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:

For ridge regression ,the optimal value of hyperparameter is 5 and for lasso the optimal value of hyperparameter is 0.0001.

Below are the changes in model when the alpha value is doubled:

1. The R2 value of the model on the train data set decreases from 0.92 to 0.91 for ridge regression. For lasso the value decreases from 0.93 to 0.92.
2. The R2 value of the model on the test data set increases from 0.756 to 0.778 for ridge regression. For lasso the value increases from 0.72 to 0.75.
3. The coefficient of the most important predictor variable increases from 0.135 to 0.25 in case of lasso regression and decreases from 0.063 to 0.053 in case of ridge regression

The most important predictor variable which is the **GrLivArea (above ground living area square feet)** still remain the same.

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:

Since our business objective is to predict the most important predictor variables, we will go with **lasso** regression as it will eliminate few of the predictor variables (coefficient is 0) and hence it will make it easier for analysis and deciding the important predictor variables. For ridge the coefficients won't be 0 and hence there may be chance we may include not important predictor variables for analysis.

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 five most important variables now will be:-

1. Pool quality is excellent
2. 1st floor area in square feet
3. Basement Finished Type 1 area in square feet
4. House whose rooftop material is made up of wooden shingles.
5. Total number of rooms above ground

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:

We can use the following methods:

1. Use K-folds cross validation technique while training the model and increase the number of folds.
2. We can use ridge/lasso regression in order to find the optimum value of hyperparameter ensuring that the model is not overfitting the data and at the same time increase robustness of the model.
3. We can remove the predictor variables having multicollinearity either using manual approach or using RFE and then train our model. As we are reducing the predictor variables, we can be sure that model will not overfit the data and hence increase robustness.

When we make a model robust and generalisable it will lower the accuracy as we prevent overfitting and thus lowering the variance but at the same time increasing bias resulting in lower accuracy. We have to decide the optimum value which balances both variance and bias.