Housing Assignment Part 2:

Question 1

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

Answer:

Optimal Value of Alpha:

- **Ridge Regression:** 5.0
- **Lasso Regression:** 0.00025

Changes in the Model with Double Alpha:
Ridge Regression:

- New alpha: 10.0
- Impact: Increased regularization, which may shrink the coefficients more and potentially reduce overfitting, but could also underfit if too high.

Lasso Regression:

- New alpha: 0.0005
- Impact: Increased regularization, which may lead to more coefficients being set to zero, further enhancing feature selection but also possibly excluding relevant features.

Most Important Predictor Variables After Doubling Alpha:
Common Predictors for Both Models:

Predictors	Ridge	Lasso	abs_value_coeff
GrLivArea	0.061522	0.261608	0.261608
OverallQual	0.083627	0.196771	0.196771
GarageCars	0.041865	0.077431	0.077431
TotRmsAbvGrd	0.057208	0.055678	0.055678
OverallCond	0.046577	0.048732	0.048732
BsmtFullBath	0.033650	0.041228	0.041228
YearRemodAdd	0.030301	0.029950	0.029950
Neighborhood_Somerst	0.021127	0.029919	0.029919
Neighborhood_Crawfor	0.038677	0.029603	0.029603
Neighborhood_NridgHt	0.028106	0.028807	0.028807

Question 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?

Answer:

- The Lasso model achieved an R2 score of 92.04% on the training set and 85.64% on the testing set.
- The Ridge model achieved an R2 score of 92.28% on the training set and 88.18% on the testing set.
- Both models perform well, with the Ridge model showing slightly better predictive performance.
- However, the Lasso model eliminates most of the features, enhancing interpretability and simplicity.

Decision: We will choose the Lasso model for its feature selection capability, which simplifies the model while maintaining good performance.

Question 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?

Answer:
After excluding the initial top five predictor variables, the new five most important predictor variables in the Lasso model are:

1. **1stFlrSF** - Coefficient: 0.279826
2. **2ndFlrSF** - Coefficient: 0.128106
3. **TotRmsAbvGrd** - Coefficient: 0.068389
4. **GarageCars** - Coefficient: 0.067377
5. **Neighborhood_Crawfor** - Coefficient: 0.045464

These variables will now play a critical role in the new model.

Question 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?

Answer:
To ensure that a model is robust and generalizable, several steps and practices need to be followed throughout the model development and evaluation process. Here are the key methods:

Ensuring Robustness and Generalizability:

Cross-Validation:
Technique: Use k-fold cross-validation, where the data is split into k subsets, and the model is trained and validated k times, each time using a different subset as the validation set and the remaining k-1 subsets as the training set.
Benefit: Provides a more reliable estimate of model performance by reducing the impact of any one particular train-test split.

Train-Test Split:
Technique: Ensure the data is split into separate training and testing sets. The model should be trained on the training set and evaluated on the testing set.
Benefit: Helps assess how well the model will perform on unseen data.

Hyperparameter Tuning:
Technique: Use grid search or random search combined with cross-validation to find the optimal hyperparameters for the model.
Benefit: Prevents overfitting by selecting the best parameters that work well across different subsets of the data.

Regularization:
Technique: Apply regularization techniques like Lasso (L1) or Ridge (L2) regression to penalize large coefficients.
Benefit: Helps to avoid overfitting by discouraging overly complex models.

Feature Selection and Engineering:
Technique: Select only relevant features and create new features based on domain knowledge.
Benefit: Improves model performance and reduces overfitting by eliminating noise and irrelevant information.

Handling Outliers and Missing Data:
Technique: Detect and appropriately handle outliers and missing values through imputation or removal.
Benefit: Ensures that the model is not unduly influenced by anomalous data points.

Model Evaluation Metrics:
Technique: Use appropriate evaluation metrics (e.g., accuracy, precision, recall, F1 score, RMSE) depending on the problem type (classification or regression).
Benefit: Provides a comprehensive understanding of model performance across different aspects.