Assignment - Advanced Regression

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 value of alpha for the ridge and Lasso regression in this model is 500. The optimal value helps us get the best coefficients for this model in which we are able to balance the Bias and Variance.

However if we double the value of alpha and then check the results, the coefficients are further reduced and pushed close to zero. In lasso regression alot more of the coefficients would end up being zero. This could also result in underfitting and the model might not perform well in training data as well.

The most important predictor variable after the changes would still remain as the Overall Quality. only the value of the coefficient is reduced but again it is reduced overall for all the coefficients.

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

The optimal value for both the Ridge and Lasso regression came out to be 500. I will use 500 as the value of alpha as this would give us the best coefficients.

Simple Linear Regression gave very bad results in the Test Data and hence the model was too complex and performed miserably on the unseen data.

Lasso regression did fairly better but there was still a huge difference between the results of the training data and the test data which meant that the model was still too complex.

However in Ridge Regression both the Training results and the test results were really good and hence I would choose the Ridge Regression.

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
As per the Lasso Model the 5 most important predictor variables were as follows:

- *GrLivArea*
- *RoofMatl_CompShg*
- *RoofMatl_WdShngl*
- *RoofMatl_WdShake*
- *RoofMatl_Tar&Grv*

However is for any reason these information is not available in the incoming data then the next 5 most important variables would be the following:

- *RoofMatl_Metal*
- *RoofMatl_Roll*
- *BsmtFinSF1*
- *OverallQual*
- *SaleType_New*

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
To make sure that the model is robust and generalisable there are 2 most import things.

First of all we need to make sure that the model performs equally well in the test data. if the model performs well on the test data we know that the model will perform well on any unseen data. This can be achieved by making sure that the model is neither too complex and not too simple. we need strike a balance between the Bias and Variance in order to achieve the same. This is where Regularization comes in the picture and helps us achieve that balance

Second most important thing is that the coefficients whould not carry extreme values as that can make the model very unstable. This is again something which regularization helps us with by penalising the coefficients it helps us reduce the extreme values of the coefficients.