Assignment – Advanced Regression Subjective Questions

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 1:

Optimal value of alpha for ridge & lasso regression are as follows:

Ridge: 8.0
Lasso: 0.0001
After doubling the alpha for Ridge:

1. There is slight decrease in R2 score for train (from **0.908389** to **0.896573**) and test (from **0.873047** to **0.867770**) with doubling of alpha.
2. The value of betas has come down with doubling of alpha.
3. The value of MSE has gone with very low margin train data but is slightly down for test data
4. With doubling of alpha, Top 10 predictors are *GrLivArea, 1stFlrSF, OverallQual_10, 2ndFlrSF, TotalBsmtSF, Neighborhood_NoRidge, BsmtFinSF1, OverallQual_9, RoofMatl_WdShngl and FullBath.*

After doubling the alpha for Lasso:

1. There is slight decrease in R2 score for train and but increase in R2 score for test with doubling of alpha.
2. The value of betas has come down with doubling of alpha.
3. The value of MSE has gone with very low margin.
4. With doubling of alpha, more features have been eliminated.
5. With doubling of alpha, Top 10 predictors are *2ndFlrSF, OverallQual_10, 1stFlrSF, OverallQual_9, TotalBsmtSF, RoofMatl_WdShngl, Neighborhood_NoRidge, BsmtFinSF1, OverallQual_8 and LotArea.*

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 2:

Both models are having similar R2 score with very less margin of difference. But Lasso model's ability to eliminate features by making it zero is one advantage of it.

Below are some of key points from my analysis:

1. Lasso R2 train score(0.923) is higher than Ridge(0.908) but Ridge R2 test score(0.873) is better than Lasso (0.866) .
2. RSS(test) is better for Ridge.
3. Lasso has 129 non-zero beta co-efficient.
4. Ridge has 270 non-zero beta co-efficient.
5. We choose Lasso Model after looking at all the metrics from above, due to fact that it has less beta making it simpler and R2 score is also reasonably good.

Question 3

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?

Answer 3:

Below are the next 5 important predictors from new Lasso model after removing the first 5 important predictor from 1st lasso model:

1. GrLivArea
2. TotalBsmtSF
3. Neighborhood_NoRidge
4. OverallCond_9
5. BsmtFinSF1

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 4:

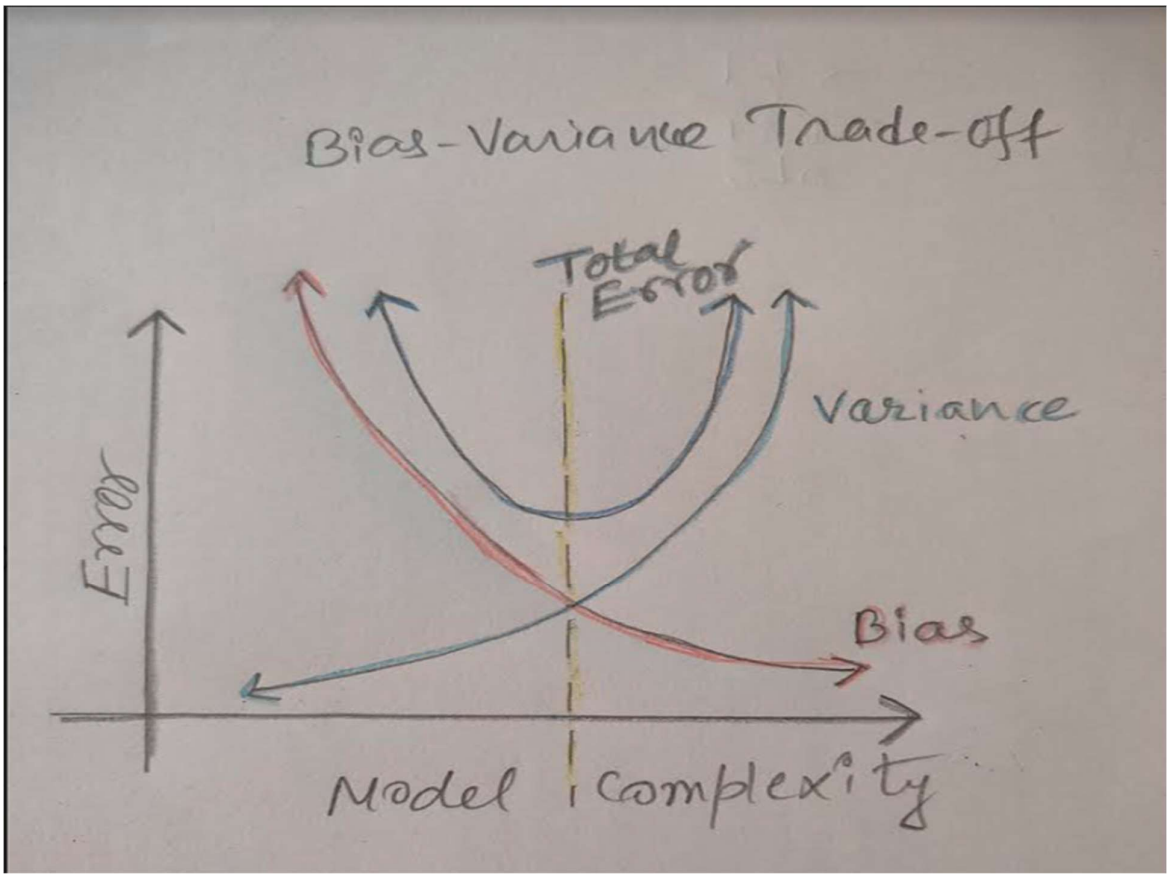
A robust model is one which is resistant to the influence of outliers or extreme values and a generalizable model is one which performs well on the unseen data also.

When any model is not over fitting and it is also as simple as possible then we can say that it is robust and generalizable.

Overfitting of the model will definitely increase the accuracy as it memorizes the training data. But it can not ensure that it will perform well with new unseen data.

A robust and generalized model may not have the best accuracy but it will have a pretty good accuracy on both train and test data and any other unseen data.

It is a balance between bias and variance as shown in the below image.


Bias-Variance Trade-Off
Total Error
Variance
Bias
Error
Model complexity