Assignment Part-II

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?

Optimal value of alpha for Lasso regression: 0.0001

Optimal value of alpha for Ridge regression: 20

Now on doubling the value of alphas:

Value of alpha for Ridge regression: 40

Value of alpha for Lasso regression: 0.0002

Ridge MSE with alpha = 20: 0.019450509001835214

Ridge MSE with alpha = 40: 0.019450509001835214

Lasso MSE with alpha = 0.001: 0.020219075353064733

Lasso MSE with alpha = 0.002: 0.02057527218331345

It has been observed that on increasing the alpha value, the Mean squared error is slightly increased for both Lasso and Ridge.

Ridge R2 score with alpha = 20: 0.8817870719700953

Ridge R2 score with alpha = 40: 0.8817870719700953

Lasso R2 score with alpha = 0.001: 0.8771160127831338

Lasso R2 score with alpha = 0.002: 0.874951181505212

With the increase in value of alpha, we can see slight decrease in the value of R2.

After doubling the value of alpha, the 5 most important variables turned out to be as mentioned below:

From Ridge Model:

1. MSSubClass

2. OverallCond
3. BsmtFullBath
4. Neighborhood_Crawfor
5. Neighborhood_NridgHt

From Lasso Model:

1. MSSubClass
2. MSZoning_RL
3. MSZoning_RH
4. MSZoning_FV
5. MSZoning_RM

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?

The optimal lambda value in case of Ridge and Lasso is as below:

- Ridge : 20
- Lasso: 0.0001

The Mean squared Error in case of Ridge and Lasso are:

- Ridge: 0.019450509001835214
- Lasso: 0.020219075353064733

The Mean squared error of Lasso and Ridge are almost equal.

Also, since Lasso helps in feature reduction (assigns zero value to insignificant features) Lasso is a better model over Ridge. Therefore the variables predicted by Lasso can be applied for choosing significant variables in predicting the price of the house.

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?

The five most important predictor variables after creating another model excluding the five most important predictor variables are:

1. LotFrontage
2. BsmtFullBath
3. Neighborhood_Crawfor
4. Neighborhood_Somerst
5. OverallCond

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?

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance Trade-off. The simpler the model the more the bias but less variance and more generalisable. Its implications in terms of accuracy is that a robust and generalisable model will perform equally well on both train and test data i.e., the accuracy does not change much for training and test data.

Bias: Bias is error in 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 poor on training and testing data.

Variance: Variance is error in model, when model tries to overlearn from the data. High variance means model performs exceptionally well on training data, but fails to perform on test data.

It is important to have balance in bias and variance to avoid overfitting and under fitting of the data.