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 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 α (Lambda):

For Ridge Regression, the optimal α was 10.0.

For Lasso Regression, the optimal α was 100.0.

Doubling the α value will increase the penalty term, potentially leading to a simpler model with smaller coefficients. This can help prevent overfitting but might also lead to underfitting if the α value is too high.

Ridge Regression

Top Predictor: Neighborhood_NoRidge with a coefficient of approximately 20,431.69

Notable Changes: The OverallQual feature, which was previously the most influential predictor, has now moved to the 7th position in terms of importance.

Lasso Regression

Top Predictor: RoofMatl_ClyTile with a coefficient of approximately -197,118.47

Notable Changes: Several features that were previously significant have remained in the top 10, although their coefficients have changed

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:

Choosing between Ridge and Lasso regression depends on various factors, including the model's performance, interpretability, and the specific needs of the business. Here are some considerations:

Performance:

Both Ridge and Lasso performed well, with R^2 scores around 0.88 and 0.90, respectively. So, from a performance standpoint, both models are quite strong.

Interpretability and Feature Selection:

Lasso Regression not only performs regularization but also feature selection. If the goal is to identify a simpler model with fewer but more important features, Lasso might be more appropriate.

Business Goal:

If the primary goal is to understand how exactly the prices vary with different features (and potentially use this information to tweak business strategies), then a model that performs feature selection (i.e., Lasso) may be more beneficial.

Complexity and Computation:

Ridge Regression is computationally less expensive than Lasso if feature selection is not a requirement. So, if the model needs to be deployed in a real-time environment, Ridge might be preferable for quicker predictions.

Given that the business goal is to understand how prices vary with different features and possibly focus on the most influential features, Lasso Regression seems like a good fit. It will allow the company to focus on fewer but more important features, thereby potentially saving costs and resources.

Question 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?

Answer:

After retraining the Lasso model without the original top five most important predictor variables, the new top five most important predictor variables are as follows:

BsmtQual_Ex

GrLivArea

KitchenQual_Gd

KitchenQual_TA

Exterior1st_BrkFace

Question 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?

Answer:

Strategies for Robustness and Generalization:

Cross-Validation

Regularization

Feature Selection

Data Augmentation

- Ensemble Methods
- Train-Test Split
- Hyperparameter Tuning
- Out-of-Sample Validation

Implications for Accuracy:

- Bias-Variance Tradeoff:
  - A highly complex model will have low bias but high variance, making it likely to overfit.
  - A too-simple model will have high bias but low variance, making it likely to underfit.
  - The goal is to find a good balance to make the model generalizable.
- Overfitting and Underfitting:
  - Overfitting leads to excellent training accuracy but poor test accuracy.
  - Underfitting leads to poor performance on both training and test data.
- Model Complexity:
  - Increasing the model's complexity (more features, less regularization, etc.) will make it fit the training data better but may reduce its ability to generalize.