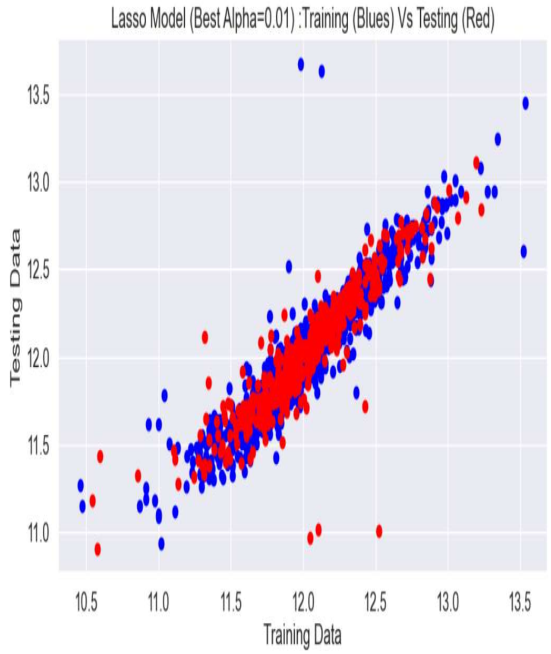
Problem Statement

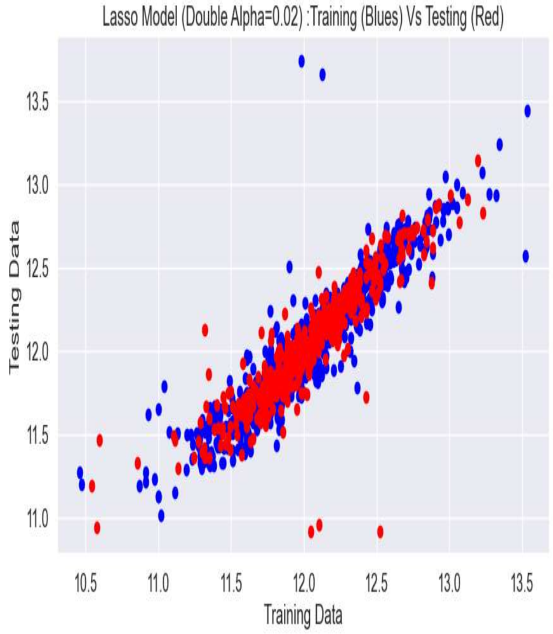
1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

➢ **<u>Lasso Regression</u>**:

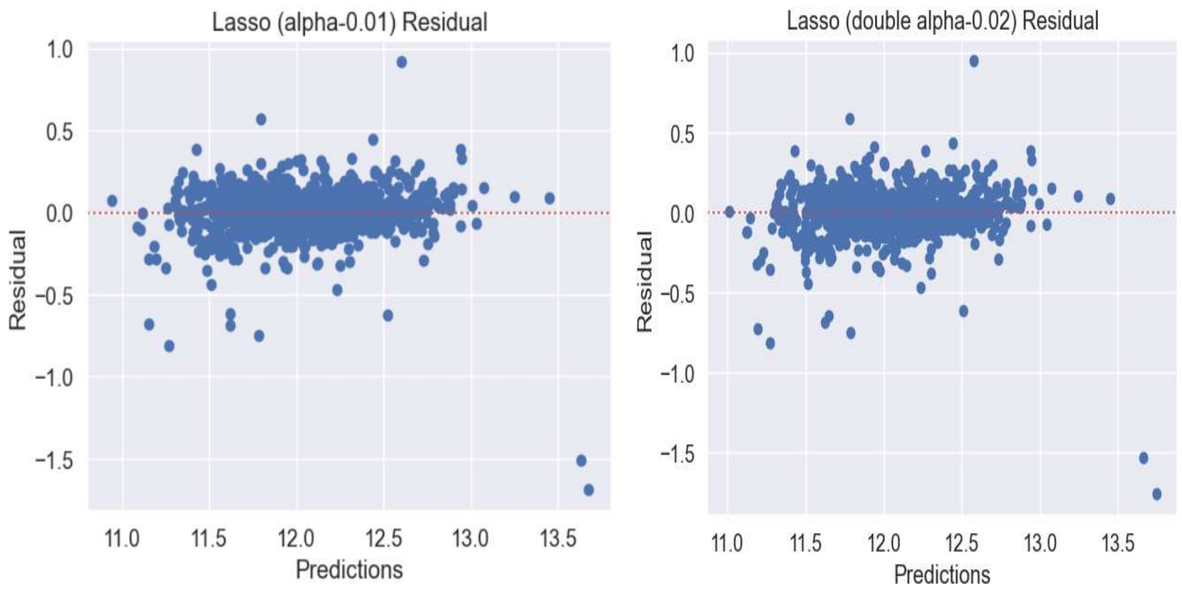
- Optimum value of alpha found to be 0.01.
  At alpha=0.01, R-square of training dataset is 0.87 and testing dataset is 0.82. The most important predictor variables are as below.
- Alpha value doubled:
  If alpha value is doubled i.e., 0.02, then R-square if training dataset is 0.86 and that of test dataset is 0.79. The testing dataset R-square value is decreased.

<u>Prediction Plot:</u>

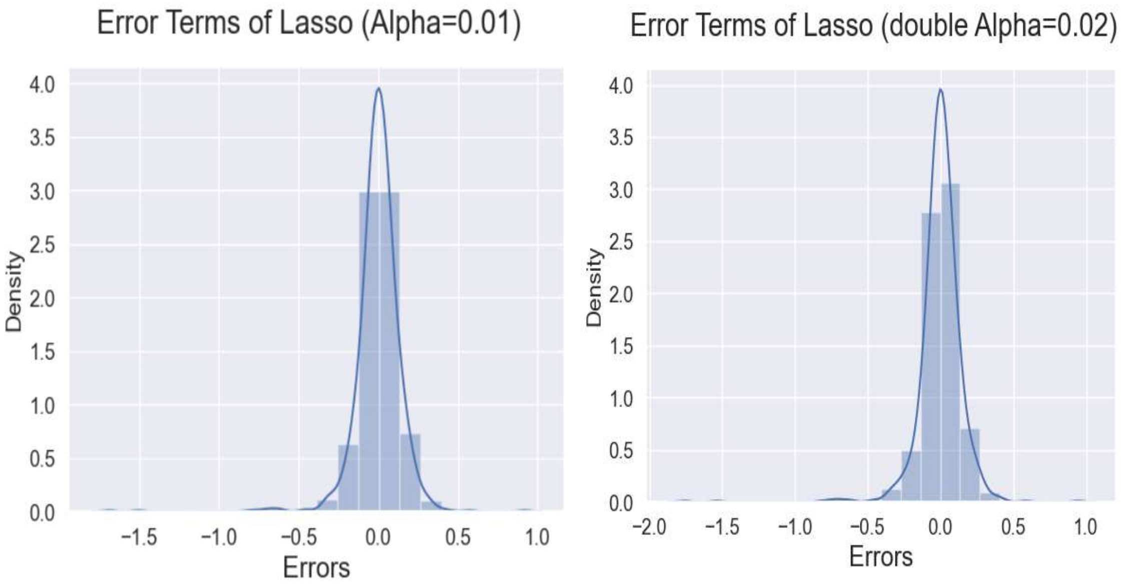




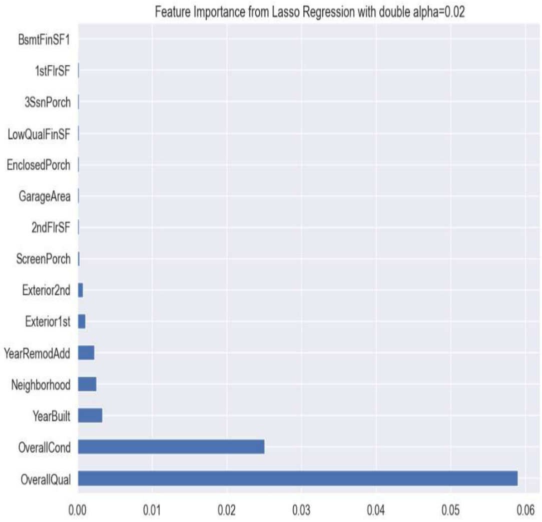
Residual Plot:

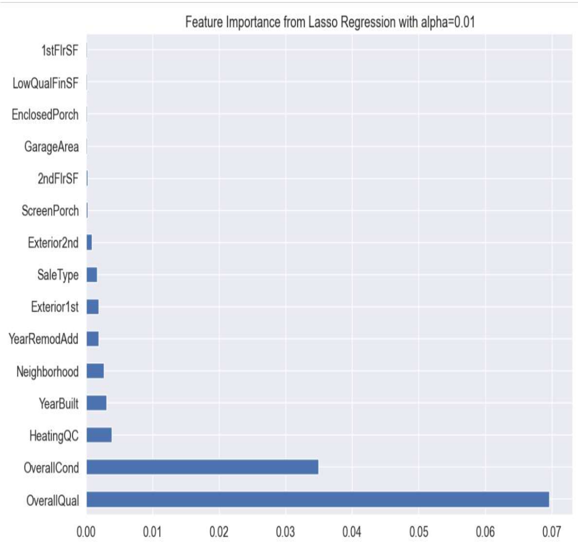


Error terms:



Below are the most important predictor variables:





Here some of the importance of predictor variables' are interchanged at different alpha values.

Top 5 feature predictors at alpha is 0.01 are ['OverallQual', 'OverallCond', 'HeatingQC', 'YearBuilt', 'Neighborhood'] &

Top 5 feature predictors at alpha is 0.02 are ['OverallQual', 'OverallCond', 'YearBuilt', 'Neighborhood','YearRemodAdd']

- **Ridge Regression**:

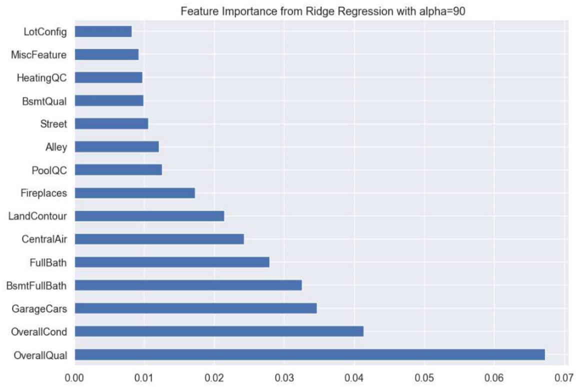
- Optimum value of alpha found to be 90.

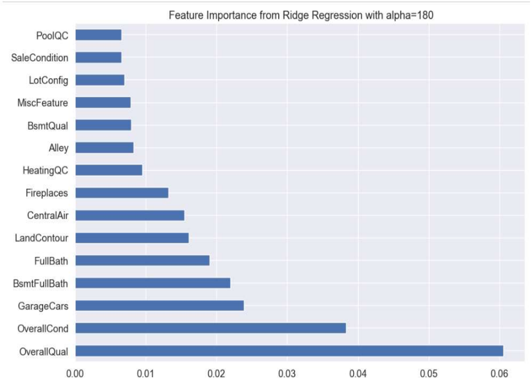
  At alpha=90, R-square of training dataset is 0.86 and testing dataset is 0.83. The most important predictor variables are as below.

- Alpha value doubled:

  If alpha value is doubled i.e., 180, then R-square if training dataset is 0.88 and that of test dataset is 0.82. The testing dataset R-square value is decreased.

Below are the most important predictor variables:





Top 5 feature predictors at alpha is 90 are ['OverallQual', 'OverallCond', 'GarageCars', 'BsmtFullBath', 'FullBath'] &
Top 5 feature predictors at alpha is 180 are ['OverallQual', 'OverallCond', 'GarageCars', 'BsmtFullBath', 'FullBath']

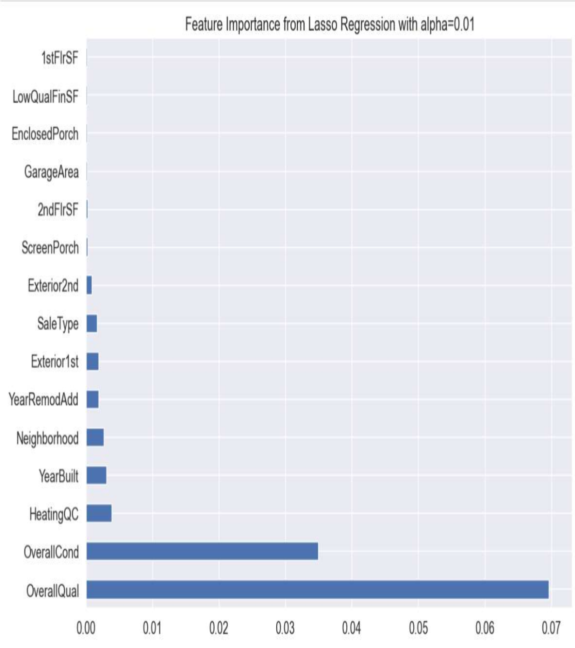
Therefore top 5 predictor features remains the same in case of Ridge Regreesion.

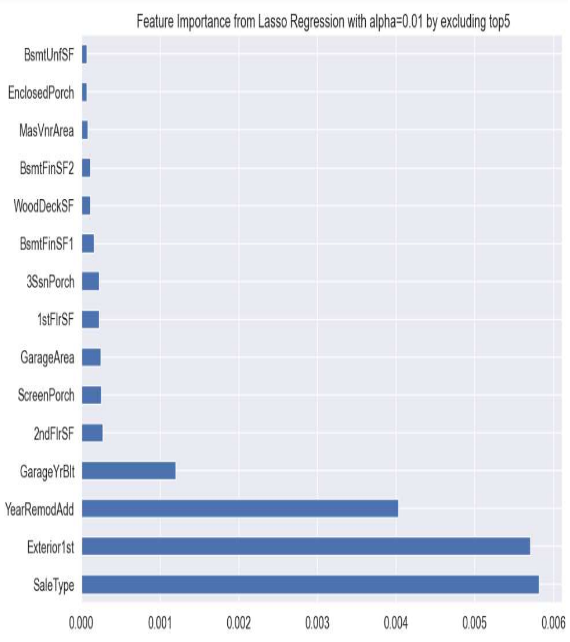
2. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

- I would prefer with Lasso Regression model, 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.
  Out of 78 predictor variables, Lasso model has relatively feature selected to 36 variables by setting rest 42 variable’s coefficient to zero. Hence 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:





- Therefore these are the top 5 predictor variables after removing ['SaleType', 'Exterior1st', 'YearRemodAdd', 'GarageYrBlt', '2ndFlrSF']

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?

- Ensuring that a model is robust and generalizable is fundamental for its successful deployment in real-world applications. To achieve this, several steps and considerations need to be taken into account
- Firstly, a 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.
- Feature engineering is another crucial aspect. It 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 is essential for fine-tuning the model's performance. This process 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.
- Ensemble methods, such as Random Forests and Gradient Boosting, can enhance model robustness by combining predictions from multiple models. They often lead to improved performance, especially when dealing with complex, noisy, or high-dimensional data.
- Balancing the bias-variance tradeoff is a delicate task. Overly complex models can have low bias but high variance, potentially leading to overfitting. Simpler models may have high bias but low variance, risking underfitting. Striking the right balance is essential for robustness.
- A robust model, which balances bias and variance, focuses on relevant information, and avoids overfitting, tends to achieve higher accuracy in practical applications. The primary goal of machine learning is to make accurate predictions on data it has never seen, making model robustness essential for real-world success.