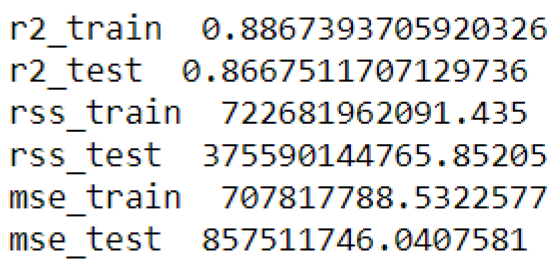
Advanced Regression - Assignment - Part II

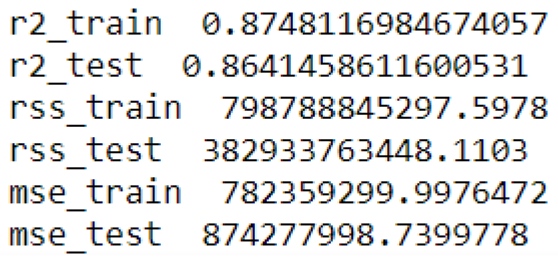
Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:
The optimal value of alpha for ridge regression was found to be 20. The optimal value of alpha for lasso regression was found to be 100.
With alpha=20 following are the metrics of Ridge regression:



```
r2_train  0.8867393705920326
r2_test  0.8667511707129736
rss_train  722681962091.435
rss_test  375590144765.85205
mse_train  707817788.5322577
mse_test  857511746.0407581
```

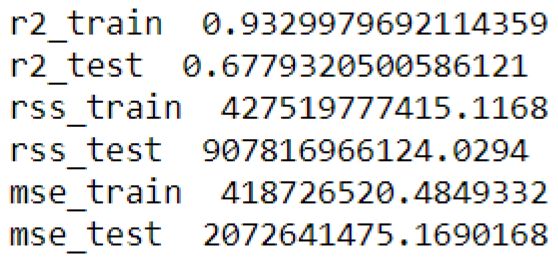
With alpha=40 following are the metrics of Ridge regression:



```
r2_train  0.8748116984674057
r2_test  0.8641458611600531
rss_train  798788845297.5978
rss_test  382933763448.1103
mse_train  782359299.9976472
mse_test  874277998.7399778
```

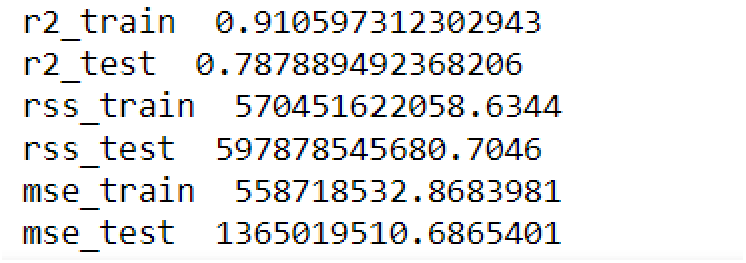
From this we can see that doubling the alpha value causes the error terms to increase and also causes a decrease in train and test R squared score. Hence model fit of alpha=40 is not as good as alpha=20

With alpha=100 following are the metrics for Lasso regression:



```
r2_train  0.9329979692114359
r2_test  0.6779320500586121
rss_train  427519777415.1168
rss_test  907816966124.0294
mse_train  418726520.4849332
mse_test  2072641475.1690168
```

With alpha=200 following are the metrics for Lasso regression:



```
r2_train  0.910597312302943
r2_test  0.787889492368206
rss_train  570451622058.6344
rss_test  597878545680.7046
mse_train  558718532.8683981
mse_test  1365019510.6865401
```

From this we can see that doubling the alpha value causes the error terms to increase, although the test errors have reduced, the training errors have increased. A positive fact here is that the R squared value has improved for the test set by doubling the alpha value.

The 5 most important predictor variables after the alpha change in Ridge regression are:

1. BsmtQual_Gd
2. Neighborhood_NoRidge
3. Neighborhood_NridgHt
4. OverallQual
5. KitchenQual_Gd

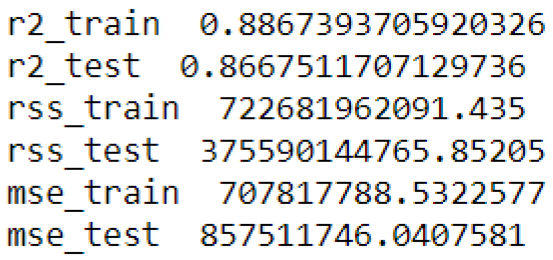
The 5 most important predictor variables after the alpha change in Lasso regression are :

1. PoolQC_Gd
2. Condition2_PosN
3. Neighborhood_NoRidge
4. Neighborhood_NridgHt
5. BsmtQual_Gd

Q2. 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?

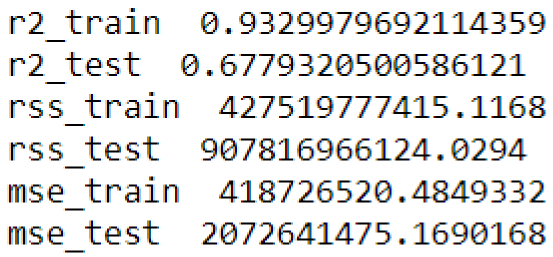
Answer:

Following are the metrics of Ridge regression



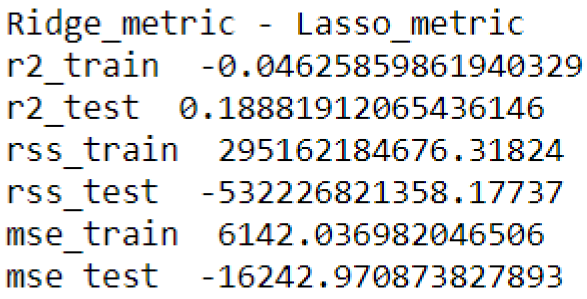
```
r2_train  0.8867393705920326
r2_test  0.8667511707129736
rss_train  722681962091.435
rss_test  375590144765.85205
mse_train  707817788.5322577
mse_test  857511746.0407581
```

Following are the metrics of Lasso regression



```
r2_train  0.9329979692114359
r2_test  0.6779320500586121
rss_train  427519777415.1168
rss_test  907816966124.0294
mse_train  418726520.4849332
mse_test  2072641475.1690168
```

On comparing the two (Ridge - Lasso)



```
Ridge_metric - Lasso_metric
r2_train  -0.04625859861940329
r2_test  0.18881912065436146
rss_train  295162184676.31824
rss_test  -532226821358.17737
mse_train  6142.036982046506
mse test  -16242.970873827893
```

We can see that Ridge regression provides better metrics than Lasso regression. Although the train values are important we are focusing more on the test scores and hence it is evident that Ridge regression gives higher R_squared score and also lesser error for the test set. Hence we can conclude that Ridge performs better.

Q3.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:

The Five most important predictor variables in the lasso model which are dropped are as follows:

1. PoolQC_Gd
2. Condition2_PosN
3. Neighborhood_NoRidge
4. Neighborhood_NridgHt
5. BsmtQual_Gd

The new lasso model is prepared with the remaining variables, and the variables with highest absolute coefficient values are considered to be most important. The five most important variables are:

1. RoofMatl_Roll
2. Electrical_Mix
3. Electrical_None
4. Neighborhood_Edwards
5. Electrical_FuseP

Q4.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:

In order to make a model robust and generalisable it is important to first split the data in order to maintain a test dataset which will act as the real world dataset. The idea being that the model will be trained and validated on the train dataset using cross-validation techniques, and then can be tested on the test dataset to understand the response and performance of the model. A model which does well during training need not do well on the test set, this becomes an important test to make sure the model developed is robust.

We can look at different metrics in order to understand the performance of the model. In our case we are solving a regression problem hence we can look at R squared score and mean square error as metrics to judge the model. A high R squared score tells us the quality of fit of the model and low mean squared error tells us how close the model is in predicting the true response.

Both these metrics have an implication on the accuracy of the model. The R squared value can tell us confidence we can have on the trained model to predict the values with high accuracy. The square mean error is inversely proportional to the accuracy, as lower error would signify higher accuracy.