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 1:

The optimal values of alpha from Ridge and Lasso regression are as follows:

• Optimal alpha for Ridge = 6.0

• Optimal alpha for Lasso = 0.001

The changes in the model, if we double the value of optimal alpha for Ridge and Lasso are as follows:

• Alpha for Ridge = 12.0

o The R2_Score for train set has decreased from 0.9358 to 0.9289

o The R2_Score for test set has also decreased from 0.8894 to 0.8875

• Alpha for Lasso = 0.002

o The R2_Score has increased from 0.9118 to 0.9289 for train data set

o The R2_Score has decreased from 0.8892 to 0.8875 for test data set

The most important predictor variables after the change is implemented are as follows:

• Ridge – Top 5 predictor variables are:

o GrLivArea

o OverallQual_Excellent

o 1stFlrSF

o Neighborhood_StoneBr

o OverallQual_Very Good

• Lasso – Top 5 predictor variables are:

o GrLivArea

o OverallQual_Excellent

o Functional_Typ

o OverallQual_Very Good

o Neighborhood_Crawfor

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:

One of the basic difference between Ridge and Lasso is that Lasso shrinks the less important feature's coefficient to zero thus, removing some feature altogether. Thereby helps in performing feature selection.

While, Ridge shrinks the coefficients to small values but not to exactly zero. Hence does not help in performing feature selection.

Therefore, the variables predicted by Lasso can be applied to choose significant variables for predicting the price of a house. So, we should go with the Lasso 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 3:

The five most important predictor variables from the Lasso regression was:

- GrLivArea
- OverallQual_Excellent
- Neighborhood_StoneBr
- OverallQual_Very Good
- Functional_Typ

Now, that these variables are not a part of dataset, we have to rebuild the model without these variables. So the new set of 5 most important variables are:

- MSZoning_FV
- MSZoning_RL
- RoofMatl_WdShngl
- MSZoning_RH
- MSZoning_RM

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:

Occam's razor is a heuristic that suggests choosing simpler machine learning models as they are expected to generalize better. We should pick the model which is simpler as:

- Simpler models are usually more ’generic’ and are more widely applicable.
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust and performs well on the unseen test data.
- Complex models tend to change wildly with changes in the training data set
- Simple models have low variance, high bias and complex models have low bias, high variance.
- Simpler models make more errors in the training set.
- Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples

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

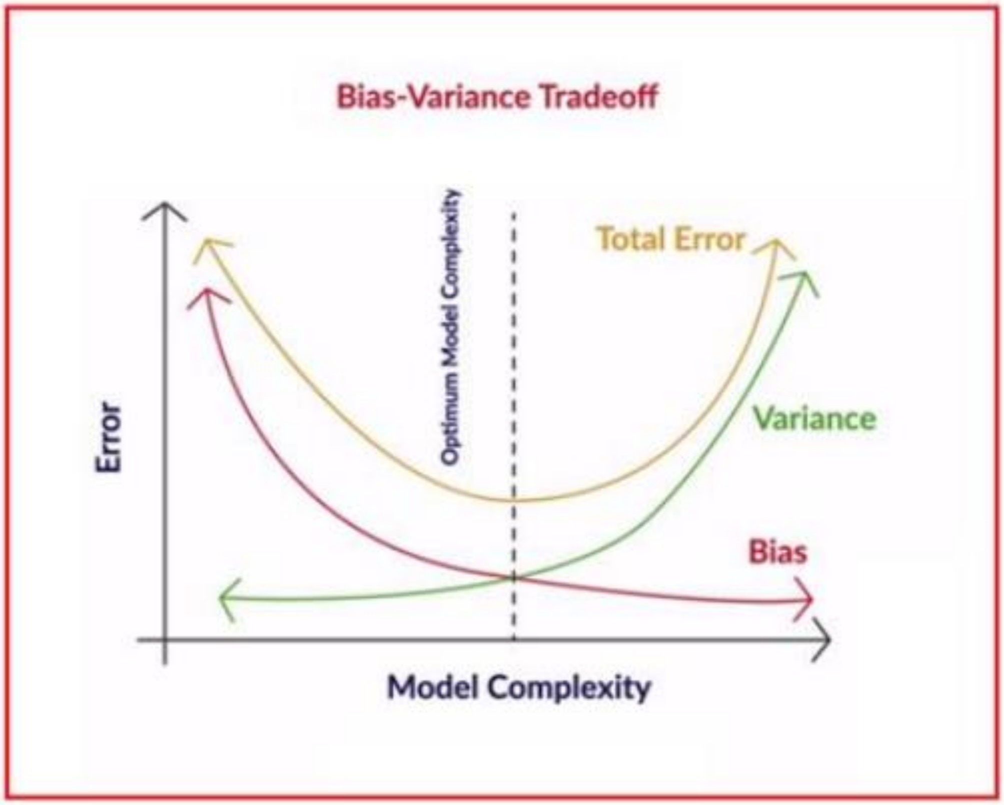
Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or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

Also, we need to take into account 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 is the model 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 naive, 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 above graph