Problem Statement- 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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

When we plot the curve between negative mean absolute error and alpha in the case of ridge regression, we observe that the error term decreases as the value of alpha increases from 0 and the training error is showing an increasing trend when the value of alpha increases. Since the test error is lowest when the value of alpha is 2, we chose to use a value of alpha equal to 2 for our ridge regression.

I have chosen to maintain a very low value for lasso regression, which is 0.01. As we increase the value of alpha, the model tries to penalize more and tries to make the majority of the coefficient value zero. It started out with a negative mean absolute error and an alpha of 0.4.

When we double the alpha value for our ridge regression, instead of using a value of alpha equal to 10, the model will apply more penalty to the curve and attempt to become more generalized, which will make the model simpler and eliminate the need to fit every piece of data in the data set. From the graph, it is clear that when alpha is 10, we experience increased error for both the test and the train data.

Similar to how increasing the lasso's alpha penalizes our model more and causes more coefficients of the variable to be reduced to zero, increasing the r2 square results in a reduction in both.

The most important variable after the changes have been implemented for ridge regression is as follows:-

1. MSZoning_FV
2. Neighborhood_Crawfor
3. SaleType_New
4. MSZoning_RL
5. Neighborhood_StoneBr
6. SaleCondition_Normal
7. MSZoning_RH
8. SaleCondition_Alloca
9. MSZoning_RM

The most important variable after the changes have been implemented for lasso regression are as follows:-

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF

5. BsmtFinSF1

6. GarageArea

7. Fireplaces

8. LotArea

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?

Regularising coefficients is crucial for increasing prediction accuracy, reducing variation, and making the model understandable.

As the penalty is the square of the magnitude of the coefficients, which is determined via cross-validation, ridge regression employs a tuning parameter called lambda. By applying the penalty, the residual sum of squares should be minimal. The coefficients with higher values are penalized because the penalty is equal to lambda times the sum of the squares of the coefficients. The variance in the model is lost when we raise the value of lambda, while the bias stays constant. In contrast to Lasso Regression, Ridge Regression incorporates all variables into the final model.

The penalty in lasso regression, which is determined via cross-validation, is the absolute value of the coefficients' magnitude. As the lambda value rises, Lasso reduces the coefficient in the direction of zero, bringing the variables exactly to zero. Lasso performs variable selection as well. When lambda is small, straightforward linear regression is performed; however, as lambda increases, shrinkage occurs and variables with a value of 0 are ignored by the model.

Q3. 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?

Those 5 most important predictor variables that will be excluded are :-

1. GrLivArea

2. OverallQual

3. OverallCond

4. TotalBsmtSF

5. GarageArea

Q4. 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?

The model should be as straightforward as feasible because it will be more reliable and generalizable even though its accuracy will suffer. The trade-off between bias and variance can also be used to understand it. The more biased the model is, the less variation it has and the more generalizable it is. It implies that a robust and generalizable model will perform equally well on both training and test data, i.e., the accuracy does not significantly differ between training and test data.

Bias: Bias is an error in the model when the model is weak to learn from the data. High bias means the model is unable to learn details in the data. The model performs poor on training and testing data.

Variance: Variance is an error in the model when the model tries to overlearn from the data. High variance means the model performs exceptionally well on training data as it is very well trained on this data but performs very poorly on testing data as it was unseen data for the model.

It is important to balance Bias and Variance to avoid overfitting and under-fitting data.

****

END