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 regression is 100 and for lasso regression it is 500.

On doubling the alpha value, i.e., for ridge, alpha = 200 and for Lasso, alpha = 500 : R2, RSS and MSE metrics are not showing much difference, but the number of feature eliminated for Lasso has been increased from 195 to 264, means 69 more feature got eliminated based on double alpha value.

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:

Ridge and Lasso Regression both helps in finding the optimal alpha value and model building. When given an option to choose from these two, I will select Lasso Regression as it would help in feature elimination and because of that model will be more robust.

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:

After analysing the coefficients data frame, the below are 5 most important columns, which got excluded by the Lasso Regression model but have sufficient high coefficient in Ridge regression:

1. SaleCondition_Partial
2. HalfBath
3. FullBath
4. TotRmsAbvGrd
5. 2ndFlrSF

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 made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable 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 weightage should not give 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.