Problem Statement

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?

<u>Lasso Regression:</u>

- Optimum value of alpha found to be 0.01.
  At alpha=0.01, R-square of training dataset is 0.86 and testing dataset is 0.79. The most important predictor variables are as below.
- Alpha value doubled:
  If alpha value is doubled i.e., 0.02, then R-square if training dataset is 0.86 and that of test dataset is 0.77.

The testing and training dataset R-square value is decreased.

Top Features identified: `'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd','BsmtFinType1'`

- **<u>Ridge Regression</u>**:
- Optimum value of alpha found to be 10.
  At alpha=10, R-square of training dataset is 0.89 and testing dataset is 0.84. The most important predictor variables are as below.
- Alpha value doubled:
  If alpha value is doubled i.e., 20, then R-square if training dataset is 0.89 and that of test dataset is 0.83.

No change to R-Square.

<u>Below are the most important predictor variables:</u>

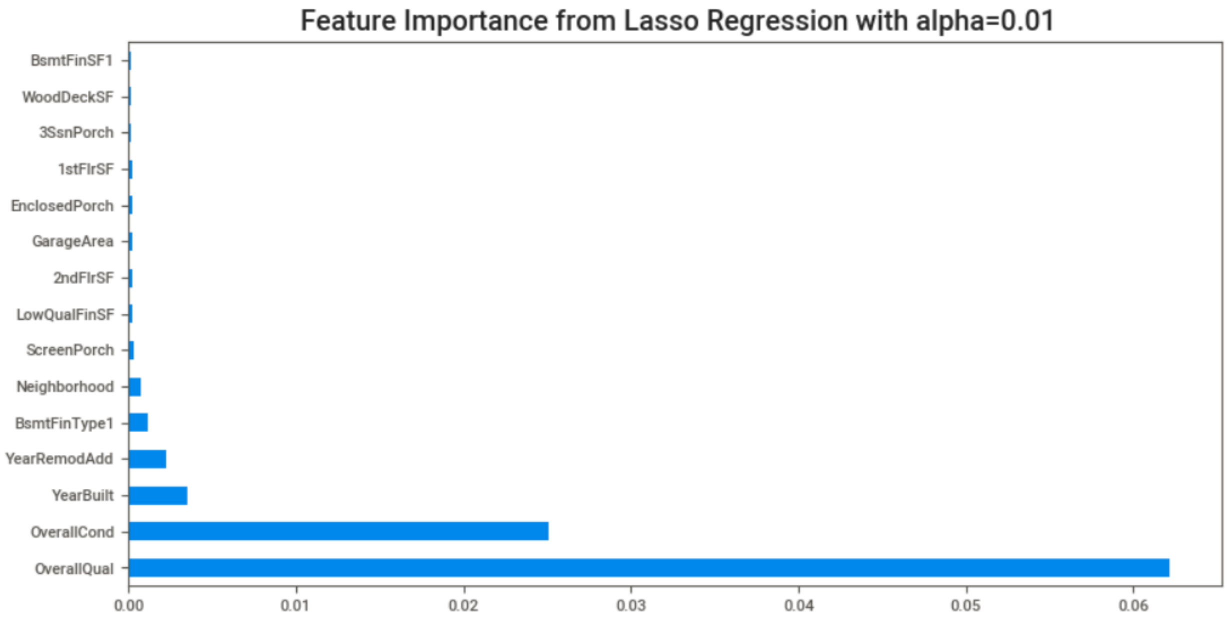
`‘OverallQual', 'OverallCond', 'BsmtFullBath', 'GarageCars', 'PoolQC'`

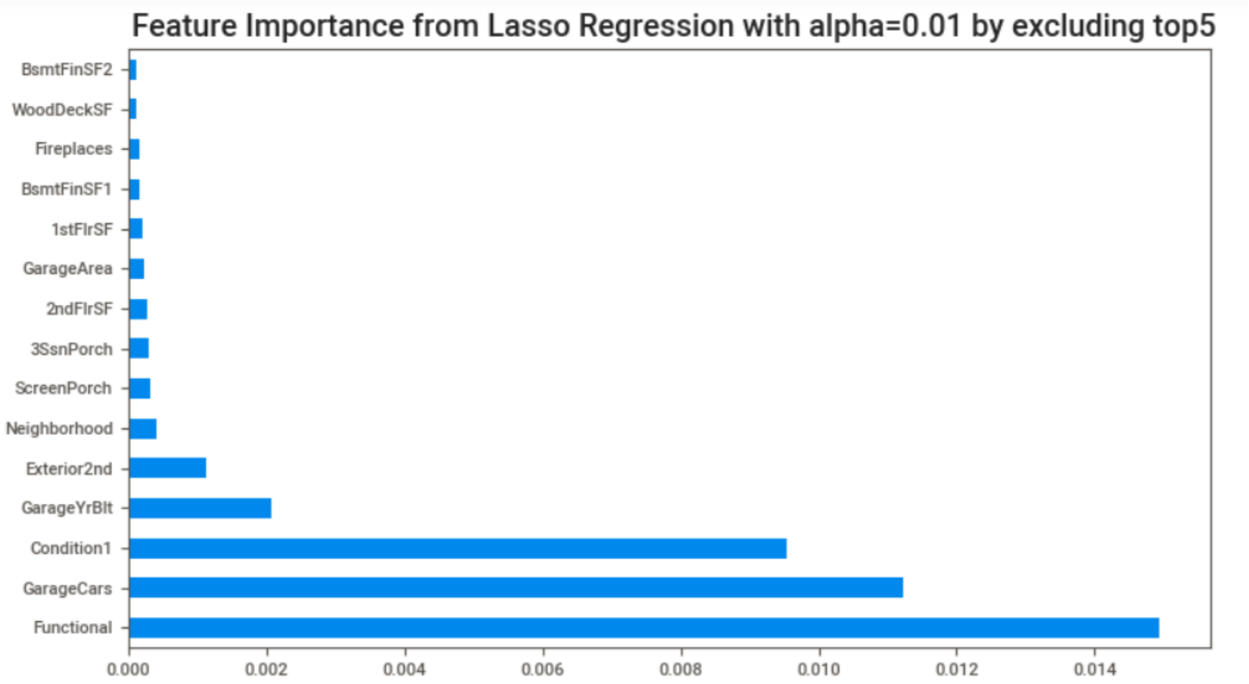
2. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Will choose Lasso Regression, since the Lasso adds an L1 regularization term to the loss function, which encourages the model to perform feature selection by setting some coefficients to zero. Lasso is generally preferred when you have a high-dimensional dataset with many irrelevant features and to identify & retain only the most important predictors. It has removed unwanted features from model without affecting the model accuracy. The model becomes simpler and reliable & preventing the model from becoming too complex and overfitting.

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?

- Difference in predictor variables after removing top 5 predictor variables are as below





Top Features identified in Lasso: `'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd','BsmtFinType1'`

Features identified after Top 5 features removal from Lasso: ‘`Functional', 'GarageCars', 'Condition1', 'GarageYrBlt', 'Exterior2nd'`

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?

- Dataset should be divided into training and testing subsets. This separation allows the model to learn patterns from the training data while also being evaluated on unseen data, assessing its ability to generalize. Moreover, cross-validation techniques, such as k-fold cross-validation, can be employed to reduce overfitting and provide a more reliable estimate of the model's generalization performance.
- Data preparation involves the selection, transformation, and scaling of features to reduce noise and emphasize important information. Regularization techniques like Ridge and Lasso can be applied to prevent overfitting by adding penalty terms to the loss function, discouraging large coefficients.
- Hyperparameter tuning involves optimizing parameters like learning rates, regularization strength, or the depth of decision trees. It can be done using methods like grid search, random search, or Bayesian optimization.
- Data preprocessing, which encompasses handling missing data, addressing outliers, and managing class imbalances, plays a vital role in model robustness. Proper treatment of

missing values, outlier detection and management, and techniques like resampling or synthetic data generation for imbalanced classes contribute to a more robust model.

- Balancing bias and variance focus on relevant information, and avoids overfitting leads to model accuracy of > 70-75%. The primary goal making model robustness essential for real-world success with accurate prediction