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 value of alpha for ridge regression is 500 and for lasso Regression is 0.001

When we double the values of the alpha variable, we observe a dip in the model performance.

The R-square value is lower for both test and train data for ridge and lasso regression.

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:

For Model simplification, I would choose Lasso model as some of the coefficients become 0, it results in easier model selection and hence easier interpretation, particularly when the number of coefficients are very large. Although from the datapoints in the comparative study table shows that the ridge regression model has lower residual sum of squares for the test data compared to the train data. That shows the robust nature of Ridge regression. In the current case it would be best to go for Lasso Regression.

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:

The top 5 predictors are : MSZoning, RoofMatl, BsmtFinSF1, KitchenQual and GarageArea

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:

A model is considered robust and generalisable if its output dependent variable performs consistently well on train and test data even if one or more of the input independent variables or features are changed. In real life scenario, it is always the case that the test data will be unseen and unpredictable. The accuracy of the model is measured by metrics like R-squared value, mean-squared error and Residual sum of squares.