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?

Ans :

The optimal lasso and ridge values are as follows: for the lasso:

0.0001 is the best alpha

for the ridge:

finest alpha = 1.0

If we double the alpha value, the accuracy changes and becomes low, and most of the coefficients become zero .

In the case of the ridge model, it becomes simple.

Most important predictive variables after the change is implanted are :

Lasso :

a. GrLivArea

b. OverallQual

c. LotArea

Ridge:

a. OverallQual

b. 1stFlrSF

c. 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?

Ans :

It is critical to regularise coefficients in order to improve prediction accuracy while also decreasing variance and making the model interpretable.

Lasso regression employs a tuning parameter known as lambda as the penalty, which is the absolute magnitude of the coefficients as determined by cross validation. Lasso shrinks the coefficient towards zero as the lambda value increases, making the variables exactly equal to 0. Lasso can also perform variable selection.

When lambda is small, the model performs simple linear regression; as lambda increases, shrinkage occurs, and variables with 0 values are ignored by the model.

Ridge regression employs a tuning parameter known as lambda as the penalty is the square of the magnitude of the coefficients as determined by cross validation. Using the penalty, the residual sum or squares should be small. The penalty is lambda times the sum of the squares of the coefficients, so the coefficients with higher values are penalised. As we increase the value of lambda, the variance in the model decreases while the bias remains constant. In contrast to Lasso Regression, Ridge Regression includes all variables in the final model.

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?

Ans:

If this occurs, we must select the next five most important predictors.

They are as follows:

1. BsmtFinSF1
2. RoofMatl_Tar&Grv
3. 1stFlrSF
4. GarageArea
5. Neighborhood_StoneBr

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?

Ans:

The model should be as simple as possible, as this will reduce accuracy while increasing robustness and generalizability. The bias-variance trade-off can also be used to understand it. The simpler the model, the greater the bias, but the lower the variance and generalizability. Its accuracy implication is that a robust and generalizable model will perform equally well on both training and test data, i.e. the accuracy does not differ significantly between training and test data.

Variance is an error that occurs in a model when it attempts to learn from data. The high variance mean model performs exceptionally well on training data because it has been very well trained on this data, but it performs extremely poorly on testing data because the model has never seen this data before.

To avoid overfitting and underfitting of data, it is critical to maintain a balance of bias and variance.

A bias in a model occurs when the model is too weak to learn from the data. A high bias indicates that the model is unable to learn details from the data. On training and testing data, the model performs poorly.