Problem Statement - Part II

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 0.001 and lasso regression is 0.01.

When we double the value of alpha for our Ridge regression, the model applies more penalty to the curve, attempting to make the model more generalized and simpler, without trying to fit every data point in the dataset. From the graph, it is evident that when alpha is set to 10, we observe higher errors for both the test and train datasets.

Similarly, when we increase the value of alpha for Lasso regression, we further penalize the model, causing more coefficients of the variables to reduce to zero, which, in turn, leads to a decrease in the R-squared value.

The most important variables after these changes have been implemented for Ridge regression are as follows: MSZoning_FV, MSZoning_RL

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:

I would use Ridge regression because is have the coefficients towards zero and makes Mean Squared Error exactly equal to 0.

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:

Those 5 most important predictor variables: 1. GarageAre 2. TotalBsmtSF

3. GrLivArea 4. OverallQual 5. OverallCond

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:

Simpler models are more versatile and can be applied to a wider range of problems. They require fewer training samples to be effectively trained and are easier to work with. Simpler models tend to be more robust, as they don't change dramatically with changes in the training data. Complex models, on the other hand, often have low bias and high variance, making them prone to overfitting and less generalizable.

To strike the right balance between simplicity and complexity, regularization can be employed. Regularization adds a penalty term to the model's cost, encouraging it to remain simple without becoming too naive.

Making a model simple relates to the Bias-Variance Trade-off: Complex models are unstable and sensitive to changes in the training data. Simpler models, which capture key data patterns, tend to remain stable even with data additions or removals.

Bias measures a model's likely accuracy on test data. A very simple model can have high bias, making it inaccurate for most test inputs. In contrast, variance refers to how much the model changes with different training data, which can impact its performance.