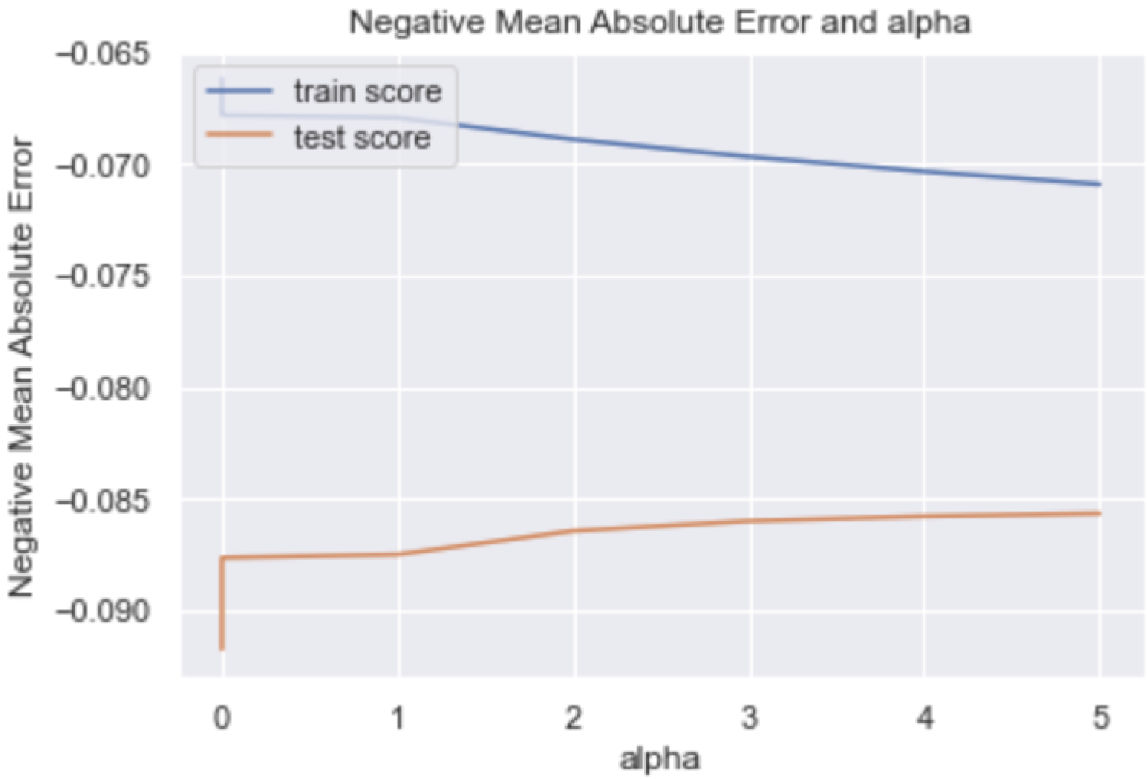
Question 1:
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

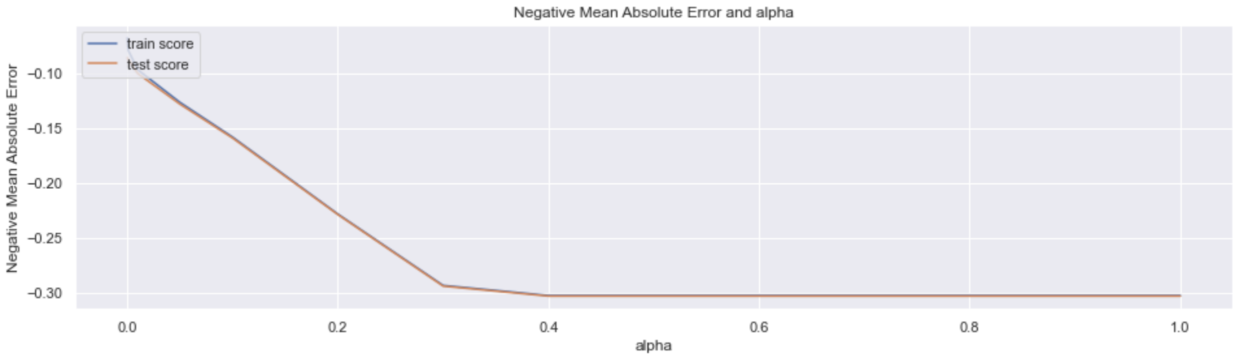
Answer 1:

Ridge regression:

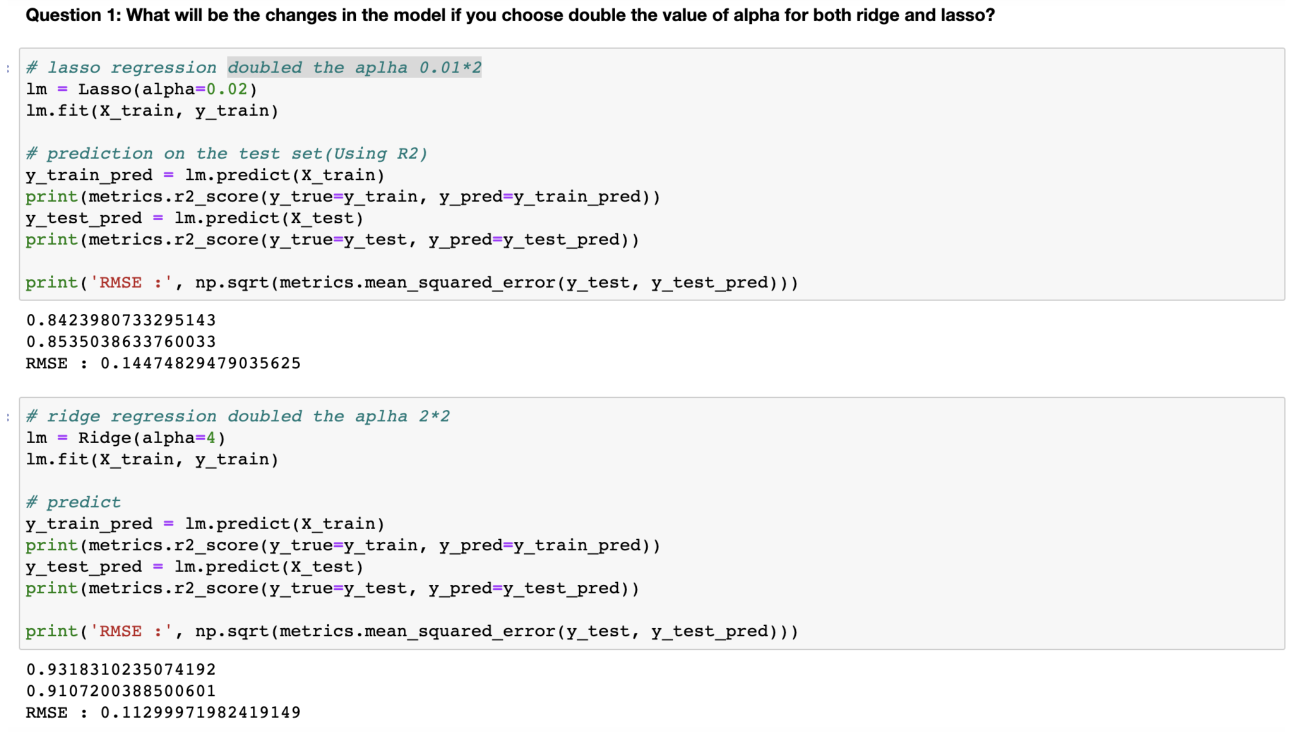


According to plot it seen when alpha value increases then -ve Mean Absolute Error is decreasing and at 2 we choose the optimum value as at this balanced optimized model
RMSE = 0.113593999122709 55
Train = 0.9340893303403607
Test = 0.9097785011189363

LASSO regression:



According to plot it seen when alpha value increases then -ve Mean Absolute Error is decreasing and at 0.01 we choose the optimum value as at this balanced optimized model
RMSE = 0.13172353619212385
Train = 0.8730262400638582
Test = 0.8786817823030547



Question 1: What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

```
# lasso regression doubled the aplha 0.01*2
lm = Lasso(alpha=0.02)
lm.fit(X_train, y_train)

# prediction on the test set(Using R2)
y_train_pred = lm.predict(X_train)
print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))
y_test_pred = lm.predict(X_test)
print(metrics.r2_score(y_true=y_test, y_pred=y_test_pred))

print('RMSE :', np.sqrt(metrics.mean_squared_error(y_test, y_test_pred)))
```

```
0.8423980733295143
0.8535038633760033
RMSE : 0.14474829479035625
```

```
# ridge regression doubled the aplha 2*2
lm = Ridge(alpha=4)
lm.fit(X_train, y_train)

# predict
y_train_pred = lm.predict(X_train)
print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))
y_test_pred = lm.predict(X_test)
print(metrics.r2_score(y_true=y_test, y_pred=y_test_pred))

print('RMSE :', np.sqrt(metrics.mean_squared_error(y_test, y_test_pred)))
```

```
0.9318310235074192
0.9107200388500601
RMSE : 0.11299971982419149
```

The most important predictor variables for ridge regression are mentioned below:

MSZoning_FV, MSZoning_RL, MSZoning_RH, MSZoning_RM, SaleCondition_Partial, Neighborhood_StoneBr, Neighborhood_Crawfor, Foundation_PConc, SaleCondition_Normal, Condition1_Norm

The most important predictor variables for Lasso regression are mentioned below:

OverallQual, GrLivArea, GarageArea, KitchenAbvGr, KitchenQual_TA, BsmtQual_TA

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 2:
It depends upon the business requirements completely , if the requirement was maximum accuracy then for sure Ridge is better and if want robust and simple then lasso will perform better .
In terms of r2 score Ridge is better but it is Always suggested that robust and simple model should be used in that case Lasso is Better.
Lasso since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables

Question 3:
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?

Answer 3:
MSZoning_RL, Foundation_PConc, OverallCond, FullBath, Fireplaces

Question 4:
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer 4:
The model accuracy may be decreased but it is as simple as possible and it will be defined by Bias-Variance trade-off. Like a Simple model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.
Bias: Bias is an error in a model, when the model is weak to learn from the data. High bias means the model is unable to learn details in the data. Model performs poorly on training and testing data.
Variance: Variance is error in model, when model tries to over learn from the data. High variance means the model performs exceptionally well on training data as it has very well trained on this data but performs very poorly on testing data as it was unseen data for the model.

And we have also kept in mind to take a balance between Bias and Variance to avoid overfitting and under-fitting of data.