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:

Alpha Values
-- Ridge Regression --> 9.0
-- Lasso Regression --> 0.001

Doubling the value of alpha in both ridge regression and Lasso regression would increase the regularization strength in both models. This would have the following effects:

Ridge Regression:

- Increased alpha in Ridge regression would lead to stronger regularization, shrinking the coefficients more aggressively towards zero.
- The model would likely become more robust to multicollinearity, as higher alpha values discourage the model from relying too heavily on any single feature.

Lasso Regression:

- Similarly, doubling alpha in Lasso regression would increase the penalty for non-zero coefficients, leading to sparser solutions where more coefficients are pushed towards zero.
- This can result in feature selection, as features with coefficients that are less impactful might be reduced to zero, effectively removing them from the model.

Overall, doubling alpha in both Ridge and Lasso regression would result in more constrained models with potentially simpler and more interpretable solutions, but it might also lead to increased bias due to stronger regularization.

The most important predictor variables after the change is implemented are:
Positive**:**

1. SaleCondition_Partial
2. SaleCondition_Normal

Negative:

1. MSSubClass
2. LotFrontage

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:

Ridge regression is typically favoured in scenarios where:

- Numerous predictors exhibit modest to moderate impacts.
- Minimizing multicollinearity among predictors is a priority.
- It's assumed that predictors contribute fairly evenly to the outcome.

Lasso regression is the go-to choice when:

- Numerous predictors exist, but only a portion are anticipated to strongly influence the outcome, resulting in a sparse model.
- Feature selection is a priority, as Lasso's tendency to zero out some coefficients aids in removing less impactful predictors from the model.
- A more interpretable model with a reduced number of predictors is desired.

In deciding between ridge and lasso regression, considerations include finding a balance between bias and variance, the interpretability of results, and the analysis's specific objectives. Opting for lasso regression is beneficial when aiming for a simpler model with feature selection. Conversely, if preserving all predictors while mitigating multicollinearity is key, ridge regression is a better fit.

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:

As the top 5 predictors are not available, so the next available predictors will hold importance.

The next top 5 predictors are:

1. SaleType_WD
2. SaleType_Oth
3. SaleType_New
4. SaleType_ConLw
5. SaleType_ConLD

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:

The goal is to strike a balance between model simplicity and accuracy, considering the Bias-Variance trade-off. Simplifying the model reduces its complexity, increasing bias but decreasing variance, leading to better generalization. A robust and generalizable model maintains consistent performance across training and test data, indicating balanced bias and variance. Bias refers to model error due to insufficient learning, resulting in poor performance on both training and testing data. In contrast, variance reflects error from overlearning, where the model performs well on training but poorly on testing data. Balancing bias and variance is crucial to prevent overfitting and underfitting of the data.