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:

For the given problem, optimal value for α_{Lasso}=0.001, α_{Ridge}=1.0.

If the alpha value is doubled then for both Lasso and Ridge, The test and train score has decreased. There is some shuffling in ranking of the coefficients with respect to the significance. After the change is implemented most important predictor variable is same as before which is GrLivArea in both lasso and ridge. But further increase in alpha changes the most important predictor variable to OverAllQual.

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:

Ovearall we will choose lasso because it is helpful in feature elimination and it can bring down the insignificant feature to zero. It is more helpful in explaining which features are not significant for business needs. So, from business and data point of view, data for lesser feature can give a good result in case of Lasso. Also, in this case Lasso gives a little better fit than Ridge.

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:

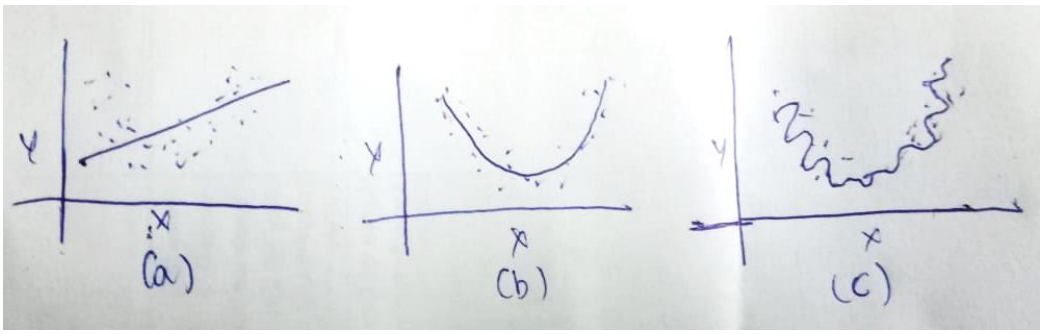
After removing the top 5 variable, new 5 predictor variable are 1stFlrSF, 2ndFlrSF, GarageArea, FullBath, Remod_Age(Remodelling year-Year of sale)

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 if the accuracy of the model does not vary much even when it is predicting data on which it was not trained.



If we observe the three model fits shown in the image above we can clearly see that the feature does not have a linear relationship with the target variable. So the fig(a) will have high bias and it will underfit and also it will have less accuracy. Similarly fig(c) has overfitted the data. It will have great accuracy with training dataset otherwise it may perform very poor, it has low bias but high variance. On the other hand fig(b) has low bias as well as low variance and will have good accuracy on both train and test data. Fig(b) is a robust model. Moreover, fig(b) is generalised with 2 degree polynomial feature instead of linear relation. Generalizing a model increase the accuracy because it will improve the fit by deciding which function best fits the model.