# Assignment Part-II

## 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 to double the value of alpha for both ridge and lasso?

What will be the most important predictor variables after the change is implemented?

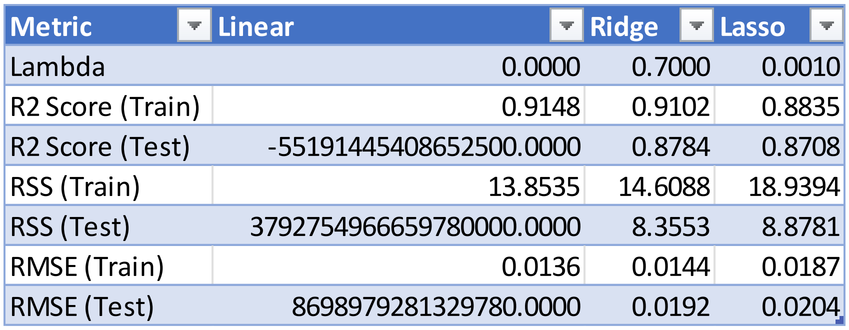
**Answer #1:**

Ridge regression: 0.7

Lasso regression: 0.001

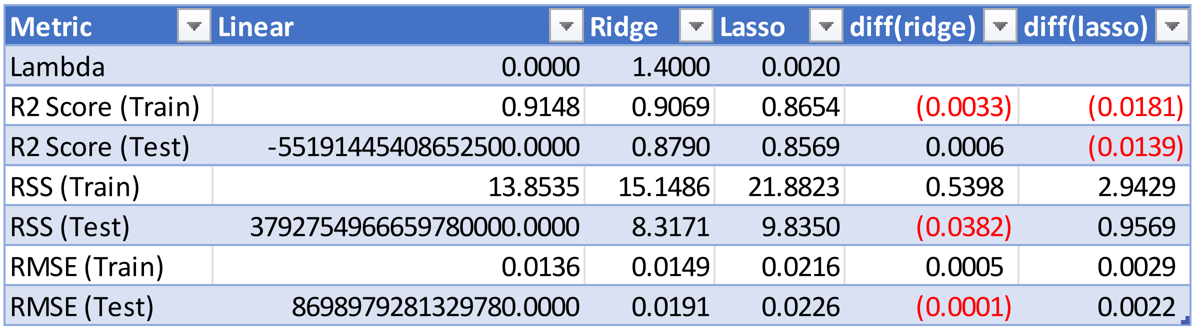
**Answer #2:**

Before doubling alpha:



| Metric | Linear | Ridge | Lasso |
|---|---|---|---|
| Lambda | 0.0000 | 0.7000 | 0.0010 |
| R2 Score (Train) | 0.9148 | 0.9102 | 0.8835 |
| R2 Score (Test) | -55191445408652500.0000 | 0.8784 | 0.8708 |
| RSS (Train) | 13.8535 | 14.6088 | 18.9394 |
| RSS (Test) | 3792754966659780000.0000 | 8.3553 | 8.8781 |
| RMSE (Train) | 0.0136 | 0.0144 | 0.0187 |
| RMSE (Test) | 8698979281329780.0000 | 0.0192 | 0.0204 |

After doubling alpha:



| Metric | Linear | Ridge | Lasso | diff(ridge) | diff(lasso) |
|---|---|---|---|---|---|
| Lambda | 0.0000 | 1.4000 | 0.0020 | | |
| R2 Score (Train) | 0.9148 | 0.9069 | 0.8654 | (0.0033) | (0.0181) |
| R2 Score (Test) | -55191445408652500.0000 | 0.8790 | 0.8569 | 0.0006 | (0.0139) |
| RSS (Train) | 13.8535 | 15.1486 | 21.8823 | 0.5398 | 2.9429 |
| RSS (Test) | 3792754966659780000.0000 | 8.3171 | 9.8350 | (0.0382) | 0.9569 |
| RMSE (Train) | 0.0136 | 0.0149 | 0.0216 | 0.0005 | 0.0029 |
| RMSE (Test) | 8698979281329780.0000 | 0.0191 | 0.0226 | (0.0001) | 0.0022 |

After doubling the alpha, we see that generally the R2 score has reduced by a small amount which means that the model isn’t optimum.

And the error has slightly increased for the training data for the ridge regression and for the lasso regression both the test and training data errors have increased.  This is given by the rows RSS and RMSE

**Answer #3:**

Ridge regression and Lasso Regression

GrLivArea:  Above grade (ground) living area square feet.

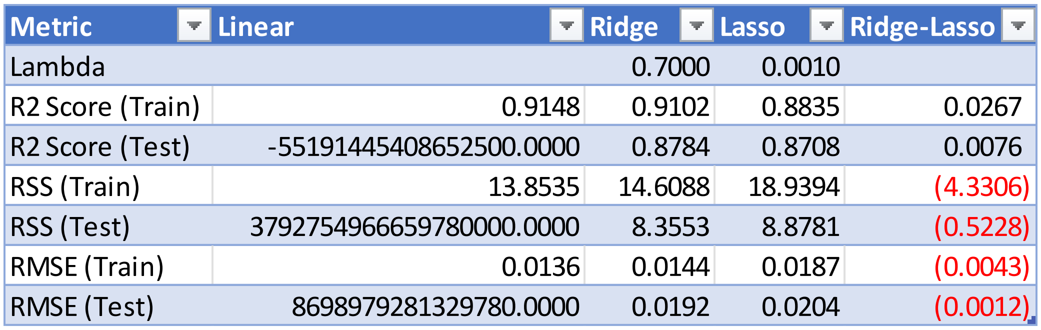
## Question 2

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**:

With the optimal value of lambda, between Ridge and lasso the r2 score on the training data reduced by ~2% and on the test data its reduced by 0.76%

The errors have increased by a small amount when using the Lasso regression.



| Metric | Linear | Ridge | Lasso | Ridge-Lasso |
|---|---|---|---|---|
| Lambda | | 0.7000 | 0.0010 | |
| R2 Score (Train) | 0.9148 | 0.9102 | 0.8835 | 0.0267 |
| R2 Score (Test) | -5519144540865250 0.0000 | 0.8784 | 0.8708 | 0.0076 |
| RSS (Train) | 13.8535 | 14.6088 | 18.9394 | (4.3306) |
| RSS (Test) | 3792754966659780000.0000 | 8.3553 | 8.8781 | (0.5228) |
| RMSE (Train) | 0.0136 | 0.0144 | 0.0187 | (0.0043) |
| RMSE (Test) | 869897928132978 0.0000 | 0.0192 | 0.0204 | (0.0012) |

The advantage of Lasso is that the model only needs to consider 78 features rather than the 186 features that it needs to cater to (after the encoding of categorical variables)

This helps in improving the performance of the model and allows for better fit for the test data as the model will not over fit, given we are ignoring variables that don't make much difference.

So, the recommendation would be to go with the *lasso regression.*

## 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:**

The five top predictors to drop:

1. GrLivArea
2. OverallQual_Very Good
3. Neighborhood_Crawfor
4. Neighborhood_NridgHt

5. MSZoning_FV

The five most important predictors after removing the above predictors are:

1. 1stFlrSF
2. GarageArea
3. FullBath_3
4. Neighborhood_StoneBr
5. Exterior1st_BrkFace

## 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 #1:**

To make sure the model is robust and generalizable we need to perform the following:

1. Perform data cleaning by
   a. Removing columns that have a majority of same values as these don't contribute to the model
   b. Remove columns that have majority null or empty values.
   c. Fill default values for rows where business logic is understood.
   d. Remove outliers in the data.
   e. Remove one of the predictors that are highly colinear to another predictor.
   f. Reduce the categorical variables when they have large number of constants by giving them ranges.
   g. Convert categorical variables to dummy variables.
2. If the errors on the response variable is very large, we will perform data transformation on the response variable to reduce the errors
3. Scale the numerical predictors so that they are in scale to the categorical variables.
4. After building the model using Linear regression and ensure that the r2 and RSS are in acceptable ranges between the training and test data
5. If not perform regularization:
   a. Prefer Lasso regression if the R2 and error values are not very different from the ridge regression as this will also reduce the features in the model.
   b. Reducing the features ensure the model doesn't overfit and will generalize for the test data
   c. This also improves the performance of the model
   d. Also regularization will also ensure that the linearity of the predictor variables to the response variable is maintained
6. Always perform residual analysis
   a. The errors to the predicted response shouldn't show any patterns.

b. The errors should be normally distributed.

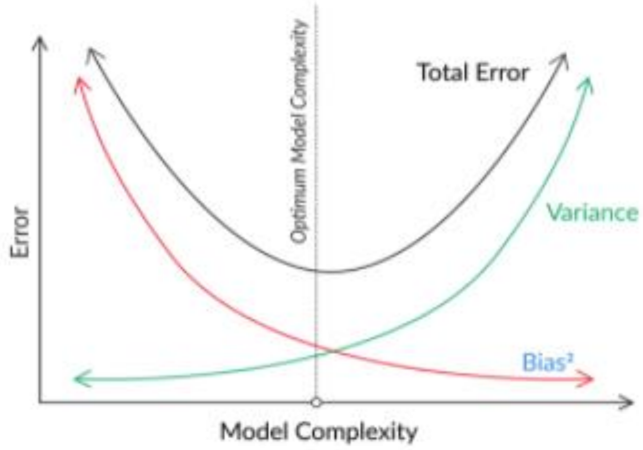
**Answer #2:**

The implication of doing this is we get the lowest total error, i.e., low bias and low variance, such that the model identifies all the patterns that it should and is also able to perform well with unseen data.

In the below graph we are trying to reduce both the Bias and Variance

If Bias is too high (variance is low) the model becomes too simple and might be too general to solve a particular problem

And if the variance is too high (Bias is low) the model becomes to complex and might end up overfitting the training data. This would mean that the model wouldn't work for the changes in the test data.



Regularization helps in fitting in various hyper parameters to find the optimum model with low bias and variance that doesn't underfit or over the data.