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 :

Optimal alpha value for both ridge and lasso regression for my model came as 100.

Five important predictor variables before changes are as follows

Ridge	Lasso
GrLivArea	RoofMatl_CompShg
OverallQual	RoofMatl_Tar&Grv
RoofMatl_CompShg(Roof material with standard shingle)	RoofMatl_WdShngl
2ndFlrSF	RoofMatl_WdShake
RoofMatl_WdShngl	GrLivArea

After the value is changed there is not much significant change in the metricsh such as r2 score and mean squared error but the weightage of feature coefficients got reduced drastically.

For lasso no of features that got eliminated (that is coefficient 0) has increased(from 58 to 78).

After the changes predictor variables are as follows

Ridge	Lasso
GrLivArea	RoofMatl_CompShg
OverallQual	RoofMatl_Tar&Grv
2ndFlrSF	RoofMatl_WdShngl
RoofMatl_WdShngl	GrLivArea
Neighborhood_NoRidge	RoofMatl_WdShake

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?

Even though accuracy wise there isn't a drastic difference between two techniques, I will choose lasso over ridge considering it does feature elimination i.e it eliminates 58 features making the model less complex and more interpretable.

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?

After removing the variables again there is not much change in the accuracy metrics and no of features eliminated are now 48 and top 5 important features are now

1. 2ndFlrSF
2. 1stFlrSF
3. OverallQual
4. GarageCars
5. Neighborhood_NoRidge

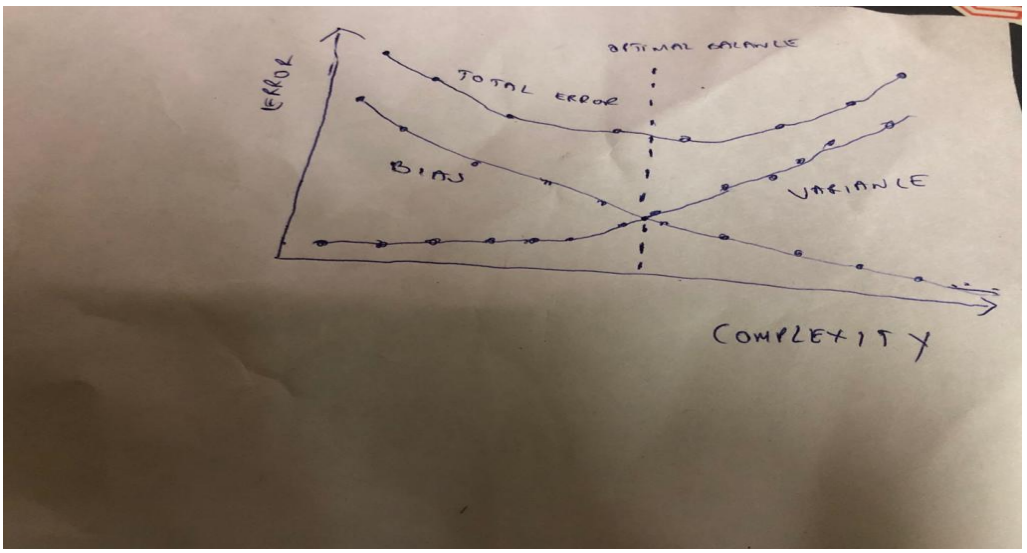
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

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?

- Model performance should not be affected by any changes to the data
- Generic model should adapt and do well on the unseen data and accuracy should be good
- As per ocam razor a good model should not be complex and it should interprettable

But point 2 & 3 has a trade off let say if the model is too simple it leads to underfitting that is it doesnt learn important patterns leading to low accuracy(**Bias)** and on the other hand if the model is too complex it overfits training data and does not generalise well on the unseen data this also leads to any changes to data will impact the performance of the model(**Variance)**

The above variation is called bias variance trade off and we should find optimal point that balances both as per the below diagram



There are few techniques that we can follow to maintain balance between model complexity

1. Use cross validation techniques during the model building process as it trains on the subset of input data and testing on previously unseen subset of the input data. This approach makes sure that all patterns are captured during training
2. Use regularization techniques like lasso or ridge which reduces complexity by penalising the coefficients.This also make sure that multicollinearity between independent variables are handled.