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:

Optimal Value of Alpha:

• The computed optimal value of alpha for Ridge Regression (Original Model): 8

• The computed optimal value of alpha for Lasso Regression (Original Model): 0.001

The changes in the model if you choose to double the value of alpha for both ridge and lasso

Ridge Regression => Original Model (alpha=8), Doubled Alpha Model(alpha=16)

Observations:

1. The ridge regression model with alpha value 8 has a slightly higher test accuracy compared to the model with double alpha value 16.
2. The mean squared error (MSE) test scores for the original and doubled alpha models show that the single alpha model has a slightly smaller MSE than the doubled alpha model.
3. Furthermore, the single alpha model performs better on both the train and test data compared to the doubled alpha model. Increasing the alpha value results in a decrease in the R2 score.

Lasso Regression => Original Model (alpha=0.001), Doubled Alpha Model(alpha=0.002)

1. The test accuracy of the lasso regression model with an alpha value of 0.001 is slightly higher compared to the doubled alpha model with an alpha value of 0.002.
2. The MSE test scores of similar data from the original dataset and the doubled alpha model show that the single alpha model has slightly smaller scores than the doubled alpha model.
3. In terms of train and test data, the single alpha Lasso Regression model performs better than the doubled alpha model.

The most important predictor variables after the change are implemented as follows

- Ridge Regression Model (doubled alpha=16)
- Lasso Regression Model (doubled alpha=0.002)

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:

Optimal Value of Alpha:

- The computed optimal value of alpha for Ridge Regression (Original Model): 8.0
- The computed optimal value of alpha for Lasso Regression (Original Model): 0.001

The Lasso Regression Model has a slightly better R2 test score in comparison to the Ridge Regression Model. Additionally, the reduction in training accuracy makes it a better option as it performs better on unseen data.

The Test set MSE of Lasso Regression is slightly lower than that of Ridge Regression Model, indicating better performance of Lasso Regression on unseen data. The Lasso method also helps in feature selection, with some coefficient values of insignificant predictor variables becoming 0.

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:

Top five features in original Lasso Model (before removing) were as follows:

```
['GrLivArea', 'MSZoning_FV', 'MSSubClass_160', 'Exterior1st_BrkComm',
 'AgeofProperty']
```

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?

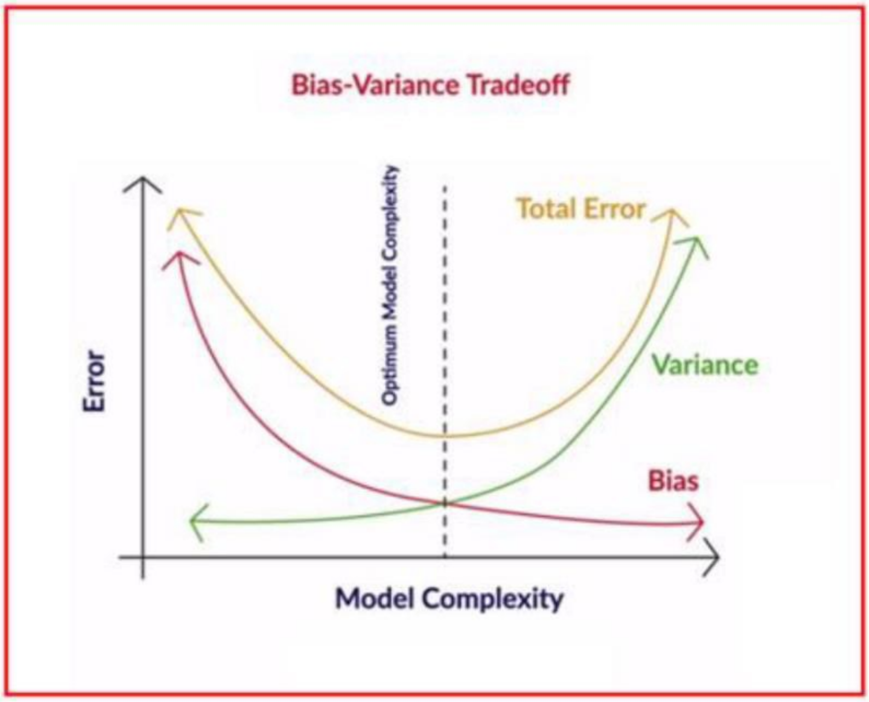
Answer:

The robustness of a model is measured by its ability to perform consistently well on data sets other than the one used for training and testing.

This can be achieved by implementing regularization techniques that control the balance between model complexity and bias. Regularization penalizes the coefficients for making the model too complex and allows only the optimal amount of complexity. A delicate balance between simplicity and complexity must be achieved to make the model robust and generalizable.

This leads to the Bias-Variance Trade-off where a complex model may be accurate but is very unstable and sensitive to any changes in the training data. On the other hand, a simpler model that abstracts some pattern from the data points is unlikely to change even if more points are added or removed.

Bias helps quantify how accurate the model is likely to be on test data, while variance is the degree of changes in the model with respect to changes in the training data.



Thus, accuracy and robustness may be at the odds to each other as too much accurate model can be prey to over fitting hence it can be too much accurate on train data but fails when it faces the actual data or vice versa.