Problem Statement - Part II

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?

   Ridge_alpha = 0.1 after doubling 0.2
   Lasso_alpha =0.0001 after doubling 0.0002
   After doubling the alpha variables, I could see slight difference in r2 and RSS values.
   The top features after doubling the alpha variables are:

   - OverallQual
   - TotalBsmtSF
   - GrLivArea
   - MSZoning_FV
   - MSZoning_RH
   - MSZoning_RL
   - MSZoning_RM
   - LandSlope_Sev
   - Neighborhood_SWIS
   - Neighborhood_StoneBr
   - Condition1_RRAe

   Predictors are being the same, only the coefficient of the variables has been changed.

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?

   On the evaluation, the r2_value for Lasso regression was better compared to Ridge regression. Also Lasso regression eliminates the variables that doesn't contribute for model prediction by making the coefficients to zero to get better r2_value.

Hence, concluding Lasso regression will be the better decision to apply for this model accuracy and also it allows feature selection.

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?

   After dropping these top 5 variables **'OverallQual', 'TotalBsmtSF', 'GrLivArea', 'MSZoning_FV', 'MSZoning_RH'** the R_squared values for the test and train data got reduced to 63.09 and 59.55 from 87.94 and 86.53.

   Now below are the top 5 features:

   - MSZoning_RL
   - MSZoning_RM
   - LandSlope_Sev
   - Neighborhood_SWISU
   - Neighborhood_StoneBr

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?

   - A robust and generalisable model shouldn't be impacted by the outliers in the training data.
   - On this case the outliers need to be treated properly by removing the outliers that doesn't make much sense in the model prediction.
   - The test accuracy is not lesser than the training score.
   - The model should be accurate for datasets other than the ones which were used during training.