Assignment – Advanced Regression (Part II – Subjective Questions)

Please limit your answers to less than 500 words per question.

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 (sometimes also referred to as lambda) for ridge and lasso regression is found using variance and bias square trade-off function. The optimal value of alpha is found where the total error on variance-bias trade-off graph is minimal.

For our Ridge regression model, the optimal value of alpha is 100.
And for Lasso regression model, the optimal value of alpha is 1000.

If the alpha is chosen as double then the penalty term will double. Thus increasing the cost function for ridge and lasso regression models. If alpha increases the variance decreases and bias increases.

The most important predictor value, after the alpha (or lambda) is doubled, is 'GrLivArea'.

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 lasso regression model will be chosen as it has eliminated some features which do not contribute to the model much. This will smoothen the linear model and reduce the number of features improving the interpretability of 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 first five most important predictor variables are:

1. GrLivArea (19069)
2. GarageArea (6475)
3. ExterQual_Gd (5241)
4. BsmtQual_Ex (4873)
5. BsmtFinType1_GLQ (4127)

The next five most important predictor variables are:

6. BldgType_1Fam (4119)
7. KitchenQual_Ex (4031)
8. Neighborhood_NridgHt (3987)
9. Neighborhood_Somerst (2988)
10. Functional_Typ (2712)

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:

The training and test score of the model should not defer much to ensure that the model is robust and generalizable. Also the data for training and testing needs to be chosen carefully to avoid any bias or influence.

If the training score is high but the testing score is low, it means that the model is learnt the training data and is an example of overfit. If the training score is lower than the test score then the model is underfit. Ridge and Lasso regression are used to improve the accuracy of the model by reducing total error.