Part 2- Subjective Question

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:

In the case of ridge regression, When we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decrease and the training errors show an increasing trend when the value of alpha increases. when the value of alpha is 2 the test error is minimum so we decided to go with the value of alpha equal to 2 for our ridge regression.

For lasso regression, I have decided to keep a very small value that is 0.01, when we increase the value of alpha the model tries to penalize more and try to make most of the coefficient value zero. Initially, it came as 0.4 in negative mean absolute error and alpha.

When we double the value of alpha for our ridge regression no we will take the value of alpha equal to 10 the model will apply more penalty on the curve and try to make the model more generalized that is making the model 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 our model, and more coefficients of the variable will be reduced to zero, when we increase the value of our r2 square also decreases.

The most important variable after the changes have 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

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

GrLivArea

OverallQual

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 to and why?

Solution:

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and make the model interpretably.

Ridge regression, uses a tuning parameter called lambda as the penalty is the square of the magnitude of coefficients which is identified by cross-validation. The residual sum of squares should be small by using the penalty. The penalty is lambda times the sum of squares of the coefficients, hence the coefficients that have greater values get penalized. As we increase the value of lambda the variance in the model is dropped and bias remains constant. Ridge regression includes all variables in the final model, unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is the absolute value of the magnitude of coefficients which is identified by cross-validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it makes the variables exactly equal to 0. Lasso also does variable selection. When the lambda value is small it performs simple linear regression and as the lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

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?

Solution:

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

GrLivArea

OverallQual

OverallCond

TotalBsmtSF

GarageArea

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 robustness of a model implies, that either the testing error of the model is consistent with the training error, or the model performs well with enough stability even after adding some noise to the dataset. Thus, the robustness (or generalizability) of a model is a measure of its successful application to data sets other than the one used for training and testing.

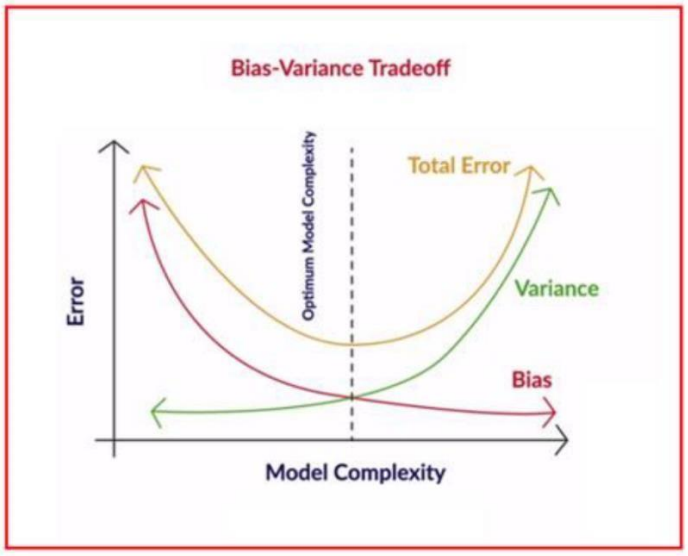
By implementing regularization techniques, we can control the trade-off between model complexity and bias which is directly connected to the robustness of the model. Regularization helps in penalizing the coefficients for making the model too complex; thereby allowing only the optimal amount of complexity to the model. It helps in controlling the robustness of the model by making the model optimal simpler. Therefore, in order to make the model more robust and generalizable, one needs 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.

Bias helps you quantify, how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there has to be enough training data. Models that are too naïve, for e.g., one that gives the same results for all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs is very high. Variance is the degree of changes in the model itself with respect to changes in the training data.

Thus, the accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph.



Thus, accuracy and robustness may be at the odds with each other as too many accurate model can be prey to overfitting hence it can be too much accurate on train data but fails when it faces the actual data or vice versa.