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 and Lasso regression during the assignment is as below:

- Ridge 8
- Lasso 0.0001

As we increase or double the above values, their will be more penalty imposed for parameters and hence the model will become less complex with more bias and if it is increased sufficiently, it will lead to underfitting.

The 10 most important variable after doubling the value for ridge regression are as follows:-

1) GrLivArea
2) MSZoning_RL
3) OverallQual
4) MSZoning_RM
5) GarageCars
6) OverallCond
7) MSZoning_FV
8) GarageType_Attchd
9) TotalBsmtSF
10) BsmtFinSF1

The most important variable after doubling the value for lasso regression are as follows:-

1) MSZoning_RL
2) GrLivArea
3) MSZoning_RM
4) OverallQual
5) MSZoning_FV
6) GarageCars
7) OverallCond
8) GarageType_Attchd
9) Foundation_PConc
10) 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:

The optimal value of lambda for Ridge and Lasso regression during the assignment is as below:

- Ridge 8
- Lasso 0.0001

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

- Ridge 0.122496
- Lasso 0.121863

The Mean Squared Error (Test) of Lasso is slightly lower than that of Ridge.

The R2 Score(Test) in case of Ridge and Lasso are:

- Ridge 0.887144
- Lasso 0.888307

The R2 Score(Test) of lasso is slightly better than Ridge and also it has less difference from R2 Score(Train).
Also, since Lasso helps in feature reduction (as the coefficient value of one of the feature 'RoofStyle_Gable' became 0), Lasso has a slightly better edge over Ridge.

Therefore, the variables predicted by Lasso can be applied to choose significant variables for predicting the price of a house.

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:

For analyzing this situation, I dropped the 5 Most important feature variables through lasso regularization in my previous regression analysis with full variables. Those 5 most important predictor variables in the lasso model, which were dropped, are :-

1. MSZoning_RL
2. GrLivArea
3. MSZoning_RM
4. OverallQual
5. MSZoning_FV

After dropping the above 5 variables, the regularization was run again for ridge and lasso.

The updated list of 5 most important feature variables for lasso regularization are as below :-

1) 2ndFlrSF
2) 1stFlrSF

3) OverallCond
4) TotalBsmtSF
5) GarageCars

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:

While selecting a final model, apart from performance/accuracy, the robustness and generalizability of the model on unseen data should be ascertained. The model should perform equally good on train dataset and the unseen test dataset. At the same time, it should not be too naïve, to have high error terms for both training and test dataset. Hence, a balance is required to maintain sufficient accuracy, along with generalizability of the model.

Given that 2 models show similar performance, a less complex model is more preferred choice, as

Less complex models are usually more 'generic', 'robust', easy to explain and hence are more acceptable

Less complex models require fewer training samples for effective training

Less complex models have low variance, high bias and complex models have low bias, high variance

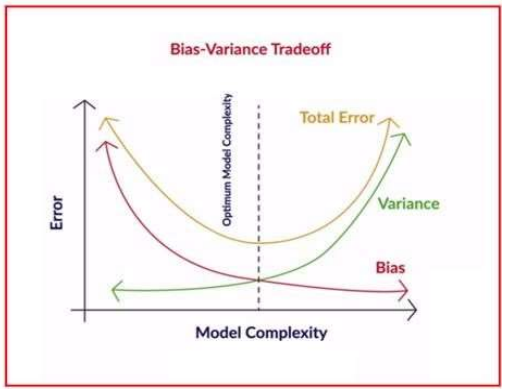
Complex models tend to change wildly with changes in the training data set

Less complex models make more errors in the training set.

Complex models may lead to overfitting. They work very well for the training samples but fail miserably when applied to unseen test data.

Hence, to make the model more robust and generalizable, we should make the model simple but not so simple that it has no applicability and not sufficient accuracy in prediction.

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. This leads to Bias-Variance Trade-off:



A simpler model that abstracts out some pattern followed by the data points, 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. 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.