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:

1. Optimal value of alpha for ridge is 500
2. Optimal value of alpha for lasso is 0.001

Upon doubling the alpha value for Ridge and Lasso regression, the changes in model performance metrics and implications for predictor variables are as follows:

Ridge:

- **R-squared:** Decreased slightly for both training (from 0.91 to 0.89) and testing datasets (from 0.92 to 0.91), indicating a minor deterioration.
- **RMSE:** Increased for both training (from 23938.82 to 27058.66) and testing datasets (from 21022.99 to 21746.45), suggesting reduced prediction accuracy.
- **Top Predictors:** The importance of OverallQual, YearBuilt_OverallQual, and GrLivArea persisted, although with slightly adjusted coefficients.

Lasso:

- **R-squared:** Decreased slightly for both training (from 0.94 to 0.93) and testing datasets (from 0.92 to 0.91), indicating a minor deterioration.
- **RMSE:** Showed mixed changes, slightly improving for the test dataset (from 20009.29 to 19405.57) and worsening for the training dataset (from 19941.14 to 20988.98).
- **Feature Selection:** The number of zeroed-out coefficients increased (from 101 to 123), proving Lasso's enhanced feature selection capability with higher alpha, leading to a simpler, more interpretable model.
- **Top Predictors:** TotalSqFt and TotalBsmtSF_1stFlrSF continued to be significant, though with reduced coefficients, underscoring their continued relevance to the model.

After doubling the alpha value, the key predictors largely remain unchanged. Variables such as OverallQual, GrLivArea, and TotalSqFt in Ridge, and TotalSqFt, TotalBsmtSF_1stFlrSF in Lasso, continue to be significant, albeit with adjusted coefficients.

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:

Given the analysis, Lasso Regression is the preferred model for several reasons:

1. It exhibits higher R-squared values (0.94 for training and 0.93 for testing) compared to Ridge Regression, indicating a better fit.
2. Lasso achieves lower RMSE scores (19941.14 for training and 20009.29 for testing), signifying more accurate predictions.
3. It demonstrates superior performance in minimizing errors, as shown by lower MSE and RSS values, making it more efficient.

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:

Upon excluding the five most important predictors in the Lasso model, the next five most important predictor variables are:

1. OverallCond
2. GarageCars_GarageArea
3. LotArea
4. SaleCondition_Partial
5. Functional

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:

To ensure a model is robust and can generalize well, we must use a variety of techniques as follows.

1. **Cross-validation**: Essential for assessing model performance across different data subsets, ensuring it generalizes beyond the training data.

2. **Regularization**: Techniques like Ridge and Lasso reduce overfitting by penalizing model complexity, fostering models that generalize better.

The implications of enhancing a model's robustness and generalizability for its accuracy are as follows

1. **Reduced Overfitting:** These strategies lead to less overfitting, potentially lowering training accuracy but improving performance on unseen data
2. **Bias-Variance Trade-off:** Aiming for robustness involves balancing bias and variance to minimize overall error and optimize prediction reliability on new data.
3. **Slight Decrease in Training Accuracy:** While efforts to enhance generalizability might slightly diminish the model’s fit to the training set, they crucially improve its utility and predictive accuracy in real-world applications.