Assignment 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:

The optimal values of alpha for ridge and lasso regression are 20 and 100. The r2 score obtained are as under:

```
r2_score in train dataset
r2_score for ridge: 0.89
r2_score for lasso: 0.9

r2_score in test dataset:
r2_score for ridge: 0.82
r2_score for lasso: 0.82
```

After doubling the alpha values for Ridge and Lasso, the prediction accuracy would remain similar except for a slight change in 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:

Though we have got similar r2 output for both test and train datasets using Ridge and Lasso, I would choose Lasso because it helps in feature elimination and thus, makes the model more robust. Through feature elimination Lasso automatically selects those features that are useful and discards those features that are not useful. The key difference between the two is that Lasso will often zero out features while Ridge reduces the weight of most in the model.

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:

The five most important predictor variables now will be:

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

Question 4:

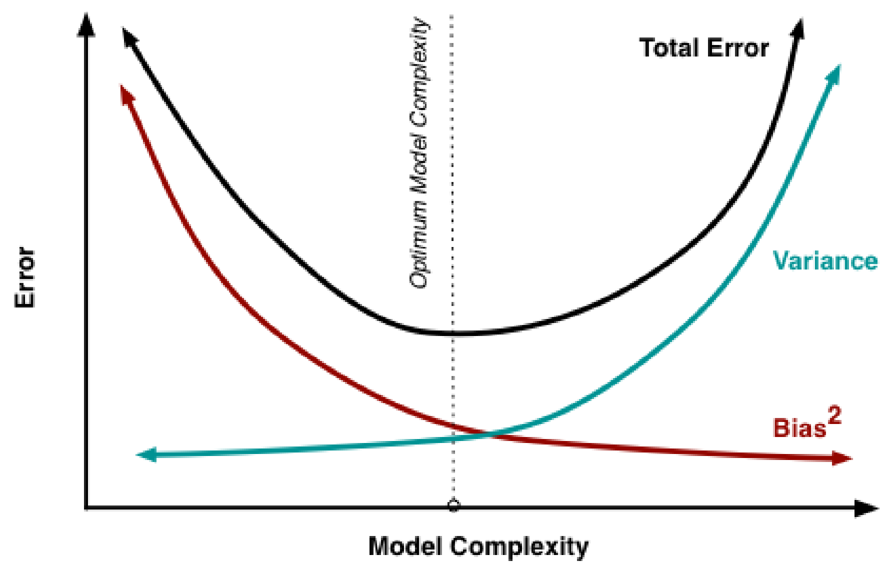
How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:

Any model that is kept simple can be said to be robust and generalizable. One preliminary approach is through proper outlier treatment at the EDA stage.

Regularization also helps to make the model simpler. Simply put, it’s a smoothing technique. It just adds an extra term to the cost function that is being used to evaluate the model. This term controls the parameters in the target function and make sure that they don’t take extreme values. In practice, regularization often lead to a slightly higher bias but significantly reduces the variance. This is what we call the **bias-variance tradeoff.**

For a model to be robust and generalizable, it should perform equally well both of train and test sets without much deviation. The accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as depicted in the graph below:



References, if any:

https://medium.com/analytics-vidhya/regularization-a-method-to-solve-overfitting-in-machine-learning-ed5f13647b91

https://towardsdatascience.com/lasso-and-ridge-regression-an-intuitive-comparison-3ee415487d18