ASSIGNMENT PART II - SUBJECTIVE QUESTIONS

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 optimal value for alpha in **Ridge Regression** is **21.315789473684212.**
     - Train Score = 0.8738, Test Score = 0.8805
     - When we double the alpha value to **42.631578947368425**:
       - Train Score = 0.8682, Test Score = 0.8802
     - Most important predictor variables (with coef_) after the change are:
       - OverallQual(0.09), Condition1_Norm(0.05), 1stFlrSF(0.05), GarageCars(0.05), Neighborhood_NridgHt(0.04)
       - The top predictor variable remains the same and 3 of 5 variables are still present in top 5 after doubling the alpha. Also, the coefficients for the vars have decreased.
       - There are top 5 variables. For all variables, please refer Jupyter Notebook.
   - The optimal value for alpha in **Lasso Regression** is **0.0010819288389513114.**
     - Train Score = 0.8710, Test Score = 0.8810
     - When we double the alpha value to **0.002163857677902623:**
       - Train Score = 0.8608, Test Score = 0.8769
     - Most important predictor variables (with coef_) after the change are:
       - OverallQual(0.1), 1stFlrSF(0.08), Condition1_Norm(0.06), GarageCars(0.05), OverallCond(0.04)
       - The top 2 variables remain the same here. Other variables are different. Also, the coefficients for the vars have decreased.
       - These are top 5 predictor variables. For all variables please refer to the Jupyter Notebook.
   - Hence there is a clear slight dip in the train and test scores after doubling the alpha values for both Ridge and Lasso.

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?
   a. Alpha value for Ridge is 21.315789473684212 and number of variables are 78.
   b. Alpha value for Lasso is 0.0010819288389513114 and number of variables are 55.
   c. I prefer going with Lasso model due to below reasons:
      i. Less number of variables which makes the model simpler than Ridge.
      ii. Lower value of alpha which means there will be less penalty to the model based on the coefficients.
      iii. Which in turn means that larger the value of alpha greater the error term.
      iv. For Lasso, the alpha value is very low that means the increase in error term is less when compared to the Ridge model.

3. After building the model, you realized 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?
    a. **Ridge**
        i. The best value of alpha after removing the top 5 important variables is 90.45226130653266. Train Score is 0.8267 and Test Score is 0.8549.
        ii. The new top 5 important features after dropping the earlier top 5 are ['1stFlrSF', 'GarageCars', 'OverallCond', 'BsmtQual', 'KitchenQual']
    b. **Lasso**
        i. The best value of alpha after removing the top 5 important variables is 0.0008080808080808081 and the Test Score is 0.8471, Train Score is 0.8594.
        ii. The top 5 important features after dropping the previous top 5 for Lasso are ['GrLivArea', 'Condition1_Norm', 'SaleCondition_Normal', 'GarageCars', 'SaleCondition_Others']
    c. Please refer to Jupyter Notebook for detailed implementation.

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?
    a. A model is said to be robust if it gives a good test score though there are some unforeseen changes in the data. i.e. the model is adaptive to some real-time changes in the training data.
    b. A model is generalizable if it is as simple as possible. A simple model is expected to perform better on the unseen datasets.
    c. There will not be any overfitting in the simple model and hence the train score might be lesser when compared to the complex models, but the test score will be better than other complex models.
    d. The model should use required number of features and required amount of data. It should not use features or data which are more than required for training purpose.
    e. Generally, the accuracy of the robust and generic model will be lesser than that of other models in on the training set. This is because of the limited resources/data the model uses to get trained whereas when we investigate the accuracy on the test data, a robust and generic model will outperform the complex models are they are adaptive to the changes with the right bias and variance trade-off in place.
        i. A complex model might have high bias and low variance or low bias and high variance whereas the generic model will have a good bias-variance tradeoff and hence is more adaptive to the real world.