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?

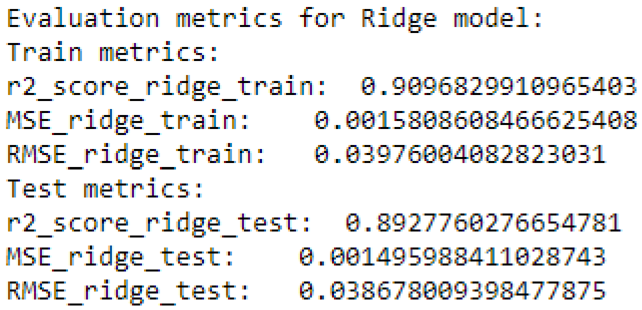
Answer:-

optimal alpha for Ridge = 0.9 with a r2 score of 0.892

optimal alpha for Lasso = 0.0001 with a r2 score of 0.891

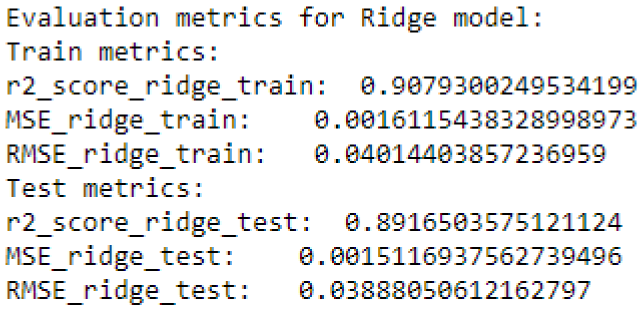
The R2 score after doubling the alpha for ridge and lasso is 0.89. There isn't much difference in the r2 score even after the alpha change.

Before change :



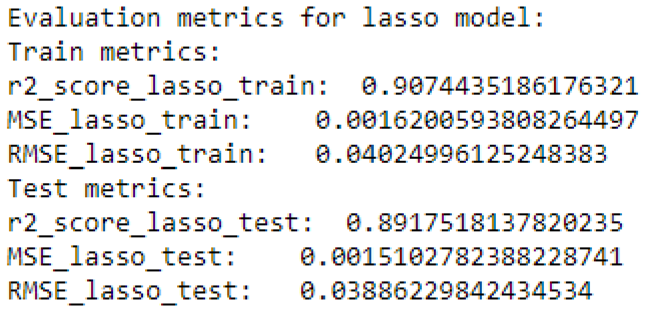
```
Evaluation metrics for Ridge model:
Train metrics:
r2_score_ridge_train:  0.9096829910965403
MSE_ridge_train:     0.0015808608466625408
RMSE_ridge_train:    0.03976004082823031
Test metrics:
r2_score_ridge_test:  0.8927760276654781
MSE_ridge_test:     0.001495988411028743
RMSE_ridge_test:    0.038678009398477875
```

After change :



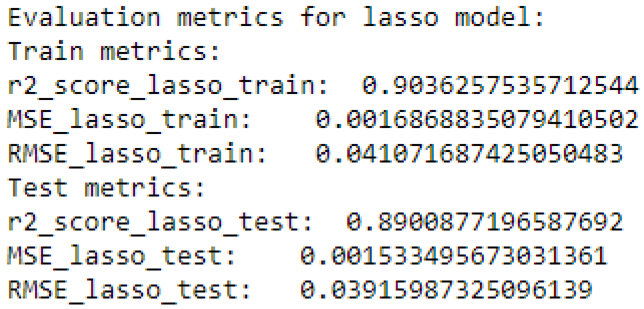
```
Evaluation metrics for Ridge model:
Train metrics:
r2_score_ridge_train:  0.9079300249534199
MSE_ridge_train:     0.0016115438328998973
RMSE_ridge_train:    0.04014403857236959
Test metrics:
r2_score_ridge_test:  0.8916503575121124
MSE_ridge_test:     0.0015116937562739496
RMSE_ridge_test:    0.03888050612162797
```

Before change:



```
Evaluation metrics for lasso model:
Train metrics:
r2_score_lasso_train:  0.9074435186176321
MSE_lasso_train:    0.0016200593808264497
RMSE_lasso_train:   0.04024996125248383
Test metrics:
r2_score_lasso_test:  0.8917518137820235
MSE_lasso_test:    0.0015102782388228741
RMSE_lasso_test:   0.03886229842434534
```

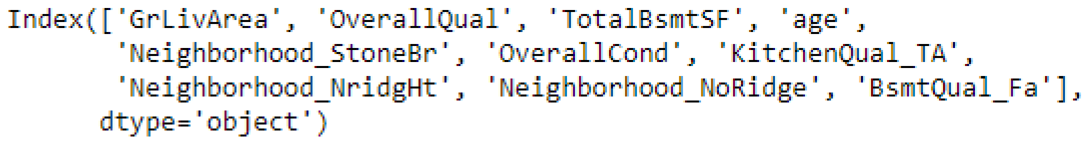
After change:



```
Evaluation metrics for lasso model:
Train metrics:
r2_score_lasso_train:  0.9036257535712544
MSE_lasso_train:    0.0016868835079410502
RMSE_lasso_train:   0.041071687425050483
Test metrics:
r2_score_lasso_test:  0.8900877196587692
MSE_lasso_test:    0.001533495673031361
RMSE_lasso_test:   0.03915987325096139
```

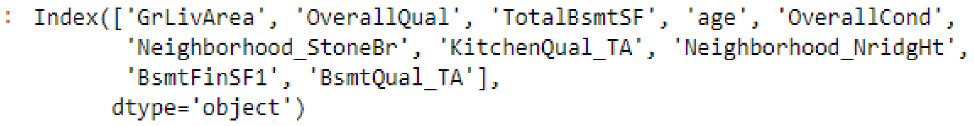
These are the top 10 important variables after doubling the alpha

For ridge:



```
Index(['GrLivArea', 'OverallQual', 'TotalBsmtSF', 'age',
       'Neighborhood_StoneBr', 'OverallCond', 'KitchenQual_TA',
       'Neighborhood_NridgHt', 'Neighborhood_NoRidge', 'BsmtQual_Fa'],
      dtype='object')
```

For lasso :

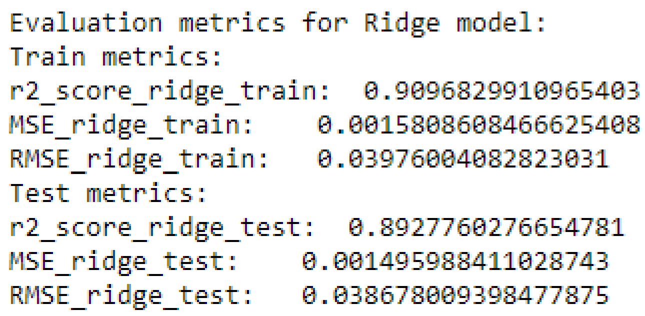


```
: Index(['GrLivArea', 'OverallQual', 'TotalBsmtSF', 'age', 'OverallCond',
         'Neighborhood_StoneBr', 'KitchenQual_TA', 'Neighborhood_NridgHt',
         'BsmtFinSF1', 'BsmtQual_TA'],
        dtype='object')
```

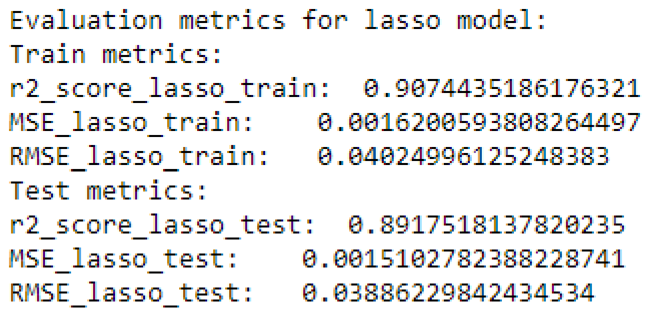
Question 2

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:



```
Evaluation metrics for Ridge model:
Train metrics:
r2_score_ridge_train:  0.9096829910965403
MSE_ridge_train:    0.0015808608466625408
RMSE_ridge_train:   0.03976004082823031
Test metrics:
r2_score_ridge_test:  0.8927760276654781
MSE_ridge_test:    0.001495988411028743
RMSE_ridge_test:   0.038678009398477875
```



```
Evaluation metrics for lasso model:
Train metrics:
r2_score_lasso_train:  0.9074435186176321
MSE_lasso_train:    0.0016200593808264497
RMSE_lasso_train:   0.04024996125248383
Test metrics:
r2_score_lasso_test:  0.8917518137820235
MSE_lasso_test:    0.001510278238822874l
RMSE_lasso_test:   0.03886229842434534
```

Both the models are comparable in their evaluation metrics. In Lasso regression some of the coefficients are zero, which provide feature selection and help the model to be less complex than ridge regression model. Hence Lasso regression model is better.

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:

These are the five most important features before removing features - **GrLivArea, OverallQual, TotalBsmtSF, age, OverallCond**

These are the five most important features after removing some features - **1stFlrSF, 2ndFlrSF, smtFinSF1, GarageArea, Neighborhood_StoneBr**.

Question 4

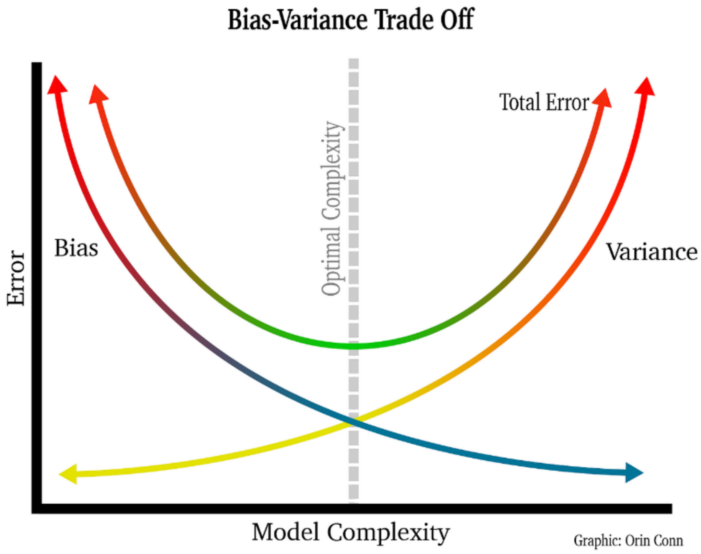
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

As per, Occam's Razor – Between the models showing similar performace, we should prefer simpler model with fewer coefficients over complex models.

- Simpler models are more generalizable and need small training set compared to complex models.
- Simple models have low variance, high bias. Complex models have high variance, low bias.
- Complex models lead to overfitting and can't be generalizable.

Hence to make the model robust and generalizable make it more simple, but not too simple that it compromises on accuracy. Regularization can be used to make the model simple. Regularization strikes balance between model complexity and generalization, ensuring robust performance on new data while avoiding overfitting.



The goal is to find the right balance between bias and variance that the model does not overfitt (high variance) or the model is too simple and leads to underfitting.