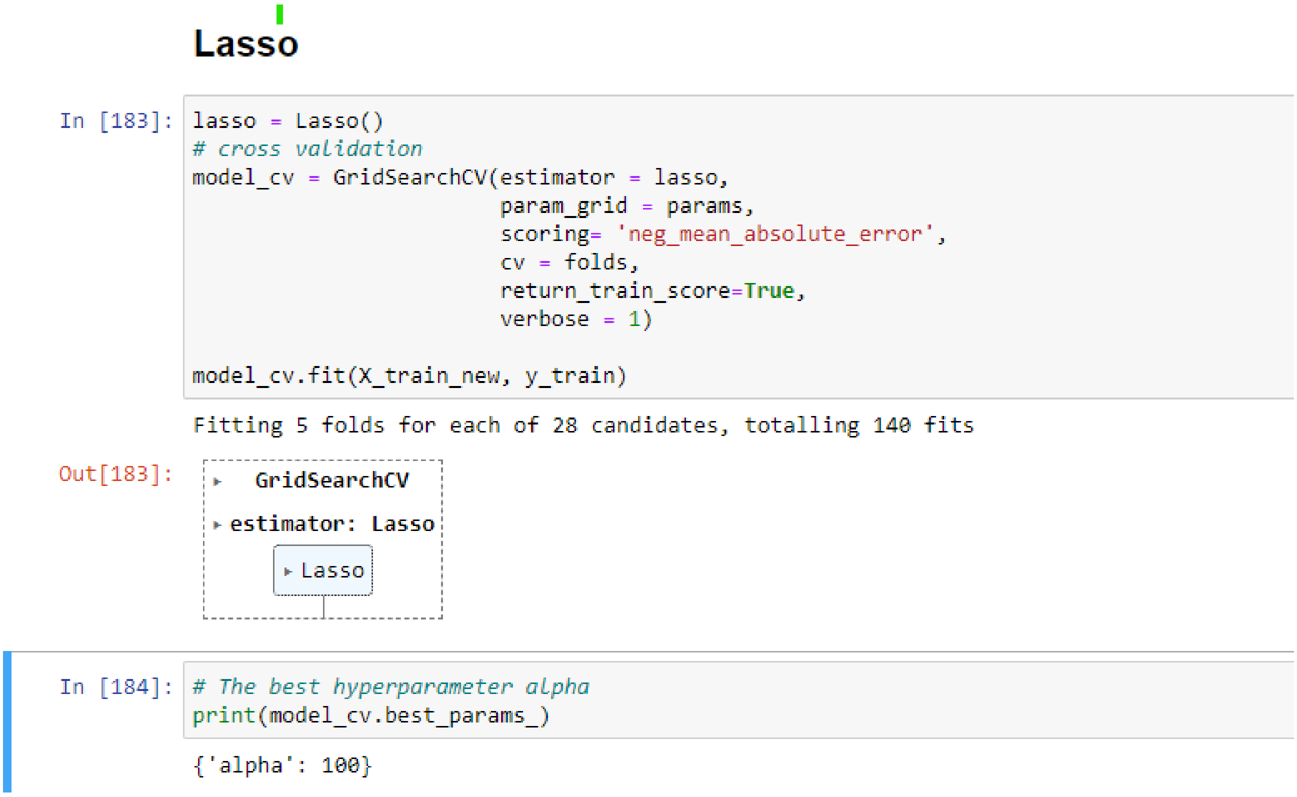
```
Fitting 5 folds for each of 28 candidates, totalling 140 fits
```

Out[179]:

In [180]:
```
# Printing the best hyperparameter alpha
print(model_cv.best_params_)
```

```
{'alpha': 10.0}
```

Optimal Value for alpha for lasso regression: 100



Lasso

In [183]:
```
lasso = Lasso()
# cross validation
model_cv = GridSearchCV(estimator = lasso,
                        param_grid = params,
                        scoring= 'neg_mean_absolute_error',
                        cv = folds,
                        return_train_score=True,
                        verbose = 1)

model_cv.fit(X_train_new, y_train)
```

```
Fitting 5 folds for each of 28 candidates, totalling 140 fits
```

Out[183]:

In [184]:
```
# The best hyperparameter alpha
print(model_cv.best_params_)
```

```
{'alpha': 100}
```

For Ridge: Coeff values will increase as alpha will increase. r2_score of train data will also drop.

For Lasso: If alpha value increases r2score is also dropped by 1% in both test and train data

Top Features: Neighborhood_NoRidge, Neighborhood_NridgHt, OverallQual, overallQual Neighborhood_Veenkar

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: Since Lasso gives feature selection option also, I'll choose Lasso. It helps to remove unwanted feature from model without affecting the model accuracy.

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 : Top 5 features are Neighborhood_NoRidge, Neighborhood_NridgHt, 2ndFlrSF, OverallQual, Neighborhood_Veenker.

If we drom these features, them model accuracy reduced drastically. From analysis we found it reduced from 80 to 55% for test data and 81 to 55% for test data.

Next top 5 features after droping 5 main predictors 1stFlrSF, MSSubClass_90, MSSubClass_120, TotalBsmtSF, HouseStyle_1Story

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 : For any model to be robust and generalisable these below features should be in given range-

1. Model accuracy - It should be > 70%: In our case We got, accuracy 80% for train data and 81% for test data. which is a good value.

2. P-value of all the features is < 0.05. We had removed all the attributes with high P-values and rebuilt the model again and again.

3. VIF of all the features are < 5 . We have also removed the variables with high VIF and re built the model.