SUBJECTIVE QUESTIONS

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 to 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

Below are the final/optimal alpha values for both the models:

- Ridge: alpha = 10
- Lasso: alpha = 0.001

After doubling the values of alpha, below are the values obtained:

1. Ridge:

With alpha = 20, below are the values obtained:

- R-squared value on train dataset - 0.91
- R-squared value on test dataset - 0.88
- RMSE value - 0.022

2. Lasso:

With alpha = 0.002, below are the values obtained:

- R-squared value on train dataset - 0.89
- R-squared value on test dataset - 0.87
- RMSE value - 0.020

The most important predictor variable after the change in alpha would be:

1. Ridge

Before (alpha = 10)	After (alpha = 20)
Ridge maximum column = Neighborhood_Crawfor Ridge maximum coefficient = 0.1035818724317891	Ridge_double max col = OverallQual Ridge_double max coef = 0.08323328821939213

2. Lasso

Before (alpha = 0.001)	After (alpha = 0.002)
Lasso maximum column = GrLivArea Lasso maximum coefficient = 0.11803117274292615	Lasso_double maximum column = OverallQual Lasso_double maximum coefficient = 0.10285704712399321

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

Below are the final alpha values for both the models:

- Ridge: alpha = 10
- Lasso: alpha = 0.001

Among the 2 models (Ridge and Lasso) considering the Mean Square Error values and R-squared values, we can conclude that the regression performed using the Ridge model has better results than Lasso.

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

After excluding the five most important predictor variables below are the five most important predictor variables:

- 2ndFlrSF
- 1stFlrSF
- MSZoning_FV
- Condition1_Norm
- YearBuilt

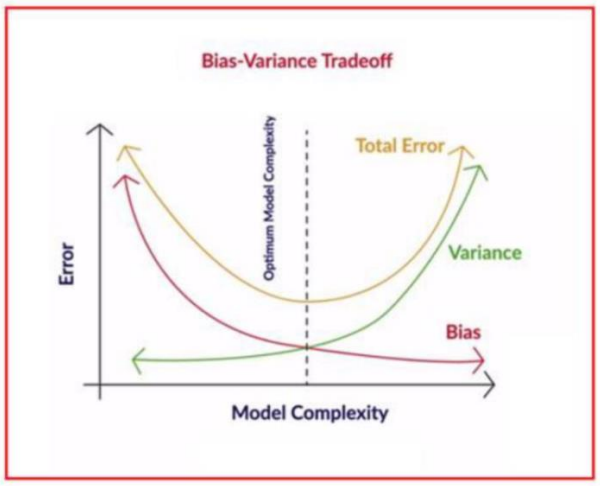
Question 4

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 testing error of a model must be consistent with the training error, or the model must perform well with sufficient stability even after adding noise to the dataset. As a result, the

model's robustness (generalizability) is a measure of how well it can be applied on data points other than the ones used for testing and training.



We can adjust the trade-off between model complexity and bias, which is directly related to the model's robustness, by using regularisation approaches. Regularization aids in penalising coefficients for overcomplicating the model, allowing only the optimum level of complexity to be used. It aids in the regulation of the model's robustness by making the model optimally simple. As a result, to make the model more generalizable and robust, a fine balance must be struck between keeping the model simple and not make it too naive to be useful. Making a model simple also results in a Bias-Variance Trade-off.

Bias allows you to determine how accurate a model is on test data. A non-simple model can make accurate prediction(s) if adequate training data is available. Models that are overly naive, such as the one that produces the same results for every test input and makes no distinction, have a high bias since their anticipated error across all test inputs is quite high. The degree of change in the model in relation to changes in the training data is referred to as variance.

As demonstrated in the graph above, the accuracy of the model may be maintained by maintaining a balance between Bias and Variance, as this reduces the total error.

A model which is complex will need to alter for every small change in the dataset, making it extremely unstable and vulnerable to changes in the training data. Even if more data points are added or deleted, a simpler model will display some pattern followed by the data points presented is unlikely to vary dramatically.

Therefore, robustness and accuracy may be at conflict, as an overly accurate model might be prone to overfitting, causing it to be overly accurate on train data but fail when dealt with actual data, or vice versa.