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?

Ans:

The optimal value of alpha for Rigde is 4.

The optimal value of alpha for Lasso is 0.001.

When I doubled the alpha values and ran the model with 8 for Rigde regression below are the observations

1. The model accuracy score of R2 has come down slightly from 0.88 to 0.87
2. The coefficient values have come down slightly
3. The most significant variable are
    a. OverallQual_9
    b. TotalBsmtSF
    c. 2ndFlrSF
    d. OverallCond_9
    e. OverallCond_5

When I doubled the alpha values and ran the model with 0.002 for Lasso regression below are the observations

2. The model accuracy score of R2 has come down slightly from 0.87 to 0.86
3. The coefficient values of variables have come down slightly
4. The most significant variable are
    a. OverallQual_9
    b. TotalBsmtSF
    c. GrLivArea
    d. 2ndFlrSF
    e. 1stFlrSF

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?

Ans:

We would prefer Lasso regression because,

1. Lasso significantly reduced the coefficient values to 0 for 15 variables out of 30, thus making the final model with only 15 features
2. This will significantly reduce the complexity of the model and computationally efficient
3. Whereas Rigde still has all features with in the final model, thus making it generally computationally extensive if we have lot many features in the model

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?

Ans:

These are the five most important features in the final Lasso model after removing the first 5 most significant variables

1. MSZoning_FV
2. MSZoning_RL
3. BsmtExposure_Gd
4. MSZoning_RH
5. MSZoning_RM

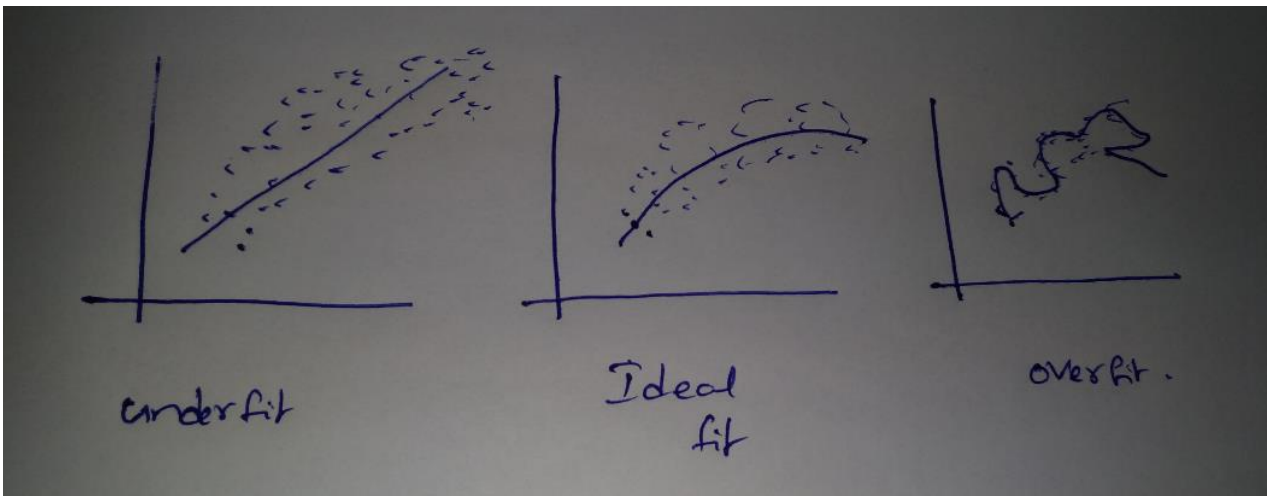
Question 4

How can you make sure that a model is robust and generalizable ? What are the implications of the same for the accuracy of the model and why?

Ans:

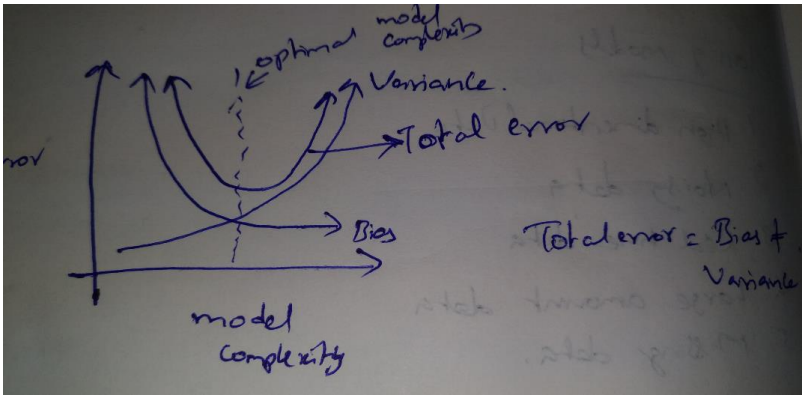
To make the model more Robust and generalized one we should not overfit or underfit our model to the training data.

For example, consider the below models.



a. The first one is using a regular Linear Regression. Since the data points are sparser and the regression line does not follow the pattern it is an underfit
b. The second one is an ideal model, because the regression line follows the pattern of the data without touching every point. This can be achieved by more degrees of polynomials
c. The last one where the regression line passes through all the data points is an Overfit. Since it passes through all the data points it tends to remember the data. This is achieved by higher degree polynomials

This can also be explained using Bias-Variance trade off



As per the above figure, the ideal model which is robust and generalized should have low bias and low variance.

For any model to be more robust and general, it should lie between underfit and overfit.

Accuracy of robust models:

Considering the above example,

a. The accuracy of underfit model will be lesser for both train and test data, because the regression line cannot fully follow the pattern
b. The accuracy of an Overfilt model will be close to 100 and R2 will be ~1. Ie., the model has remembered almost all the data points of training data and hence performs poorly on Test data. The accuracy of this model on test will be very bad and it cannot be used for any other dataset.
c. But for Ideal model, the accuracy ranges between 70-90, this will follow the pattern without remembering the data points and hence performs better on any other data set.