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

Answer:
Lasso regression: For lasso regression, initially we have seen that Negative Mean Absolute Error is quite low at approx alpha = 0.4 and after that it stabilizes So we need to choose a low value of alpha to balance the trade-off between Bias-Variance and to get the coefficients of smallest of features and thus decided to keep very small value i.e. 0.01.

Riidge regression:- For ridge regression, we have seen that Negative Mean Absolute Error stabilizes at alpha = 2 when we plot the curve between negative mean absolute error and alpha . Here the test error is minimum so we decided to go alpha equal to 2.

In this case of our model, if we choose to double the value of alpha, it won’t affect much, as I have implemented in the jupiter file. But if we increase/double the value of alpha on a broader prospect for ridge regression i.e. alpha equal to 10, the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set . Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases

Important variables after the changes has been implemented for ridge regression are as follows:-

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

Important variables 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
9. LotArea
10. LotFrontag

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: Generally there is a typical difference between lasso and ridge regression which makes them unique in their own ways. But it is important in our prediction that we should regularize the coefficients and make the model more interpretable.

In Ridge regression,as we increase the value of lambda the variance in the model is dropped and bias remains constant and also it includes all variables in the final model unlike Lasso Regression. On the contrary, in lasso regression, as the lambda value increases Lasso shrinks the coefficient towards zero and it makes the variables exactly equal to 0, which results in variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model

Hence, I will select alpha value same as ridge regression, as it will consider the required variable, without the notion of eliminating the features, which might be useful in model building.

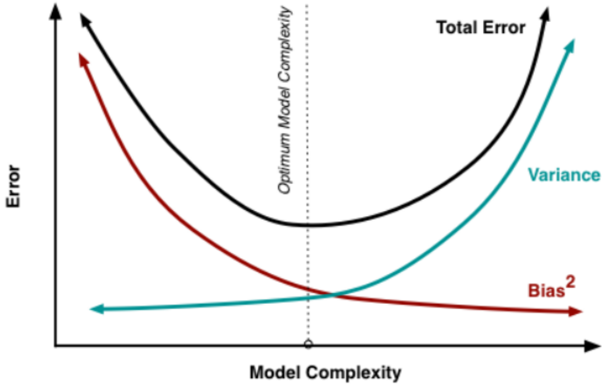
Question 3: After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer: If we will not consider the first five most important predictors, then we can consider the next 5 important predictor variable, which are as follows:

1. BsmtFinSF1 : Type 1 finished square feet
2. LotArea : Lot size in square feet
3. GarageArea : Size of garage in square feet
4. Fireplaces : Number of fireplaces
5. LotFrontage : Linear feet of street connected to property

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?

Answer: To get the clear picture of a model, being robust and generalisable, it must be as simple as possible. Although its accuracy will decrease, it will be considered on the basis of broader prospects. We can also understand this concept using the Bias-Variance trade-off.



The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is the error 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 poorly on training and testing data.

Variance: Variance is the error when a model tries to over learn from the data. High variance means the model performs exceptionally well on training data as it has 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 have balance in Bias and Variance to avoid overfitting and under-fitting of data.