Advanced Regression assignment Q&A

Q1: 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?

Ans:

- Ridge: {'alpha': 6.0}
- Lasso: {'alpha': 0.0001}

When we double the alpha value in our Ridge regression, the model imposes a greater penalty on the curve. This encourages the model to prioritize generalization, aiming for simplicity rather than attempting to fit every data point in the dataset. This tendency is evident from the graph and it also leads to increased errors for both the test and train datasets.

Likewise, by increasing the alpha value in Lasso regression, we intensify the penalization on our model, resulting in more coefficients of the variables being pushed to zero. As we raise the alpha value, the R-squared value, a measure of the model's explanatory power, also decreases.

Q2: 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?

Ans: I would prefer to use Lasso regression since the dataset is large, and this approach will assist in selecting predictors along with regularization.

Q3: 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?

Ans: Below are the five most important predictor variables.

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

Q4: 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?

Ans: The model's simplicity enhances robustness and generalizability, even though accuracy may decline. This concept aligns with the Bias-Variance trade-off. Simpler models possess more bias but less variance, leading to greater generalizability. This implies that a model with these traits performs consistently well on both training and test data. Bias refers to errors stemming from a weak model's inability to learn from data, causing poor performance on both data types. On the other hand, variance signifies errors arising from a model that overlearns data, excelling on training but failing on unseen testing data. Maintaining balance between bias and variance is crucial to prevent overfitting and underfitting.