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?

The optimum value for lambda in ridge regression is 5

The optimum value for lambda in lasso regression is 100

If we double the value of alpha for Ridge Regression-

Ridge regression is extension of linear regression where the loss function is modified to minimize the complexity of the model. This modification is done by adding a penalty parameter that is equivalent to the square of the magnitude of the coefficients.

Loss function = OLS + alpha * summation (squared coefficient values)

A low alpha value can lead to over-fitting, whereas a high alpha value can lead to under-fitting.
So, if the alpha value is 0, it means that it is just an Ordinary Least Squares Regression model. So, the larger is the alpha, the higher is the smoothness constraint.

So, the smaller the value of alpha, the higher would be the magnitude of the coefficients and vice-versa.

If we double the value of alpha for Lasso Regression-

In Lasso, the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients.

Loss function = OLS + alpha * summation (absolute values of the magnitude of the coefficients)

In the above loss function, alpha is the penalty parameter we need to select.
The higher the alpha, the most feature coefficients are zero. That is, when alpha is 0, Lasso regression produces the same coefficients as a linear regression. When alpha is very large, all coefficients are zero.

Most Important predictors after increasing the value of alpha to double its value –

GrLivArea, Neighborhood_NoRidge, Neighborhood_Crawfor, SaleType_New, Neighborhood_StoneBr, Neighborhood_NridgHt

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?

The optimum value for lambda in ridge regression is 5

The optimum value for lambda in lasso regression is 100

Lasso Regression with lambda value as 100 will be selected as Lasso results in feature elimination by making the coefficients of insignificant variables as 0, which provides the features/variables which are significant in predicting the house price.

The Ridge regression can't zero out coefficients thus, we either end up including all the coefficients in the model or none of them.

Lasso method overcomes the disadvantage of Ridge regression by punishing high values of the coefficients and by setting them to zero if they are not relevant. Therefore, we end up with fewer features included in the model than what we started with, which is a huge advantage.

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?

Exterior1st_BrkFace

RoofMatl_WdShngl

BsmtExposure_Gd

SaleType_New

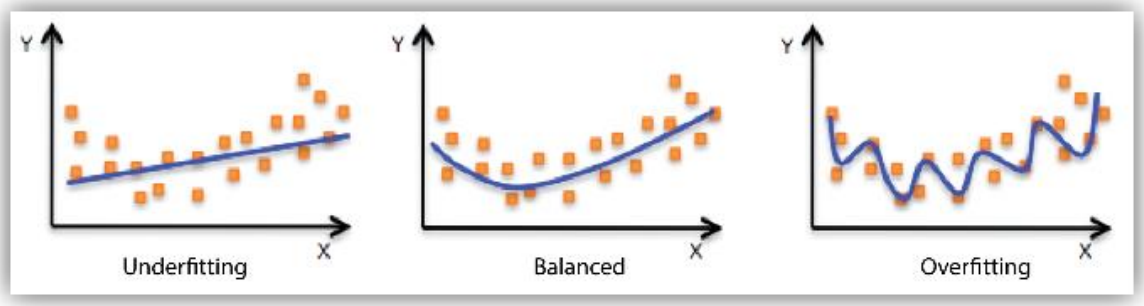
Functional_Typ

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?

Model robustness can be understood as - If a model has a testing error(on a new test set) equal to the training error, then the model is said to be robust i.e the model generalises well and doesn't overfit. A model will not be called robust when this model failed to generalise to the new dataset.

Generalization is a term used to describe a model's ability to react to new data. That is, after being trained on a training set, a model can digest new data and make accurate predictions. A model's ability to generalize is central to the success of a model. If a model has been trained too well on training data, it will be unable to generalize. It will make inaccurate predictions when given new data, making the model useless even though it is able to make accurate predictions for the training data. This is called overfitting. The inverse is also true. Underfitting happens when a model has not been trained enough on the data. In the case of underfitting, it makes the model just as useless and it is not capable of making accurate predictions, even with the training data.



The figure demonstrates the three concepts discussed above. On the left, the blue line represents a model that is underfitting. The model notes that there is some trend in the data, but it is not specific enough to capture relevant information. It is unable to make accurate predictions for training or new data. In the middle, the blue line represents a model that is balanced. This model notes there is a trend in the data, and accurately models it. This middle model will be able to generalize successfully. On the right, the blue line represents a model that is overfitting. The model notes a trend in the data, and accurately models the training data, but it is too specific. It will fail to make accurate predictions with new data because it learned the training data too well.

Extremely complex models don't generalize well since they are prone to change with small changes in the input data. Extremely simple models are likely to fail in predicting hence are prone to make errors and less accurate. Hence, we should always select a model which is just complex enough to understand the variance in the data without much inaccuracy at the same time not too complex to overfit. This can be achieved using regularization. Regularization is the process of deliberately simplifying models to achieve the correct balance between keeping the model simple and yet not too naïve.