Advance Regression Assignment Part – II
Answers of Subjective question by Anirudhya Bhattacharya

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 Value of alpha for ridge and lasso regression are:
    - Optimal Value of lambda for Ridge: 10
    - Optimal Value of lambda for Lasso: 0.0001

The tuning parameter lambda(alpha) helps us determine how much we wish to regularize the model. When we double alpha below changes will happen:

Ridge:

- It will decrease the coefficients of the variables.
- It will increase the bias square and decrease the variance. Which means the regularization will be higher.

Lasso:

- It will decrease the coefficients of the variables and push the coefficients more towards zero and results in setting more coefficients to exactly zero.
- With an increase in the value of lambda, variance reduces with a slight compromise in terms of bias.

After these changes are implemented,
The most important predictor variable for Ridge is MSZoning_RM with a coefficient of 0.21266.
The most important predictor variable for Lasso is MSZoning_RM with a coefficient of 0.52634.

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:

During my assignment the optimal value of lambda for ridge and lasso came out to be 10 and 0.0001.

Ridge regression:
R2 Score for training Data: 0.9288422154518519
R2 Score for test Data: 0.9156456903563702

Lasso regression:
R2 Score for training Data: 0.918916698503848
R2 Score for test Data: 0.917062109753026

I have got good scores for both Ridge and Lasso, but I will choose to finalize with Lasso as it also pushes some coefficients to be exactly 0 and thus performs variable selection. This variable selection results in models that are easier to interpret.

Question 3

After building the model, you realized 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:

Initially the top 5 variables and their coefficients were:

GrLivArea	0.29664
OverallQual	0.23678
GarageCars	0.10639
OverallCond	0.10597
BsmtFinSF1	0.09307

Once we dropped these variables and run the model again the new top 5 variables came out to be as below

MSZoning_RM	0.63543
Utilities_NoSeWa	0.49349
MSZoning_RH	0.39281
LowQualFinSF	0.31413
2ndFlrSF	0.29081

In my submitted notebook the coding part has been explained in detail.

Question 4:

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

A model is considered to be robust if the model is stable, i.e. does not change drastically upon changing the training set. The model is considered generalizable if it does not overfits the training data and works well with unseen data.

Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e. **the accuracy does not change much for training and test/unseen data**.