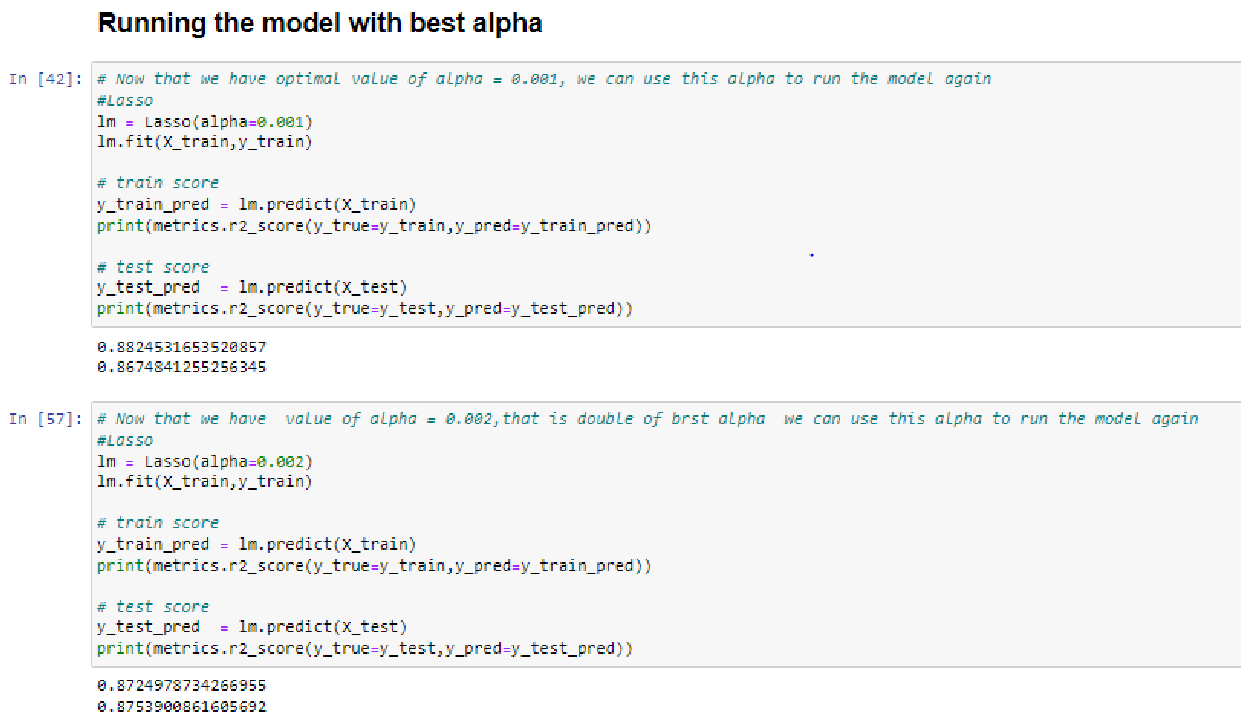
Question 1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer.

1. The optimal value of alpha for ridge and lasso regression
    a. Ridge Alpha 5
    b. lasso Alpha .001

2. Changes in model on doubling Alpha :
    a) For Lasso

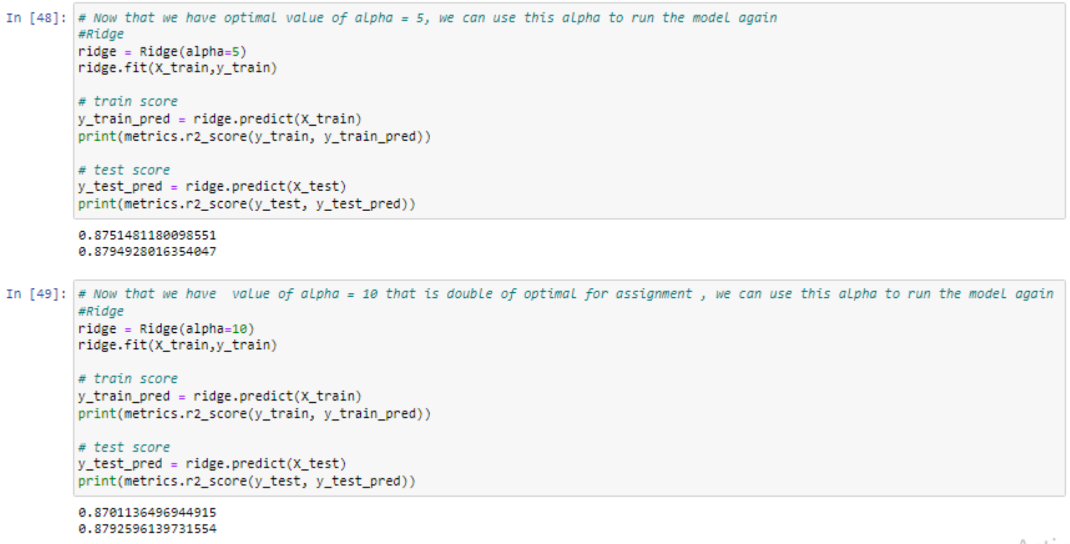


Running the model with best alpha

```
In [42]: # Now that we have optimal value of alpha = 0.001, we can use this alpha to run the model again
         #Lasso
         lm = Lasso(alpha=0.001)
         lm.fit(X_train,y_train)

         # train score
         y_train_pred = lm.predict(X_train)
         print(metrics.r2_score(y_true=y_train,y_pred=y_train_pred))

         # test score
         y_test_pred  = lm.predict(X_test)
         print(metrics.r2_score(y_true=y_test,y_pred=y_test_pred))

         0.8824531653520857
         0.8674841255256345

In [57]: # Now that we have  value of alpha = 0.002,that is double of brst alpha  we can use this alpha to run the model again
         #Lasso
         lm = Lasso(alpha=0.002)
         lm.fit(X_train,y_train)

         # train score
         y_train_pred = lm.predict(X_train)
         print(metrics.r2_score(y_true=y_train,y_pred=y_train_pred))

         # test score
         y_test_pred  = lm.predict(X_test)
         print(metrics.r2_score(y_true=y_test,y_pred=y_test_pred))

         0.8724978734266955
         0.8753900861605692
```

Here on doubling the alpha for lasso we see that r2 of training set has decreased however for test it has increased.

b) For Ridge :



```
In [48]: # Now that we have optimal value of alpha = 5, we can use this alpha to run the model again
         #Ridge
         ridge = Ridge(alpha=5)
         ridge.fit(X_train,y_train)

         # train score
         y_train_pred = ridge.predict(X_train)
         print(metrics.r2_score(y_train, y_train_pred))

         # test score
         y_test_pred = ridge.predict(X_test)
         print(metrics.r2_score(y_test, y_test_pred))

         0.8751481180098551
         0.8794928016354047

In [49]: # Now that we have  value of alpha = 10 that is double of optimal for assignment , we can use this alpha to run the model again
         #Ridge
         ridge = Ridge(alpha=10)
         ridge.fit(X_train,y_train)

         # train score
         y_train_pred = ridge.predict(X_train)
         print(metrics.r2_score(y_train, y_train_pred))

         # test score
         y_test_pred = ridge.predict(X_test)
         print(metrics.r2_score(y_test, y_test_pred))

         0.8701136496944915
         0.8792596139731554
```

In case of ridge there is not much effect on R2 of test and train data however , if we see it closely the r2 scores have reduced at little for both test and train data sets.

3. After comparing both the model we can see that the below Features are best explaining the DataSet

- BsmtFinType1
- ExterCond
- KitchenAbvGr
- BsmtCond
- OverallCond
- GrLivArea
- OverallQual
- LotArea
- MasVnrArea

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. Lasso can set some coefficients to zero, thus performing variable selection, while ridge regression cannot. Lasso tends to do well if there are a small number of significant parameters and the others are close to zero (when only a few predictors actually influence the response).

Ridge works well if there are many large parameters of about the same value (when most predictors impact the response).

In our Assignment R2 score while using ridge is .88 & while in case of lasso is .87. Here both R2 are comparable . However lasso penalizes more on dataset and make more coefficient equal to zero its better to use Lasso . As it gives relatively less and accurate parameters to focus on and predict accurate sales price.

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. five most important predictor variables

1. 11stFlrSF-----------First Floor square feet
2. GrLivArea-----------Above grade (ground) living area square feet
3. Street_Pave---------Pave road access to property
4. RoofMatl_Metal------Roof material_Metal
5. RoofStyle_Shed------Type of roof(Shed)

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. 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. 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. In addition to that EDA also is a very important that as if data not cleaned properly may effect the results and even cause error making prediction more cumbersome.