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:

Ridge: 2.0

Lasso: 0.0001

If we double the alpha value for

1. Lasso: (alpha – 0.0002)

   R2_score(train) 0.9298

   R2_score(test) 0.8657

The r2_score reduces from train - 0.9407 to 0.9298, but r2_score for test - 0.8572 increases to 0.8657 (with alpha = 0.0001)

2. Ridge: (alpha 4.0)

   R2_score (train) 0.9407

   R2_score (test) 0.8572

The r2_score increases from train 0.9277 to 0.9407 and decreases for test - 0.8613 to 0.8572 (alpha = 2.0)

Most important predictor variable from lasso is : GrLivArea

Most important predictor variable for ridge is: GrLivArea

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:

Based on the alpha/lambda values I have got from the model, Lasso will be a better option as it shrinks the coefficients of the variables and this helps in feature elimination making the model simpler considering r2_score is similar for both the regression models. Lasso will also help in dealing with multicollinearity.

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:

*Details on model creation as a part of notebook:*

Five most important predictor variables after creating another model are:

| | |
|---|---|
| OverallQual_9 | (coeff: 0.141) |
| Neighborhood_Crawfor | (coeff: 0.136) |
| CentralAir | (coeff: 0.110) |
| OverallQual_8 | (coeff: 0.099) |
| MSZoning_FV | (coeff: 0.089) |

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:

Ensuring that a model is robust and generalizable involves following steps:

1. Cross-Validation: It involved dividing the dataset into multiple subsets and train the model on different combinations of these subsets.
2. Regularization: Lasso and Ridge regression (Regularization methods) can help prevent overfitting by adding a penalty term to the loss function that the model is trying to minimize. Lasso also helps in feature elimination which further helps to make model simpler.
3. Feature Scaling: Using feature scaling we ensure that all features are on a similar scale, Algorithms perform better when numerical input variables are on a similar scale and this also helps in speeding up the learning by optimizing algorithm coverage faster.
4. Hyperparameter Tuning: This involves adjusting the parameters of the model to find the values that produce the best performance on a validation set.

A model that is robust and generalizable will have higher accuracy on unseen data, as it has been trained to perform well under a variety of conditions and will not overfit. This means that slight modifications to the test data will not result in significant change in model accuracy.