Problem Statement - 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: The Optimal value of alpha for ridge is 20 and for lasso it is 0.001.

After choosing the double value of alpha, the value for ridge is 40 and for lasso it is 0.002.

Once we double the alpha values in the Ridge and Lasso, there is a small change in the coefficient values. The new model is created and demonstrated in the Jupyter notebook. Below are the changes in the coefficients.

For Ridge model: There are no changes in the coefficient value as we double alpha value for ridge model but R2 value for both train and test data slightly dropped.

For Lasso model: The coefficient values are decreasing as we double the alpha value as well as R2 value for both train and test data slightly dropped.

Sinch the alpha values are small, we have not seen a huge changes in the model after doubling the alpha

Most important predictor variables: GrLivArea, OverallQual, SaleType_New, OverallCond, TotalBsmtSF.

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: We will choose Lasso regression as it is giving a feature selection option . It also removed the unwanted features from the model without affecting the model accuracy which makes the model generalized and simple and accurate.

According to my code the optimum lambda value in case of Ridge and Lasso is as follows:-

Ridge – 20
Lasso – 0.001

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

Ridge - 0.1200835212248011

Lasso - 0.11912028637034187

The Mean Squared Error of both the models are almost the same.Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), should be used as the final model.

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: The five most important predictor variables in the current lasso model is:

1. GrLivArea

2. OverallQual

3. SaleType_New

4. OverallCond

5. TotalBsmtSF

After removing the above variables from the dataset we build a Lasso model in the Jupyter notebook.

The R2 value of the new model without the top 5 predictors dropped in both train (0.903779485555622) & test (0.9061665307821379) data.

The Mean Squared Error increases to 0.12311186095571727

The Top new 5 predictor variables are:-

1. Neighborhood_StoneBr

2. 1stFlrSF

3. 2ndFlrSF

4. Neighborhood_NridgHt

5. Neighborhood_Crawfor

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 make model robust and generalisable 3 features are required:

1. Model accuracy should be > 70-75% and it is coming more than 90%(Train) and 90%(Test) which is correct.

2. P-value of all the features is < 0.05

3. VIF of all the features are < 5

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

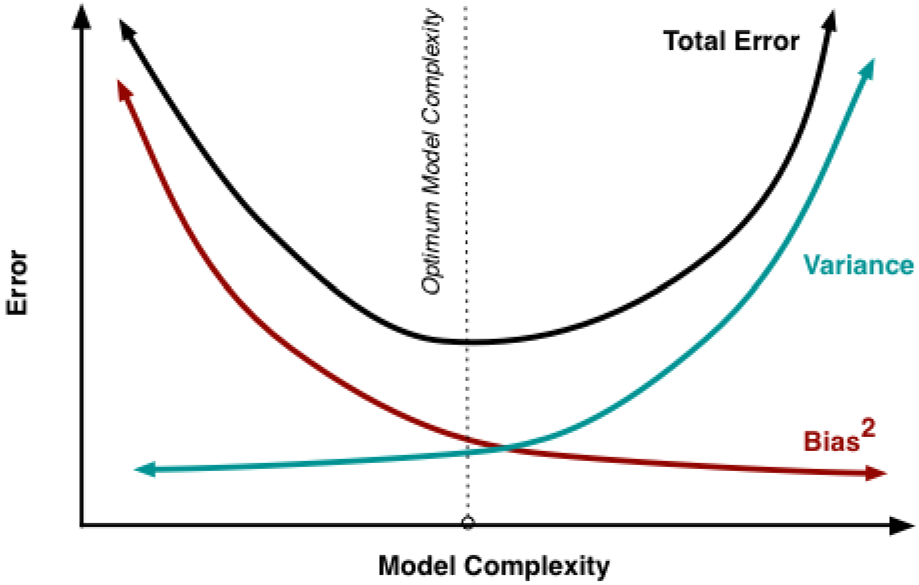
Also, Making a model simple leads to Bias-Variance Trade-off:

- 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 quantifies how accurate the model is likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.

Variance refers to the degree of changes in the model itself with respect to changes in the training data. Thus 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.


Optimum Model Complexity
Total Error
Variance
Bias^2
Error
Model Complexity