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 1-

Ridge regression – When we plot the curve between negative mean absolute error and alpha as we see that s the value of alpha increases from 0 the error term decreases and the train error is showing decreasing trend when alpha increases.

For lasso regression I have decided to keep very small value that is 0.01, when we increase the value of alpha the model penalize more and make most of coefficient 0.

When the value of alpha is doubled for ridge regression now we will take the value of alpha 10 the model will apply more penalty on the curve and try to make the model more generalized that is making model simpler and no thinking to fit every value of data set.

When we increase the value of alpha for lasso, we try to penalize more our model and more coefficients of the variable will be reduced to 0.

For Ridge regression, the most important variable after the changes has been implemented are as follows: -

1. Neighborhood\_\_Crawfor
2. MSZoning\_RH
3. MSZoning\_FV
4. MSZoning\_RM
5. MSZoning\_RL
6. Neighborhood\_Stonebr
7. SaleCondition\_Partial
8. GrLiv\_Area
9. Exterior1st\_BrkFace
10. SaleComditoin\_Normal

For Lasso regression, the most imp. Variables after the changes are as follows:-

1. GrLivArea
2. OverallCond
3. TotalBsmtSF
4. OverallQual
5. LotArea
6. Firplaces
7. GarageArea
8. LotFrontage
9. BsmntFinSF1

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 2 -

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and making the model interpretably.

Ridge regression, uses a tuning parameter called lambda as the panelty is square of magnitude of coefficients which is identified by crosss validation. Residual sum or squares should be smaller by using the penalty. The panelty is lambda times sum of squares of the coefficients, hence the coefficients have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remain constant. Ridge regression includes all variables in final model unlike lasso regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficients towards zero.

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 3 –

The 5 most important predictor variables are: -

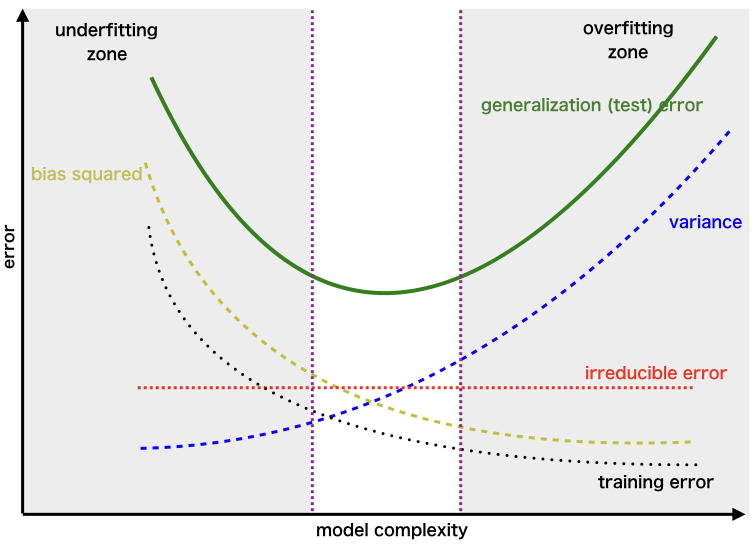
1. GarageArea
2. OverallQual
3. GrLivArea
4. OverallCond
5. TotalBsmntSF

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 4 –

The model should be as simple as possible, though it’s accuracy will decrease but it will be more robust and generalized. It can be seen using the Bias-Variance Trade off relationship. The simpler the model the more the bas but less variance and more generalizable.



**Bias**: Bias is the error in model when the model is weak to learn from the data. High bias mean model is unable to learn details in the data. Model performs poor on both train and test data.

**Variance**: Variance is the error in the model when the model perfectly learns from trin data. High variance means model perform exceptionally well on training data as it is very well trained on this but performs very poor on test data. This is also called overfitting.

It is important to have balance in Bias and Variance to avoid overfitting and underfitting.