

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 – 5.0

The optimal value of alpha for Lasso – 0.0004

If we double the value of alpha the MSE value decreasing in both Ridge and lasso methods, even though MSE decreasing we have to set better trade off for good fit.

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:

As we know it is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model more reliable.

Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum of squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the 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?

Answer:

The 5 Most predictor variables are mentioned below,

1. MSZoning_RL
2. GrLivArea
3. MSZoning_RM
4. OverallQual
5. MSZoning_FV

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 needs to be designed as robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalise so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used typically as 3-5 standard deviation. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.