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 optimum value of alpha for Lasso Regression is: 0.001.

The optimum value of alpha for Lasso Regression is: 0.2.

After increasing the value of alpha 2 times for both Ridge and Lasso we found the change in value of R2 for both test and train data.

Change for Lasso:

R2 for train data: 0.7940437405946943
R2 for test data: 0.7767492746365372
Accuracy decreases both for test and train data, which means it shows a tendency of under fitting .

Change for Ridge:

R2 for train data: 0.8365620624367285
R2 for test data: 0.7858831858288889
There is a minute change is noticed in R2 score for both test and train data. Accuracy goes down.

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: Lasso regression is recommended over Ridfge regression as it serves the purpose of feature elimination. As the dimensionality is reduce the model will become more robust.

Question 3

After building the model, you realized 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: In that case, we need to go for feature selection once again. We can run Recursive Feature Selection (RFE), or we can do it manually. This is because when RFE select features it works as a combines relations, there may be certain amount multicolinearity and other factors. If we do it manually, we need to check, P-value, Multicolinearity, variance etc.

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 said to be robust and generalized when both its bias and variance are found to be low.

As model should not be impacted by outliers in the training data, it needs to be made robust and generalized one. Moreover, it should be generalized so that the test accuracy is almost parallel to the train – score. It is the prior to make the model accurate for the unseen dataset. Outliers should not be prioritized much so that accuracy can be high for training data and accuracy falls for testing data. In order to get rid of the said problem, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Sometime any types of variation is treated as outliers and removed. This step may bring harm to the model, rather keeping them may help increase the accuracy of the predictions made by the model. Confidence intervals can be used (normally 3-5 standard deviations) which would help standardize the predictions made by the model.