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?

Solution:

(1) Optimal Value of Alpha:

• The computed optimal value of alpha for Ridge Regression (Original Model):10

• The computed optimal value of alpha for Lasso Regression (Original Model): *0.001*

(2) Changes in the model, if you choose double the value of alpha for both ridge and lasso regression:

Original Model (alpha=10), Doubled Alpha Model (alpha=20)

Ridge Regression

Alpha = 10

Metric	Train	Test
R-Squared	0.9454069223696316	0.8896634254960011
RMSE	0.015579464291382315	

Alpha = 20

Metric	Train	Test
R-Squared	0.9426081292764921	0.8885756103571907
RMSE	0.01573306319715681	

Observations:

• The test accuracy of the ridge regression model (alpha=10) is slightly higher in comparison to the test accuracy of the doubled alpha model (doubled alpha=20).

• MSE test scores comparing similar data of the original dataset and doubled alpha model gives us an idea that it is slightly smaller for the single alpha model than the doubled alpha model.

Lasso Regression

Alpha = 0.001

Metric	Train	Test
R-Squared	0.9380137784840006	0.8870946822311135
RMSE	0.015942169442860383	

Alpha = 0.002

Metric	Train	Test
R-Squared	0.9295064143216281	0.8796840555215417
RMSE	0.01698854590250255	

Observations:

• The test accuracy of the lasso regression model (alpha=0.001) is slightly higher in comparison to the test accuracy of the doubled alpha model (doubled alpha=0.002).

• MSE test scores comparing similar data of the original dataset and

(3) The most important predictor variables after the change is implemented. Top 10 features are as follows:

Ridge Regression

1. ('Age', -0.053),
2. ('MSSubClass_2-STORY PUD - 1946 & NEWER', -0.039),
3. ('MSSubClass_1-STORY 1945 & OLDER', -0.033),
4. ('Neighborhood_MeadowV', -0.028),
5. ('Neighborhood_OldTown', -0.027),
6. ('BldgType_Twnhs', -0.026),
7. ('SaleType_WD', -0.023),
8. ('MonthSold_November', -0.022),
9. ('Neighborhood_Gilbert', -0.021),
10. ('Neighborhood_IDOTRR', -0.021)

Lasso

1. ('Age', -0.06),
2. ('MSZoning_RM', -0.025),
3. ('Remod_Age', -0.012),
4. ('KitchenAbvGr', -0.011),
5. ('BsmtCond', -0.01),
6. ('SaleType_WD', -0.005),
7. ('MSSubClass_2-STORY PUD - 1946 & NEWER', -0.004),
8. ('LotShape', -0.003),
9. ('ExterCond', -0.003),
10. ('LandContour', -0.002)

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?

Solution:

Optimal Value of Alpha:

- The computed optimal value of alpha for Ridge Regression (Original Model): 10.0
- The computed optimal value of alpha for Lasso Regression (Original Model): 0.001

Ridge Regression

Alpha = 10

Metric	Train	Test
R-Squared	0.9454069223696316	0.8896634254960011
RMSE	0.015579464291382315	

Lasso Regression

Alpha = 0.001

Metric	Train	Test
R-Squared	0.9380137784840006	0.8870946822311135
RMSE	0.015942169442860383	

- The R2 test score on the Ridge Regression Model is slightly better than that of Lasso Regression Model. Moreover, the training accuracy is slightly reduced; hence, making the model an optimal choice as it seems to perform better on the unseen data.
- The MSE for Test set (Ridge Regression) is slightly lower than that of the Lasso Regression Model; implies Ridge Regression performs better on the unseen test data

Question 3

Solution:

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?

Solution

The top 5 Features for the Lasso Model are

1. Age
2. MSZoning_RM
3. Remod_Age
4. KitchenAbvGr
5. BsmtCond

The Top variables if the above original variables are removed are as follows

1. SaleType_WD
2. MSSubClass_2-STORY PUD - 1946 & NEWER
3. LotShape
4. ExterCond
5. LandContour

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?

Solution:

The testing error of a model must be consistent with the training error, or the model must perform well with sufficient stability even after adding noise to the dataset. As a result, a model's robustness (or generalizability) is a measure of how well it can be applied to data sets other than the ones used for training and testing.

We can adjust the trade-off between model complexity and bias, which is directly related to the model's robustness, by using regularization approaches. Regularization aids in penalizing coefficients for overcomplicating the model, allowing only the optimum level of complexity to be used. It aids in the regulation of the model's robustness by making the model optimally simple. Therefore, in order to make the model more robust and generalizable, one need to make sure that there is a delicate balance between keeping the model simple and not making it too naive to be of any use. Also, making a model simple leads to Bias-Variance Trade-off

➢ A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.

➢ A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.