Assignment-based 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?

A. The choice of the optimal alpha value for Ridge and Lasso regression depends on the specific data and problem you are addressing. Alpha is a hyperparameter that regulates the extent of regularization in these regression methods. It is typically determined through techniques like cross-validation, aiming to select the alpha that minimizes the model's prediction error on a validation dataset. In Ridge regression, increasing alpha intensifies regularization, resulting in a simpler model with smaller coefficient values. When you double the alpha value in Ridge, you augment the regularization strength, leading to even smaller coefficient values and enhanced resistance to multicollinearity (correlations between predictor variables).

Lasso regression also sees increased regularization as alpha grows, but Lasso has a distinctive property: it can drive some coefficients to precisely zero. This property makes Lasso valuable for feature selection. When you double the alpha in Lasso regression, you amplify the regularization, and more coefficients are likely to become zero. This leads to a simpler and more sparse model. After increasing alpha, the most important predictor variables will be those that the model retains. Generally, increasing alpha in both Ridge and Lasso penalizes model complexity and steers the model towards focusing on the most influential features. Consequently, variables with small coefficients or those that have limited relevance to the target variable are more likely to be set to zero in Lasso or reduced significantly in magnitude in Ridge. Identifying the most important predictor variables after adjusting alpha typically involves examining the model's coefficients. Coefficients significantly different from zero are indicative of important predictors. Additionally, feature selection techniques or visualization of feature importance scores can help identify influential variables. Keep in mind that the specific impact of doubling alpha on coefficients and variable importance will vary depending on your dataset, emphasizing the importance of employing cross-validation and proper model evaluation procedures for your specific problem.

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. we determined the optimal value of lambda for ridge and lasso regressions

- Ridge Regression optimal value is 20
- Lasso Regression optimal value is 100

To choose those values we run GridSearchCV to find the optimal threshold, we are given set of possible values to search with the threshold that gives less mean_absolute_error based on that we divided training data into 5 folds to generalize the model on test data for that optimal threshold.

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?

A. the top five most important predictors are

- 'PoolQC_Gd',
- 'RoofMatl_WdShngl',
- 'Neighborhood_NoRidge',
- 'GrLivArea',
- 'Neighborhood_StoneBr'

After removing those features from the model and training a new model based on the remaining features, the top fine most important features are

- '2ndFlrSF',
- 'LotShape_IR3',
- 'KitchenQual_Gd',
- 'KitchenQual_TA',
- '1stFlrSF'

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.
Train-Test Split: As mentioned previously, always split your data into training and testing sets. This allows you to assess how well your linear model generalizes to unseen data.

Cross-Validation: Utilize k-fold cross-validation, especially in cases where the dataset is limited. Cross-validation helps you estimate the model's performance on different data subsets and ensures that the model's performance is not overly influenced by a particular split.

Feature Scaling: Standardize or normalize your features to ensure that they are on a similar scale. Scaling helps linear models converge faster and can improve their generalization.

Regularization: Consider using Ridge or Lasso regression, both of which are linear models with built-in regularization. Ridge regression adds an L2 penalty term to the loss function, and Lasso regression adds an L1 penalty term. These penalties help prevent overfitting and improve generalization.

Hyperparameter Tuning: Optimize the hyperparameters of your linear model. For Ridge and Lasso, this includes tuning the regularization strength (alpha or lambda) to find the right balance between model complexity and generalization.

Residual Analysis: Analyze the residuals (the differences between actual and predicted values) to check for patterns or biases in the model's errors. This can guide further improvements in the model's generalization.