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?

Solution:

The optimal value of alpha for ridge and lasso regression are 0.05 and 0.0001 according to my analysis.

And the results are as follows:

For ridge: (50 features using RFE):

Train R2 score= 0.9124102634286256

Test R2 score= 0.8501917065778409

Train MSE= 0.0014505222915650261

Test MSE= 0.002706505985834966

Ridge after doubling the alpha:

Train R2 score= 0.9084965912668707

Test R2 score= 0.8680177897418132

Train MSE= 0.0015153343224572203

Test MSE= 0.002384451714438089

For Lasso: (50 features using RFE)

Train R2 score= 0.8832913793550109

Test R2 score= 0.8829237614407246

Train MSE= 0.001932743064313415

Test MSE= 0.0021151535286954384

Lasso after doubling the alpha (0.0002):

Train R2 score= 0.8769726268360212

Test R2 score= 0.8819085780579621

Train MSE= 0.0020373842213993096

Test MSE= 0.002133494301688721

We can’t see any difference after doubling the alpha in both the cases.

The most important predictor variables are now: (0.0002 for lasso)

- overallqual
- grlivarea
- garagearea
- fireplaces
- bsmtqual
- fullbath
- bsmtfullbath
- bsmtexposure
- heatingqc
- exterior1st_BrkFace

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?

Even though the lasso Ridge model has higher train accuracy ~91%, it performs comparatively poorer ~85% also the main predictor suggested by the model are

Roof material categories, (top 6 are them only)

I will choose the doubled value of lambda in Lasso algorithm, which is 0.0002,

Because the predictor variables predicted in this model are believable and make more sense in business case scenario, also it has almost same ~88% scores in both test and train cases, that too after doing CV of 10.

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?

Okay, now for this

the top variables if the model is built on the new data which has the actual top 5's

- totalbsmtsf – Total square feet of basement area
- fullbath - Full bathrooms above grade
- roofmatl_WdShngl- roof material Wood Shingles
- fireplaces - Number of fireplaces
- overallcond- Rates the overall condition of the house

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?

Yes, the model can further be made more generalizable, by putting harsher penalties, for example in this case, increasing alpha more. The implication of this action will be, model may tend to underfit and become unable to recognise the underlying patterns of the data. In simple terms, decrease in Train accuracy, Test accuracy may increase up to certain limit, and then it will also start decreasing.

One more thing can be done, PCA can be used to make new features so that they won't lose the information and as well as we don't have to deal with the dimensionality curse, but this process also has its own disadvantage of losing interpretability, (original features will get transformed).